
Text Mining and Gene Expression Analysis

Towards Combined Interpretation of High Throughput Data

Katrin Fundel



München 2007

Text Mining and Gene Expression Analysis

Towards Combined Interpretation of High Throughput Data

Katrin Fundel

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Katrin Fundel
aus Friedrichshafen

München, den 18.04.2007

Erstgutachter: Prof. Dr. Ralf Zimmer

Zweitgutachter: Prof. Dr. Oliver Kohlbacher

Tag der mündlichen Prüfung: 13.09.2007

Contents

Summary	xiii
Zusammenfassung	xv
1 Introduction	1
Part I Text Mining	9
2 Background: Text Mining	11
2.1 Fundamentals in Text Mining	11
2.2 Text Mining Tasks	13
2.3 Text Mining in Bioinformatics	15
2.4 Evaluation	17
3 Nomenclature of Biological Objects	21
3.1 Introduction and Literature Review	21
3.2 Gene and Protein Name Dictionaries	25
3.2.1 Generation of gene and protein name dictionaries	26
3.2.2 Curation	28
3.3 Analysis of Gene and Protein Name Dictionaries	30
3.3.1 Size of Gene Name Dictionaries	31
3.3.2 Ambiguity	33
3.3.3 Overlap between Data Sources	37
3.3.4 Relevance of Ambiguities for Mining MEDLINE	40
3.4 Hierarchical Synonym Dictionaries	42
3.4.1 Generation of Hierarchical Synonym Dictionaries	42
3.4.2 Evaluation	43
3.5 Other Dictionaries	46
3.5.1 Non-Gene and Non-Protein Synonym Dictionaries	46
3.5.2 Abbreviation Dictionary	47
3.5.3 Interaction Term List	48
3.6 Applications of Synonym Dictionaries	49
3.6.1 Literature Mine Browser (LiMB)	49

3.6.2	The ProThesaurus, BeThesaurus, and LiMB Web Services	50
3.6.3	The ProTag Client Applications	51
3.6.4	The ProThesaurus Wiki	52
3.7	Chapter Summary	53
4	Gene and Protein Name Identification	55
4.1	Introduction and Literature Review	56
4.2	The Exact Matching Approach	58
4.2.1	Match Detection	58
4.2.2	Rule-Based Postfilter	58
4.2.3	SVM-Based Postfilter	59
4.3	The ProMiner Approach	60
4.3.1	Principles	61
4.3.2	Match Detection	62
4.3.3	Extensions for BioCreAtIvE	64
4.4	The Combined Approach	65
4.4.1	Gene Name Detection	65
4.4.2	Extended Rule-Based Postfilter	65
4.4.3	Disambiguation between and within dictionaries	66
4.5	Evaluation	69
4.5.1	The BioCreAtIvE challenge	69
4.5.2	Evaluation Settings	70
4.5.3	Evaluation Results	72
4.5.4	Discussion of the Individual Approaches	74
4.6	Conclusions	82
4.7	Chapter Summary	84
5	Gene and Protein Relations	87
5.1	Introduction and Literature Review	87
5.2	RelEx - Relation Extraction Utilizing Dependency Parse Trees	90
5.2.1	The RelEx Workflow	90
5.2.2	Evaluation	97
5.3	Large-Scale Network Generation, Analysis, and Applications	101
5.3.1	Large-Scale Network Generation	102
5.3.2	Comparing RelEx Relations with HPRD Interactions	103
5.3.3	Comparing RelEx Relations with Y2H and Literature PPI Data . .	105
5.3.4	Using RelEx for Network Expansion	109
5.3.5	Network Schemes: A Means for Exploiting Context	111
5.4	Characterization of Gene/Protein Interactions	114
5.4.1	Data Preparation and Classification Approach	114
5.4.2	Evaluation	116
5.5	Conclusions	117
5.6	Chapter Summary	118

Part II Gene Expression Data Analysis	119
6 Background: Gene Expression Data Analysis	121
6.1 Microarrays – Biological Background	121
6.2 Microarray Technology	123
6.3 Microarray Expression Data Analysis Overview	125
6.4 Osteoarthritis	128
7 Gene Expression Data Analysis	133
7.1 Introduction and Literature Review	133
7.2 Data Sets	134
7.2.1 GPC Four-Class Data Set	135
7.3 Analyzing the Effects of Primary Data Processing	136
7.3.1 Normalization	137
7.3.2 Differential Expression	141
7.3.3 Number of Regulated Genes	147
7.3.4 Robustness Analysis	148
7.4 Deriving Reliable Gene Signatures for Microarray Classification	150
7.4.1 The StabPerf Approach	151
7.4.2 Application on Osteoarthritis Data	153
7.5 Chapter Summary	155
8 Gene Expression in Osteoarthritis	157
8.1 Analysis of Gene Expression in Osteoarthritis	157
8.1.1 Expression Levels of Genes Relevant for Anabolism	158
8.1.2 Differential Expression between Sample Groups	158
8.1.3 Clustering Analysis	159
8.2 Comparison of Microarray Platforms	161
8.3 Analysis of Osteoarthritis Models	163
8.4 IL1-Stimulation Time Series Analysis	164
8.5 Chapter Summary	166
Part III Integrated Data Analysis and Conclusions	167
9 Background: Integrated Gene Expression Data Analysis	169
9.1 Integration with Manually Compiled Data	169
9.2 Integration with Large-Scale Networks	170
9.3 Integration with Text Mining	171
10 Text Mining applied for the Interpretation of Gene Expression Data	173
10.1 ConceptMaker	173

11 Conclusions	179
11.1 Contributions of this thesis	179
11.2 Perspectives for Future Research	181
Abbreviations	186
Bibliography	189

List of Figures

1.1	Overview of the thesis structure	2
3.1	Example entry of a synonym dictionary in XML representation	26
3.2	Size of gene name dictionaries	32
3.3	Ambiguity within gene name dictionaries	34
3.4	Ambiguity between gene names and English and domain-related terms	37
3.5	Overlap between different data sources	39
3.6	The Literature Mine Browser (LiMB)	49
3.7	The ProThesaurus, BeThesaurus, and LiMB Web Services	50
3.8	The ProTag client application	51
3.9	The ProThesaurus Wiki	52
4.1	Workflow of the gene name identification systems applied for BioCreAtIvE	57
4.2	Evaluation results of all BioCreAtIvE submissions	72
4.3	Results for yeast in the BioCreAtIvE I (2004) gene normalization task	74
4.4	Results for mouse in the BioCreAtIvE I (2004) gene normalization task	75
4.5	Results for fly on the BioCreAtIvE I (2004) challenge data set	79
4.6	Results for human in the BioCreAtIvE II (2006) gene normalization task	81
5.1	RelEx workflow	91
5.2	Example of phrase structure and dependency parse tree	92
5.3	Example of RelEx rule 1	94
5.4	Examples of RelEx rules 2 and 3	95
5.5	Evaluation of RelEx on LLL-challenge data set	100
5.6	Comparison of manually annotated relations, HPRD, and RelEx relations	104
5.7	Schematic diagram of the analyzed interaction networks	106
5.8	Overlap of RelEx networks with public Y2H and literature PPI maps	107
5.9	Venn diagrams of RelEx networks and public Y2H and literature PPI maps	108
5.10	Manually compiled regulation network (IL1 pathway)	109
5.11	MMP13 interaction network	111
5.12	Network schema for exploiting context in text-mining networks	112
6.1	The central dogma of molecular biology	121
6.2	Articular capsule of normal and osteoarthritic joint	128

6.3 Chondrocytes are responsible for balanced turnover of extracellular matrix	131
7.1 Raw data distribution and normal probability plot for the four-class data set	138
7.2 Boxplot and group-level plot	139
7.3 Effects of normalization: Group-level and volcano plots	140
7.4 p-value combination: Motivation	142
7.5 Comparison of different methods for gene p-value determination	144
7.6 p-value combination: Exemplary results for the GPC four-class data set	145
7.7 p-value combination: Exemplary results for the Affymetrix two-class data set	146
7.8 Effect of normalization on the number of significantly regulated genes	148
7.9 Leave-one-out robustness analysis	149
7.10 Subset sampling robustness analysis	150
7.11 Overview of StabPerf	152
8.1 Overrepresentation analysis of osteoarthritis gene expression data	159
8.2 Heatmap and dendrograms of clustered osteoarthritis related samples	160
8.3 Comparison of Microarray Platforms and Experiments: PCA	161
8.4 Comparison of Microarray Platforms and Experiments: Clustering	162
8.5 Analysis of osteoarthritic models: Clustering of Affymetrix data sets	164
8.6 Analysis of osteoarthritic models: PCA of Affymetrix data sets	165
8.7 IL1-stimulation time-series experiment: Expression profiles	165
10.1 The ConceptMaker approach	175

List of Tables

1.1 Examples of public databases containing biological information	6
3.1 Example entries of gene and protein name information in Swiss-Prot	22
3.2 Data sources used for gene and protein name dictionary generation	27
3.3 Token classes used for curation of gene and protein name dictionaries	29
3.4 Inter-species ambiguity	36
3.5 Relevance of inter-dictionary ambiguities for mining MEDLINE	41
3.6 Results of evaluation of hierarchical gene and protein name dictionary	44
3.7 Analysis of false negatives of gene and protein name dictionary	45
3.8 Analysis of false positives of gene and protein name dictionary	45
4.1 Definition of token classes in ProMiner	61
4.2 Examples for alternative concepts	68
4.3 ProMiner parameter settings applied for BioCreAtIvE I (2004)	71
4.4 Results in BioCreAtIvE I (2004) and II (2006) gene normalization tasks	73
4.5 Types of errors and examples of false positive matches	76
4.6 Examples of false negative matches	77
4.7 Effect of rule-based postfilter: Examples	78
5.1 Expressions used for effector-effectee detection	96
5.2 Evaluation results of RelEx	99
5.3 Characteristics of human literature-derived networks	102
5.4 Results of RelEx large-scale application and comparison against HPRD	104
5.5 Characteristics of Y2H and literature-curated networks	106
5.6 Characteristics of the hull of MMP13 in large-scale text-mining networks	110
5.7 Results of applying network schemes for context analysis	113
5.8 Annotation schema for gene/protein interactions	115
5.9 Attributes for interaction classification	115
5.10 Evaluation of feature sources for interaction characterization	116
6.1 Marker genes for cartilage degradation	131
7.1 Results of StabPerf applied on osteoarthritis data	154

Summary

Microarrays can capture gene expression activity for thousands of genes simultaneously and thus make it possible to analyze cell physiology and disease processes on molecular level. The interpretation of microarray gene expression experiments profits from knowledge on the analyzed genes and proteins and the biochemical networks in which they play a role. The trend is towards the development of data analysis methods that integrate diverse data types. Currently, the most comprehensive biomedical knowledge source is a large repository of free text articles. Text mining makes it possible to automatically extract and use information from texts.

This thesis addresses two key aspects, biomedical text mining and gene expression data analysis, with the focus on providing high-quality methods and data that contribute to the development of integrated analysis approaches.

The work is structured in three parts. Each part begins by providing the relevant background, and each chapter describes the developed methods as well as applications and results.

Part I deals with biomedical text mining:

Chapter 2 summarizes the relevant background of text mining; it describes text mining fundamentals, important text mining tasks, applications and particularities of text mining in the biomedical domain, and evaluation issues.

In Chapter 3, a method for generating high-quality gene and protein name dictionaries is described. The analysis of the generated dictionaries revealed important properties of individual nomenclatures and the used databases (Fundel and Zimmer, 2006). The dictionaries are publicly available via a Wiki, a web service, and several client applications (Szugat *et al.*, 2005).

In Chapter 4, methods for the dictionary-based recognition of gene and protein names in texts and their mapping onto unique database identifiers are described. These methods make it possible to extract information from texts and to integrate text-derived information with data from other sources. Three named entity identification systems have been set up, two of them building upon the previously existing tool ProMiner (Hanisch *et al.*, 2003). All of them have shown very good performance in the BioCreAtIvE challenges (Fundel *et al.*, 2005a; Hanisch *et al.*, 2005; Fundel and Zimmer, 2007).

In Chapter 5, a new method for relation extraction (Fundel *et al.*, 2007) is presented. It was applied on the largest collection of biomedical literature abstracts, and thus a comprehen-

hensive network of human gene and protein relations has been generated. A classification approach (Küffner *et al.*, 2006) can be used to specify relation types further; e.g., as activating, direct physical, or gene regulatory relation.

Part II deals with gene expression data analysis:

Gene expression data needs to be processed so that differentially expressed genes can be identified. Gene expression data processing consists of several sequential steps. Two important steps are normalization, which aims at removing systematic variances between measurements, and quantification of differential expression by p-value and fold change determination. Numerous methods exist for these tasks.

Chapter 6 describes the relevant background of gene expression data analysis; it presents the biological and technical principles of microarrays and gives an overview of the most relevant data processing steps. Finally, it provides a short introduction to osteoarthritis, which is in the focus of the analyzed gene expression data sets.

In Chapter 7, quality criteria for the selection of normalization methods are described, and a method for the identification of differentially expressed genes is proposed, which is appropriate for data with large intensity variances between spots representing the same gene (Fundel *et al.*, 2005b). Furthermore, a system is described that selects an appropriate combination of feature selection method and classifier, and thus identifies genes which lead to good classification results and show consistent behavior in different sample subgroups (Davis *et al.*, 2006).

The analysis of several gene expression data sets dealing with osteoarthritis is described in Chapter 8. This chapter contains the biomedical analysis of relevant disease processes and distinct disease stages (Aigner *et al.*, 2006a), and a comparison of various microarray platforms and osteoarthritis models.

Part III deals with integrated approaches and thus provides the connection between parts I and II:

Chapter 9 gives an overview of different types of integrated data analysis approaches, with a focus on approaches that integrate gene expression data with manually compiled data, large-scale networks, or text mining.

In Chapter 10, a method for the identification of genes which are consistently regulated and have a coherent literature background (Küffner *et al.*, 2005) is described. This method indicates how gene and protein name identification and gene expression data can be integrated to return clusters which contain genes that are relevant for the respective experiment together with literature information that supports interpretation.

Finally, in Chapter 11 ideas on how the described methods can contribute to current research and possible future directions are presented.

Zusammenfassung

Mit Microarrays kann die Genexpressionsaktivität vieler tausender Gene gleichzeitig erfasst werden; dies ermöglicht die Analyse von Zellphysiologie und Krankheitsprozessen auf molekularer Ebene. Für die Interpretation von Genexpressionsdaten ist Fachwissen über die untersuchten Gene und Proteine und die biochemischen Netzwerke, in denen diese eine Rolle spielen, von Nutzen. Die Entwicklung geht in Richtung von Analysemethoden, die verschiedene Datentypen integrieren. Die gegenwärtig umfassendste biomedizinische Informationsquelle ist eine große Sammlung von Freitext Artikeln. Text Mining ermöglicht es, Information aus Texten automatisch zu extrahieren und zu verwenden.

Die vorliegende Arbeit geht zwei Schwerpunkte an, biomedizinisches Text Mining und Genexpressionsdatenanalyse. Das Hauptaugenmerk liegt dabei auf der Bereitstellung qualitativ hochwertiger Methoden und Daten, die zur Entwicklung integrierter Analyseverfahren beitragen.

Die Arbeit ist in drei Teile gegliedert. Jeder Teil beginnt mit einer Beschreibung des relevanten Hintergrundwissens, und jedes Kapitel beschreibt die entwickelten Methoden sowie Anwendungen und Ergebnisse.

Teil I behandelt biomedizinisches Text Mining:

Kapitel 2 fasst den nötigen Hintergrund zu Text Mining zusammen. Es werden allgemeine Text Mining Grundlagen, wichtige Text Mining Aufgaben, Anwendungen und Besonderheiten von Text Mining im biomedizinischen Bereich, und Themen hinsichtlich der Evaluierung von Text Mining Methoden beschrieben.

In Kapitel 3 wird eine Methode zur Erstellung qualitativ hochwertiger Gen- und Protein-Wörterbücher vorgestellt. Die Analyse der erzeugten Wörterbücher zeigt wichtige Eigenschaften der verschiedenen Nomenklaturen und der verwendeten Datenbanken ([Fundel and Zimmer, 2006](#)). Die Wörterbücher sind über ein Wiki, einen Web-Service und verschiedene Anwendungsprogramme allgemein verfügbar ([Szugat et al., 2005](#)).

In Kapitel 4 werden Methoden zur wörterbuchbasierten Erkennung von Gen- und Proteinnamen und deren Abbildung auf eindeutige Datenbank-Bezeichner beschrieben. Diese Methoden ermöglichen die Informationsextraktion aus Texten und das Zusammenführen der so gewonnenen Informationen mit Daten aus anderen Quellen. Es wurden drei Systeme zur Identifikation benannter Objekte in Texten entwickelt, von denen zwei auf dem bereits existierenden Programm ProMiner ([Hanisch et al., 2003](#)) aufbauen. Alle erzielten sehr gute Ergebnisse in den BioCreAtIvE Evaluierungen ([Fundel et al., 2005a](#); [Hanisch et al., 2005](#);

[Fundel and Zimmer, 2007](#)).

Kapitel 5 führt eine neue Methode zur Relationsextraktion ein ([Fundel et al., 2007](#)). Diese wurde auf einen umfangreichen Satz biomedizinischer Literatur-Abstracts angewendet und dadurch wurde ein umfassendes Netzwerk humaner Gen- und Proteinrelationen erstellt. Ein Klassifikationsansatz ([Küffner et al., 2006](#)) erlaubt die nähere Spezifikation von Relationen, z. B. als aktivierende, direkt physikalische, oder genregulatorische Relationen.

Teil II behandelt die Analyse von Genexpressionsdaten:

Genexpressionsdaten müssen prozessiert werden damit differentiell exprimierte Gene identifiziert werden können. Die Prozessierung von Genexpressionsdaten besteht aus mehreren nacheinander ausgeführten Schritten. Zwei wichtige Schritte sind die Normalisierung, die darauf abzielt, systematische Unterschiede zwischen Messungen zu beseitigen, und die Quantifizierung differentieller Expression durch die Bestimmung von p-value und Expressionsunterschied (fold change). Es stehen zahlreiche Methoden für diese Arbeitsschritte zur Verfügung.

Kapitel 6 beschreibt den nötigen Hintergrund zur Genexpressionsdatenanalyse und gibt einen Überblick über die biologischen und technischen Prinzipien von Microarrays sowie die wichtigsten Schritte der Datenprozessierung. Es folgt eine kurze Einführung in Osteoarrose; diese steht im Mittelpunkt der analysierten Genexpressionsdatensätze.

In Kapitel 7 werden Qualitätskriterien für die Auswahl von Normalisierungsmethoden vorgeschlagen und eine Methode zur Identifikation von differentiell exprimierten Genen wird vorgestellt, die für Daten mit großen Intensitätsunterschieden zwischen spots, die ein Gen repräsentieren, geeignet ist ([Fundel et al., 2005b](#)). Zudem wird ein System diskutiert, das eine geeignete Kombination aus Feature-Auswahl Methode und Klassifikator wählt und so Gene identifiziert, die gute Klassifikationsergebnisse erzielen und gleichzeitig ein einheitliches Verhalten in verschiedenen Teilmengen der Proben zeigen ([Davis et al., 2006](#)). Die Analyse verschiedener Genexpressionsdatensätze aus dem Bereich Osteoarrose wird in Kapitel 8 beschrieben. Dieses Kapitel beinhaltet die biomedizinische Analyse relevanter Krankheitsprozesse und unterschiedlicher Krankheitsstadien ([Aigner et al., 2006a](#)) sowie einen Vergleich verschiedener Microarray-Plattformen und Osteoarrose-Modelle.

Teil III behandelt integrierte Ansätze und bildet so die Verbindung zwischen den Teilen I und II:

Kapitel 9 gibt einen Überblick über verschiedene Arten integrierter Analysemethoden; der Schwerpunkt liegt dabei auf Ansätzen, die Genexpressionsdaten mit manuell generierten Daten, großen Netzwerken oder Text Mining zusammenführen.

In Kapitel 10 wird eine Methode zur Identifikation von Genen betrachtet, die einheitlich reguliert sind und einen kohärenten Literaturhintergrund haben ([Küffner et al., 2005](#)). Diese Methode ist ein Beispiel dafür, wie die Identifikation von Gen- und Proteinnamen in Texten zur automatischen Analyse experimenteller Daten beitragen kann.

Abschließend werden in Kapitel 11 Ideen aufgezeigt, wie die beschriebenen Methoden zur aktuellen Forschung beitragen können und es werden Zukunftsperspektiven diskutiert.

Chapter 1

Introduction

New experimental techniques and increased automatization allow researchers to collect detailed and comprehensive biological data in short time. For example, microarrays are routinely used for monitoring gene expression levels of all genes of a genome simultaneously. The resulting data needs to be processed, analyzed, and interpreted to produce useful scientific knowledge. Computer programs support large-scale data processing and analysis. The interpretation of biological data requires extensive background knowledge on the investigated systems. Scientific literature represent an extremely important source for biological information; a comprehensive collection of publications provides public access to free text articles.

The size of the experimental data sets to be analyzed and the amount of available literature suggests automatic means for data analysis. One trend in current research is the development of integrated data analysis methods that exploit various data sources and thus can derive new interpretations, insights, or hypotheses. Automatic exploitation of biomedical publications is a challenging task as the scientific language is characterized by numerous technical terms, multi-word terms, abbreviations, and a high level of ambiguity.

Numerous experimental techniques focus on the analysis of genes and proteins as these are crucial for the biochemical machinery in a cell. Extraction of information on genes and proteins is thus of primary importance for understanding biochemical processes.

The development of the methods presented here started in the context of a large research project (“Leitprojekt Diagnose und Therapie der Osteoarthrose”) funded by the Federal Ministry of Education and Research with the goal of elucidating the pathomechanisms involved in osteoarthritis, a multifactorial degenerative joint disease. In the project, several gene expression data sets have been generated. One of the aims of the bioinformatics part of the project was to develop integrated data analysis methods that make use of experimental data and existing knowledge to generate hypotheses on disease causes and options for treatment, and to uncover new aspects of the disease mechanisms. Generally, integrated data analysis approaches implement multi-step procedures and strongly depend on the performance of the individual underlying methods.

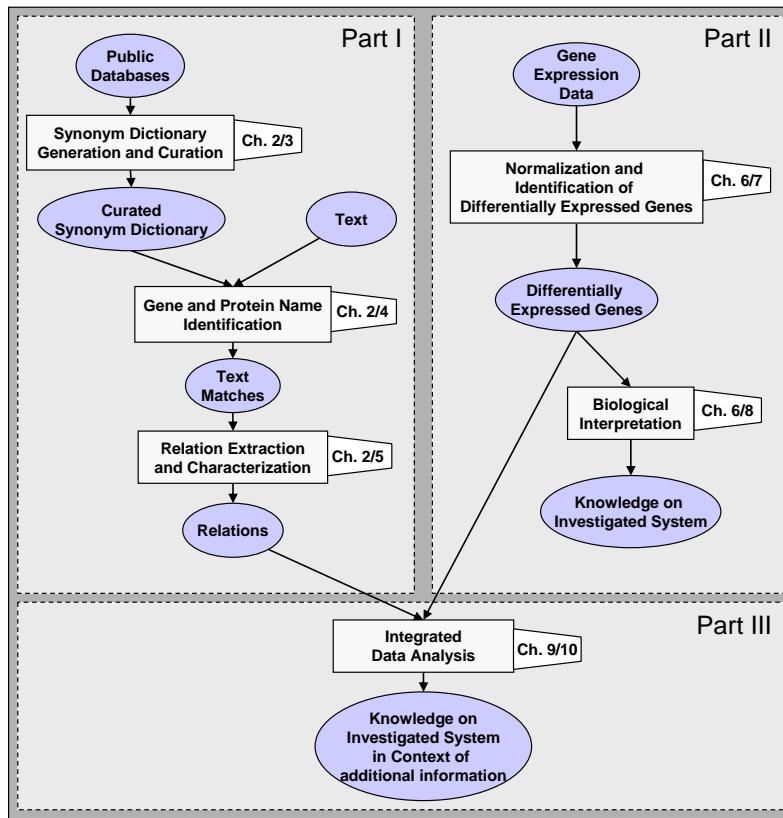


Figure 1.1: Overview of the thesis structure: Part I explores biomedical text mining, Part II gene expression data analysis, and Part III brings together the two previous parts and describes an integrated data analysis method. Chapter 1 contains the introduction, Chapter 11 the conclusions, and Chapters 2, 6, 9 provide the relevant background for the parts I, II, and III, respectively. The other chapters describe the developed methods and present applications and results.

In this work, two important aspects that form the basis of many integrated data analysis methods, namely text mining and gene expression data analysis, are addressed and an example of an integrated data analysis method is presented. The structure of this thesis (Figure 1.1) reflects these core topics:

Part I is centered around text mining, which is the main topic of this work. It presents the developed methods that exploit free texts to generate data that is suitable for integrated data analysis methods.

- *Generation of dictionaries for biomedical named entity identification.* In Chapter 3, a method for the generation of high-quality gene and protein name dictionaries is described. The analysis of several dictionaries pinpoints characteristics of the respective organism nomenclatures.
- *Gene and protein name identification* is a fundamental step for integrating text data with data derived from other sources. In Chapter 4, three dictionary-based systems

for gene and protein name identification are described. Their independent evaluation in the BioCreAtIvE challenges demonstrated very good performance.

- *Relation extraction* is important for the generation of networks from texts. Networks represent a model of dependencies between proteins which are useful for integrated analysis with other types of data (e.g. Hanisch *et al.* (2002); Sohler *et al.* (2004); Sohler and Zimmer (2005); Herrgard *et al.* (2006); Tan *et al.* (2007)). In Chapter 5, a method is described that extracts relations from free texts, and results of its application on the largest repository of biomedical publication abstracts are presented. Furthermore, an approach for relation characterization is described.

Part II focuses on gene expression data analysis, the second important topic of this work. Many integrated data analysis methods start from a set of differentially expressed genes.

- *Identification of differentially expressed genes.* In Chapter 7, an analysis of the effects of primary data processing on the identification of differentially expressed genes is described. Quality criteria for the selection of normalization methods are presented, and a method for the identification of differentially expressed genes is proposed which is appropriate for data with large intensity variances between spots representing the same gene.
- *Biomedical interpretation.* Chapter 8 describes the biomedical analysis of one gene expression data set dealing with osteoarthritis and the comparison of different microarray platforms and osteoarthritis models.

Part III presents an example for an integrated data analysis approach that is based on the methods and results presented in parts I and II. This part thus establishes the connection between the two previous parts.

- *Integrated data analysis.* In Chapter 10, a method is described that integrates text data with gene expression data and thus identifies genes which share a coherent literature background and are significantly regulated. This method builds on the above approach for gene and protein name identification and requires genes to be assigned with a value that describes the level of differential expression.

Finally, Chapter 11 discusses the accomplishments of this work and presents possible future extensions and directions.

In the following, a short overview of general aspects of bioinformatics relevant to this thesis is given. The following chapters reflect the structure summarized above. Figure 1.1 provides an overview of the work and shows how the individual chapters are linked to each other. Each part starts by providing the relevant background. Each chapter describes the used and developed methods as well as applications and results.

General Trends in Bioinformatics

The last years of biomedical, biotechnological and bioinformatics research have been characterized by several trends described by *omics*-terms (for a review see e.g. [Joyce and Palsson \(2006\)](#)). These trends focused on the study of specific types of data or biochemical entities. *Genomics* arose by the possibility of sequencing entire genomes. Genomics studies are generally based on the hypothesis that many, if not all, biological phenomena can be explained by genomic sequences. Thus, predisposition for diseases, responsiveness to therapies and drugs, as well as phenotypes in general are assumed to be caused by genetic sequence and respective variants alone. *Transcriptomics* deals with the analysis of the transcriptome; that is, the set of transcribed genes. Gene transcription reflects gene activity, which changes in response to external stimuli, developmental stage, diseases, etc. Expression levels of thousands of genes can routinely be quantified by microarray measurements. *Proteomics* then focused on the idea that not only the genome, but the entire set of proteins of an organism is relevant for describing and analyzing its properties and function. This view approaches the holistic view of biological processes, but it is currently experimentally much more difficult to analyze as the experimental techniques for analyzing whole proteomes are not yet as well established as the techniques for analyzing whole genomes. Protein arrays are being developed, but in contrast to nucleotide arrays, they are not yet state of the art. Mass spectrometry is an alternative for analyzing whole proteomes, but this technique is far more laborious and expensive than microarrays and therefore can not be used as easily at large scale. *Metabolomics* then enlarged the scope further, stating that besides the genome and the entire set of proteins all other kinds of small molecules and metabolites need to be considered for understanding and explaining the entire organism. This approach seems promising, yet is experimentally even more difficult to accomplish. With mass spectrometry, numerous metabolites can be identified quantitatively, but due to the significant costs and labor-intensity rather few such data is available.

Pharmacogenomics deals with investigating relations between genomic sequences and pharmaceuticals. These relations are important to know when aiming at personalized medication. An example is the analysis of single nucleotide polymorphisms (SNPs) for predicting whether a medication will show the desired effect.

Systems biology aims at describing biological phenomena as complete systems; that is, as quantitative models of genes, proteins, and metabolites. Finally, all modeled processes shall be integrated in a system model. The system to be described can be a cell, a cellular compartment, an organ or tissue, an entire organism, etc. Such a model is intended to describe the systems behavior in response to an effector such as a mutation or an external stimulus.

Biological Data and Data Resources

The initial data obtained by researchers doing laboratory experiments is usually not available to the public. When researchers decide to publish their results, this is generally in processed form and added with supplementary information. Negative results are rarely

published. The way data is presented and made available as well as the amount of detail of provided additional information depends on the researcher and the publisher. Some results are presented as images, due to the fact that a number of techniques such as immunoblotting, histologic staining, protein migration studies with fluorescence markers directly provide that kind of result. A large part of biomedical knowledge is available as unstructured natural language text, mainly as research papers in scientific journals.

During the last years, an increasing amount of data has been organized in databases which ensure that different data sets are organized in a uniform way. Thus, queries can be formulated to extract subsets of data and automated methods can be used for data analysis. Many databases are made publicly available via the Internet and can be queried online or downloaded as local copy. The Database issue of the journal *Nucleic Acids Research* appears yearly and gives an overview on most relevant databases (Galperin, 2007). Biological databases can be classified according to the contained data. Primary databases contain experimental results as submitted by laboratory researchers (e.g. gene sequences). Secondary databases additionally contain annotations of the primary data (e.g. functional annotations). Tertiary databases integrate different kinds of data, such as gene and protein sequences added with annotations on function, localization, etc. Databases generally have a certain focus (for examples see Table 1.1). The content of most databases is compiled by manual literature curation; that is, scientists read the literature and enter pieces of information into the database.

MEDLINE is the prevailing online source for abstracts of biomedical research publications. It is maintained by the National Library of Medicine (NLM) and the National Institute of Health (NIH) and hosted at the National Center of Biotechnology Information (NCBI). MEDLINE contains over 16 million abstracts and additional information (February 2007). It is freely accessible and can be searched through the PubMed interface¹ by boolean queries for PubMed Identifier (pmid), words in the title and abstract (e.g. gene names), author, journal, MeSH-terms, etc. Data can also be downloaded via ftp; an XML-format allows straightforward extraction of certain sections of the individual entries.

Most biomedical research articles are available as full text via the Internet. In the last years, articles are increasingly offered for free, either immediately after acceptance (e.g. BMC Bioinformatics by BioMedCentral), or after a certain delay (e.g. Bioinformatics by Oxford Journals). At the NCBI, PubMedCentral integrates links to free full text articles for the abstracts contained in MEDLINE and thus provides seamless information access and high visibility. The full text articles are generally provided in HTML or pdf format; the styles differ between journals and thus automatic analysis of full text articles requires significantly higher parsing effort than analysis of abstracts derived from MEDLINE.

Manually compiled information is only of high value if the community has a common understanding of the used terms. Furthermore, given the increasing amount of information contained in databases, it is essential to interpret data computationally.

Controlled vocabularies specify terms and their meaning and thus provide a means to define

¹<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>

Database Focus	Examples
General databases on genes and proteins	Swiss-Prot (Bairoch et al., 2005) Entrez Gene (formerly LocusLink, Maglott et al. (2005)) GeneCards (Rebhan et al., 1998)
Organism specific databases on genes and proteins	Human gene nomenclature database HUGO, Genew (Povey et al., 2001; Eyre et al., 2006) Mouse Genome Informatics (MGI) (Blake et al., 2003; Eppig et al., 2005) Rat Genome Database (RGD, de la Cruz et al. (2005)) FlyBase (Drysdale et al., 2005) Saccharomyces Genome Database (SGD, Balakrishnan et al. (2005))
Synonyms	GPSD (Pillet et al., 2005)
Abbreviations	Stanford Biomedical Abbreviation Server (Chang et al., 2002) Acronym Resolving General Heuristics (Wren and Garner, 2002)
Molecular interactions	MINT (Zanzoni et al., 2002) BIND (Gilbert, 2005; Alfarano et al., 2005)
Protein-protein interactions	DIP (Database of Interacting Proteins) (Xenarios et al., 2001; Salwinski et al., 2004)
Regulatory pathways	TRANSPATH (Schacherer et al., 2001; Krull et al., 2006) KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000; Kanehisa et al., 2002)
Metabolic pathways	MetaCyc (Karp et al., 2000; Caspi et al., 2006)
Diseases	Online Mendelian Inheritance in Men (OMIM, Hamosh et al. (2005))
Protein structures	Protein DataBase (PDB, Berman et al. (2007)) SCOP (structural classification of proteins, Andreeva et al. (2004)) CATH (Greene et al., 2007)

Table 1.1: Examples of public databases containing biological information.

a *common language*. Controlled vocabularies contain a set of precisely defined vocabulary terms, each representing a *concept* in the domain and assigned with a detailed definition or description. Application of controlled vocabularies makes it possible to share information and easily analyze data computationally.

An *Ontology* is a controlled vocabulary for which each concept is assigned to a node in a directed acyclic graph (DAG). Thus, an ontology defines relationships between concepts: Each term may be a *child* of one or more *parents*. An ontology ([Gruber, 1993](#)) has two pragmatic purposes: To facilitate communication between people and organizations and to improve interoperability between systems.

Examples of biomedical controlled vocabularies and ontologies include the Medical Subject Headings, the Unified Medical Language System, and Gene Ontology:

The Medical Subject Headings (MeSH)² has been developed at the National Library of Medicine (NLM) for indexing articles for MEDLINE/PubMed. MeSH contains medical terms such as diseases, diagnostic techniques, cellular compartments, or cell types. A large

²<http://www.nlm.nih.gov/mesh/>

number of PubMed abstracts has been manually annotated with MeSH terms which indicate the main themes of the corresponding article.

The Unified Medical Language System (UMLS, [Bodenreider \(2004\)](#)) currently represents the largest thesaurus in the biomedical domain. It contains a metathesaurus that integrates information from numerous biomedical thesauri, and the SPECIALIST lexicon, which contains information that is useful for term variant generation. The metathesaurus is structured in a uniform way which facilitates data extraction. The concepts are categorized by semantic types and linked to each other by hierarchical and non-hierarchical relationships (parent, child and sibling relations). MetaMap ([Aronson, 2001](#)) is a program to discover metathesaurus concepts referred to in texts; it is one of the foundations of NLM's Indexing Initiative System which is being applied to both semi-automatic and fully automatic indexing of the biomedical literature.

Gene Ontology (GO, [Ashburner and Lewis \(2002\)](#)) provides a controlled vocabulary and ontologies for the description of eucaryotic genes and gene-products. GO is maintained by the Gene Ontology Consortium which is a joint project of three model organism database committees: FlyBase ([Drysdale et al., 2005](#)), Mouse Genome Informatics ([Blake et al., 2003](#); [Eppig et al., 2005](#)), and the Saccharomyces Genome Database ([Balakrishnan et al., 2005](#)). GO is structured in three categories: *Biological process* refers to a biological objective to which the gene or gene product contributes, *cellular compartment* describes the location in the cell where a gene product is active, and *molecular function* defines the biochemical activity of a gene product. The combination of several GO-terms from the three categories provides a precise description of a gene product. Within the three categories, the terms of the controlled vocabulary are linked to each other by relationships of the types “*is a*” or “*part of*”. The hierarchical organization makes it possible to systematically narrow or widen a query and thus to adapt the granularity of a search. Many of the organizations that maintain organism-specific databases make use of Gene Ontology for annotating newly discovered or characterized genes and proteins. Furthermore, general annotation approaches such as GOA ([Camon et al., 2004](#)) attempt to annotate genes for multiple organisms in a consistent manner and integrate the resulting data in a publicly available database.

The ontologies described above have been used for a wide variety of biomedical applications. Gene Ontology is commonly used for the interpretation of gene expression data via overrepresentation analysis (see also Section 9.1), but has also been used for determining the semantic similarity between gene products ([Lord et al., 2003b,a](#)). By exploring the relationships between UMLS and GO, the Genestrace system ([Cantor et al., 2005](#)) could infer relationships between disease concepts and gene products.

Part I

Text Mining

Chapter 2

Background: Text Mining

This chapter contains a short introduction to text mining. First, frequently used notions and fundamental tasks of text mining are explained (Sections 2.1 and 2.2). Then, specific applications and issues of text mining in the biomedical domain are discussed (Section 2.3). Finally, aspects concerning the evaluation of text mining approaches are summarized (Section 2.4).

2.1 Fundamentals in Text Mining

Text Mining describes the automated process of analyzing natural language text with the goal of discovering information and knowledge. A number of terms describe specific aspects of automatic text analysis:

This work deals with *Information Extraction (IE)*, *Named Entity Recognition (NER)*, and *Named Entity Identification (NEI)* approaches; these concepts are therefore described in more detail in separate sections below.

Natural Language Processing (NLP) deals with all aspects of automatically processing written and spoken language. NLP often refers to approaches that focus on grammatical and syntactic analysis of texts. *Question Answering* returns a phrase, sentence, or summary together with pointers to the underlying documents in response to a question. *Information Retrieval (IR)* tackles the task of finding documents that contain a specific information in a large set of documents. The desired information is usually expressed by the user as a *query*. Different types of queries can be used. A boolean query is a boolean combination of terms (e.g. key words). A similarity query is usually based on the vector-space model, in which query and texts are represented as vectors and texts are selected based on the vector similarity. More flexible approaches that depend less on the explicit query terms are *Latent Semantic Indexing* and *Probabilistic Models*. The most familiar IR example in the biomedical domain is PubMed, which retrieves biomedical abstracts from the MEDLINE database.

Text Categorization is centered around labeling texts with thematic categories from a predefined set of category-tags. *Text Classification* is applied to group documents according

to their content. *Indexing* is the process of determining a set of terms or words for a document that should be used when matching that document to a query. For indexing, it is important to distinguish *content words* (i. e. words that express specific semantic concepts) from *function words* (i. e. words that contain only grammatical information).

Linguistic Notions

Morphology is the knowledge of how words are formed; it is required to understand complex words such as *dephosphorylation*. *Syntax* describes the rules for correct combination of phrases. So far, there is still no comprehensive analysis of the English syntax available. Knowledge of syntax is needed to understand why sentences like “A inhibits activation of B and C” are ambiguous: Here, “activation of” might refer either to B only, or to B and C. *Semantic* describes the meaning of words, expressions or phrases. Numerous words have multiple meanings; for example, *a plane* can be used for an airplane, for a landscape, or in the mathematical sense.

Basic Text Processing

Typical text mining systems commonly apply the text processing operations tokenization, part of speech tagging, and (shallow) parsing.

Tokenization is the process of breaking the text up into its constituents – the *tokens*. Different levels of tokenization include segmenting texts into sections, such as sentences, words, or syllables. Most applications in bioinformatics require fragmentation of texts into words and sentences. In specialized language word boundaries can be debatable; e. g., subtype specifier of protein names can be considered as constituents of the corresponding name or as independent and therefore separate words. Sentence splitting can also pose problems as the common rule of a full stop followed by a space and the following sentence beginning with an upper case letter does not necessarily apply for biomedical texts. Nomenclature guidelines sometimes require genes or mutants to be spelled in lower case letters, even at the beginning of a new sentence. For example, names/symbols in upper case often design the dominant form, while lower case names describe the recessive alleles like in the sentence “*bif* displays strong genetic interaction with *msn*.” which begins with a mention of the recessive form of the Bifocal gene. Other difficulties result from abbreviations.

Part-of-Speech Tagging annotates words with word-categories in form of part-of-speech (POS) tags, which reflect the syntactic role or grammatical class of the corresponding word. The sets of tags varies between tools, the most common set contains Noun, Verb, Article, Adjective, Preposition, and Number. A tag is generally assigned to a word based on its context in the sentence. Most taggers are based on rules or Hidden Markov Models (HMMs); they are language-specific and generally achieve high accuracy. *Stemming* is used to determine the stem of a word; that is, the main part of the word truncated for the parts indicating word form, plurality, tense, etc. For example, *phosphorylat* is the stem of the words *phosphorylation*, *phosphorylates*, *phosphorylated*.

Parsing determines the complete syntactic structure of a sentence. Given a sentence, a

syntactic parser produces a syntax tree in which the leaves correspond to individual words and the internal nodes describe syntactic structures, such as noun phrase, verb phrase, or prepositional phrase (for more details on parsing see also Section 5.2.1). *Shallow Parsing* is a coarser process of breaking sentences in phrases. A *phrase* groups syntactically related words together and is annotated with a tag (e. g. Noun Phrase, Adjective Phrase, Conjunction Phrase). Shallow Parsing is generally faster and more robust than full parsing and is therefore frequently used as preprocessing step.

2.2 Text Mining Tasks

Information Extraction

Information Extraction (IE) is concerned with extracting pertinent information from large volumes of text according to the user's needs. The extracted information is provided with pointers back to the literature from which it has been derived. Often, the task is to find entities, attributes, facts, relations, or events in unstructured texts. IE often combines NLP, lexical resources, and semantic constraints.

IE has been extensively applied and proven successful in the newswire domain. For example, results from various evaluations show that information extraction systems can identify and classify names of persons, organizations, and locations at accuracies exceeding 90%, and binary relations among these entities at over 75% accuracy (Hirschman *et al.*, 2002b). Systematic common evaluations provide neutrally selected benchmark data sets, which alleviates rapid methodology and system improvements. Prominent evaluation events are the *Message Understanding Conferences (MUC)* (for an overview see Hirschman (1998)), *Knowledge Discovery in Databases (KDD)* and *Text REtrieval Conferences (TREC)*, which are annually organized with varying challenge tasks.

Information extraction bares numerous issues for common English texts and even more for biomedical texts. General issues are: resolution of *Coordination* (i. e. linking of structures by coordinating words such as *and* and *or*), *Anaphora* (i. e. references to previously mentioned entities by words such as *it*, *they*), and *Negation resolution*. Frequently, these issues, especially anaphora resolution and negation resolution, are ignored by approaches for biomedical information extraction.

Named Entity Recognition

Named Entity Recognition (NER), a subdiscipline of Information Extraction, is centered around recognizing specific entities, events, or facts in texts and extracting the relevant text fragments. Basically, all techniques proposed for recognizing named entities use some form of character-by-character or word-by-word pattern to identify entities. The patterns can be designed by hand or automatically learned from examples previously annotated by hand. Text is then scanned for exact or close matches to the predefined patterns, or evaluated with respect to a statistical model.

In the biomedical domain, named entity recognition is also frequently referred to as *tagging*. Gene and protein names can be recognized by hand-crafted detection rules which reflect known regularities and naming conventions of the respective entities; for example, “*a ‘p’ followed by a number*” (Fukuda *et al.*, 1998; Seki and Mostafa, 2003). The YAPEX system (Franzen *et al.*, 2002; Eriksson *et al.*, 2002) uses hand-written rules and information from an off-the-shelf syntactic parser; thus, it makes use of combined lexical and syntactic information, heuristic filters and a local dictionary. *Context based* approaches require a dictionary of sentence contexts that suggest entity names. For example, a text fragment matching the pattern “*<x> phosphorylates <y>*” indicates that *<x>* and *<y>* are candidate protein names. Proux *et al.* (1998) applied a series of sieves of lexical, morphological, and semantic analysis to extract from a sentence those words that are potential gene symbols or names.

Learning-based methods apply general machine learning methods (e.g. Hidden Markov Models) for learning patterns of characters or words, or they train part-of-speech taggers or shallow parsers for the recognition of specific entities. These methods generally do not need predefined dictionaries or rule sets, but they require annotated training data.

Involved systems combine linguistic and statistical information (e.g. Frantzi *et al.* (1998)), make use of statistical methods, decision trees, and shallow parsing for term candidate identification and classification (Nobata *et al.*, 1999), apply lexical rules based on part of speech tags (Tanabe and Wilbur, 2002), statistical models of gene names (Chang *et al.*, 2004), combinations of various machine learning methods (Zhou *et al.*, 2004), support vector machines (e.g. Kazama *et al.* (2002); Takeuchi and Collier (2003); Hakenberg *et al.* (2005)), markov models (Wren *et al.*, 2005b), or conditional random fields (Settles, 2005). Named entity recognition is generally evaluated by the left and right boundary of the entity description, by one of these criteria, or by the overlap of a detected entity with the correct text fragment (for an overview see Tsai *et al.* (2006)). The BioCreAtIVe gene mention finding evaluation (Task 1A, Yeh *et al.* (2005)) and Coling BioNLP (JNLPBA, Kim *et al.* (2004)) challenges evaluated biomedical named entity recognition and showed widely differing results for systems participating in both challenges (Dingare *et al.*, 2005; Tsai *et al.*, 2006) which is, at least partly, due to differing annotation schemes.

Named Entity Identification

Named Entity Identification (NEI) focuses on the detection and identification of entities in texts. Here, in extension to named entity recognition, the identity of a recognized entity has to be determined; that is, each text fragment needs to be mapped to a unique identifier that represents the respective concept or entity.

Lexicon-based approaches for NEI make use of large gene or protein name dictionaries, or likely components of the entity names, and perform an exact or approximate text look-up. The performance of these approaches largely depends on the quality of the dictionary. For more details and a literature review on named entity identification see Section 4.1.

Alternatively, named entity identification can be conceived as a three step process consisting of (1) term recognition, which has been described in the previous section, (2) term

classification or term categorization (i.e. the assignment of terms to broad biomedical classes, such as genes, proteins, or mRNA), and (3) term mapping, which links terms to identifiers. For a review on methods for each of the three steps see [Krauthammer and Nenadic \(2004\)](#). These steps can be merged; for example, dictionary-based term recognition directly provides links to the respective identifiers. Finally, some methods combine manual or dictionary-based and learning-based approaches.

2.3 Text Mining in Bioinformatics

The technical means for conducting biological large-scale experiments involving thousands of genes and proteins are established and such experiments are routinely performed. Their interpretation remains an important problem. For example, gene expression can be measured for large gene sets, and subsets of genes with correlated expression patterns can be extracted by statistic analysis and clustering of measured data. Yet, similar expression patterns do not necessarily imply involvement in the same biological process, and functional relationships cannot be determined from cluster data alone.

Published literature contains information that can be used for a more detailed analysis. Due to the amount of data to be analyzed it becomes tedious or even impossible to read and analyze all published literature dealing with all genes returned from crude data analysis of such experiments. Here, Text Mining provides help by automatically preselecting documents and analyzing the contained information. For reviews on text mining in the biomedical domain see [Shatkay and Feldman \(2003\)](#); [Jensen et al. \(2006\)](#)

Major bioinformatics conferences such as *Intelligent Systems for Molecular Biology (ISMB)* and *Pacific Symposium on Biocomputing (PSB)* responded to the growing interest in text mining in bioinformatics by publishing papers on the topic since the early 1990s and by devoting entire sessions to the field since the late 1990s. The NLP and bioinformatics domains both dispose of sound scientific achievements and the effort of growing together is certainly useful for both. As this integration is an ongoing process, and the underlying domains also continue to evolve, significant scientific progress is expected in the upcoming years.

Examples for specific literature mining tasks in bioinformatics are:

- Extraction of keywords and functional annotation of proteins ([Andrade and Valencia, 1998](#))
- Generating gene summaries (e.g. sequence information, phenotypes, interactions) ([Ling et al., 2006](#))
- Predicting the subcellular localization of proteins ([Stapley et al., 2002](#)), in conjunction with protein sequence based features ([Hoglund et al., 2006](#))
- Annotation of enzyme classes with disease-related information ([Hofmann and Schomburg, 2005](#))
- Finding protein-protein interactions (see Chapter 5)

- Assisting BLAST searches ([Chang and Lin, 2001](#))
- Detection of remote homologs by combination of PSI-BLAST and analysis of Swiss-Prot annotations ([MacCallum *et al.*, 2000](#))
- Discovering protein similarity ([Sarkar and Rindflesch, 2002](#))
- Assisting microarray data interpretation ([Masys, 2001; Masys *et al.*, 2001](#))
- Pathway discovery ([Blaschke *et al.*, 1999; Krauthammer *et al.*, 2002](#))
- Nucleic acid and peptide sequence identification in texts ([Wren *et al.*, 2005b](#))
- Discovery of themes or gene groups with similar functionality within gene lists ([Pehkonen *et al.*, 2005](#))
- Ranking of documents by their relevance with respect to gene queries ([Sehgal and Srinivasan, 2006](#)) or Swiss-Prot medical annotation ([Dobrokhотов *et al.*, 2003](#))
- Generation of hypotheses for the explanation of experimental or clinical data ([Swanson, 1986; Smalheiser and Swanson, 1998; Weeber *et al.*, 2001; Srinivasan and Libbus, 2004](#))

Most of these specific tasks require a mapping between entities and articles containing information about these entities. Manually curated public databases generally contain references to the articles where the curated information has been obtained from. These references can be used for generating a mapping between entities and articles. Alternatively, named entity identification can be applied and thus, more comprehensive mappings can be generated.

Biomedical Language

Biomedical language significantly differs from common English language. This is largely due to the descriptive nature of biomedical sciences and the important number of technical terms. A standardization effort has only emerged during the last decades. Biomedical language is characterized by *synonymy*; that is, most biological objects and concepts are represented by more than one term. For example, genes and proteins have five to ten synonyms on average. *Ambiguity* also occurs frequently; that is, a term frequently refers to several entities/concepts. Abbreviations and acronyms (i. e. abbreviations that are formed by combining the first, and sometimes other, letters of the principal words) are very common in biomedical texts. Abbreviations and acronyms are frequently defined as required by the author, and especially prone to ambiguity. Multi-word units play an important role in describing biomedical concepts, and spelling variants occur frequently. Generally, it is difficult to map biomedical text to standard ontologies or thesauri in an automated way due to the numerous spelling variants and due to nested terms.

Biomedical language is constantly changing. As new entities are discovered, new terms are introduced for them. Sometimes, names are changed when more knowledge on individual entities is accumulated. For example, when it becomes evident that a protein forms part of a family, the protein is usually assigned with the family name and a specific subtype

identifier such as a number, letter, or Greek letter.

Interestingly, the amount of information which is available in the literature for individual genes follows an extreme power law distribution and the impact of a gene in the scientific literature is not correlated to its centrality in protein-interaction networks ([Hoffmann and Valencia, 2003a,b](#)).

Scientific articles in the biomedical domain are generally structured according to a well-defined schema: An article contains a *title* and an *abstract*. The full text article additionally contains the following sections: *Introduction*, *Materials and Methods*, *Results*, *Discussion*, and *Conclusions*. The information content and occurrences of gene symbols and names varies between the different text sections ([Shah et al., 2003](#)): Abstracts contain the highest ratio of keywords per total of words. Besides the abstract, the introduction and discussion appear as appropriate places when searching for gene and protein names and interactions. The Methods section is generally most different from all other sections, best suited for looking for technical data, measurements and chemicals, and least suited for searching for genes, proteins, and interactions between them. Information density is highest in abstracts, but the information coverage is much greater in full texts than in abstracts, with highest information coverage in the results section, and 30–40% of the information mentioned in each section is unique to the section ([Schuemie et al., 2004](#)).

The difference between common English language and biomedical language is also reflected in the performance of information extraction approaches: Recall and precision for identifying person, organization, and location names in news stories have been reported in the range of 93–95%, while the values for identifying biological names are in the 75–80% range ([Hirschman et al., 2002a](#)). Possible explanations for this divergence, besides the ones mentioned above, are given by (1) the small number of shared training and test sets for setting up systems and measuring progress in the biomedical domain, (2) experience, which is significantly smaller for text mining in the biomedical domain than in the news domain, and (3) the task definition. In contrast to news articles, annotation of biomedical text needs profound background knowledge and thus needs to be done by expert scientists who often perceive the linguistic task as somewhat artificial. Biomedical text annotations are often debatable and annotation guidelines are sometimes unclear which results in lower inter-annotator agreement.

2.4 Evaluation

Evaluation requires a *gold standard*, which is, in the best case, data that is manually annotated by domain experts. Gold standard data sets are often constructed by running an automated system against a set of input texts and then having domain experts analyzing the result and correcting the systems output. Having multiple experts annotate the data makes it possible to determine the *inter-annotator agreement*, which indicates the upper limit accuracy of an automated system. Creating a gold standard is a tedious task. Therefore, proprietary gold standards are often of moderate size. Larger public gold standards are of high value for the development and evaluation of text processing systems.

Evaluation Corpora

Most of the named entity recognition and identification systems published so far have been evaluated on data sets assembled and annotated by the individual authors. Only recently carefully curated data sets have become publicly available. These can be used as common gold standard to directly compare approaches.

The GENIA corpus ([Kim et al., 2003](#)) is a hand-annotated corpus of 2000 abstracts on human blood cell transcription factors. It is split into sentences, fully tokenized, part-of-speech tagged and contains almost 100 000 annotations for various biological objects such as genes, gene products, cell types, cell lines. Containing 18 545 sentences and 39 373 named entities it is the largest corpus of its type currently available. A subset of this corpus annotated with a reduced set of biological object has been used for the BioNLP named entity recognition challenge ([Kim et al., 2004](#)).

The YAPEX data set ([Franzen et al., 2002](#)) consists of about 200 MEDLINE abstracts. A part of these abstracts was derived from the GENIA corpus and re-annotated.

GENETAG ([Tanabe et al., 2005](#)) is a corpus for gene/protein named entity recognition containing 20 000 MEDLINE sentences and approximately 24 000 genes. A subset of 15 000 sentences from this set was used for the gene mention finding evaluation of the first BioCreAtIvE challenge.

MEDSTRACT ([Pustejovsky et al., 2001](#)) is a corpus for acronym recognition. It consists of 100 MEDLINE abstracts annotated with 168 manually marked occurrences of acronyms. In the last years, several assessments have been set up to evaluate systems on a blind prediction basis. Generally, the used data sets are made publicly available after the challenge evaluation. Thus, for example, the corpora of the BioCreAtIvE assessments ([Hirschman et al. \(2005b\)](#), see also Section 4.5.1) and the Learning Language in Logic (LLL05) shared task ([Nédellec \(2005\)](#), see also Section 5.2.2) became available.

[Cohen et al. \(2005b\)](#) described the design and data of six biomedical text corpora and general aspects of evaluation corpora. [Wilbur et al. \(2006\)](#) proposed new annotation guidelines based on a subcategorization of annotations in five qualitative dimensions; the application of their guidelines results in 70–80% inter-annotator agreement.

Annotation Schemes

Text annotation can be performed in various ways and levels of detail. Several annotation schemes are commonly used for annotating biomedical texts. For Named Entity Recognition, the *B-I-O tags* are frequently applied: B-tags mark words that represent the beginning of a term, I-tags are used for words inside a term, and O-tags are used for words outside terms. The tags can be complemented with labels indicating the respective entity class; for example, a B-GENE tag denotes a word at the beginning of a gene name. This schema has been applied, for instance, for the subset of the GENIA corpus prepared for the JNLPBA BioNLP-challenge ([Kim et al., 2004](#)).

The basic annotation schema for named entity identification contains entity identifiers mapped to text identifiers, eventually supplemented by the respective text fragments rep-

resenting gene names as found in the text. This schema has been applied, for instance, for the gene mention normalization task of the BioCreAtIvE challenges ([Hirschman *et al.*, 2005a](#)).

Inter-annotator Agreement

Manual tagging of text corpora is not only labor-intensive, but also a non-trivial task, as in many cases several of results are possible and acceptable. For example, given the text passage “...the phosphorylated human protein A ...”, the words *phosphorylated* and *human* may be considered as part of the protein name or as optional modifiers. Precise annotation guidelines generally lead to increased annotation consistency, yet, they can also introduce some artificial bias.

When preparing an annotated corpus, it is useful to have the corpus annotated by two or more annotators. Thus, a consensus annotation can be generated. This increases annotation quality and consistency and makes it possible to determine the inter-annotator agreement. Inter-annotator agreement represents an upper limit of the performance of an automated system and reflects the difficulty of the task as well as the level of detail of the annotation guidelines. Estimates of inter-annotator agreement indicate that experts agree on whether a name refers to a gene, protein, or mRNA only 77% of the time ([Hatzivassiloglou *et al.*, 2001](#)) and experts agree on whether a word is a gene or protein 69% of the time ([Krauthammer *et al.*, 2000](#)).

Evaluation Measures

Several measures are used for the evaluation of text mining results. The measures are based on the number of correct outputs (*true positives*, TP), incorrect outputs (*false positives*, FP), the number of results that should have been output but were not (*false negatives*, FN), and the number of results that were correctly not output (*true negatives*, TN).

Definition 2.1 *Recall (Sensitivity)* describes the fraction of correct outputs (TP) to the total number of correct results ($TP+FN$); that is, the proportion of known positives identified by the system:

$$\text{Recall } R(TP, FN) = \frac{TP}{TP + FN}$$

Definition 2.2 *Precision* measures the fraction of correct outputs (TP) to the total number of outputs ($TP+FP$); that is, how often a system is correct when it makes a positive prediction:

$$\text{Precision } P(TP, FP) = \frac{TP}{TP + FP}$$

Definition 2.3 *Specificity* is the proportion of true negatives (TN) of all negative cases in the population ($TN+FP$); that is, how often a system is correct when it makes a negative prediction:

$$\text{Specificity } S(TN, FP) = \frac{TN}{TN + FP}$$

Specificity is used for general classification tasks where negative outcome is of interest; e.g., for evaluation of medical tests. It differs from precision, especially for differing class sizes.

Definition 2.4 The *F-measure* is the weighted harmonic mean between precision and recall:

$$F\text{-measure } F_\alpha(P, R) = \frac{1}{\frac{1}{1+\alpha}(\frac{1}{P} + \frac{\alpha}{R})} = \frac{(1+\alpha) \cdot P \cdot R}{\alpha \cdot P + R}$$

where the weighting factor $\alpha \geq 0$ can be used to shift the weight towards precision or recall; the weight is balanced for $\alpha = 1$.

Generally, there is a trade-off between recall and precision: High precision can be typically achieved at lower recall, and vice versa. Depending on the precise task, precision or recall may be more important. Frequently, the aim is to achieve both, high precision as well as high recall. In these cases, the F-measure provides a means to take precision and recall into account. If α is set to one, high and balanced values of precision and recall yield a high F-measure.

Definition 2.5 *Total accuracy* measures the fraction of correct answers with respect to the total number of test cases:

$$\text{Accuracy } A(TP, FP, FN, TN) = \frac{TP + TN}{TP + FP + TN + FN}$$

In text mining, special attention has to be put on the precise definition of the individual instances to be evaluated. For example, for named entity identification, a tuple of abstract identifier and gene identifier, a tuple of sentence identifier and gene identifier, or each occurrence of a gene in an article can represent an instance.

Chapter 3

Nomenclature of Biological Objects

One of the most important tasks in biomedical text mining is the recognition and identification of biological objects, especially genes and proteins, in texts. This requires knowledge on the characteristics of biological object names and a mapping between object names and identifiers. Often, this information is compiled in a *dictionary*, which contains for each object a unique identifier and a definition and/or alternative names.

In the following, methods to generate comprehensive high-quality gene and protein name dictionaries are presented. Therefore, data is extracted automatically from public databases and subsequently subjected to automatic curation (Section 3.2). Gene name dictionaries and the underlying databases are characterized by a systematic analysis (Section 3.3, [Friedel and Zimmer \(2006\)](#)).

Methods for the generation of hierarchical gene name dictionaries are presented; these were developed together with Caroline Friedel and Cornelia Donner (Section 3.4, [Donner \(2003\)](#); [Friedel \(2003\)](#)). Furthermore, non-gene and non-protein dictionaries, an abbreviation dictionary, and an interaction term list are generated (Section 3.5).

Finally, several applications are described which make the generated dictionaries publicly available; these have been developed by and with Joannis Apostolakis, Daniel Gütler, and Martin Szugat (Section 3.6, [Szugat et al. \(2005\)](#); [Gütler \(2006\)](#)).

The derived dictionaries allow gene and protein name identification with high recall and precision, as will be shown in the next chapter (Chapter 4).

3.1 Introduction and Literature Review

Genes and proteins are often named for their function (e. g. *growth hormone*), homology or similarity (e. g. *Rho-like protein*), phenotype (e. g. *wingless*) or localization (e. g. *HIV-1 envelope glycoprotein gp120*). This can lead to quite long designations such as “*Basic salivary proline-rich protein 4 allele S precursor*”. Thus, gene and protein names are often descriptive and do not represent proper names in the strict sense (see also examples in Table 3.1). Frequently, genes are named according to common rules; for example, genes in

a family are typically named by addition of Greek letters as prefixes or postfixes. Most genes/proteins are referred to by several names (*synonymy*), and a name can be associated with several genes/proteins (*homonymy*) which causes *ambiguity*. The complex naming leads to significant usage of abbreviations, which results in additional synonyms and often introduces important ambiguity. Furthermore, gene symbols and names can overlap with English words, such as the gene names *leg*, *white*, and *key*. The exchange of knowledge on objects requires consistent names or identifiers for each object. Several communities provide nomenclature paradigms. Yet, the generation and assignment of names to newly identified genes and proteins is not strictly standardized and standards are not strictly enforced. Thus, researchers are free to define, assign and use names as required in particular in scientific papers. In fact, the percentage of genes that are cited predominantly by their official name is increasing only slowly (e.g. from 35% in 1994 to 44% in 2004, [Tamames and Valencia \(2006\)](#)), which indicates that the guidelines are generally not supported by the scientific community.

Entry name	ATS2_HUMAN
Protein name	ADAMTS-2 [Precursor]
Synonyms	EC 3.4.24.14 A disintegrin and metalloproteinase with thrombospondin motifs 2 ADAM-TS 2 ADAM-TS2 Procollagen I/II amino propeptide-processing enzyme Procollagen I N-proteinase PC I-NP Procollagen N-endopeptidase pNPI
Gene name	Name: ADAMTS2 Synonyms: PCINP, PCPNI
Entry name	MMP9_HUMAN
Protein name	Matrix metalloproteinase-9 [Precursor]
Synonyms	EC 3.4.24.35 MMP-9 92 kDa type IV collagenase 92 kDa gelatinase Gelatinase B GELB Contains 67 kDa matrix metalloproteinase-9 82 kDa matrix metalloproteinase-9
Gene name	Name: MMP9 Synonyms: CLG4B

Table 3.1: Example entries of gene and protein name information in Swiss-Prot (ATS2_HUMAN, MMP9_HUMAN).

Gene and protein name dictionaries contain gene and protein identifiers, symbols and

names in a uniform format. They are useful for manual literature research as well as for automatic approaches focusing on gene name identification in the context of information retrieval or information extraction.

Two major classes of approaches have been applied to gather information on gene and protein nomenclature and to compile gene and protein dictionaries. The first class of approaches extracts synonym gene and protein terms directly from texts. These approaches do not depend on predefined dictionaries or structured data sources like databases. They can detect names which are classified as obsolete by the official nomenclature committee, newly introduced names not yet covered in databases, and spelling errors. Generally, these approaches exhibit high recall but modest precision. Gene name synonyms have been extracted by manually defined patterns in which synonyms commonly occur (Yu *et al.*, 2002). An integrated method consisting of unsupervised, partially supervised, supervised machine-learning techniques, and a manual knowledge-based approach achieved 80% recall at 8% precision or 30% recall at 23% precision (Yu *et al.*, 2003). The structure of symbol co-occurrence networks was exploited by Cohen *et al.* (2005a); they started with seed pairs of synonym names, extracted patterns surrounding co-occurrences of a pair from the texts, matched these patterns against texts to extract further name pairs, added the newly detected pairs to the co-occurrence network and ranked the pairs according to a clustering coefficient based quality measure. Shi and Campagne (2005) selected high-frequency terms and applied support vector machines (SVMs) for classifying terms by their context as protein versus cell, protein versus process, and protein versus interaction. Distributional clustering methods which group words based on the contexts they appear in have been shown to aid in the construction of term lists (Sandler *et al.*, 2006).

The second class of approaches makes use of structured data sources such as databases (for examples see Table 1.1). Several public databases contain gene and protein names and assign unique identifiers to genes and proteins. The databases generally contain high quality information that is often obtained from manual annotation and curation. Yet, data is presented for human users and not for text mining applications and thus generally limited to a subset of the possible naming variants.

For well-studied organisms several general or organism specific databases can be consulted. The format of the databases differs; e.g. Entrez Gene (Maglott *et al.*, 2005) provides tab-separated files, which are straightforward to parse, while Swiss-Prot (Bairoch *et al.*, 2005) files, the protein names are contained in different fields: the *Gene name* field, which is easy to parse automatically, and the *description* field, that contains long forms and is more difficult to parse due to nested parentheses which sometimes contain separate synonyms of varying specificity and sometimes contain subtypes, specifications of or additions to previous synonyms (e.g. *(Na(+)/I(-)-symporter)* or *Amyloid beta A4 protein precursor (APP) (ABPP) (Alzheimer's disease amyloid protein homolog) [Contains: Soluble APP-alpha (S-APP-alpha); C99; Beta-amylid protein 42 (Beta-APP42); C83; P3(42); P3(40); Gamma-CTF(59) (Gamma-secretase C-terminal fragment 59); C31]*). The extensive usage of Swiss-Prot thus requires more elaborate parsing.

Several studies dealing with ambiguity in biomedical nomenclatures have been accom-

plished in the last years. Most of these focus on the detection and analysis of abbreviations (Liu *et al.*, 2002a; Adar, 2004; Schwartz and Hearst, 2003; Yu *et al.*, 2002), or on the compilation of databases containing mappings between abbreviations and the corresponding long forms (e.g. Chang *et al.* (2002); Wren and Garner (2002), for an overview see Wren *et al.* (2005a)). This is due to the omnipresence of abbreviations in the biomedical domain and to the significant problems they entail. Abbreviations frequently have numerous different meanings which can belong to the same or distinct semantic fields (e.g. protein names, experimental techniques, cell lines, or others). Furthermore, authors frequently define their own abbreviations and names which are then more or less only valid for the document they are contained in and possibly closely related documents.

Weeber *et al.* (2003) studied the ambiguity of human gene symbols; they showed that gene symbols from LocusLink overlap with abbreviations and that many of the corresponding occurrences in MEDLINE abstracts are not related to the corresponding gene. Hirschman *et al.* (2002a) investigated the problems encountered when identifying biological names in texts; they describe the challenge of recognizing fly gene names in detail.

The first study known to us aiming at a systematic comparison of gene nomenclatures from different organisms (Tuason *et al.*, 2004) analyzed the nomenclatures of mouse, fly, worm, and yeast. The authors evaluated ambiguity within and across nomenclatures and with general English by exact matching of symbols and names, and applied an NLP system for analyzing recall and ambiguity of matching the derived mouse dictionary against a set of MEDLINE abstracts. Chen *et al.* (2005) analyzed eukaryotic gene name ambiguity in terms of intra-species ambiguity, ambiguity with general English and medical terms and across species. This work focused on the comparison of a large number of organisms with respect to differences in ambiguity between official gene symbols and aliases, and they analyzed author preferences for symbols or full names.

Benchmarking of gene and protein name dictionaries is an important issue. A complete gold-standard would contain all names used in any data source (database or free text) for a given gene or protein and assign these to unique object identifiers. This would make it possible to determine coverage and ambiguity or recall and precision for gene or protein name dictionaries. Currently, there is no such large-scale gold-standard available.

The gene normalization task of BioCreAtIVe (Hirschman *et al.*, 2005a) was a first independent assessment for gene name normalization; this can be considered as a small-scale benchmark for gene and protein name dictionaries in that the participants were required to recover numerous gene names from text and return unique identifiers (see Chapter 4).

The work presented in the following focuses on the usage of publicly available databases for the generation of synonym dictionaries. Various databases are integrated to generate comprehensive dictionaries. The application of an automatic curation procedure makes it possible to generate high-quality dictionaries and to tune their content versus recall or precision. The detailed analysis of the gene name dictionaries for various organisms and from several databases provides insights into the characteristics of the individual nomenclatures that play an important role for gene name identification.

3.2 Gene and Protein Name Dictionaries

Gene and protein name dictionaries are *synonym dictionaries* where the individual objects represent genes/proteins. Synonym dictionaries are lists of synonyms (names) mapped to a unique object identifier. Genes and the derived proteins usually carry the same name; therefore, we use the terms *gene* and *protein* interchangeably.

Definition 3.1 A *synonym dictionary* d consists of a set of objects $\text{objects}(d)$, where each object $o \in \text{objects}(d)$ is a tuple $(\text{identifiers}(o), \text{synonyms}(o))$ containing a unique set of $\text{identifiers}(o)$ from one or more databases and a set of $\text{synonyms}(o)$.

The set of all synonyms of a dictionary d is:

$$\text{synonyms}(d) = \bigcup_{o \in \text{objects}(d)} \text{synonyms}(o)$$

The set of objects assigned to a synonym s is:

$$\text{objects}(s) = \{o \in \text{objects}(d) | s \in \text{synonyms}(o)\}$$

The format of the synonym dictionary is described by the following Backus-Naur Form ([Naur, 1960](#)):

```

<object> ::= <identifier> ":" <synonyms>
<identifier> ::= <object_identifier> "[" "<identifier>"]
<object_identifier> ::= <database_key> "@" <database_name>
<synonyms> ::= <synonym> "[" "<synonyms>"]
<database_key> ::= string
<database_name> ::= string
<synonym> ::= string
  
```

where *database_name* is a short name for the database the data was obtained from and *database_key* is an identifier for the object assigned by the organization maintaining the respective public database, and *synonym* is a name derived from any of the source databases or generated by the curation procedure. In the case of a *gene and protein name dictionary*, each object represents a gene or protein; *database_name* corresponds to HUGO, Swiss-Prot, MGD, SGD, etc., and *synonym* is a gene or protein name. An example entry of a gene and protein name dictionary is given in the following:

IL1B@HUGO|IL1B_HUMAN@SWISSPROT|3553@ENTREZGENE:IL1B|IL-1beta|
interleukin 1, beta|interleukin-1|Interleukin-1 beta precursor|Catabolin

For certain applications, it is useful to include additional information such as GO-annotations and synonym origin for the entries in a synonym dictionary. To store such data, an XML schema for synonym dictionaries has been developed by Martin Szugat in the Bioschemas project¹ (see Example in Figure 3.1).

```

<entry organism="Human" name="BCL2">
    <synonym name="BCL2">
        <evidence source="HUGO" accNumber="BCL2"/>
        <evidence source="ENTREZGENE" accNumber="596"/>
    </synonym>
    <synonym name="Bcl-2">
        <evidence source="HUGO" accNumber="BCL2"/>
        <evidence source="ENTREZGENE" accNumber="596"/>
    </synonym>
    <go:annotation accNumber="0005783">
        <go:evidence source="ENTREZGENE" code="IEA" accNumber="596"/>
    </go:annotation>
    <go:annotation accNumber="0000074">
        <go:evidence source="ENTREZGENE" code="TAS" accNumber="596"/>
    </go:annotation>
    <cr:database entry="BCL2" name="HUGO"/>
    <cr:database entry="596" name="ENTREZGENE"/>
</entry>
```

Figure 3.1: Example entry of a synonym dictionary in XML representation showing a human gene object with two synonyms, two gene ontology annotations with evidence codes, two database references, and the corresponding evidence sources (databases) for the data.

3.2.1 Generation of gene and protein name dictionaries

Several public databases are available for the construction of gene and protein name dictionaries. For the generation of large scale gene and protein name dictionaries it is advantageous to use data sources such as databases that support large scale queries or can be downloaded and that are straightforward to parse. For merging of synonym dictionaries obtained from different databases, mappings between the database identifiers are required. As databases evolve (i. e. entries are added, modified and removed), the gene and protein name dictionaries need to be regularly reconstructed.

Generally, every gene within an organism has a unique identifier within each database and a set of associated names. For many organisms, the gene and protein names overlap; this can be due to ortholog genes that are assigned with the same name. Thus, entries within

¹<http://bioschemas.sourceforge.net/>

gene and protein name dictionaries are organism specific.

Two major types of data sources that contain gene and protein names can be defined: *organism-specific databases* and *general databases*. The organism-specific databases are mostly maintained by the organizations that are the authorities for gene annotation for the respective organism; they provide high-quality, mostly manually annotated information for one organism. General databases integrate data for many organisms; they contain either manually annotated (e.g. Swiss-Prot) or automatically compiled entries (e.g. Entrez Gene).

Here, a set of frequently used model organisms has been selected: Human, Mouse, Rat, Fly, and Yeast. For each of these, gene and protein name dictionaries have been generated from respective organism specific data sources and two general databases (see Table 3.2)

Organism	Organism-specific data source	General data sources
Human (<i>H. sapiens</i>)	HUGO (Eyre et al., 2006)	
Mouse (<i>M. musculus</i>)	Mouse Genome Informatics (MGI) (Eppig et al., 2005, 2007)	Swiss-Prot (Bairoch et al., 2005)
Rat (<i>R. norvegicus</i>)	Rat Genome Database (RGD) (de la Cruz et al., 2005; Twigger et al., 2007)	
Fly (<i>D. melanogaster</i>)	FlyBase (Drysdale et al., 2005; Crosby et al., 2007)	Entrez Gene (Maglott et al., 2005, 2007)
Yeast (<i>S. cerevisiae</i>)	Saccharomyces Genome Database (SGD) (Balakrishnan et al., 2005; Nash et al., 2007)	

Table 3.2: Data sources used for gene and protein name dictionary generation. The general data sources are used for all organisms.

An automatic procedure has been set up that downloads the latest data sets from each of the databases, parses the relevant information and compiles primary gene and protein name dictionaries. All entries from the organism specific databases are extracted for all organisms except for fly, for which only entries that are assigned to *Drosophila melanogaster* are extracted, as for most applications, only this species is of interest. From Swiss-Prot and Entrez Gene, all entries assigned to the respective organism are extracted. From Entrez Gene, only entries of type *protein-coding* are considered. After extraction, all symbols, aliases, and names are treated equivalently as synonyms for the object in question.

Mappings between database identifiers of the different databases are also extracted from the downloaded files. These mappings are used for generating the *combined dictionaries*. Entries are merged if the corresponding identifiers are directly mapped to each other in one of the considered databases. For each organism under consideration, a combined dictionary is generated by joining the entries from the corresponding organism specific database and the general databases Swiss-Prot and Entrez Gene. Entries from different databases are merged into a single entry if the corresponding identifiers are mapped to each other in any of the three databases.

3.2.2 Curation

Gene and protein name dictionaries obtained from public databases often contain synonyms that are unspecific and thus not appropriate for automatic text mining approaches. Furthermore, they frequently lack useful synonyms such as spelling variants or abbreviations. This is due to the fact that the public databases are generally designed for the human user who recognizes terms that are not gene names but additional information and can easily infer spelling variants and abbreviations.

Thus, *curation* of synonym dictionaries is necessary for removing inappropriate and unspecific terms (i. e. terms that do not clearly specify a single gene) and for adding missing but required synonyms. Here, curation is performed by a rule-based approach. The individual curation steps are fully automated; they can be applied individually and thus the curation procedure can be adapted to the synonym dictionary that needs to be curated. Most of the curation rules are generally applicable and thus, usually only very few parameters or rules are changed when it is applied to a different synonym dictionary. The curation procedure consists in sequential application of several rules. In the following, these curation rules are grouped into three sequential steps, the first two being concerned with the addition of spelling variants and basic pruning according to general simple rules. The third step consists in extensive token class and rule-based expansion and pruning of synonyms; this is explained in the next section. The distinction into three steps is relevant for the evaluation of text-mining results in the first BioCreAtIVe challenge, where the effect of the individual curation steps has been analyzed in detail (see Section 4.5.4).

Addition of Spelling Variants and Basic Pruning

In a **first step**, synonyms consisting solely of digits and/or non-alphanumeric characters and synonyms of a length below a threshold length (generally two characters) are removed. Subtype specifiers are expanded to equivalent other specifiers ($a \Leftrightarrow \alpha$). Non-alphanumeric characters at the beginning or end of a synonym are removed. Different spelling variants such as the insertion of a hyphen or space between alphabetic characters and digits are added ($Igf\ 1 \Leftrightarrow Igf\text{-}1 \Leftrightarrow Igf1$). Synonyms of length less than six characters are added in upper case and with the first character in upper case.

Eventually, organism specific expansion is performed. For example, yeast synonyms as defined in the synonym dictionary are often mentioned in texts with the extension p ; thus each synonym is additionally expanded by p ($SOH6 \rightarrow SOH6p$). The rules for such organism specific expansions must be deduced from a given training set (e. g. the above rule for yeast synonyms was obtained from analysis of the BioCreAtIVe training set) or by manual analysis of a set of texts if no annotated training set is available.

In a **second step**, synonyms matching common English words are removed. This step can be ignored for organisms which have many valid protein names that are common English words as it is the case for fly. Synonyms containing subtype specifiers are expanded by the synonym without subtype specifier if there is only one subtype mentioned in the synonym dictionary (aminoacylase 1 → aminoacylase).

Rule-Based Expansion and Pruning

The **third step of curation** accomplishes further expansion and pruning based on rules and lists. The tool used for this purpose was implemented and provided by D. Hanisch ([Hanisch et al., 2003](#)). The tool depends on comprehensive sets of lists and rules (see below), which have been compiled by analysis of matching statistics.

In the *expansion phase*, new synonyms are added to the existing ones. The expansion is based on rules and lists. A list of frequent abbreviations and long names is used for expanding every occurrence of a common abbreviation in the synonym dictionary to the corresponding long name and reducing long names to abbreviations (IL \leftrightarrow interleukin). The applied abbreviation list contains 112 entries.

Inappropriate synonyms are detected and removed in the *pruning phase* by using token-class based regular expressions. A *token* can be any sequence of letters and/or numbers. A *token class* is a group of words which have a similar meaning or usage. Examples of token classes are given in Table 3.3.

Token class	Examples
Description	tRNA, Ser, Tyr, binding, finger, activating, factor, transcription, ...
Specification	class, family, form, group, isoform, subfamily, chain, type, ...
Organism	avian, bovine, chicken, drosophila, feline, human, mouse, rabbit, ...
Function	kinase, protease, dehydrogenase, reductase, oxidoreductase, ...
Similarity	homolog, hypothetical, like, orphan, putative, related, similar, ...
Common words	if, and, as, for, in, of, or, with, non, ...
Measuring unit	kDa, Da, mg, ...

Table 3.3: Examples of token classes used for curation of gene and protein name dictionaries. The token classes are combined in regular expression for pruning of unspecific synonyms.

These token classes are combined in regular expressions, such as “*a number followed by a measuring unit*”, “*one description*”, “*an expression starting with a similarity token*”, or “*a common word followed by a number*”. Synonyms that are matched exactly by one of these regular expressions are removed. For example, 22 kDa is removed by the regular expression “*a number followed by a measuring unit*” and If 1 is removed by the pattern “*a common word followed by a number*”. The lists of words belonging to a token class and the rules for combining them in regular expressions were compiled based on analysis of synonyms provided in Swiss-Prot ([Bairoch et al., 2005](#)) and HUGO ([Povey et al., 2001; Eyre et al., 2006](#)) and their matching statistics against MEDLINE abstracts. The lists and rules used during the third curation step are general and hence usually do not need to be adapted when applied to new synonym dictionaries. The standard curation procedure makes use of 507 terms assigned to 17 token classes which are combined in 55 regular expressions. Ambiguous synonyms (i.e. synonyms belonging to more than one protein) can be chosen to be pruned from the dictionary. Objects which have no synonym left are removed from the synonym dictionary.

The curation is largely independent of the synonym dictionary to be curated since the individual curation steps are of general character. Nevertheless, the system can easily be adapted to cover specific problems of synonym dictionaries, such as missing synonyms that are frequently used in texts and which can be deduced from the synonyms in the dictionary by application of rules.

3.3 Analysis of Gene and Protein Name Dictionaries Derived from Public Databases

A large number of public databases organize information on genes and proteins. They represent useful resources for generating gene and protein name dictionaries. In this section, gene and protein name dictionaries derived from distinct databases are compared with respect to the following features: (1) size of the gene name dictionaries; (2) ambiguity of gene names within a dictionary and between dictionaries, that is, with respect to gene objects of different species; (3) ambiguity of the combined dictionaries with general English words and with non-gene and non-protein, but domain-related terms; (4) ambiguity of gene name dictionaries after extensive curation; and (5) degree of overlap of gene names for a given species covered in different public databases. The ambiguity analyses provide information on the degree of difficulty of accurate text searches and hence the effort that has to be spent on contextual filtering and disambiguation. The overlap analysis investigates the relevance of joining information from different data sources.

Several studies ([Hirschman et al., 2002a](#); [Weeber et al., 2003](#); [Tuason et al., 2004](#); [Chen et al., 2005](#)) investigated some of the above points. The work presented here extends these by analyzing ambiguity within and between dictionaries by using three different definitions for term equivalence, which reveals some properties of the analyzed nomenclatures; by evaluating different public data sources for extracting dictionaries separately, which allows an individual rating of the different data sources; and by analyzing the degree of overlap of gene names contained in different data sources.

Gene name dictionaries

For each of the five organisms human, mouse, rat, fly, and yeast, five synonym dictionaries have been compiled and analyzed: three dictionaries have been derived from a single data source each (Entrez Gene, Swiss-Prot, and an organism specific database as listed in Table 3.2), one combined dictionary (in which the entries derived from the three data sources are merged according to the downloaded mappings), and one curated dictionary (see Section 3.2.2).

Lexicon of common English words and domain-related non-gene and non-protein terms

As lexicon of common English words, a lexicon of words from the Wall Street Journal

(WSJ) and Brown corpus has been used as provided with Brill's part of speech tagger ([Brill, 1992](#)); it contains 93 694 entries.

The lexicon for domain-related non-gene and non-protein terms has been derived from the Unified Medical Language System ([Bodenreider, 2004](#)) as described in Section [3.5.1](#); it contains 1 062 223 entries.

Term equivalence

Gene symbols and long names show quite variable properties; while the distinction between upper and lower case letters can be important for whether a short gene name refers to one gene or another, the spelling of long names is usually much more flexible. Therefore, three different definitions of term equivalence are applied:

Definition 3.2 *Term equivalence* \sim_e , $e \in \{\text{exact}, \text{mixed}, \text{norm}\}$:

- **exact:** Two terms s, s' are equivalent if they are equal to each other in a case sensitive way: $s \sim_{\text{exact}} s'$
- **mixed:** Two terms s, s' are equivalent if they are equal to each other, where the case of letters is only considered if the name consists solely of letters and is of length less than six and the case of letters is ignored otherwise: $s \sim_{\text{mixed}} s'$
- **norm:** Two terms s, s' are equivalent if they are equal to each other when the case of letters is ignored and after any sequence of non-alphanumeric characters has been replaced by a single placeholder: $s \sim_{\text{norm}} s'$

A term is *normalized* by converting it to lower case and replacing any sequence of non-alphanumeric characters by a single placeholder.

3.3.1 Size of Gene Name Dictionaries

Definition 3.3 The *size(d)* of a gene name dictionary d is quantified by the number of objects (i. e. distinct genes) in the dictionary ($\#\text{objects}(d)$) and the number of distinct synonyms according to the equivalence \sim_e ($\#\text{_esynonyms}(d)$, i. e. the number of equivalence classes in $\text{synonyms}(d)$):

$$\text{Size}(d) = (\#\text{objects}(d), \#\text{_esynonyms}(d))$$

Gene name dictionaries vary significantly in their size, between different organisms as well as between different data sources (Figure [3.2](#)). Interestingly, the number of entries in the different databases for a given organism varies significantly. For example, Flybase contains approximately 15 times the number of objects contained in Swiss-Prot for *Drosophila*, which is, at least in part, due to the transgenes (i. e. genes that have been introduced into *Drosophila*) in FlyBase. Some of the differences between the databases might be explained by their different scope and objectives; for example, Swiss-Prot contains data that is manually compiled, which explains the smaller number of objects.

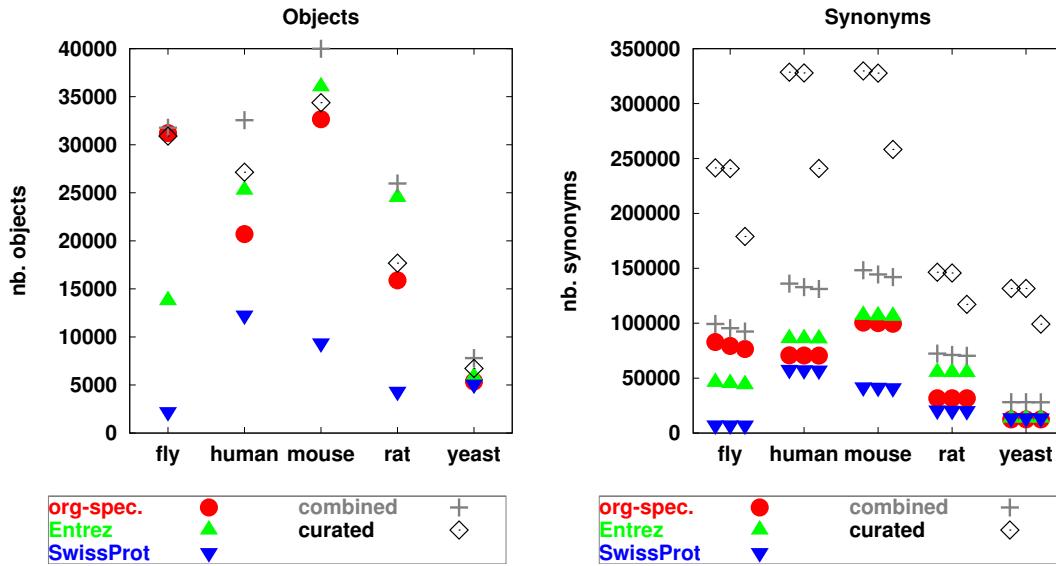


Figure 3.2: Size of gene name dictionaries: Number of objects (left panel) and synonyms (right panel) for gene name dictionaries compiled from different data sources (organism-specific database: Yeast: SGD, Fly: FlyBase, Mouse: MGI, Rat: RGD, Human: HUGO; *combined* is the merged dictionary from the organism-specific database, Swiss-Prot and Entrez Gene; *curated* is additionally expanded and pruned). The three marks for each dictionary in the right panel correspond to the three definitions of equivalence: exact, mixed, and norm, from left to right.

It is important to note that the difficulty of extracting relevant gene names from the data files varies significantly between the different data sources. Most files have a format that is easy to parse, with defined separators between distinct names. Some databases use individual conventions for representing special characters, Greek letters or formatting of name parts (e.g. *&bgr;-Cop* for *beta'-Cop* in FlyBase, or *Cyp11b2^{m1}* for *Cyp11b2^{m1}* in RGD). In order to generate dictionaries applicable for named entity recognition systems, these formatting conventions need to be accounted for. Swiss-Prot is the database which is most difficult to parse among the databases analyzed here. This is due to the choice of parentheses as separators between long names. Given the fact that long names frequently contain parentheses, this entails the necessity of a more involved parser than required for other data sources. Furthermore, protein long names are contained in the description section together with further information. For example, when a protein has several functional domains which have individual names, this is specified by an expression of the format *entire_protein /Includes: domain_1_name; domain_2_name/*.

Figure 3.2 also shows that the curation procedure leads to a modest decrease in the number of objects, which is due to merging of objects that have a significant number of synonyms in common, and to a large increase in the number of synonyms, which is mainly due to the addition of spelling variants.

3.3.2 Ambiguity

Gene names frequently have various meanings. The definition of synonym ambiguity makes it possible to define the degree of ambiguity of a synonym dictionary which estimates the number of synonyms in the dictionary that have various meanings. Ambiguity is defined within a dictionary, between two dictionaries, and between a dictionary and a lexicon of general terms:

Definition 3.4 A synonym s of a object o in a dictionary d is said to be **ambiguous** if it is equivalent to (according to the definition of term equivalence \sim_e , where $e \in \{\text{exact}, \text{mixed}, \text{norm}\}$)

- a second synonym s' of a different object o' :
 $\exists s \in \text{synonyms}(o) \exists s' \in \text{synonyms}(o') : s \sim_e s' \wedge o \neq o'$
 $\Rightarrow s$ is ambiguous within dictionary d : $s \in \text{ambiguous}(d)$
- a second synonym s' of a different dictionary d' :
 $\exists s \in \text{synonyms}(d) \exists s' \in \text{synonyms}(d') : s \sim_e s' \wedge d \neq d'$
 $\Rightarrow s$ is ambiguous between dictionaries d and d' : $s \in \text{ambiguous}(d, d')$
- an entry l of a lexicon L of general terms:
 $\exists s \in \text{synonyms}(o) \exists l \in L : s \sim_e l$
 $\Rightarrow s$ is ambiguous between dictionary d and lexicon L : $s \in \text{ambiguous}(d, L)$

otherwise the synonym is **unique**.

Definition 3.5 The **degree of ambiguity** DoA_{comp} , $comp \in \{\text{intra}, \text{inter}, \text{lex}\}$, is the quotient of the number of ambiguous synonyms in a set X_A and the total number of synonyms in X_T .

$$DoA_{comp}(X) = \frac{\#_e(\text{ambiguous}(X_A))}{\#_e(X_T)}$$

Depending on the data to be analyzed, three variants of the degree of ambiguity are applied:

- **Intra-dictionary degree of ambiguity** DoA_{intra} for a dictionary d :
 $\text{ambiguous}(X_A)$ is the set of synonyms that are ambiguous within d : $X_A = d$;
 X_T is the set of all synonyms for all objects in d : $X_T = \text{synonyms}(d)$.
- **Inter-dictionary degree of ambiguity** DoA_{inter} for two dictionaries d, d' :
 $\text{ambiguous}(X_A)$ refers to synonyms that are ambiguous between the two dictionaries:
 $X_A = (d, d')$;
 X_T is the union of the synonyms of the two dictionaries:
 $X_T = \text{synonyms}(d \cup d')$.
- **Dictionary-lexicon degree of ambiguity** DoA_{lex} for a dictionary d and a lexicon L :
 $\text{ambiguous}(X_A)$ is the set of synonyms that are ambiguous between d and L : $X_A = (d, L)$;
 X_T is the set of synonyms in d (here, the denominator is independent of the number of synonyms in the lexicon): $X_T = \text{synonyms}(d)$.

Intra-Species Ambiguity

The degree of intra-dictionary ambiguity (DoA_{intra}), that is, the fraction of synonyms that are assigned to more than one object within a gene name dictionary, varies significantly between different organisms (Figure 3.3). For the combined dictionaries, yeast shows the lowest and human the highest ambiguity. The obtained results agree with previous studies (Tuason *et al.*, 2004; Chen *et al.*, 2005) which investigated the intra-species ambiguity of dictionaries combined from organism-specific databases and LocusLink (which is now Entrez Gene); only for fly the authors of these studies obtained significantly higher ambiguity (>12%) than obtained here (1.8–4.4%). This might be due to the fact that, here, entries from FlyBase have been restricted to those specifically assigned to *Drosophila melanogaster*. The intra-dictionary degree of ambiguity for a given organism varies significantly between the different data sources. For example, the human dictionary derived from HUGO has a DoA_{intra} of 1.68–1.83% while the human Entrez Gene dictionary has a DoA_{intra} of 3.16–3.32%, even though the number of synonyms is similar for the two dictionaries.

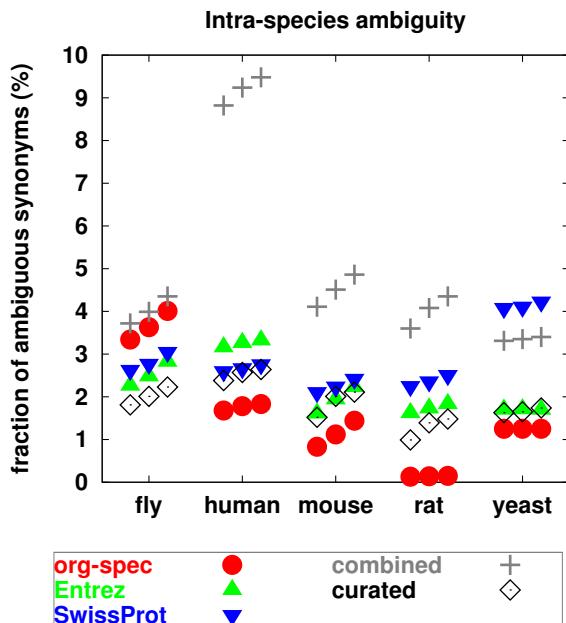


Figure 3.3: Ambiguity within gene name dictionaries according to the three definitions of equivalence: exact, mixed, and norm, from left to right. The ambiguity within gene name dictionaries derived from different data sources and for different organisms varies significantly. Combined dictionaries generally show relatively high ambiguity. Curation reduces ambiguity.

The applied definition of term equivalence has an important effect on the results. For yeast the difference between the three measures is very small, which indicates that gene names are clearly distinct from each other. For all other organisms the difference between the

three measures is significantly larger, indicating that case and exact spelling can distinguish between one gene or another.

Furthermore, the results show that for yeast and fly, the DoA_{intra} of the combined dictionary corresponds to the highest DoA_{intra} of the individual dictionaries. For human, mouse, and rat the DoA_{intra} of the combined dictionaries are significantly higher than the DoA_{intra} of the individual dictionaries. Thus, after joining the dictionaries, numerous gene names are assigned to more than one object. In fact, a number of objects contain several gene names in common (as indicated by merging step during curation). In these cases, the applied mappings appear deficient.

For all organisms, curation reduces the number of ambiguous terms, which is due to removal of unspecific synonyms and to merging of objects sharing a large number of equivalent synonyms. The DoA_{intra} of the considered dictionaries after curation is 1–2.6%.

Inter-Species Ambiguity

The degree of inter-species ambiguity (DoA_{inter} , Table 3.4) between mouse, rat, and human is significantly higher than between yeast or fly and any of the other organisms. One explanation for this fact is that mouse, rat, and human are much closer related to each other than to yeast, and homologs in different organisms often carry the same name (Chen *et al.*, 2005). The highest degree of ambiguity is found between human and mouse, ranging between 15% and 25% for the different measures. The nomenclature guidelines from MGD and RGD explicitly state that “genes that are recognizable orthologs of already-named human genes should be given the same name and symbol as the human gene”; and also the HUGO guideline states “that homologous genes in different vertebrate species should where possible have the same gene nomenclature” and that “the agreement between human and mouse gene nomenclature for many homologous genes should be continued and extended to other vertebrate species where possible”. Generally, the committees for the nomenclatures of rat, human, and mouse genes coordinate their work increasingly. This brings about co-assignment of nomenclatures to ortholog genes, mappings between orthologs by cross-references, and thus an increasing unification of the individual nomenclatures.

The curation has diverse effects on the inter-species ambiguity: for human, mouse, and rat, curation leads to an increase in inter-species ambiguity, for other pairs of organisms curation leads to a decrease in inter-species ambiguity. An increase in inter-species ambiguity is due to the expansion of synonyms (e.g. expansion of abbreviations) which can result in equivalent synonyms that were not present originally in the two lists to be compared. A decrease in inter-species ambiguity is caused by removal of unspecific synonyms, but also by the increase in total number of synonyms emerging from the expansion of abbreviations and addition of spelling variants.

		Human	Mouse	Rat	Yeast
combined	Fly	1.4/1.9/2.4	1.6/1.9/2.3	1.1/1.4/1.7	0.9/1.3/1.4
	Human		15.1/22.5/24.8	8.5/12.8/14.3	2.3/2.5/2.5
	Mouse			13.5/13.9/14.1	1.2/2.0/2.1
	Rat				1.0/1.7/1.8
curated		Human	Mouse	Rat	Yeast
	Fly	0.9/1.6/1.9	1.0/1.6/1.8	0.8/1.3/1.5	0.5/1.0/1.1
	Human		13.6/24.8/25.5	9.5/16.4/17.3	1.9/2.1/2.1
	Mouse			17.4/18.7/18.0	0.9/1.8/1.7
	Rat				0.6/1.4/1.4

Table 3.4: Inter-species ambiguity: Degree of ambiguity between combined and curated gene name dictionaries of different organisms. The three numbers in each field correspond to the three definitions of equivalence: exact, mixed, and norm, from left to right; numbers are percentages.

Ambiguity with General English Lexicon and Domain-Related Terms

The degree of ambiguity between the dictionaries and a lexicon of common English words, or domain-related non-gene and non-protein terms (Figure 3.4) shows some important organism-specific gene name characteristics. Yeast has the lowest ambiguity with common English words as well as with domain-related terms (0.01–0.3%/0.09–0.4% for combined/curated dictionary). Fly has the highest DoA_{lex} with common English words (0.55–2.4%). This is due to phenotypic descriptions and abbreviations thereof which are frequently used as fly gene names. For example, *We* is the abbreviation and valid symbol for a gene named *Washed eye* in FlyBase; in this case, the abbreviation as well as the words of the long name are perfect English words. The gene nomenclature guidelines for FlyBase are relatively loose (Tuason *et al.*, 2004): “Gene names must be concise, unique, and not have been previously used for a *Drosophila* gene, should allude to the genes function, mutant phenotype or other relevant characteristic; furthermore gene names should be inoffensive.” No format is proposed for the symbols, and no restrictions about ambiguities with English words or other terms are made. The guideline favors the usage of descriptive names, which might be useful for an immediate functional classification of genes by a researcher when reading scientific articles, but brings about significant disadvantages for literature search and automatic text processing.

The degree of ambiguity with the lexicon of common English words agrees with previously reported results (Tuason *et al.*, 2004; Chen *et al.*, 2005), even though these were obtained with a different lexicon of English words (the Moby lexicon project²). The degree of ambiguity with UMLS-terms was estimated to be significantly higher (7–28% for fly, human, mouse, and rat) by Chen *et al.* (2005). This might be due to their expansion of the set of UMLS terms by adding abbreviations extracted from UMLS.

Generally, the percentages of ambiguity may seem rather small. Yet, for example the 2.4%

²<http://www.dcs.shef.ac.uk/research/ilash/Moby/>

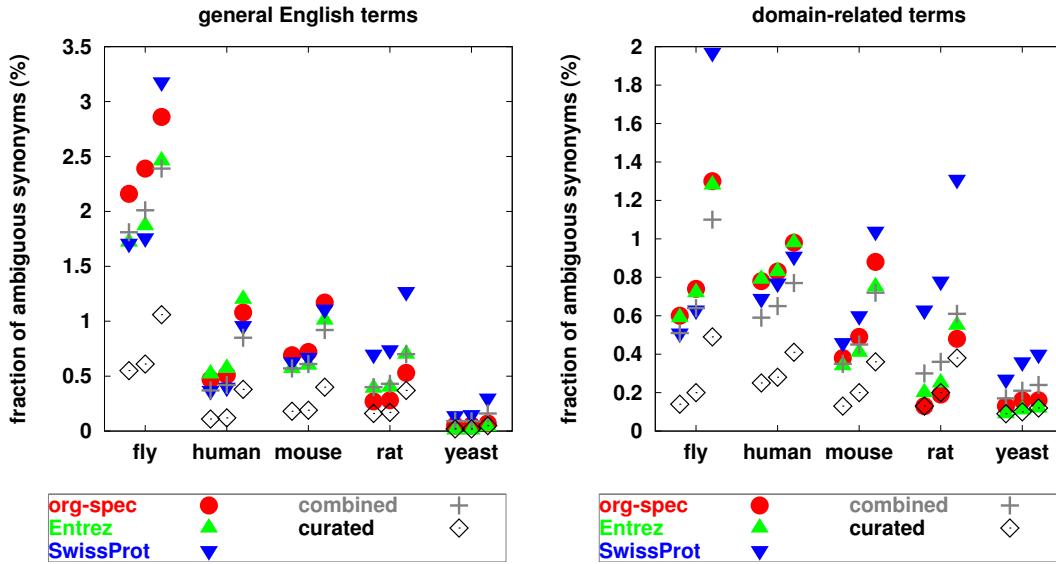


Figure 3.4: Ambiguity between gene name dictionaries and general English terms (left panel) and domain-related non-gene and non-protein terms (right panel) according to the three definitions of equivalence: exact, mixed, and norm, from left to right. Fly shows highest ambiguity with general English terms. All dictionaries show higher ambiguity for normalized gene names (*norm*) than for exact gene names.

of fly synonyms in the combined list ambiguous with common English words correspond to a total number of 2208 synonyms. Several of these ambiguous synonyms are similar to English words which are frequently contained in scientific articles (e.g. *We* (2 655 352), *gel* (298 680), *fold* (251 172), *inactive* (54 429), numbers indicate MEDLINE abstracts that contain the respective term (March 2007)); these synonyms hamper manual and automatic literature search. Furthermore, only gene names that directly match entries of a lexicon have been analyzed here. Gene names may not match to an individual entry of a lexicon, but to a combination of several entries; for example, the gene names *Washed eye* and *legless* were not found as such in the used English lexicon, but represent combinations of common English words that are contained in the lexicon. Such gene names are critical for detection if their occurrence in texts fulfills standard English syntactic rules. For some of these gene names, additional methods allow safe detection. For example, for the synonym *legless*, part-of-speech tagging allows to decide on whether an occurrence refers to a gene name (when tagged as noun) or to a phenotype description (when tagged as adjective).

3.3.3 Overlap between Data Sources

The degree of overlap of two dictionaries d, d' indicates how similar the content of two dictionaries is, given a mapping between the dictionary entries.

Definition 3.6 A *mapping* between two dictionaries d, d' is a set of tuples (o, o') where $o \in \text{objects}(d)$ and $o' \in \text{objects}(d')$.

Definition 3.7 The *degree of overlap* $\text{DoO}(d, d')$ of two dictionaries d, d' is determined as follows: Given a mapping map between the dictionaries d and d' , the degree of overlap $\text{DoO}(d, d')$ is defined as the quotient of the equivalent synonyms assigned to a pair of objects $(o, o') \in \text{map}$ and the union of synonyms assigned to o or $o' \in \text{map}$:

$$\text{DoO}(d, d') = \frac{\#_e(S_{eq})}{\#_e(S_{tot})}$$

$$S_{eq} = \bigcup_{(o, o') \in \text{map}} \{s \mid s \in \text{synonyms}(o) \wedge s' \in \text{synonyms}(o') \wedge s \sim_e s'\}$$

$$S_{tot} = \bigcup_{(o, o') \in \text{map}} \text{synonyms}(o) \cup \text{synonyms}(o')$$

For each organism, mappings between database identifiers are obtained from the three relevant databases (Swiss-Prot, Entrez Gene, and organism-specific). All direct mappings between database identifiers are used to generate pairs of objects (o, o') from the two dictionaries under consideration. Objects that cannot be mapped to an object of the second dictionary are ignored. For each pair of objects, the number of equivalent synonyms is determined. The fraction of the total number of equivalent synonyms to the total number of distinct synonyms belonging to the considered objects represents the degree of overlap. The degree of overlap of synonyms between different data sources (Figure 3.5) is highly variable (between 11% and 83%). Particularly, the overlap between organism-specific databases and Entrez Gene is significantly higher than the overlap between the other pairs of databases. Swiss-Prot appears to be rather dissimilar to the other databases.

These results strengthen the hypothesis that it is necessary to combine entries from several data sources in order to generate a dictionary that is as complete as possible.

The values depend on the applied definition of equivalence. The selected definition has little effect on the overlap between organism specific databases and Entrez Gene, indicating that the gene names in these databases are more or less identical. Important differences between the applied definition can be observed in the comparison between organism-specific databases and Swiss-Prot; e.g., for mouse, the overlap is only 18% when exact identity is required, but 25% when gene names are normalized (*norm*). This indicates that numerous gene names in these databases are not exactly identical, but their normalized forms are the same and thus they are very similar.

The differences in overlap are presumably due to the structures and strategies of the organizations that maintain the databases. The organizations maintaining the organism-specific databases are the authorities for official nomenclature and genome annotation. The individual nomenclature committees and organizations were rather separated from each other when they started gathering information and first set up databases for making information publicly available. Yet, in the last years, they started to increasingly coordinate and

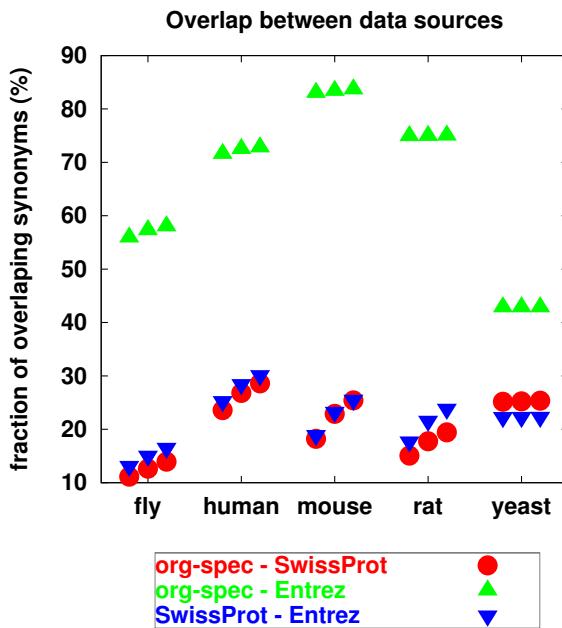


Figure 3.5: Overlap between different data sources according to the three definitions of equivalence: exact, mixed, and norm, from left to right. The overlap between gene name dictionaries compiled from different data sources varies for different organisms and pairs of databases. The organism-specific databases and Entrez Gene show highest overlap for all organisms.

integrate their work by agreeing common nomenclature guidelines, exchanging data and enforcing co-assignment and co-curation of gene and protein annotation. Today, model organism databases and general sequence repository resources such as NCBI Entrez Gene exchange data on a regular basis to reflect the official nomenclature. As a consequence, the number of cross-links between databases has already increased significantly. Swiss-Prot also works together with model organism databases. Entrez Gene and Swiss-Prot are historically separated as they have a different focus. NCBI established Entrez Gene as a database for gene-specific information with a focus on completely sequenced genomes or genomes that have an active research community to contribute gene-specific information. Swiss-Prot is a manually annotated protein sequence database; besides nomenclature, the annotation includes protein structure, function, and associated diseases. Swiss-Prot forms part of the UniProt Knowledgebase, which is concerned with integrating information to provide a central, stable, comprehensive, fully classified, richly and accurately annotated protein sequence database with extensive cross-references to other data sources. With the advent of the UniProt project, one can expect that Swiss-Prot/UniProt and Entrez Gene will increasingly share nomenclature and that the mapping between databases will be increasingly complete and unambiguous. This will facilitate the generation of gene name dictionaries and text mining applications.

3.3.4 Relevance of Ambiguities for Mining MEDLINE

For text-mining applications it is important to know, besides the degree of ambiguity of synonym dictionaries, whether ambiguous synonyms occur frequently in texts. Here, the relevance of inter-species ambiguities for mining MEDLINE abstracts is estimated as follows: All pair-wise combinations of the organisms human, mouse, and rat are considered. The curated gene/protein name dictionaries are matched against approximately 7 million MEDLINE abstracts (from 1990 or later) by ProMiner ([Hanisch et al., 2003, 2005](#)) (see Section 4.3). This returns the abstracts in which a gene name was found. Then, the number of abstracts that contain synonyms that are ambiguous between the respective pair of organisms is determined. This indicates how often inter-species ambiguities occur when mining MEDLINE abstracts.

Generally, semantically ambiguous words are only used with one sense in any given discourse or throughout a limited unit of text (the *one sense per discourse* hypothesis). For example, the word *bank* will rarely describe a financial institute and a seating within the same article. In accordance with this hypothesis, the (simplifying) assumption made here is that ambiguous synonyms that occur within a same abstract share the same context and thus refer to the same organism. Here, abstracts are expected to (mostly) deal with single organisms. The disambiguation task is reduced to selecting for each abstract one of the two organisms under consideration.

For giving an estimate of the relevance of ambiguities for mining MEDLINE, a simple disambiguation strategy is applied: (1) If the abstract contains, besides the ambiguous synonym(s), further synonyms that are assigned to only one of the considered organisms, the ambiguous synonyms are also assumed to refer to this organism. For example, *BNIP3H* is a synonym of the human gene *BCL2/adenovirus E1B 19-kDa protein-interacting protein 3-like* (BNI3L_HUMAN) in Swiss-Prot, but not of the corresponding gene in mouse (BNI3L_MOUSE); thus, if *BNIP3H* occurs in an abstract, one can hypothesize that the article deals with human and thus presumably all ambiguous synonyms within this abstract also refer to human. (2) If an abstract contains, besides ambiguous synonym(s), only one of the corresponding organism names, the ambiguous synonyms are assumed to refer to this organism. As organism names also have synonyms (e.g. human, *Homo sapiens*, *H. sapiens*), an organism name dictionary was compiled from UMLS (Section 3.5.1) and matched against abstracts according to the mixed equivalence described above.

The fraction of abstracts that could be disambiguated by this strategy is estimated by the number of abstracts that contain, besides the ambiguous synonyms, either unique synonyms or a corresponding organism name compared to the number of abstracts that contain a gene/protein name of the respective pair of organisms.

Table 3.5 shows the results of this relevance analysis. Out of approximately 7 million abstracts, 2.2–2.8 million abstracts were found to contain at least one gene/protein name of the considered pairs of organisms. Approximately 58–65% of the latter contain protein names that are ambiguous between the two synonym dictionaries. Between 34% and 52% of the abstracts contain either a gene/protein name or an organism name unique to one of the considered organisms. Therefore, the basic organism disambiguation strategy con-

	nb. found abstracts	% amb. abstracts	% amb.+ unique synonym	% amb.+ unique organism	% amb.+ unique synonym or organism
human–mouse	2 761 987	60.5	23.1	37.8	46.5
human–rat	2 238 212	64.5	27.2	43.5	52.1
mouse–rat	2 532 682	58.2	24.2	17.1	33.7

Table 3.5: Relevance of inter-dictionary ambiguities for mining MEDLINE (amb.: ambiguous). The column *nb. found abstracts* contains the number of MEDLINE abstracts (from within a set of approximately 7 million abstracts) that contain at least one gene/protein name of the respective organisms. The values in the other columns are percentages of the values in the column *nb. found abstracts*.

sisting in assigning the organism that is indicated by direct mention of the organism or unique gene/protein name(s) can cover at most this percentage of abstracts. Consequently, the remaining 12.4% (human–rat) to 24.5% (mouse–rat) of the abstracts containing a gene name contain neither unique synonyms nor organism names that could easily be used for disambiguation. Thus, this fraction of abstracts containing a gene name cannot be disambiguated by the described simple strategy and thus require for other disambiguation methods.

Clearly, these numbers can only represent a rough estimate as several assumptions were made for this analysis; for example, MEDLINE abstracts do not necessarily refer to only one organism, and a mention of an organism does not necessarily imply that the gene names refer to this organism. Disambiguation accuracy is not evaluated here; this would imply comparison against a benchmark data set.

Yet, the results show that inter-species ambiguity is not only a problem inherent to synonym dictionaries but also directly affects text mining of MEDLINE abstracts. Ambiguous synonyms occur often in abstracts, inter-species disambiguation is not trivial, and thus more involved disambiguation approaches are definitively required.

The problem of how to handle synonyms ambiguous for different organisms within a named entity recognition system cannot generally be solved as a solution depends on the task at hand. Organism disambiguation is important when the aim is to get information for certain individual proteins within a given organism. When it becomes important to integrate information on a higher level, say, to find genes relevant for a certain human disease, it might be preferable to also integrate ortholog genes or transgenes from model organisms. In this case, it would be preferable not to exclude objects from related organisms. For this application it would further be useful to ensure orthology between the corresponding objects.

3.4 Hierarchical Synonym Dictionaries

Publications often contain information on gene or protein groups such as *Matrix Metalloproteases (MMPs)*, *Collagen*, or *Bone Morphogenic Protein (BMP)*, without mentioning a specific member of the group. These gene or protein groups can be structural families, functionally similar proteins, a complex consisting of various proteins, or the proteins implicated in a common regulatory or signaling event or pathway. Texts might contain information concerning all members of the respective group or individual members that are specified in a way that is not recognized by a named entity identification approach. Public databases contain some information with respect to higher-level groupings of genes and proteins, such as annotation with Gene Ontology terms or references to pathways and complexes. For example, [Koike and Takagi \(2004\)](#) generated a family name dictionary based on the InterPro ([Mulder et al., 2007](#)) family hierarchy and manual construction of the remaining hierarchy based on sequence similarities. Yet, the databases generally do not contain comprehensive information on gene groups as it would be required for text mining. The extraction of group terms directly from gene and protein synonym dictionaries makes it possible to derive group terms that are appropriate for text mining, with various levels of specificity and irrespectively of the type of group.

In the following, an approach for generating hierarchical synonym dictionaries and a corresponding benchmark is presented. It has been developed together with and implemented by Caroline Friedel and Cornelia Donner as part of their bachelor theses ([Friedel, 2003](#); [Donner, 2003](#)).

3.4.1 Generation of Hierarchical Synonym Dictionaries

Hierarchical synonym dictionaries expand on standard synonym dictionaries by additionally containing objects and synonyms for gene and protein families and groups. The lowest level of the hierarchy consists in the standard gene and protein synonym dictionaries as described in Section 3.2. The higher levels contain name groups with different levels of generalization. A *mapping* links the group objects to the respective constituents.

The principle steps of the heuristics applied for generating hierarchical synonym dictionaries from standard synonym dictionaries are described in the following:

(1) **Extraction of group terms:** For every synonym containing a *object specifier*; i. e., ending with one of the following expressions

- a number
- a number followed by a letter (A-F)
- a roman number
- a letter (A-F) preceded by a space or hyphen and followed by any number of digits,

the substring pruned for the object specifier is extracted as *group term*. Initially, every group term is assigned to a separate *group object*. The group term is also used as identifier

for the respective group object. The pair of original object identifier and group identifier is added to the mapping. For each original gene or protein object, all generated group terms are gathered as a set of *alternative group terms*.

(2) **Curation** consists in filtering and merging steps. Group terms below a threshold length are removed (e.g. three characters). Terms that are too general as group designations are filtered, such as *kinase* or *protein*. Group objects are then merged with the aim of pooling group terms that refer to the same set of genes/proteins. First, group objects are merged if the respective group terms were derived from the same original objects. The terms of a combined group object are then checked for their alternative group terms. Alternative group terms that are shared by at least 40% of the members of the combined group are added to this group object. Next, group objects are merged if they share at least 70% of their synonyms (i.e. group terms) or respective original object identifier. The shortest group term is used as identifier for the combined group object. Group objects are curated similarly to single objects (Section 3.2.2), yet regular expressions for synonym pruning are not applied as these would remove many group synonyms.

(3) **Ambiguities between groups** are reduced with the aim of assigning the ambiguous terms to the most specific group objects. An ambiguous synonym that is used as group identifier for one of the respective groups is assigned only to this group. An ambiguous synonym of a limited length (here: 7 characters) that occurs as suffix of another synonym assigned to one of the respective group objects is assigned to only this group (e.g. *Bcl* is assigned to the group of *Apoptosis regulator Bcl*). Furthermore, an ambiguous term ending with a number is removed from a group object given that the group object contains further terms that are identical to the ambiguous term except for the final number. Again, group objects are merged if they share 70% of their terms. Ambiguous synonyms ending with a letter (A-F) or a number followed by a letter are assigned only to the group objects of the lower hierarchy level, other ambiguous synonyms are assigned only to the group object of the higher hierarchy level.

(4) **Ambiguities between groups and single objects** are reduced by removing the synonyms which are ambiguous between single and group objects from the single objects provided the single object is mapped to the group object.

This procedure finally returns a hierarchical synonym dictionary that contains the original gene and protein objects as well as the newly generated group objects, and a mapping between group objects and their corresponding original gene or protein objects.

3.4.2 Evaluation

A Benchmark for Gene/Protein and Hierarchical Synonym Dictionaries

A hierarchical synonym dictionary has been generated by application of the above procedure on a human synonym dictionary derived from HUGO ([Eyre et al., 2006](#)) and Swiss-

Prot (Bairoch *et al.*, 2005). ProMiner (Hanisch *et al.* (2003, 2005), see Section 4.3) was applied for named entity identification with the hierarchical synonym dictionary.

200 MEDLINE abstracts have been annotated manually with 746 occurrences of 486 distinct gene objects and 219 occurrences of 121 distinct group objects. Two thirds of the abstracts have been selected randomly and one third has been selected specifically for relevance of gene groups. These were abstracts in which ProMiner identified at least five distinct objects and for which the results with the standard synonym dictionary and the hierarchical synonym dictionary differed.

The evaluation of hierarchical synonym dictionaries requires for specific criteria: If a synonym occurrence is ambiguous between a single object and a group object either match is accepted; i. e., the occurrence is counted once (56 cases). If the text mentions a specific member of a group and the automatic search returns the specific member as well as the group, the match of the group is ignored (96 cases). If a specific member of a group is not matched due to unusual or complicated grammatical constructs, the match of the group term is also accepted (15 cases). All other matches of group objects are evaluated without special treatment (123 cases).

Evaluation Results

The evaluation results (Table 3.6) show that the expansion for gene and protein groups increases recall at similar precision, which leads to a remarkable increase in F-measure.

Synonym dictionary	TP	FN	FP	Recall	Precision	F-measure
standard dictionary	656	211	107	0.76	0.86	0.81
hierarchical dictionary	781	86	121	0.9	0.87	0.88

Table 3.6: Results of evaluation of standard gene and protein name dictionary and the hierarchical synonym dictionary expanded for gene and protein groups on the manually annotated benchmark set (TP: true positives, FN: false negatives, FP: false positives).

The detailed analysis of the results on the benchmark set indicate similar categories of errors for the group synonym dictionaries as for the standard synonym dictionaries (see also Chapter 4). The most important category of false negative errors of the standard synonym dictionary (Table 3.7) corresponds to group terms; usage of the hierarchical synonym dictionary clearly increases recall.

Some of the false negative errors are caused by deficiencies of the matching strategy applied by ProMiner such as missing enumeration resolution. Others could be corrected by different parameter settings, improved ProMiner term lists, or improved synonym dictionaries. With the hierarchical synonym dictionary, some group terms are not found due to ambiguous synonyms (3 cases) or parentheses in synonym occurrences (2 cases). Furthermore, some standard objects are not found anymore (5 cases) as the respective synonyms are similar to group synonyms.

Most of the false positive matches (Table 3.8) are caused by correct matching of terms that have various meanings. The addition of group terms removes some false positive matches

Type of error	occurrences	Example
gene/protein groups	123	growth hormone
Missing Synonym	29	PKCepsilon, NRAS
Ambiguous Synonym	19	SMN1
Enumerations	19	ERK1/2
unclear	10	Ang II, CCK A receptor
Parentheses	9	Bcl-x(L), cyclooxygenase (COX)-1
Standard English	2	Bad

Table 3.7: False negative results of the evaluation of the standard gene and protein synonym dictionary on the manually annotated benchmark set.

Type of error	occurrences	Example
semantic ambiguity	63	<i>PtK1</i> vs. PtK1 tissue cells
match within longer expression	16	<i>GSH-S</i> -transferase, <i>c-Jun</i> NH2-terminal kinase
unspecific synonym	14	<i>motor protein, serine/threonine kinase</i>
insertions	6	transcription factor-1 vs. transcription factor IDX-1
permutation	5	<i>alpha 2-M</i> vs. <i>alpha M beta 2</i> <i>IL-2 receptor</i> vs. <i>IL-1 receptor II</i>
token-class weighting	2	<i>diabetes-associated peptide</i> vs. diabetes associated

Table 3.8: False positive results of the evaluation of gene and protein synonym dictionaries on the manually annotated benchmark set. Synonyms of gene or protein objects are marked in *italics*.

(e.g. if a group term achieves a better match score than a standard term) but introduces others (e.g. if a group synonym is unspecific and overlaps with a non-gene term). Most approaches for named entity identification return at each position the longest string that can be identified. Biological objects are often described by nested terms and gene and protein names frequently form part of a longer biological expression (e.g. cell type or mutation). In these cases, context analysis is required to resolve the correct meaning and thus improve performance. Unspecific synonyms are undesired for individual gene and protein objects, yet interesting for gene and protein groups. Thus, the curation of synonym dictionaries could be improved to better distinguish between specific gene and protein objects and various levels of generality of group terms. Permutations of synonym-constituents, insertions and token-class based weighting of the synonym-constituents in rare cases also caused errors.

In summary, the addition of group terms to the synonym dictionary leads to significant improvement of named identity identification. The proposed approach for group term and object generation is based on the application of a set of heuristic rules. The automatic expansion of standard synonym dictionaries is thus straightforward. The evaluation showed that the performance can be increased further. The generation rules can easily be expanded;

for example, by working on terms with subtype descriptors by letters other than A-F, or Roman numbers. Finally, group synonyms could be derived from corresponding protein family or class databases such as InterPro (Mulder *et al.*, 2007).

3.5 Other Dictionaries

Gene and protein name dictionaries make it possible to link text-derived information with experimental data for genes and proteins and thus clearly play the most important role for the approaches described in this thesis. Synonym dictionaries for non-gene and non-protein objects and terms (e.g. organisms, body parts, tissues) provide a means for exploiting context in gene and protein named entity identification and relation extraction (Chapters 4 and 5).

Applications include organism disambiguation, restricting abstracts to those dealing with specific organisms/body parts/tissues, or analysis of common contexts given a set of abstracts (e.g. Section 5.3.5). The abbreviation dictionary is used for inter-dictionary disambiguation (Section 4.4.3), and an interaction term list is used for restricting relations extracted from texts to specific interaction types (Section 5.2). The construction of these dictionaries is described in the following.

3.5.1 Non-Gene and Non-Protein Synonym Dictionaries

Synonym dictionaries for non-gene and non-protein objects have been derived from the Unified Medical Language System (UMLS)³ (Bodenreider, 2004). UMLS is a project initiated by the National Library of Medicine; it integrates terminologies from various data sources including public databases and medical dictionaries and makes data easily accessible via unified file formats, querying tools and a public web interface. UMLS includes a metathesaurus and a semantic network. The metathesaurus defines concepts and relationships between concepts. Each concept is assigned to a semantic type and complemented with one or more synonyms. Relationships between semantic types are defined in the semantic network.

Based on the data contained in the metathesaurus, synonym dictionaries for the following objects have been generated:

- non-gene and non-protein objects
- organisms
- body parts
- tissues
- cell types/cell lines
- diseases

The general *non-gene and non-protein dictionary* has been compiled by integration of all concepts that are not assigned to a semantic type related to genes or proteins (“Amino

³<http://umlsinfo.nlm.nih.gov/>

Acid, Peptide, or Protein”, “Enzyme”, “Amino Acid Sequence”, “Gene or Genome”, “Receptor”, “Hormone”). This dictionary is used for the analysis of synonym dictionary ambiguities (Section 3.3) and inter-dictionary disambiguation in named entity identification (Section 4.4.3). For the other synonym dictionaries, the metathesaurus data has been restricted to the respective semantic types.

The names of non-gene and non-protein objects exhibit less variability than gene and protein names; thus, these dictionaries do not require extensive adding of spelling variants. The synonym dictionaries are curated with a reduced set of curation rules.

For giving a rough estimate of the quality of these synonym dictionaries, named entity recognition performance of the cell-types/cell-lines dictionary has been evaluated on the test corpus of the Bio-Entity Recognition Task at BioNLP/NLPBA 2004 (Kim *et al.*, 2004). This corpus corresponds to a subset of the GENIA (Version 3.02) corpus (Kim *et al.*, 2003) and contains 404 MEDLINE abstracts annotated with BIO-tags. Exact matching of the dictionary resulted in a recall of 77%, precision 72%, F-measure 74% with the right boundary criterion (i. e. a named entity is correctly detected when its right boundary is matched exactly), and recall 81%, precision 76%, F-measure 79% with partial match criterion (i. e. a named entity is correctly detected when one of its words is matched).

The synonym dictionaries for non-gene and non-protein objects have been searched against the entire MEDLINE. The resulting mapping between abstract identifiers and non-gene/non-protein object identifiers is used for the network schemes approach (Section 5.3.5).

3.5.2 Abbreviation Dictionary

Biomedical language is characterized by frequent usage of abbreviations and acronyms (i. e. abbreviations that are formed by combining the first, and sometimes other, letters of the principal words); they make text shorter and easier to read. Many abbreviations are widely adopted, such as PCR, SDS-PAGE, mRNA, HIV. Others are individually defined by the authors and are used in a single article. Thus, every year, thousands of new abbreviations and acronyms appear in the biomedical literature (Adar, 2004; Chang *et al.*, 2002). Generally, when an abbreviation is used for the first time in an article, the phrase is spelled out and followed by the abbreviation in parentheses as in the following example:

Both particulate-phase (PP) constituents including nicotine, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), and N'-nitrosonornicotine (NNN), two tobacco-specific nitrosamines (TSNA), and gas-vapor-phase (GVP) constituents including carbon monoxide (CO), isoprene (IP), acetaldehyde (AA), and ethylene, were studied.
(PMID:17169864)

Most systems for the extraction of abbreviation/long-form pairs are rule-based and take advantage of this convention (Yoshida *et al.*, 2000; Liu *et al.*, 2002a; Nenadic *et al.*, 2002; Yu *et al.*, 2002, 2003). Schwartz and Hearst (2003) extracted the shortest corresponding long form for a given acronym by a straightforward approach and achieved high performance (precision 99%, recall 84% on MEDSTRACT).

Other approaches made use of linguistic preprocessing such as shallow parsing ([Pustejovsky et al., 2001](#)), co-occurrence statistics ([Okazaki and Ananiadou, 2006](#)), rules and statistics ([Zhou et al., 2006](#)), or machine learning ([Chang et al., 2002](#)).

A number of databases for biomedical acronyms and abbreviations exist (e.g. [Rimer and O'Connell \(1998\)](#); [Chang et al. \(2002\)](#); [Zhou et al. \(2006\)](#)); most of them offer online query functionality. For constructing an abbreviation dictionary that is appropriate for text mining, it is advantageous to use data that can be downloaded.

Here, the data compiled by [Gaudan et al. \(2005\)](#) has been used as this had been compiled from the entire MEDLINE; it covers more than 186 000 abbreviations and 623 000 abbreviation/long-form pairs and is available for download. Furthermore, the authors evaluated their dictionary in a disambiguation approach and achieved high performance.

Additionally, a straightforward rule-based approach to extract acronym definitions from texts has been implemented. This can recognize spelling variants and thus compensate for deficiencies of the available dictionaries. This rule-based approach has been applied for the postfilter for gene and protein name identification (see [Section 4.2.2](#)). For inter-dictionary disambiguation in the named entity approach ([Section 4.4.3](#)), the rule-based approach is applied to the texts under investigation and the resulting matches are combined with the dictionary compiled by [Gaudan et al. \(2005\)](#) to construct a comprehensive abbreviation dictionary.

3.5.3 Interaction Term List

An interaction term list is a synonym dictionary where the individual objects represent groups of interaction terms. An object can, for example, group words that have the same stem or words assigned to the same interaction type (e.g. upregulation, chemical reaction, physical interaction). An interaction term list has been compiled by an iterative procedure: A small set of terms was compiled manually. Then, gene names and interaction terms were matched against MEDLINE abstracts. Sentences containing two gene names and an interaction term were collected. From these sentences, patterns were derived by retaining the words between the gene names and the interaction term and replacing the gene names and interaction term by wildcards. The most frequent patterns were then matched against the abstracts. The words that matched at the position of the initial interaction term were sorted by frequency. The top-candidates were inspected manually and, if approved, added to the interaction term list.

Here, the focus is on physical and genetic interactions. By the above procedure a comprehensive set of more than 1000 interaction terms has been compiled. The terms have been grouped by their word stems (175 stems) to yield a term list in synonym dictionary format. The entries of the interaction term dictionary have further been grouped by their meaning (e.g. up-regulation, down-regulation, neutral interaction, chemical reaction). Depending on the application different groupings can be used. The interaction term dictionary is used for the relation extraction approach presented in [Section 5.2](#).

3.6 Applications of Synonym Dictionaries

The synonym dictionaries derived by the procedures described above map names to unique database identifiers and contain information which is useful for numerous applications such as automatic text mining, manual literature research, or report writing. The Internet is an important means for making various kinds of information available to the public.

In the following, several tools are described which make synonym dictionaries accessible and directly usable to the public. These are the *Literature Mine Browser* (*LiMB*, [Göttler \(2006\)](#)) and *LiMB web service*, which have been developed and implemented by Daniel Göttler and Joannis Apostolakis, and the *ProThesaurus and BeThesaurus Web Services*, the *The ProTag client applications* ([Szugat et al., 2005](#)), and the *Prothesaurus Wiki*, which have been developed with and implemented and set up by Martin Szugat.

3.6.1 Literature Mine Browser (LiMB)

The Literature Mine Browser (*LiMB*, [Göttler \(2006\)](#)) is a web-based tool for performing named entity identification based on synonym dictionaries and visualizing the results. It applies exact matching of synonyms derived from a synonym dictionary. Its focus is on result visualization (Figure 3.6); the user can easily browse and curate the results by inspecting match statistics, marked sentences or abstracts. By recording the manual curation steps, information on inappropriate synonyms or required postfilters can be gathered. Multiple users can work and perform search runs in parallel. Data sources as well as results are stored in a database. LiMB proved to be useful during the preparation of the BioCreAtIVe challenge; the inspection of matching results of different versions of the synonym dictionaries against the training data suggested improvements of the curation procedure.

Figure 3.6: The Literature Mine Browser (*LiMB*, [Göttler \(2006\)](#)) supports gene and protein name identification, visualization, and manual curation of results.

3.6.2 The ProThesaurus, BeThesaurus, and LiMB Web Services

Web Services are programs that run on a web server and provide their functionality to the public via the Internet. Web Services use standardized public interfaces according to the SOAP specification (Box *et al.*, 2000). Client applications can call remote programs implemented by a web service by sending XML data to the corresponding web server. The functionality of a web service can thus be integrated directly into a proprietary application.

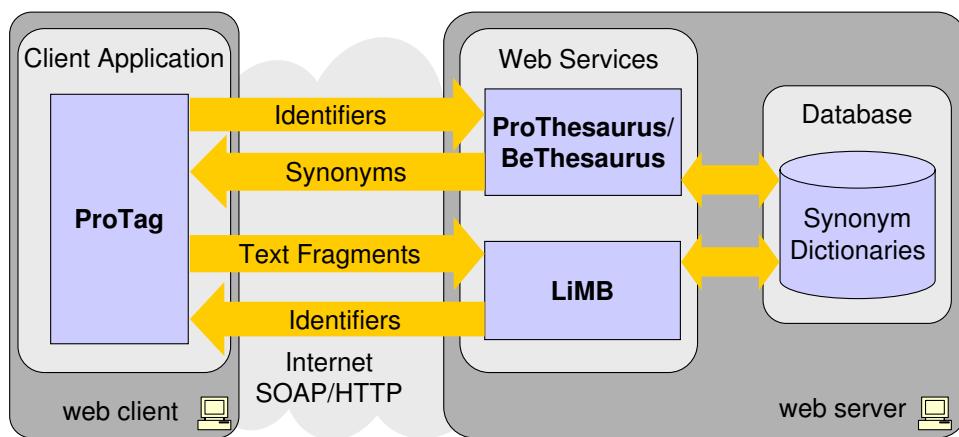


Figure 3.7: The ProThesaurus, BeThesaurus, and LiMB Web Services (Szugat *et al.*, 2005) exchange data with the ProTag client application via the Internet. The web services accept identifiers and text fragments and return synonyms and identifiers, respectively. The required identifiers and synonyms are stored in a database.

Three web services for providing gene and protein identifiers and names and named entity identification have been set up (Figure 3.7, Szugat *et al.* (2005)):

ProThesaurus is a Biological Name Service that maps gene and protein database identifiers to gene and protein names and vice versa. It can be queried with a synonym or identifier and returns the corresponding identifier or set of synonyms, respectively.

BeThesaurus provides functionality similar to ProThesaurus, yet it additionally accepts updates of synonyms. Thus, users can make propositions for additional synonyms or modify existing synonyms. The changes are marked in the database, which makes it possible to cross-check the propositions before making them available via the standard ProThesaurus Web Service.

The **LiMB** web service is a Biological Mark-up Service; that is, a web service for tagging synonyms in free text. This service accepts text fragments, matches these against the synonym dictionaries stored in a database, and returns the identifiers of the found genes and proteins. It is based on the same text matching machinery as the LiMB tool (Section 3.6.1).

3.6.3 The ProTag Client Applications

Microsoft Office applications are frequently used for standard biological data analysis tasks. For example, Excel is used for analyzing expression data and Word is used for writing reports. Often, it is required to know alternative gene or protein names or to map a given synonym to an official database identifier.

The *ProTag Add-in* (Figures 3.7 and 3.8, left panel) is a client application that integrates into the Microsoft Office applications Excel, Word, and PowerPoint and identifies biological objects by means of the ProThesaurus, BeThesaurus, and LiMB web services. The ProTag Add-in makes use of smart tags, a feature of Microsoft Office programs that makes it possible to annotate the content of a document with additional information while the document is being edited or viewed. The respective fragments are then underlined and marked by a tag which may offer actions through a context menu.

The ProTag Add-in automatically tags text fragments representing biological objects (i. e. gene or protein names or database identifiers) in a document: Once the user stops to edit or scroll the document, the currently visible part of the document content is passed to ProTag. ProTag contacts the Mark-up service LiMB, which matches the text against synonyms and database identifiers and returns a list of identifiers, and the Biological Name Service (ProThesaurus), which retrieves synonyms for database identifiers. Then, ProTag adds a smart tag to each matched text fragment. A mouse click on the smart tag displays a context menu which offers actions on the text fragment, such as retrieving synonyms and inserting them into the document as a comment.

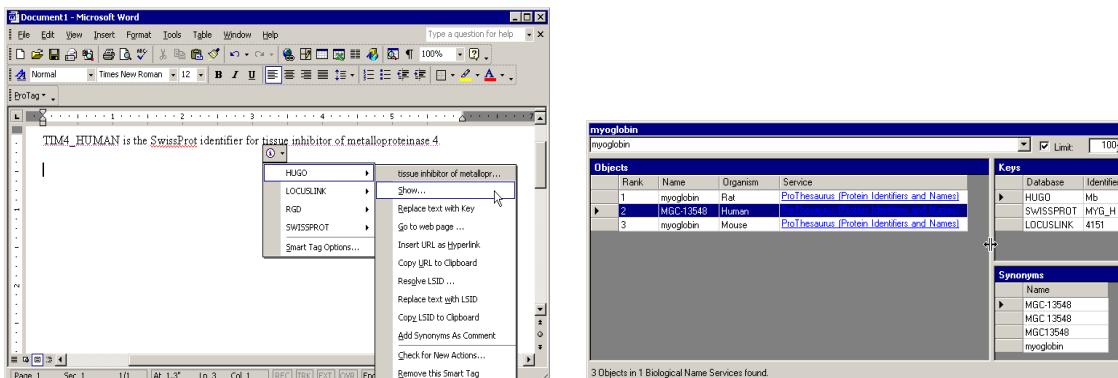


Figure 3.8: Interfaces of the ProTag client applications (Szugat *et al.*, 2005). Left panel: ProTag Add-in in Microsoft Word. Right panel: ProTag standalone client application.

The ProTag Add-In for Excel makes it possible to use the ProThesaurus functionality in Excel formulas. For example, given gene identifiers in a column of an Excel spreadsheet, a function call can be entered as formula in a second column and thus the corresponding synonyms can be retrieved.

The ProTag standalone client application (Figure 3.8, right panel) runs in the background; it accepts user input from the clipboard and queries the web services for synonyms and identifiers. The ProTag application also enables the user to propose new synonyms for

database identifiers and update the synonyms for a given identifier. The changes are submitted via the BeThesaurus Web Service and marked in the database. The user can select to query BeThesaurus in addition to ProThesaurus and thus make use of the latest updates made by the BeThesaurus users.

3.6.4 The ProThesaurus Wiki

Synonym dictionaries need to be maintained to be up to date and comprehensive. New data needs to be integrated, errors need to be detected and removed, and new objects and synonyms need to be added to the dictionary. This implies that the lists need to be updated on a regular basis in order to integrate new databases and database updates. A Wiki⁴ allows users to edit and comment entries in a standardized, comfortable and well-known way (see also: Wikipedia⁵).

Figure 3.9: The ProThesaurus Wiki makes the content of gene and protein name synonym dictionaries available to the public. The user can browse, search, and edit the entries, and search all synonyms of an object in MEDLINE via PubMed or in the Internet via Google.

The entries of the curated synonym dictionaries are available via the ProThesaurus Wiki⁶. Updates and comments are immediately publicly available. The curation of the information maintained in the Wiki together with the automated generation procedure needs to be done regularly.

⁴<http://www.mediawiki.org>

⁵<http://en.wikipedia.org/wiki/Wiki>

⁶<http://prothesaurus.bio.ifi.lmu.de/>

Every object of a synonym dictionary corresponds to an entry in the Wiki. Every entry contains links to the respective entries of the source databases as well as a query facility that allows the user to query MEDLINE via PubMed and Google with all synonyms of a gene/protein simultaneously. The Wiki supports browsing the dictionaries as well as for searching specific entries. This can easily be done by many users simultaneously over the web. Registered users may simultaneously edit and comment entries in a quite flexible way via a convenient and easy to use interface (Figure 3.9). The modifications made by users are periodically checked and it is decided whether to remove or keep the suggested edits. If necessary, the suggestions are discussed with the respective registered users. This way, it can be expected that the curated dictionaries continuously improve and the most current entries are available to the community. Of course, the actual content relies on the cooperation of many users and editors.

3.7 Chapter Summary

Most biological objects, such as genes and proteins, are assigned with several names or identifiers. For applications such as manual literature search, automated text mining, named entity identification, gene/protein annotation, and linking of information from different data sources, it is required to know the names and symbols for a given gene or protein. Various organism-specific or general public databases organize knowledge about genes and proteins. These databases can be used for deriving gene and protein name dictionaries. Gene and protein name dictionaries represent a means for compiling identifiers, symbols, and names for gene and proteins.

In this chapter, a method to automatically derive gene and protein name dictionaries from public databases has been presented. An automatic curation procedure leads to high quality gene and protein name dictionaries. The resulting dictionaries form the basis for the gene and protein name identification approaches presented in the next chapter (Chapter 4). The detailed analysis of gene and protein name dictionaries revealed important differences between the databases the dictionaries were derived from as well as between the organism nomenclatures ([Fundel and Zimmer, 2006](#)). The number of genes/proteins and synonyms covered in individual databases varies significantly for a given organism. All dictionaries show an important yet varying degree of within-dictionary ambiguity. The between-dictionary ambiguity reflects the degree of relationship of the organisms. The degree of ambiguity of gene names with common English words and domain-related non-gene terms varies for the respective organisms and reflects the nomenclature guidelines. Despite considerable efforts of co-curation, the overlap of synonyms in different data sources is rather moderate. The combination of data from several databases returns gene and protein name dictionaries that contain considerably more used names than dictionaries obtained from individual data sources. Curation increases size and decreases ambiguity of the dictionaries. Hierarchical synonym dictionaries provide a means to recover gene or protein groups, or genes/proteins that are not fully specified in a text. A method has been presented that generates hierarchical synonym dictionaries by application of heuristic rules to the stan-

dard gene and protein name dictionaries ([Friedel, 2003](#); [Donner, 2003](#)).

Dictionaries for non-gene and non-protein objects (here: organisms, body parts, tissues, cell types/cell lines, and diseases) and abbreviations have been compiled, and an interaction term list has been compiled. These are useful for context filtering and analysis, inter-dictionary disambiguation, and interaction extraction, respectively.

The gene and protein name dictionaries obtained from the combination of several data sources and subjected to the curation procedure are publicly available via several tools. LiMB ([Güttler, 2006](#)) makes use of the synonym dictionaries for text mining and focuses on user-friendly visualization of results. The ProThesaurus and BeThesaurus web services support automatic querying. The ProTag client applications enable users to query the synonym dictionaries via the web services from within Microsoft Office applications and by a standalone tool ([Szugat *et al.*, 2005](#)). The ProThesaurus Wiki enables users to query synonym dictionaries, search MEDLINE and the Internet with synonyms, and to update the Wiki content.

In the next chapter (Chapter 4), methods for high quality gene and protein identification based on the synonym dictionaries derived here will be presented.

Chapter 4

Gene and Protein Name Identification

Scientific articles are one of the main sources for biomedical information; an important part of biological knowledge is only available in free text. Due to the enormous amount of literature, it becomes necessary to exploit texts and extract information automatically. The identification of gene and protein names is one of the most important tasks in biomedical text mining. Often, this is required as a preprocessing step; for example, when aiming at involved information extraction, relation detection, or integration of text data with data from other sources.

In this chapter three modular systems for gene and protein name identification are presented (Figure 4.1); all of them rely on synonym dictionaries, which can be generated by the methods described in the previous chapter (Section 3.2).

The exact matching approach (Figure 4.1 a, Section 4.2, Fundel *et al.* (2005a)) directly evaluates the applied synonym dictionary. Various postfilters can be applied for increasing precision. This approach has been developed together with Joannis Apostolakis and Daniel Güttler.

The ProMiner system (Figure 4.1 b, Section 4.3, Hanisch *et al.* (2005)) expands on the tool ProMiner (Hanisch *et al.*, 2003) which identifies gene and protein names by approximate matching of synonyms. Here, it has been expanded by specific preprocessing of the synonym dictionary and postfiltering; thus, it has been adapted to the naming conventions of yeast, mouse, and fly. The expansion is joint work with Daniel Hanisch, Theo Mevissen, and Juliane Fluck.

The combined system (Figure 4.1 c, Section 4.4, Fundel and Zimmer (2007)) integrates the matching results of exact matching and ProMiner and implements inter-dictionary and intra-dictionary disambiguation.

All three approaches have been evaluated in the BioCreAtIvE challenge evaluations (Section 4.5) and showed very good performance.

4.1 Introduction and Literature Review

Gene and protein name identification is concerned with finding occurrences of a gene or protein name in a text and returning the corresponding unique identifier for the gene or protein together with the detected text fragment. Most genes/proteins have multiple names; gene names show high variability, are often ambiguous and can overlap with English words (see Section 3.3). Therefore, gene name identification is a difficult task. Gene and protein names of different organism vary significantly. Accordingly, names of different organisms exhibit varying degrees of difficulty for text mining.

Several approaches have been proposed to tackle gene and protein name identification. Machine learning, such as support vector machines and hidden Markov Models (HMMs) ([Bunescu et al., 2003](#)), or HMMs for gene mention tagging followed by normalization according to various filters ([Morgan et al., 2004](#)) has successfully been applied. Other methods focus on linguistic techniques (e.g. [Tanabe and Wilbur \(2002\)](#)).

Numerous methods make use of gene and protein name dictionaries which can be extracted from databases, ontologies, and other data sources ([Ono et al., 2001](#); [Hanisch et al., 2003](#); [Koike and Takagi, 2004](#); [Tsuruoka and Tsujii, 2004](#)).

Some methods rely on the combination of dictionaries and linguistic methods, such as ProtScan, that combines a dictionary based approach and a specialized tokenization approach ([Egorov et al., 2004](#)). BLAST ([Altschul et al., 1997](#)), a tool for DNA and protein sequence comparison, has also been used for matching gene and protein names against texts ([Krauthammer et al., 2000](#)). An overview of biological named entity extraction and a lexical matching exercise that depicts specific problems of fly synonyms genes has been presented by [Hirschman et al. \(2002a\)](#).

Disambiguation, that is identification of the correct meaning of an expression out of a set of possible alternatives, plays an important role in gene name identification. Several studies concerning disambiguation of abbreviations as well as words or even longer expressions have been presented. Most of them are based on machine learning, which requires annotated training data (e.g. [Hatzivassiloglou et al. \(2001\)](#); [Liu et al. \(2002b\)](#)).

The BioCreAtIvE¹ (Critical Assessment of Information Extraction systems in Biology) challenge evaluation ([Hirschman et al., 2005b](#)) is a community-wide effort for evaluating text mining and information extraction systems applied to the domain of biomedical literature. The first challenge was conducted in 2004, it consisted of three tasks: Task 1A evaluated the recognition of gene and protein names in texts. Task 1B has been set up to assess the ability of automated systems to identify names of genes and gene products and normalize them by association of a unique identifier for each gene/gene product ([Hirschman et al., 2005a](#)). The focus was on yeast, mouse, and fly. Task 2 contained several subtasks which concerned the assignment of GO-terms based on text analysis.

The second BioCreAtIvE challenge was organized in 2006. Amongst other tasks, this challenge evaluated human gene name normalization.

¹<http://biocreative.sourceforge.net/>

The selected organisms are of high general interest: They are among the experimentally most intensively studied organisms. Yeast, mouse, and fly are frequently used as model organisms to elucidate pathways and molecular interactions that might play a role in human diseases. As many scientific publications deal with these organisms, a reliable gene/protein name identification method would be a significant advance for information retrieval and extraction.

Clearly, the BioCreAtIvE challenge represents a substantial progress for the domain as it enables researchers to evaluate their systems on a blind prediction basis and for an independent test set. The challenge and the provided data sets make it possible to compare approaches. Yet, the first evaluations also suffered from limitations; for example, the test sets of 250–262 abstracts are still small compared to the over 16 million citations in MEDLINE, and the annotations are questionable in a number of cases.

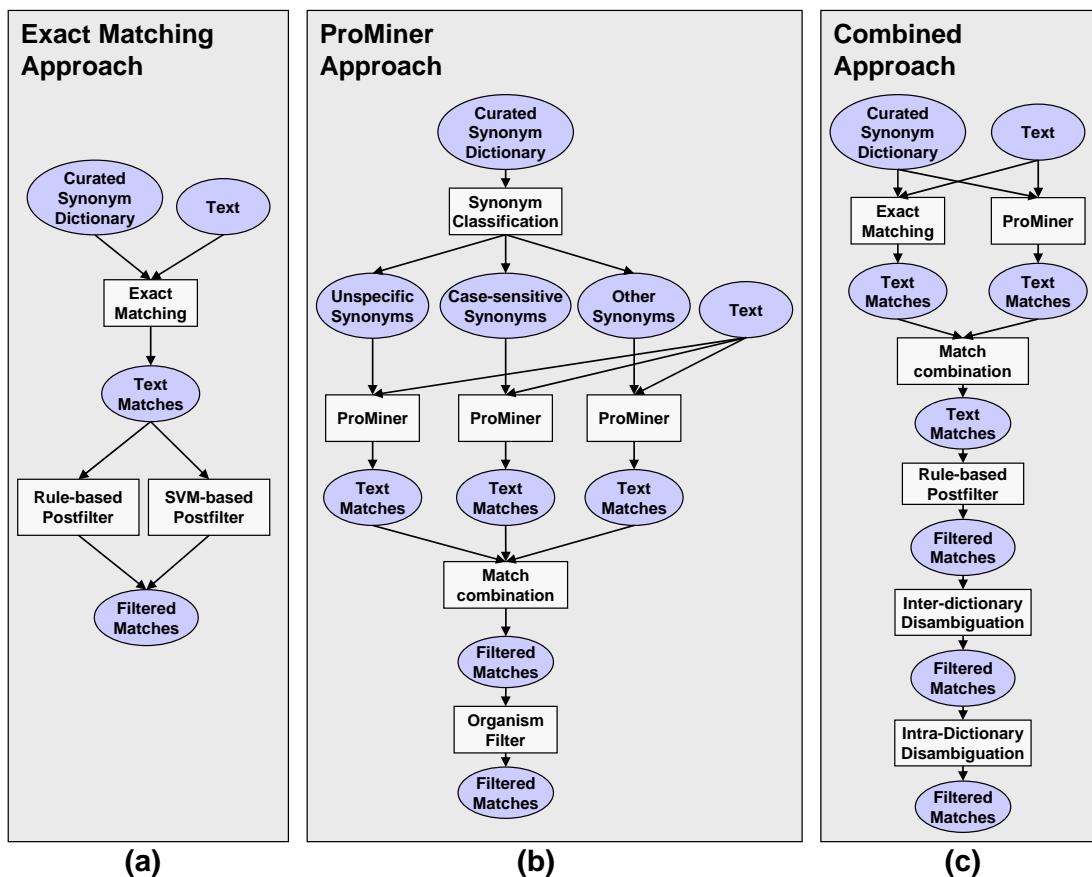


Figure 4.1: Workflow of the gene name identification systems applied for the BioCreAtIvE gene normalization tasks. All systems make use of extensively curated synonym dictionaries. The systems search synonyms against texts and return, for each found gene or protein, the unique identifier together with the detected text fragment.

The goal of this work was to develop methods for gene name identification that are applicable to large collections of text, achieve good performance with balanced recall and precision, and can be customized to meet the requirements imposed by specific organism nomenclatures.

To this end, the systems described in the following have been set up. The BioCreAtIvE challenges were an ideal scenario to evaluate the approaches. With the exact matching approach applied in the first challenge, the quality of the applied synonym dictionaries could be demonstrated. For ProMiner, the parameters were investigated, and thus the ProMiner system has been adapted for the application with yeast, mouse, and fly synonyms. By using comparable synonym dictionaries, the exact matching approach could be compared with the ProMiner approach in terms of recall and precision, as well as runtime and ease of use. In the second challenge, the combined system has been applied. Here, the focus was on the evaluation of the proposed approach for disambiguation.

4.2 The Exact Matching Approach

The exact matching approach (Figure 4.1 a) performs an exact text search for synonyms from a dictionary against text and thus directly evaluates the applied synonym dictionary; postfilters can be applied for increasing precision. It has been developed together with Joannis Apostolakis and Daniel Gütter ([Fundel et al., 2005a](#)).

4.2.1 Match Detection

Synonyms as defined in the synonym dictionary are searched in texts by exact text matching. The search is case insensitive only if the synonym contains numbers or if the synonym length is above a certain threshold (here: 5 characters). When several synonyms of different length can be matched at a certain text position only the longest match is reported.

4.2.2 Rule-Based Postfilter

A rule-based postfilter has been set up in order to implement basic context sensitivity. It checks occurrences of synonyms for nearby occurrence of modifiers (e. g. *cells*, *domains*, *cell type*, *DNA binding site*) which indicate that a text fragment does not refer to a gene or protein.

Short synonyms in parentheses often overlap with definitions of abbreviations differing from the assumed protein as in the following examples: “... mapped by fluorescence in situ hybridization (FISH) ...”, “... developing mouse submandibular gland (SMG) ...”. *Fish* and *SMG* are valid mouse gene names, but in the text, these terms do not refer to genes. The meaning of such occurrences is clarified by checking the words ahead of parentheses corresponding to the letters of the synonym. If no significant overlap of these words with the alternative names of the assumed protein is found the match is discarded. For example, an alternative name for *Fish* is “five SH3 domains”, and an alternative name for *SMG* is

“small nuclear ribonucleoprotein polypeptide G”. Both alternative names have no overlap with the respective text fragments and the matches are therefore removed from the result set.

4.2.3 SVM-Based Postfilter

Fly synonyms show a significant overlap with common English words, body parts and phenotype descriptions (see Section 3.3.2) and therefore require context dependent analysis. A postfilter which is based on support vector machines (SVM) (Chang and Lin, 2001) has been set for context-dependent pruning of matches.

First, the curated fly synonym dictionary is matched against MEDLINE abstracts. Occurrences of multi-word synonyms are always accepted. Occurrences of single-word synonyms are subjected to the SVM and classified as true or false hits. The SVM uses the following features:

- surface clues (i. e. orthographic properties of the matched synonym): synonym length; whether it contains non-characters, numbers, Greek numbers, capitals, lower-case letters, numbers and letters; whether it consists entirely of capitals, lower-case letters; whether it has a capital after a non-capital; whether the first letter is upper case followed by only lower case letters
- part-of-speech tags (Brill, 1992) of the matched synonym and directly adjacent words
- prefix and suffix of the synonym (the first and last 2 and 3 letters)
- all substrings of length 3 of the synonym

The feature value for the synonym length corresponds to the number of characters of the synonym. All other features are encoded as binary values (e.g. one feature is defined for each possible substring of length three; for all substrings that appear in the considered synonym the corresponding feature value is set to one and all other substring feature values are set to zero).

Furthermore, *scores* are used that indicate how often a word is found close to a correct synonym match. Six categories of words are used: nearest verbs, nearest nouns, and words adjacent to a synonym match; occurrences before and after a match are considered separately.

Scores for nouns and verbs have been determined from the 5 000 abstracts of the fly training set of the BioCreAtIvE challenge (Section 4.5.1). Each sentence that contains a synonym was analyzed and the closest verb and noun (Brill, 1992) before and after the synonym match has been extracted. The correct occurrences were used as positive samples and the false occurrences (false positives) as negative samples. For these verbs and nouns a score has been calculated as described below.

A second set of scores is based on a search of mouse synonyms against approximately 700 000 MEDLINE abstracts. In this data set, words appearing adjacent to synonym occurrences are extracted irrespective of their grammatical class. Since no standard of truth

is available for this data set, every match is assumed to be a positive sample and the adjacent words are extracted. All sentences in which no synonym has been matched are considered as negative samples and all words from these sentences are used for estimating the background word frequency.

The motivation for using scores obtained from searching fly and mouse synonyms against two different sets of abstracts is to exploit more information than given in the annotated training data.

Scores are calculated as:

$$Score_{w,i} = \frac{\frac{Occ_{i+}^w}{tot_{i+}}}{\frac{Occ_{i+}^w}{tot_{i+}} + \frac{Occ_-^w}{tot_-}}$$

where

w	: word (token consisting solely of letters, length ≥ 2 , for the BioCreAtIvE fly set only noun or verb, for the large MEDLINE mouse set of any word class)
$i \in \{\text{before, after}\}$: relative position of word w with respect to the synonym match
Occ_{i+}^w	: number of occurrences of word w at position i in positive samples
Occ_-^w	: number of occurrences of word w in negative samples
tot_{i+}	: total number of words found at position i in positive samples
tot_-	: total number of words found in negative samples

These scores are used as SVM feature values: The directly adjacent words and the closest verbs and nouns before and after a synonym match are extracted. For each category, the score of the word is used as value for the corresponding feature, it is zero if no score is defined for the word.

The SVM uses a linear kernel. The training data for the SVM has been compiled as follows: The fly synonym dictionary has been searched against the BioCreAtIvE training data for fly (Section 4.5.1) by exact matching. 10 000 occurrences of single word synonyms have been compared against the annotation provided by the BioCreAtIvE organizers. Occurrences that are correct according to the BioCreAtIvE annotation are used as positive training samples, and occurrences that are not supported by the BioCreAtIvE annotation are used as negative training samples.

For the prediction, the curated synonym dictionary is matched against the abstracts of the test set. Occurrences of multi-word synonyms are accepted directly. Every match of a single-word synonym is classified by the SVM as positive or negative. A single word synonym is only accepted for an abstract if at least one match of this synonym in the abstract is classified as positive. All occurrences of multi-word synonyms and the accepted single-word synonyms are reported as final result.

4.3 The ProMiner Approach

ProMiner ([Hanisch et al., 2003](#)) implements approximate matching; it has initially been designed for human gene and protein names. The ProMiner framework (Figure 4.1 b) com-

piled for the identification of yeast, mouse, and fly synonyms includes synonym classification, match disambiguation, and organism filtering in extension of the standard ProMiner approach. The extensions have been developed with and implemented by Daniel Hanisch, Theo Mevissen, and Juliane Fluck, partly in response to insights obtained during synonym dictionary development (Section 3.2) and benchmark analyses (Section 3.4). The extended system makes use of the abbreviation dictionary (Section 3.5.2).

In the following, the necessary principles of the ProMiner approach and the extensions for BioCreAtIVe are described.

4.3.1 Principles

Gene names often consist of multiple words and exhibit numerous spelling variants, insertions, deletions, and word permutations. For example, the protein *Interleukin-1 beta* is also described as *Interleukin type 1 beta*. ProMiner implements approximate matching to make allowance for these flexibilities. The match algorithm implemented in ProMiner is based on *token classes*. A *token* corresponds to a sequence of letters or numbers. Each token is intended to represent a word or constituent of the synonym that should not be split up further. ProMiner identifies individual tokens by splitting up the original synonym at specific delimiters, which can be defined as program parameters. Thus, the synonym *Collagen 9-A* is split up in four tokens *Collagen*, *9*, *-*, *A*. A *token class* represents a set of tokens which have similar significance for occurrence detection (Table 4.1). Different token classes have different weights according to their relevance for the protein name. For example, tokens of the class *Modifier* (class contains tokens: *inhibitor*, *ligand*, *antagonist*, etc.) are important (*X* and *X inhibitor* are different objects) and thus have a high weight whereas tokens of the token class *Description* (contains: *chain*, *component*, *product*, etc.) have a low weight. The default token class is *standard*; that is, all tokens that are not explicitly classified are considered as *standard tokens*. This weighting scheme makes it possible to recognize a multi-word synonym even if certain less relevant parts of it are missing in the text. The individual weights are defined in the ProMiner parameter file and can be adapted for each search run individually.

Token class	Description	Examples
Modifier	Semantic-modifying tokens	activator, antagonist, kinase, inhibitor, receptor
Description	Annotating tokens	fragment, molecule, precursor, product, type
Specifier	Numbers and Greek letters	1, 2, 3, alpha, beta, gamma, a, b, c
Delimiter	Separator tokens	(.)!?:/;:
Standard	Standard tokens	TNF, BMP, IL, Collagen

Table 4.1: Definition of token classes in ProMiner. Each token class is assigned with a weight which reflects the relevance of the respective terms for a gene name. All token classes except for standard tokens are explicitly defined by term lists. All tokens that are not contained in one of these term lists are classified as standard tokens.

4.3.2 Match Detection

ProMiner requires as input the text to be searched and a synonym dictionary in the format described in Section 3.2. For obtaining optimal matching results, the ProMiner token classes, the token class weight parameters, and the search parameters can be adapted for each specific synonym dictionary.

At startup, ProMiner preprocesses the synonym dictionary. Each synonym is split up into individual tokens. The individual tokens are then analyzed for the corresponding token classes. A synonym is dropped if the sum of token weights is below a given threshold or if it corresponds to an entry of an exclusion list. All remaining synonyms are then organized in a specific structure for effective text matching. For each synonym S consisting of tokens (s_1, s_2, \dots, s_n) , the *Maximum Score* $M_{max}(S)$ is defined as:

$$\text{Maximum Score : } M_{max}(S) = \sum_{s \in S} c(s),$$

where $c(x)$ is the weight of the token class that token x is assigned to.

For text matching, ProMiner processes one token of text at a time. Text matching is generally case insensitive. If the current token corresponds to a token of a synonym, the corresponding synonym is added to the set of *candidate solutions*. For each subsequent token of text, ProMiner tries to expand the matches. Let T be the sequence of text tokens (t_1, \dots, t_k) that have been analyzed k steps after a synonym S had been added to the set of candidate solutions. The candidate solution is evaluated at the respective text position by two scores:

$$\text{Match Score : } MS(S, T) = \sum_{u \in U} c(u) , \text{ where } U = \{S \cap T\}$$

$$\text{Mismatch Score : } MMS(S, T) = \sum_{v \in V} c(v) , \text{ where } V = \{v | v \in S \cup T \wedge v \notin S \cap T\}$$

Each of the two scores is compared against a respective threshold. The *Boundary Score* controls the end of the extension of a candidate match: Match expansion is stopped if the Mismatch Score rises above $M_{max} \cdot \text{Boundary Score}$, or if a delimiter token that does not form part of the candidate solution is encountered in the text. The *Acceptance Score* controls whether a sufficient fraction of the synonym tokens have been matched in the text so that the candidate solution is considered as a valid match. A candidate solution is accepted as valid match and reported if the following criteria are fulfilled:

$$MS(S, T) \geq M_{max}(S) \cdot \text{Acceptance Score}$$

$$MMS(S, T) \leq M_{max}(S) \cdot \text{Boundary Score}$$

The Boundary and Acceptance Score are defined as percentages, values close to 100% return only matches that are nearly identical to the synonyms in the synonym dictionary while lower scores allow more flexibility in matching. The scores are parameters of a search run and are the same for all synonyms.

Internal Term Lists

ProMiner can be customized by a number of internal term lists. These lists make it possible to define certain synonym or token specific matching options.

A list of *synonymous tokens* contains tokens that are treated as synonymous for matching (e.g. Arabic and Roman numbers). A list of *case sensitive tokens* is maintained for avoiding errors due to case-insensitive matching, which is especially important for terms that overlap with common English words (e.g. WAS, KILLER, BIG). Terms contained in this list are only reported as match if the match is case sensitive.

Basically, ProMiner does not consider the order in which tokens of a synonym are found in a text. *Permutation options* can be used to restrict permutation: a general threshold defines the number of tokens below which no permutation is allowed, and a *permutation control list* contains synonyms that may not be permuted. This is an important feature as the basic ProMiner search algorithm does not consider the order of tokens within a synonym; yet, in numerous cases the order is relevant (e.g. subtype specifier, EC-numbers) for assignment of the correct object.

For most applications, the lists can be used as provided. Yet, in particular cases and for obtaining optimal search results, they need to be adapted to the synonym dictionary under investigation. For example, a fly synonym dictionary shows significantly higher overlap with common English words than a human dictionary, and thus requires case sensitivity for more synonyms.

Disambiguation

The ProMiner search strategy encounters two levels of ambiguities at a given text position; it needs to resolve the synonym and the object to which a given text fragment refers.

It is assumed that a given text position refers to at most one synonym. The matching algorithm of ProMiner can handle several synonym candidates at a given text position. If several synonym candidates overlap at a given text position, ProMiner applies a sequence of rules to decide which candidate to report: (1) the candidate with the higher match score, (2) the candidate with the higher fraction of matches, or (3) the longer candidate. The candidate that first fulfills one of the rules is accepted as potential occurrence. If the candidates do not differ in all three criteria, they are all accepted.

Ambiguous synonyms (i.e. synonyms that refer to more than one object within the dictionary) are resolved according to a *disambiguation threshold*. For each ambiguous synonym match, ProMiner analyzes the current text unit, which generally is a MEDLINE abstract, for occurrences of further synonyms of the respective objects. If a further synonym is found for an object assigned to the ambiguous synonym, this is interpreted as evidence for the synonym under investigation to refer to the respective object. The set of candidate objects is accordingly narrowed down to the set of objects for which additional evidence can be found. The disambiguation threshold refers to the maximum acceptable number of candidate objects that a synonym refers to. The candidate objects are only reported for the synonym match if their number does not exceed the disambiguation threshold.

4.3.3 Extensions for BioCreAtIVe

For the BioCreAtIVe I evaluation, extensions for synonym classification, match disambiguation, and organism specificity have been added to the standard ProMiner approach.

Synonym Classification

The ProMiner parameters apply to all synonyms in the used dictionary. To search synonyms with different parameters, the synonym dictionary is split (*synonym classification*):

Unspecific synonyms are detected by analyzing the occurrence frequency of all stemmed words (Porter, 1980) in MEDLINE. One-word synonyms occurring more frequently than a cutoff are classified as unspecific synonyms. Furthermore, synonyms detected via specific regular expressions during curation are assigned to this class. Unspecific synonyms are complemented with context words (e.g. gene, protein, transcript) and only exact matches are accepted for these expanded synonyms.

Case-sensitive synonyms are defined as the set of synonyms that can only be distinguished from a synonym of another object or an entry of an external dictionary when the case of letters is considered. These synonyms are searched in case-sensitive manner.

The remaining synonyms are searched with standard search parameters.

Disambiguation

An *external dictionary* has been compiled; this contains biological processes and cellular component names derived from the Gene Ontology (Ashburner and Lewis, 2002), fly body parts from FlyBase (Drysdale *et al.*, 2005), and cell types from UMLS (Bodenreider, 2004) (Section 3.5.1). Additionally, a dictionary of abbreviations which do not refer to gene or protein names has been compiled from two sources. First, the Biomedical Abbreviation Server (Chang *et al.*, 2002) was queried for short uppercase synonyms from the gene name dictionaries. Second, abbreviations and the corresponding long forms were extracted from MEDLINE (Section 3.5.2). The long forms of all abbreviations are checked against the gene name dictionary in order to prune long forms which correspond to gene names of the considered organism.

The external dictionaries are searched together with the gene and protein name dictionary and ambiguities are resolved as described in Section 4.3.2; that is, at any position only the match yielding the maximum score will be reported. Thus, when the gene name *furrow* and the term *morphogenic furrow* describing a fly body part match a text fragment, the fly body part attains a higher score and is reported for the given position.

For synonyms that are ambiguous within the set of gene or protein terms, only the objects with most additional synonym occurrences within the abstract are reported and the disambiguation threshold is applied.

Organism Filter

The organism filter was set up to filter synonym occurrences that refer to organisms other than those of interest. It is based on the NCBI taxonomy² that contains organism names as well as generalizations organized in a hierarchy. A set of relevant organisms needs to be specified by the user. All other organisms from the NCBI taxonomy are considered as irrelevant.

Organism names as well as the generalizations are searched in the texts. Synonym occurrences for which the given abstract only contains irrelevant organism names or generalizations thereof are removed.

4.4 The Combined Approach

BioCreAtIvE I (2004) demonstrated very good performance for exact matching ([Fundel et al., 2005a](#)) and for ProMiner ([Hanisch et al., 2005](#)). The results of exact matching and ProMiner on the BioCreAtIvE II (2006) training data showed an overlap of above 90%. The combined system has thus been set up with a focus on postfiltering and disambiguation (Figure 4.1 c, [Fundel and Zimmer \(2007\)](#)).

4.4.1 Gene Name Detection

The combined approach makes use of exact matching (Section 4.2) and ProMiner (Section 4.3) and merges the synonyms found by either one of the methods into one set of matches. If a match detected by one method is a substring of a match detected by the other method only the longer match is considered. These matches are then subjected to rule-based postfiltering and disambiguation.

4.4.2 Extended Rule-Based Postfilter

The rule-based postfilter described in Section 4.2.2 has been extended significantly. Matches are filtered in the following cases:

- If all occurrences in an abstract are immediately preceded by an organism other than the one of interest.
- If all occurrences in an abstract are preceded or followed by the word *receptor* and none of the synonyms of the respective object contains the word *receptor*.

Furthermore, matches are excluded if one of the first two occurrences in an abstract fulfills the following criteria:

- The match contains a word indicating non-gene meaning (pathway, binding site, region, domain, cell, family, related, syndrome, disorder, etc.) nearby and none of

²<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy>

the occurrences has a word indicating gene meaning (gene, protein, kinase, factor, expression, etc.) nearby.

- The match is followed by a type specifier (Roman or Arabic number or Greek letter) and none of the synonyms of the respective object contains this type specifier.
- If the synonym resembles a chromosomal location (a *p* or *q* followed by a number) and the match has a word indicating chromosomal context (chromosome, region, band, deletion, insertion, etc.) nearby or forms part of a longer expression indicating chromosomal context (e.g. *6p21.3-p22*).
- If the synonym overlaps with a chemical element and the match is followed by a “+” or “-” (e.g. *Ca2+*, *Na(+)*).
- The synonym is identical to the three-letter code of an amino acid and another three-letter code for an amino acid is found.
- The synonym represents a sequence of one-letter code amino acids and at least one of the respective amino acids is also found in three-letter code or full name.

Finally, enumerations are resolved: If a synonym ends on a type specifier (Roman or Arabic number, single Latin or Greek letter) and a match of this synonym is followed by further type specifiers separated by special characters (;/, etc.) or enumeration word (*to*, *and*, *or*), the type specifier of the synonym is replaced in turn by each one of the separated specifiers. If the resulting expression matches a synonym in the synonym dictionary, the respective match is added to the result set.

4.4.3 Disambiguation between and within dictionaries

Disambiguation deals with resolving the correct meaning of an ambiguous term given the context in which it occurs. Numerous gene names are ambiguous within and between dictionaries and with English and domain-related terms (Section 3.3.2).

In the following, two types of disambiguation are distinguished:

- *Inter-dictionary disambiguation* is applied for synonyms *s* that are ambiguous between two dictionaries *d* and *d'*: $s \in \text{ambiguous}(d, d')$ (see Definition 3.4), where *d* is a gene name dictionary and *d'* is a dictionary of alternative concepts (see below). The task is to identify the correct semantic class of the term (i.e. whether it refers to a gene or not), and the correct object within the respective class.
- *Intra-dictionary disambiguation* is applied for synonyms *s* that are ambiguous within a gene name dictionary *d*: $s \in \text{ambiguous}(d)$ (see Definition 3.4). The task is to identify the correct object (i.e. gene) for the text fragment.

The dictionary-based disambiguation approach is based on the following principle: Given an occurrence of an ambiguous term, the similarity between text fragments that surround the occurrence and the alternative names of the possible genes/non-genes as contained in dictionaries is determined and the object with the synonym that is most similar to the text

fragments is returned.

Terms are represented as feature vectors for estimating the similarity between them. The individual features represent word-stems (Porter, 1980) or 3-grams (i.e. all substrings of a length of three characters derived from all words contained in MEDLINE) and are weighted by their inverse document frequency:

$$idf(s) = \log(N/n) + 1$$

where N : number of MEDLINE abstracts and n : number of MEDLINE abstracts containing string s ; s : word-stem or 3-gram.

The similarity $cossim$ between two terms a and b , which are represented as vectors u and v , is then measured as the cosine of the two feature vectors:

$$cossim(a, b) = \cos(u, v) = \frac{uv}{\|u\| \cdot \|v\|}$$

For *inter-dictionary disambiguation*, a general *alternative concept dictionary* d_{AC} has been compiled. Here, an *alternative concept* is a term that does not describe a specific gene or protein. Therefore, the general non-gene and non-protein dictionary (Section 3.5.1) is combined with the abbreviation dictionary (Section 3.5.2).

Starting from this combined dictionary, a mapping of alternative concepts that is specific for the applied gene name dictionary is derived.

A *mapping of alternative concepts* m_{AC} is a set of tuples (sf, lf) derived from a dictionary of alternative concepts d_{AC} where sf is a term that is equivalent to a synonym $s \in \text{synonyms}(o)$ and lf does not surpass a threshold similarity thr_{AC} to any of the synonyms $s'(o)$:

$$m_{AC} = \{(sf, lf) \in d_{AC} | \exists o \exists s \in \text{synonyms}(o) : sf \sim_{norm} s \wedge \forall s' \in \text{synonyms}(o) : cossim(lf, s') < thr_{AC}\}$$

where: \sim_{norm} : Two names are equivalent if they match each other in a case insensitive way and after any sequence of non-alphanumeric characters has been replaced by a single placeholder (see Definition 3.2). The threshold thr_{AC} can be used to tune the maximum acceptable similarity for a term for representing a concept other than the compared gene name (here: 0.5). Frequently, the terms sf are short forms or abbreviations, whereas the terms lf generally are long forms that describe the alternative concept.

By the above procedure, a set of terms that are used as gene names but also have other meanings is compiled and a mapping of these terms to descriptions of the alternative meanings is generated (for example entries see Table 4.2).

Synonyms s of a gene name dictionary d are classified into three sets: The first set contains synonyms that are equivalent (according to \sim_{norm}) to terms contained in the alternative concept mapping (S_{inter}); the second set contains synonyms that are ambiguous within the gene name dictionary and are not contained in the first set (S_{intra}); and the third set

sf	Assigned gene	Alternative concepts
ADSS	adenylosuccinate synthase	Alzheimer's Disease Symptomatology Scale
AF	antisecretory factor 1	Articular Disease Severity Score
AGPS	alkylglycerone phosphate synthase	abdominal fat, atypical fibroxanthoma
CCl4	chemokine (C-C motif) ligand 4	acute gallstone pancreatitis
cocoa	calcium binding and coiled-coil domain 1	Carbon tetrachloride
COPD	coatomer protein complex, subunit delta	cacao plant, Theobroma cacao
sar	sarcosine dehydrogenase	chronic obstructive pulmonary disease
TAT	tyrosine aminotransferase	scaffold attachment region
		thematic apperception test

Table 4.2: Examples for short forms (sf), assigned gene names, and alternative concepts as used for disambiguation between gene names and non-gene terms.

contains the rest of the synonyms; that is, unique synonyms (S_{uniq}):

$$\begin{aligned} S_{inter} &= \{s \in \text{synonyms}(d) | \exists sf \in m_{AC} : s \sim_{norm} sf\} \\ S_{intra} &= \{s \in \text{synonyms}(d) \setminus S_{inter} | \exists o' \neq o \exists s' \in \text{synonyms}(o') : s \sim_{norm} s'\} \\ S_{uniq} &= \text{synonyms}(d) \setminus (S_{inter} \cup S_{intra}) \end{aligned}$$

While the meaning of synonyms contained in the set S_{uniq} is clear, the meaning of occurrences of synonyms contained in the sets S_{inter} and S_{intra} has to be resolved. The sense of terms that are equivalent to gene names and non-gene terms (S_{inter}) is resolved by inter-dictionary disambiguation. Gene names that are ambiguous for different genes (S_{intra}) are subjected to intra-dictionary disambiguation.

Inter-dictionary and intra-dictionary disambiguation are based on the same principle: An abstract that contains an occurrence of a synonym to be resolved is compared with the alternative names of the respective alternative concepts or genes and the alternative concept or gene to which the text is most similar is reported. The approach can be considered as a variant of a k-nearest neighbor (kNN) approach with k=1.

Let $LF(s)$ be the set of long forms which have a short form sf that is equivalent to s :

$$LF(s) = \{lf \in m_{AC} | sf \sim_{norm} s\}$$

Let $O(s)$ be the subset of objects that s or a synonym s' that is equivalent to s is assigned to:

$$O(s) = \{o' \in O | \exists s' \in \text{synonyms}(o') : s' \sim_{norm} s\}$$

Then, let T be the set of noun phrase chunks (Smith *et al.*, 2004; Ngai and Florian, 2001) extracted from the text (e.g. an abstract) containing a match to be disambiguated. The objects/long forms for which the similarity with the text is maximized are determined:

$$O_{fin}(T, s) = \left\{ o \in \text{objects}(s') | (t', s') = \arg \max_{\substack{t \in T, \\ s'' \in (\text{synonyms}(O(s)) \cup LF(s))}} \text{cossim}(t, s'') \right\}$$

The synonym match is assigned to the objects O_{fin} . The match is reported only if the maximum similarity is above a threshold (thr_{sim} , here: 0.5) and achieved by not more than a defined number of objects (thr_{obj} , here: 1). Thus, the match is pruned if the similarity is maximized for an alternative concept, if the maximum similarity is achieved by a gene object but is below the threshold thr_{sim} , or if the synonym cannot be disambiguated to a single object. The parameters thr_{sim} and thr_{obj} can be used to tune the stringency of the approach.

4.5 Evaluation

4.5.1 The BioCreAtIvE challenge

The gene normalization task of the BioCreAtIvE challenge evaluation represents an ideal scenario for independent evaluation of the presented named entity identification approaches. The task was to automatically identify gene names in a set of abstracts and normalize them by association of unique database identifiers. The participants had to return unique object identifiers together with the relevant text fragments and the respective article identifier. The first challenge dealt with the three important model organisms mouse, fly, and yeast. The second challenge concentrated on human genes. For each of the organisms, the following data has been provided by the challenge organizers:

- Synonym dictionary: A gene name dictionary which has been derived from the respective model organism database. This dictionary defined the unique identifier to be returned for evaluation. The participants were allowed to extend and modify the synonym dictionary in any way.
- Training set: A set of 5000 abstracts automatically annotated with gene identifiers. The annotations have not been checked manually and are thus not necessarily fully correct and complete.
- Development test set: A set of manually annotated abstracts (250 for mouse, 108 for fly, 110 for yeast in BioCreAtIvE I (2004), 282 for human in BioCreAtIvE II (2006)). The annotation quality for this set can be assumed to be similar to the final test set. Thus, this set can be used to estimate the performance on the final test set.

The evaluation has been performed on a *test set* of 250 MEDLINE abstracts for each of the organisms in BioCreAtIvE I (2004) and 262 MEDLINE abstracts in BioCreAtIvE II (2006). The participants were allowed to submit a maximum of three result sets (runs) for each organism. This made it possible to apply different parameter settings. Evaluation was done by the BioCreAtIvE organizers in terms of Recall (R), Precision (P), and F-measure (F) (see Definitions 2.1, 2.2, and 2.4).

4.5.2 Evaluation Settings

The BioCreAtIvE Challenge has been conceived as an ideal scenario for evaluating gene name identification. Our main goals in participating at BioCreAtIvE I (2004) were: (1) evaluation of synonym dictionary quality and (2) evaluation of advanced text matching strategies. For the first goal, we participated with the exact matching approach described in Section 4.2 (team 24). For the second goal, we participated with the ProMiner system described in Section 4.3 (team 16). Our main goal in participating at BioCreAtIvE II (2006) was to evaluate the disambiguation approach (see Section 4.4) implemented in the combined system (team 34).

Parameter Settings

The **exact matching approach** has initially been designed as a straightforward approach for evaluating the synonym dictionary quality. The postfilter for mouse has been set up for introducing basic context sensitivity, as this was found to be relevant for certain mouse gene names. For fly, no results were submitted to the challenge evaluation; the results presented in the following have been obtained in a post-evaluation of the challenge. The presented results of the exact matching approach on the BioCreAtIvE corpora have been obtained with curated synonym dictionaries for the respective organism and the following parameters:

- Yeast: exact matching only.
- Mouse: Run 1: exact matching only,
Run 2 (RF_+): exact matching followed by rule-based postfiltering.
- Fly (SVM_+): exact matching followed by SVM-based postfiltering.

The **ProMiner system** can be tuned by various parameters (see Section 4.3). The investigated parameters and respective settings were:

- Disambiguation threshold: D_i , $i \in 1, \dots, 5$
- Organism filter: O_+ : applied / O_- : not applied
- Significance of “-” at the end of a synonym: S_+ : synonym matches ending on a “-” are removed; S_- : synonym matches ending on “-” are reported

Appropriate parameter settings were determined by analysis of the training and development test set. Combinations of parameter settings were chosen so as to bias the result towards higher recall or precision or a balanced result (Table 4.3).

The **combined system** makes use of the combined results of exact search and ProMiner and focuses on postfiltering and disambiguation (see Section 4.4). We submitted three runs which apply different postprocessing steps for evaluation in BioCreAtIvE II (2006):

Run No.	Focus	Yeast	Mouse	Fly
1	Recall	$D_3 O_- S_-$	$D_3 O_- S_-$	$D_3 O_+ S_-$
2	Precision	$D_1 O_- S_-$	$D_1 O_+ S_+$	$D_1 O_+ S_+$
3	Balanced	$D_3 O_- S_-$	$D_5 O_- S_+$	$D_3 O_+ S_+$

Table 4.3: ProMiner parameter settings applied for the runs submitted for BioCreAtIvE I (2004). D : disambiguation threshold, $O(+/-)$: organism filter, $S(+/-)$: Significance of a dash at the end of a synonym match.

- Run 1 (*inter₊intra₊*): Application of the full postprocessing pipeline; that is, rule-based postfilter, inter-dictionary and intra-dictionary disambiguation. This run evaluates the full pipeline.
- Run 2 (*inter-intra₊*): Application of rule-based postfilter and intra-dictionary disambiguation, but no inter-dictionary disambiguation. This run evaluates the relevance and performance of inter-dictionary disambiguation by comparison against run 1.
- Run 3 (*inter-amb_-*): Application of rule-based postfilter, pruning of all ambiguous gene names, no disambiguation. This run represents a baseline as it only makes use of the extended rule-based context filter. Terms which are ambiguous with the dictionary of alternative concepts are left in the result set. Ambiguous gene names are pruned from the result set.

Generation and Curation of Synonym Dictionaries

For BioCreAtIvE I (2004), the synonym dictionaries for yeast and mouse were created on the basis of the lists provided by the BioCreAtIvE organizers. These lists were originally extracted from the corresponding organism specific databases, Saccharomyces Genome Database (SGD, [Christie et al. \(2004\)](#)) and Mouse Genome Database (MGD, [Blake et al. \(2003\)](#)). The primary fly synonym dictionary was extracted directly from FlyBase ([Drysdale et al., 2005](#)) and provided by Daniel Hanisch and Juliane Fluck.

The provided lists have been curated to cover additional, frequently used synonyms and remove unspecific and inappropriate synonyms according to the procedure described in Section 3.2.2. The standard curation procedure has been adapted to BioCreAtIvE as follows: For yeast, each synonym is added with the extension “p”. For fly, common words are not removed. For all organism, synonyms that produced many false positive but no true positive matches in the training data are removed. The results on the provided hand-curated training set were analyzed manually and some obvious but missing synonyms are added (about 15 synonyms). For the exact matching approach, ambiguous synonyms are pruned from the dictionaries.

The synonym dictionaries used for exact matching and ProMiner were similar, but not identical. For yeast and fly, the synonym dictionaries were curated by the same procedure yet with slightly different parameters and term lists. For mouse, the synonym dictionary

was the same except for ambiguous synonyms which were pruned for the exact matching approach but not for the ProMiner system.

For BioCreAtIvE II (2006), a human gene and protein name dictionary has been compiled from HUGO ([Eyre et al., 2006](#)), Swiss-Prot ([Bairoch et al., 2005](#)) and Entrez Gene ([Maglott et al., 2005](#)) as described in Section [3.2.1](#) and subsequently curated as described in Section [3.2.2](#).

4.5.3 Evaluation Results

General Performance

The results of our systems in the BioCreAtIvE gene normalization tasks (Table [4.4](#)) are among the best achieved results. The results are discussed in detail in the following sections.

Differences between Organisms

The overview of all results submitted for the BioCreAtIvE I (2004) and II (2006) gene normalization tasks (Figure [4.2](#)) indicates differences between the organism nomenclatures: For yeast, the achieved F-measures are generally higher than for the other organisms, and the distribution of F-measures is most narrow. For fly, the maximum F-measure is comparable to mouse and human, but the range of attained values is much broader than for the other organisms.

The results in BioCreAtIvE demonstrate the different degrees of difficulty for protein name identification for several organisms. Yeast has a quite precise nomenclature comprising mainly specific single word synonyms; mouse has many multi word protein names, and fly has synonyms that overlap with standard English words and anatomic descriptions (see also Section [3.3](#)).

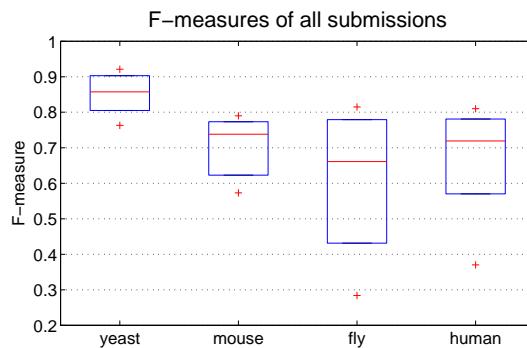


Figure 4.2: Evaluation results of all BioCreAtIvE I (2004) and II (2006) gene normalization task submissions (boxes indicate 25% and 75% percentiles, red lines mark medians, and plus signs mark the minimum and maximum of achieved F-measures).

	Parameters	R	P	F
Yeast				
BioCreAtIvE I (2004)	Exact	0.878	0.917	0.897
	ProMiner $D_3O_-S_-$	0.848	0.951	0.897
	ProMiner $D_1O_-S_-$	0.84	0.966	0.899
	Mouse			
	Exact	0.796	0.735	0.764
	Exact RF_+	0.781	0.764	0.773
	ProMiner $D_3O_-S_-$	0.79	0.752	0.771
	ProMiner $D_1O_+S_+$	0.746	0.809	0.776
	ProMiner $D_5O_-S_+$	0.814	0.766	0.79
Fly				
BioCreAtIvE II (2006)	Exact SVM_+	0.737	0.802	0.768
	ProMiner $D_3O_+S_-$	0.841	0.728	0.781
	ProMiner $D_1O_+S_+$	0.8	0.831	0.816
	ProMiner $D_3O_+S_+$	0.834	0.744	0.787
Human				
	Combined $inter_{+}intra_{+}$	0.815	0.792	0.804
	Combined $inter_{-}intra_{+}$	0.847	0.723	0.780
	Combined $inter_{-}amb_{-}$	0.789	0.739	0.763

Table 4.4: Results of our systems in the BioCreAtIvE I (2004) and II (2006) gene normalization task (R: Recall, P: Precision, F: F-measure, bold font: best result of all participants, *Exact*: Exact matching approach, RF_+ : rule-based filter, SVM_+ : SVM-based postfilter, D : disambiguation threshold, $O(+/-)$: organism filter, $S(+/-)$: significance of a dash at the end of a synonym match, *Combined*: Combined approach, including the extended rule-based filter, $inter_{+}$: inter-dictionary disambiguation, $intra_{+}$: intra-dictionary disambiguation, amb_{-} : pruning of ambiguous synonyms; fly results of exact matching have been obtained as post-evaluation).

When analyzing these results, one has to bear in mind that yeast, mouse, and fly were analyzed in the first BioCreAtIvE challenge, where the number of participants and submissions was small (in total 8 groups, yeast: 15 submissions from 7 groups, mouse: 16 submissions from 7 groups, fly: 11 submissions from 6 groups) in comparison to the second challenge (54 submissions). Furthermore, the second challenge was conducted three years after the first; accordingly, the systems and underlying data sources evolved.

Morgan *et al.* (2007) described some baseline experiments to characterize the data set used in the second challenge and to compare it against those of the first challenge. They state that human may be easier than mouse because it has more terms per identifier and fewer identifiers in total. On the other hand, their results also show that ambiguity is higher for human than for mouse, yeast and fly, and precision of their baseline approach for human is comparable to fly, but lower than for mouse and yeast.

4.5.4 Discussion of the Individual Approaches

Exact Matching Approach

For yeast and mouse, the exact matching approach achieved results close to the best submitted overall results (Table 4.4, Figures 4.3 and 4.4). The results for fly, which were obtained as post-evaluation (Figure 4.5), were also close to the best submitted results.

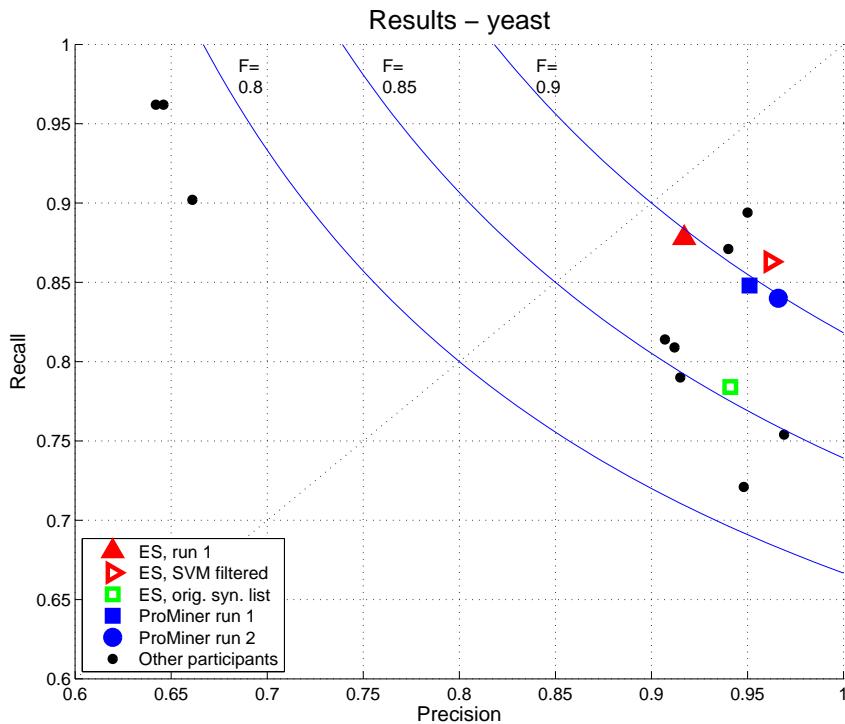


Figure 4.3: Results for yeast in the BioCreAtIvE I (2004) gene normalization task and the impact of synonym dictionary curation. The submitted results (Exact search (ES), run 1; ProMiner run 1, 2) have been obtained with the fully curated synonym dictionary.

For yeast, the difference in F-measure to the best result is 2.4 percentage points (pp). This difference is mainly due to lower precision (3.3 pp), but also recall is somewhat lower (1.6 pp). For mouse, the best result has been achieved with ProMiner. This result has been obtained with essentially the same synonym dictionary as applied for the exact matching approach, the only difference being that the dictionary for ProMiner contained ambiguous synonyms, which were removed for exact matching. The difference between ProMiner and the exact matching approach in F-measure is 2.6/1.7 pp. Some examples of errors in the identification of mouse gene names are listed in the Tables 4.5, 4.6, and 4.7. The errors in the yeast results are similar.

The results of the exact matching approach show that a straightforward approach for protein name identification can be successful. Exact matching of curated synonyms achieves

good recall and precision for yeast and mouse; the evaluation results are only marginally inferior to those of the best evaluated methods.

Curation and exact matching of fly synonyms results in low precision (Figure 4.5). This pinpoints a limit of the exact matching approach; here, context dependent filtering is indispensable. The results after application of the SVM-based postfilter show that this limit can be overcome by additional application of more involved techniques.

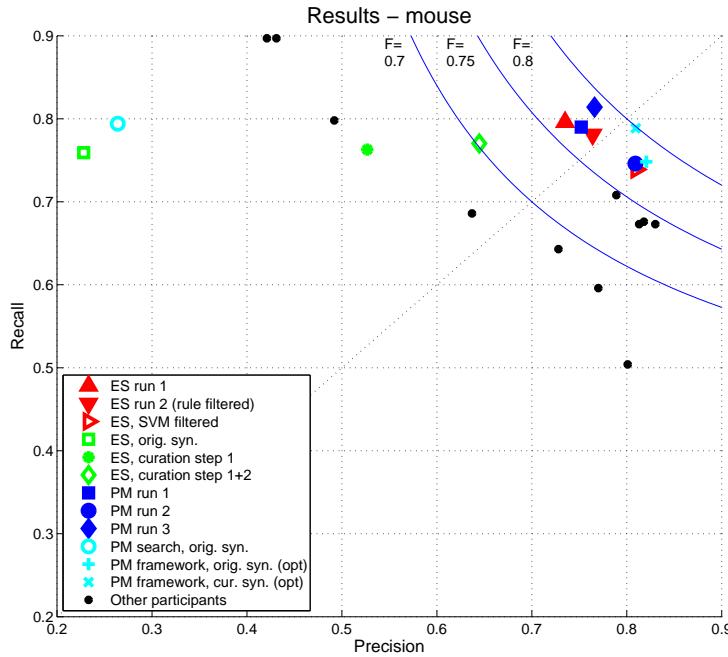


Figure 4.4: Results for mouse in the BioCreAtIvE I (2004) gene normalization task and the impact of curation: Submitted runs, runs performed with dictionaries obtained from intermediate curation steps and different ProMiner parameter settings (ES: Exact search; PM: ProMiner; cur. syn.: fully curated synonym dictionary; orig. syn.: original synonym dictionary as provided by the organizers; PM search: approximate search alone; PM framework: approximate search plus filtering and disambiguation; opt: optimal parameter setting as determined in a post-evaluation).

Curation of synonym dictionary The results obtained with the original non-curated and the final fully curated synonym dictionaries for yeast, mouse, and fly (Figures 4.3, 4.4, and 4.5) show the impact of curation. The results of the fully curated dictionaries of yeast and mouse were those we submitted to BioCreAtIvE.

The curation of the yeast synonym dictionary increases recall significantly while precision decreases slightly (Figure 4.3). For mouse, Figure 4.4 also shows results for intermediate curation steps. A synonym dictionary obtained from steps 1+2 of the curation procedure

Type of error	Examples
overlap with English words	<i>striated</i> muscle, <i>killer</i> cells, <i>Low</i> effectiveness, ...
wrong organism	Mutations in the human <i>doublecortin</i> ...
no direct mention of gene	... inhibits <i>BMP2</i> -mediated induction of ...
description of different object	... the <i>androgen receptor</i> antagonist cyproterone acetate ...
synonym has different meanings	... transgenic mice are <i>growth retarded</i> , is required for normal <i>cardiac morphogenesis</i> ...

Table 4.5: Types of errors and examples of false positive matches. The synonyms and matches (marked in *italics*) are correct but the context reveals that they should not have been reported for BioCreAtIvE gene normalization task.

already yields results which are comparable to those submitted by other groups. The additional execution of the third curation step, namely the removal of inappropriate synonyms based on regular expressions of tokens and the expansion of acronyms and long names yields a further increase in recall and precision. The complete curation procedure significantly increases precision and also slightly improves recall of the mouse synonym dictionary.

For the exact matching approach, all ambiguous synonyms were removed by the curation procedure. Two scenarios have been analyzed for estimating the effect of not removing ambiguous synonyms from the mouse synonym dictionary. If all ambiguous synonyms were retained and all identifiers to which they belong were reported, 24 additional correct matches and 133 false matches would be obtained (R: 84.0%, P: 61.2%, F: 70.8%). If the ambiguous synonyms were disambiguated to the correct objects, which would be the ideal case, 24 additional correct matches and no additional false matches would be obtained (R: 84.0%, P: 74.6%, F: 79.0%).

For fly (Figure 4.5), curation significantly increases precision and slightly decreases recall. The F-measure obtained with the fully curated list is still low (43.1%), which is due to the low precision (29.1%) caused by matches of synonyms which resemble common words and descriptions; a problem that is addressed and largely eliminated by the SVM-based postfilter.

Error analysis The *false positive* matches for yeast and mouse were caused by ignoring the context of the matches; valid synonyms were correctly matched against the text, but the text fragment does not refer to the given gene or protein. Examples for the different categories of false positive matches are listed in Tables 4.5 and 4.7. The rule-based postfilter removes several false positive matches and thus slightly increases precision (see Table 4.7 for examples). Several false positives originate from phenotype descriptions (e.g. *growth retarded*). Detailed grammar or semantic analysis would be required to distinguish between such descriptions and the gene being associated with the phenotype. Other false positive matches have keywords close-by that clearly indicate that the match should not be reported because it refers to a different organism or it is not in the focus of interest (e.g. “human *doublecortin*” or “*BMP2*-mediated”). These matches can be filtered by an extended rule-based postfilter as applied for BioCreAtIvE II (2006) (Section 4.4.2).

Synonym(s)	Occurrence in text	Type of error
Lpa1, Lpa2, Lpa3	lpa(1-3)	enumeration
Pkcb, Pkce	PKC beta, PCK-epsilon	different spelling
retinoic acid receptor, alpha	retinoic acid receptor-alpha	different spelling
interferon gamma	gamma-interferon	inversion
Braf2, Braf-rs1	Braf	ambiguity
peroxisome proliferator activated receptor gamma	peroxisome proliferating antigen receptor gamma	not evident

Table 4.6: Examples of false negative matches: Most similar synonyms in synonym dictionary, occurrence in text, and type of error.

The *false negative* matches (see Table 4.6 for some examples) are mainly caused by missing synonyms, different spellings of synonyms, and ambiguous synonyms. These deficiencies can be tackled by adding more spelling variants during curation, by retaining ambiguous synonyms, or by applying a flexible matching approach such as ProMiner. In some cases genes are mentioned by expressions which have no clear relation to any of the given synonyms. These cases indicate that the content of the respective database used for dictionary generation is not sufficient. The provided synonym dictionary, which has been curated (Section 4.5.2) for generating the submitted results, was derived from a single data source. The analysis of false negative matches of yeast showed that long names of some proteins were used in abstracts while the synonym dictionary contained only the corresponding short names. Some of these long names could have been extracted from the description fields of the *Saccharomyces Genome Database* or *Swiss-Prot*. This observation indicates that the combination of several data sources for dictionary generation, as performed for BioCreAtIvE II (2006), is expected to entail an increase in recall.

The *rule-based postfilter* has been applied on mouse results. It increases precision by 2.9 pp and decreases recall by 1.5 pp. The approach is thus in principle useful but its performance is limited. The filter rules were defined after a crude manual analysis of the results on the training set. The analysis of false positive matches (see examples in Table 4.5) suggest further rules: All matches with a close-by occurrence of words indicating a passing mention (“...-mediated”, “...-activated”, etc.) could be removed; matches that co-occur with organism names other than the organism of interest could be disapproved; part-of-speech tagging could be used to prune matches that are tagged as adjective, and thus descriptions such as “*striated muscle*” could be removed.

Several of these observations were taken into account for improving the rule-based postfilter for the second BioCreAtIvE challenge (Section 4.4.2). Manual compilation of rules is labor intensive, but makes it possible to generate specific rules for certain classes of synonyms or objects. For example, an occurrence of a synonym that is followed by the word *receptor* is presumably acceptable if the respective object is known to be a receptor, but not otherwise, as in the latter case the match will most likely not refer to the assumed protein but to its receptor.

Match in context	Other synonym for identified object	RF
fluorescence-activated cell sorter (<i>FACS</i>)	fatty acid Coenzyme A ligase, long chain 2	y
<i>HEK</i> cells	Eph receptor A3	y
NTera 2(<i>NT2</i>) cell line	zinc finger protein 263	y
polymorphonuclear (<i>PMN</i>) infiltration	progressive motor neuropathy	y
diethylnitrosamine (<i>DEN</i>)	denuded	y
chromosome 2p16- <i>p21</i>	cyclin-dependent kinase inhibitor 1A (P21)	n
<i>E. coli</i> plasmid <i>pCR1</i>	mannosidase 1, alpha	n
area <i>CA1</i> of the hippocampus	carbonic anhydrase 1	n
<i>Eph</i> family of receptors	Eph receptor A1	n
<i>All-trans</i> retinoic acid	retinol dehydrogenase 2	n

Table 4.7: Effect of rule-based postfilter: Examples of false positive matches (marked in *italics*) of the exact matching approach, alternative synonyms for the wrongly identified objects, and the effect of the rule-based postfilter on these matches (y: filtered by rule-based postfilter (RF); n: not filtered).

The identification of fly gene names highlights the limits of the exact matching approach. The nomenclature of fly requires for context-dependent filtering of matches. Here, a SVM has been used to filter matches. This makes use of a number of commonly used features, such as surface clues, part-of-speech tags, and substrings. Furthermore, we exploit the capability of our system to recognize mouse synonyms with satisfying accuracy and speed in large corpora. We determine scores for words appearing close to synonym matches in a large set of MEDLINE abstracts. These scores indicate the frequency of occurrence of the word close to synonym matches. Some examples of the top-ranked words are: interactor, protooncogene, costimulates (category *word directly after synonym match*); heterodimer, transcripts, corepressor (category *noun after match*); exerts, suppresses, encodes (category *verb after match*). These words are strong indicators of a gene/protein mention.

This approach includes information beyond the provided training sets. The SVM-based postfilter proves to be very effective in filtering matches of fly synonyms; it increases precision by 51.1 pp and F-measure by 33.7 pp compared to the exact matching of the curated synonym dictionary without postfiltering.

The analysis of the filtered matches of the evaluation data set showed that most synonyms were either never (e. g. modulo, rough, snake, forked) or always (e. g. to, for, key, gel, lines) filtered. This is almost always correct according to the annotation of the organizers. In some cases context is crucial for correctly classifying results; for example, the word *torpedo* in “...the signals transduced by the *torpedo* product ...” describes a fly gene, whereas in “...the mature Drosophila AChE is closely homologous to that of *Torpedo* AChE.” it describes an organism. These mentions were correctly classified by the SVM-based postfilter. The filter also has a positive effect on the matches of yeast and mouse synonyms (Figures 4.3 and 4.4). A significant advantage of this filtering approach compared to the rule-based postfilter is its independence of manually generated rules and its general applicability.

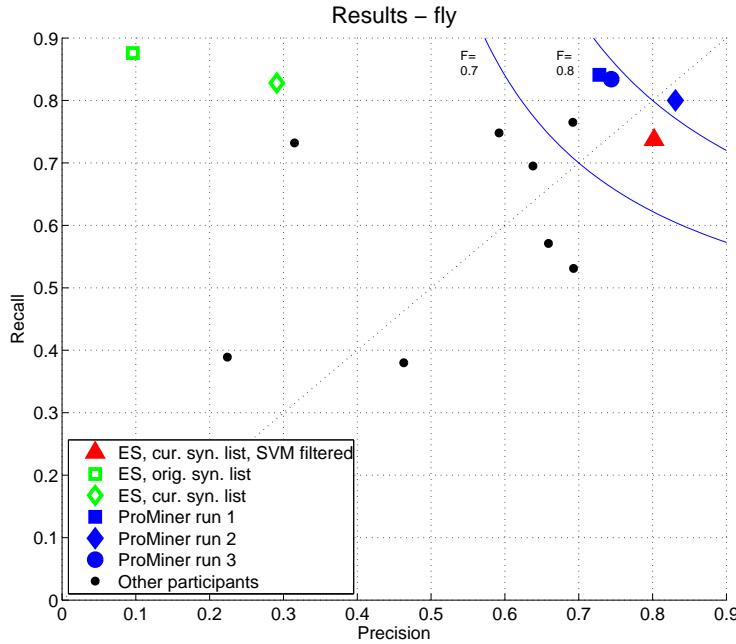


Figure 4.5: Results for fly on the BioCreAtIvE I (2004) challenge data set. Results of the exact matching approach have been obtained as post-evaluation of the BioCreAtIvE assessment (ES: Exact search; orig. syn. list: synonym dictionary as provided and used by the ProMiner team).

Evaluation of ProMiner

The ProMiner framework showed very good performance (Table 4.4). For fly and mouse, setting the disambiguation threshold to one (D_1) and treating a dash as significant (S_+) returned the best results. The organism filter yielded increased precision and unchanged recall for fly. For mouse, the organism filter reduced recall without gaining precision. For *yeast*, most approaches exhibited good performance. Due to the relatively well-defined nomenclature and gene names that are straightforward to recognize, simpler approaches are sufficient to achieve good results and ProMiner shows no significant advantages. For *mouse*, ProMiner achieved the best result of all submissions. The impact of the organism filter was unclear from the training set; only post-evaluation revealed that disambiguation to 1 and no organism filter (O_-) resulted in a further gain of performance ($F: 0.80$). For the non-curated synonym dictionary, the basic ProMiner search with no additional filtering and disambiguation (Figure 4.4, PM search, orig. syn.) returns slightly better results than exact matching. This is due to approximate matching and the ProMiner internal scoring function that eliminates poor matches. The full ProMiner framework includes extensive filtering and disambiguation. With optimal parameter setting this system shows good results even when using the non-curated synonym dictionary ($F: 0.78$, PM frame-

work, orig. syn. (opt)). By using the curated synonym dictionary with the same settings (PM framework, cur. syn. (opt)), the F-measure increases further (F: 0.80). This indicates that even for an advanced approach such as ProMiner the synonym dictionary curation has an important effect on the search result.

The detailed analysis of the results revealed that disambiguation failed in more complicated cases that required decisions for individual genes instead of entire abstracts. In some cases, disambiguation failed because of missing synonyms in the external dictionaries. In some other cases, the gold standard was doubtful as abstracts described organisms other than mouse.

For *fly*, ProMiner achieved the best results of all participants. The best result was obtained with disambiguation threshold one (D_1), the organism filter (O_+) and treating the dash as significant (S_+). The detailed analysis of the results showed that *questionable* and *case sensitive* synonyms are important for achieving high recall and precision (for details see [Hanisch et al. \(2005\)](#)). The organism filter improved results for fly as only matches of a specific subspecies were of interest.

Evaluation of the Combined Approach

We submitted three runs for evaluation in the BioCreAtIvE II (2006) human gene/protein normalization task (Table 4.4). The interquartile ranges (Figure 4.2) indicate that the best submission overall achieved an F-measure of 81.0% and 25% of the submissions achieved an F-measure of above 77.1%. Run 1 of the combined approach was thus ranked in the top-quartile, close to the best overall result.

Figure 4.6 shows the results of the combined approach. The detailed results of all participants were not yet available at the time of writing this thesis. Application of inter-dictionary and intra-dictionary disambiguation (run 1) compared to no inter-dictionary disambiguation and ignoring ambiguous synonyms (run 3) results in an increase in precision (5.3 pp), recall (2.6 pp) and F-measure (4.1 pp). Run 2 yielded highest recall, while, compared to run 1 and 3, precision is slightly lowered.

Again, the synonym dictionary is highly relevant. The original synonym dictionaries provided by the organizers yield low precision (ES orig syn, PM orig syn, 12–35%). Curation leads to a slight increase in recall and an important increase in precision. For the challenge, we used a combined dictionary derived from HUGO, Swiss-Prot and Entrez Gene as described in Section 3.2. Curation was tuned towards recall (e. g. by setting the minimum length for a synonym to two characters, by allowing synonyms consisting of a single letter and a number). Compared to the curated original dictionary (CS cur syn, R: 86%), the combined dictionary achieved significantly higher recall (CS comb syn, 91%), at similar precision (38%). The rule-based filter improves precision (51%) at slightly decreased recall (89%). Applying inter-dictionary disambiguation for deciding whether a term refers to a gene or alternative concept and retaining ambiguous synonyms (CS comb.syn RF *inter*₊) further improves precision (67%) and decreases recall (84%). Similarly, ignoring ambiguous synonyms (Run 3) leads to increased precision, i.e. an important fraction of the false posi-

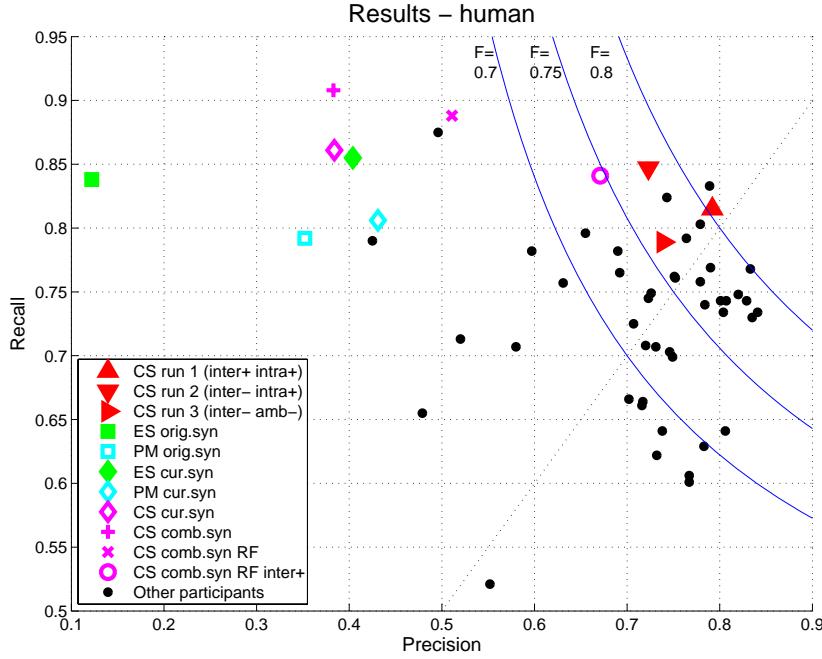


Figure 4.6: Results for human in the BioCreAtIvE II (2006) gene normalization task (ES: Exact Search, PM: ProMiner, CS: Combined System, orig syn: original synonym dictionary as provided by the organizers, cur syn: curated original synonym dictionary, comb syn: curated combined synonym dictionary derived from HUGO, Swiss-Prot and Entrez Gene as described in Section 3.2, RF: extended rule-based filter, *inter₊*: inter-dictionary disambiguation, *intra₊*: intra-dictionary disambiguation, *amb₋*: ambiguous synonyms pruned from dictionary).

tives passing the rule based filter are ambiguous synonyms. The proposed disambiguation procedure yields precision close to using only unique synonyms (72% in run 2 vs. 74% in run 3), but significantly higher recall (85% vs. 79%).

The disambiguation procedure itself has an accuracy of 84%. The inter-dictionary disambiguation procedure and the disambiguation procedure together have an accuracy of 83%. For providing a more detailed description of the relevance of inter-dictionary disambiguation and disambiguation for gene and protein name identification, two hypothetical scenarios have been analyzed: For the first scenario, no inter-dictionary disambiguation is applied, and the performance of the disambiguation procedure is compared against ideal disambiguation: The disambiguation procedure (i. e. run 2, CS comb syn RF *inter₋ intra₊*) achieves R: 85% and P: 72%. If no inter-dictionary disambiguation was applied and the ambiguous synonyms were ideally disambiguated, the result would be R: 88.8% and P: 69.1%. The difference in recall (3.8 pp) is due to deficiencies of the disambiguation approach. The difference in precision (2.9 pp) is due to terms which are filtered by the threshold criteria in the disambiguation procedure. This indicates that an important fraction of terms among

the ambiguous synonyms do not refer to one of the respective genes in the given contexts. In the second scenario, inter-dictionary disambiguation is applied only for deciding whether an occurrence refers to a gene or non-gene term; that is, the threshold similarity is set to zero and any number of objects yielding maximum similarity is accepted (i. e. $thr_{sim}=0$, $thr_{obj} = \infty$). If ambiguous synonyms were ideally disambiguated, the result would be R: 86.9% and P: 76.2% (F: 81.2%). If no disambiguation is applied and all possible objects for an ambiguous synonym are reported, the result is R: 86.9% and P: 64.5% (F: 74%). The difference in precision of 11.7 pp indicates that disambiguation is important for this data. Together, these results confirm that inter-dictionary disambiguation and disambiguation indeed play an important role in gene normalization. We applied a dictionary-based approach for context-dependent disambiguation. Importantly, this approach relies solely on the information contained in the mapping of alternative concepts and gene name dictionary. Thus, the approach does not require annotated training data which is labor intensive to generate for each ambiguous term.

4.6 Conclusions

Three systems for named entity identification have been presented. All are based on searching entries from a synonym dictionary against texts. Their modular structure allows customization with respect to specific applications or specific organism nomenclatures.

The BioCreAtIvE gene normalization challenges (2004 and 2006) were an ideal scenario for evaluating the presented approaches on a blind prediction basis and for independent test sets. In the first challenge, both the exact matching approach and the ProMiner framework showed very good performance. In the second challenge, the combined system proved its very good performance.

The results for yeast indicated that for organisms which have a stringent nomenclature, an approximate matching approach as implemented in ProMiner does not improve the results significantly compared to the exact matching approach. For mouse, the application of exact matching and ProMiner with essentially the same synonym dictionary indicated a slight difference in terms of recall and precision.

The exact matching approach does not need to be adapted for specific synonym dictionaries in terms of parameter tuning or internal lists, which eases straightforward application. Furthermore, it is fast and easy to maintain. For example, the exact search of the yeast synonym dictionary against the 5 000 abstracts of the training set including analysis and report of results takes about 45 seconds on a standard machine. The exact search tool is implemented in Perl; it has less than 750 lines of code and is easy to adapt to different input and output formats. Yet, this approach clearly requires high-quality dictionaries.

ProMiner offers more flexibility in terms of searching. It can be tuned towards recall or precision, depends less on the synonym dictionary curation, implements elementary synonym disambiguation and basic context dependent filtering, but is more difficult to set up and handle. ProMiner has been specifically designed for large scale text mining; it works very efficiently on large text corpora such as the entire MEDLINE. Yet, it takes relatively

long time on small text sets as it extensively preprocesses the synonym dictionary (i. e. tokenization of synonyms, analysis for token classes, and organization in a specific data structure). For example, ProMiner needs about 1.5 minutes for preprocessing the yeast synonym dictionary and 3.5 minutes for the search against the 5 000 abstracts of the training set including filtering and report of results.

In the second challenge, a combined approach that integrates exact matching and ProMiner has been applied. The combined system implements disambiguation with respect to non-gene and non-protein terms as well as within the applied gene name dictionary. The presented disambiguation approach is dictionary based. Importantly, it does not require annotated training data, which is generally labor-intensive to generate, yet achieves good performance.

All investigated approaches make use of gene name dictionaries. The applied dictionaries need to be as complete and correct as possible, even though, due to internal term lists, ProMiner is more tolerant to deficiencies than exact matching. The curation approach described in Section 3.2.2 implements the relevant steps for generating high quality synonym dictionaries. As the curation step is independent of the search, an iterative curation procedure can be established. For example, statistics on search results on an annotated training set can indicate terms that should be removed from the dictionary. This can be done by adding the respective terms to the curation lists.

All presented approaches make use of postfilters. The separation between matching and filtering provides flexibility in the kind of filter applied, and also makes it possible to tune the final result towards recall or precision. The postfilter can be selected in compliance with the characteristics of the synonyms to be searched. For example, fly results without postfiltering are unsatisfactory because of low precision. The usage of the SVM-based postfilter proved to be very effective.

BioCreAtIvE clearly demonstrated the varying degrees of difficulty for identifying gene names of the investigated organisms. The evaluation results of all participants were best for yeast and inferior for mouse and fly.

The BioCreAtIvE organizers estimated the quality of the initial gold standard by inter-annotator agreement. On 30 abstracts, the disagreement was 9% for yeast, 13% for fly, and 31% for mouse ([Colosimo et al., 2005](#)). These numbers roughly reflect the varying nomenclature stringency of the investigated organisms. Interestingly, the disagreement for fly was significantly lower than for mouse, which is not reflected in the submitted results. This can be explained by the fact that a human reader can easily distinguish gene names from common English words, whereas this distinction is difficult for an automatic approach.

The performance values of the presented approaches are already roughly comparable to the inter-annotator agreement of biological experts. Thus, it can be assumed that further performance increases can only be achieved when annotation guidelines are made more precise.

4.7 Chapter Summary

Significant parts of biological knowledge are available only as unstructured text in articles of biomedical journals. By automatically identifying gene and gene product (protein) names and mapping these to unique database identifiers, it becomes possible to extract and integrate information from articles and various data sources.

In this chapter, three systems for named entity identification are described, all of which make use of synonym dictionaries that map database identifiers for each gene/protein to a set of synonyms. The methods for deriving high-quality dictionaries are described in Section 3.2. All systems have been independently evaluated in the BioCreAtIvE challenges and achieved very good results.

The exact matching approach ([Fundel *et al.*, 2005a](#)) applies exact text search of synonyms from an extensively curated dictionary against texts. The approach is straightforward and efficient. The BioCreAtIvE evaluation showed high recall and precision with F-measures of 0.897 for yeast and 0.76/0.77 for mouse. For fly, an F-measure of 0.768 was determined in a post-evaluation. The results of the exact matching approach directly reflect the quality of the applied synonym dictionaries. Depending on the synonym properties, it can be crucial to consider context and to filter erroneous synonym occurrences. This is especially important for fly, which has many gene names that resemble common English words, and thus gene name identification is a challenging task. Rule-based and SVM-based postfilters have been shown to bring about a significant increase in precision, especially the latter proved to be very effective for the identification of fly synonyms.

The ProMiner system ([Hanisch *et al.*, 2005](#)) implements named entity identification based on approximate string matching of synonyms from a dictionary. Here, it has been expanded and adapted for the application with yeast, mouse, and fly synonyms. For example, external dictionaries have been included and it has been complemented by postfilters. In BioCreAtIvE, ProMiner achieved highest F-measures for mouse (0.79) and fly (0.82) among all submitted results, and was also among the top performing results for yeast (0.899). ProMiner can be tuned to meet the requirements imposed by the particularities of certain synonym dictionaries or towards higher recall or precision via numerous parameters. Due to the approximate matching and internal term lists, it depends less on the quality of the synonym dictionary than the exact matching approach. Nevertheless, a carefully curated synonym dictionary leads to a significant increase in performance.

The combined system builds on the results obtained by exact matching and ProMiner. It implements extended rule-based postfiltering and inter-dictionary and intra-dictionary disambiguation ([Fundel and Zimmer, 2007](#)). The evaluation results of this system on human genes in the second BioCreAtIvE challenge (F-measure: 0.804) are among the best submitted results and underline the relevance of context-dependent postfiltering and disambiguation.

High quality named entity identification represents one of the requirements for advanced text mining approaches such as relation extraction or the integrated analysis of texts with data derived from other sources. In the next chapter (Chapter 5), a method for the extrac-

tion of relations between biomedical entities from texts is presented; this method is not limited to, but currently focused on relations between genes and proteins. In Chapter 10, a method for analyzing gene expression data together with text data is described. Both methods rely on the named entity identification approaches described above.

Chapter 5

Gene and Protein Relations

Besides information on individual biological objects such as genes or proteins, information on relationships between biological entities is of high interest for understanding biological observations. Knowledge on relations or interactions can be used for generating network models of regulatory or metabolic pathways; these are useful for advanced data analysis and for understanding cellular and biochemical processes.

In this chapter ReEx, a system for relation extraction, is presented (Section 5.2, Fundel *et al.* (2007)). It builds on the named entity identification methods presented in Chapter 4 and achieves high recall and precision. Section 5.3 discusses its large-scale application, analyzes characteristics of the generated comprehensive gene and protein network, and presents exemplary applications. Finally, an approach for the automatic characterization of relations is presented (Section 5.4, Küffner *et al.* (2006)); this has mainly been developed by Robert Küffner and Timo Duchrow.

5.1 Introduction and Literature Review

Relation extraction approaches take as input texts and return relations according to the type of relations defined by the user; some approaches additionally provide pointers into the underlying literature where evidence for a relation was found. Various approaches for extracting relations from texts have been applied in the biomedical domain.

The *co-occurrence-based* or *bibliometric approaches* rely on the hypothesis that entities which are repeatedly mentioned together are somehow related. Co-occurrences are typically searched in abstracts or sentences (Jenssen *et al.*, 2001; Ding *et al.*, 2002; Jelier *et al.*, 2005). Generally, co-occurrence search extracts large numbers of relations, but a large fraction of the co-occurrences do not describe a relation of interest. For reducing the number of non-relevant relations, the result of a co-occurrence search can be restricted in various ways: (1) by imposing a minimum number of co-occurrences, (2) by defining a cutoff p-value for the statistical significance of an observed number of co-occurrences based on a binomial distribution, or (3) by restricting the search to sentences or abstracts containing

an interaction keyword or some other specific term, which corresponds to a *tri-occurrence search*. Generally, these approaches cannot determine the type and direction of a relation, and all documents in which a co-occurrence was found are of the same value; that is, text fragments cannot be ranked by their relevance for a given relation.

One of the first large-scale analyses of relation extraction in the biomedical domain has been done by [Jenssen et al. \(2001\)](#). They matched a predefined list of gene names against the entire MEDLINE and extracted co-occurrences. The resulting gene network is represented as graph that contains genes as nodes and an edge between two nodes if the corresponding genes co-occur in at least one abstract.

Information extraction techniques consider specific word sequences and/or resolve assertions. These approaches generally achieve higher precision than bibliometric systems at the cost of lower recall. Information extraction approaches can be classified into two broad categories: They are either based on *manually defined rules*, which can be straightforward pattern-based rules or imply detailed natural language processing (NLP) for linguistic text analysis, or based on *machine learning*.

A large number of rule-based methods that apply *pattern matching* against sentences have been presented: The most straightforward approaches require an interaction keyword to occur with a pair of gene and protein names; eventually in a defined order or within a defined range of distances ([Blaschke et al., 1999](#)). Some approaches are restricted to specific interaction words ([Thomas et al., 2000](#)), or specific prepositions ([Leroy and Chen, 2002](#)). A reliability score can be assigned to individual patterns which quantifies the confidence in a relation ([Blaschke and Valencia, 2001](#); [Blaschke et al., 2002](#)). Several of the rule-based approaches focus on specific types of relations, such as phosphorylation events, and extract specific information, such as agent, target, and phosphorylation site ([Hu et al., 2005](#); [Narayanaswamy et al., 2005](#)). Technically more involved systems implement rules in more flexible models such as cascaded finite state automata ([Saric et al., 2006](#)).

Different levels of NLP can be applied: One can make use of POS-tags for defining patterns ([Ono et al., 2001](#)), apply noun-phrase chunking and define generic patterns on closed-class words ([Leroy et al., 2003](#)), or apply patterns with particular grammatical structure and prespecified lists of verbs and nouns ([Domedel-Puig and Wernisch, 2005](#)).

In summary, most of the approaches that match patterns against sentences are rather straightforward and effective; that is, fast to compute and thus applicable on large datasets. The approaches vary in the level of detail of the patterns. The distinction to the NLP-based methods is not always clear-cut as pattern-based approaches often imply NLP, such as POS-tagging and/or chunking, as preprocessing step. The performance depends on the gene and protein name detection and the coverage of the applied patterns. Generally, these approaches achieve significantly higher precision than co-occurrence-based approaches at the cost of lower recall. Manual compilation of rules is labor intensive as exhaustive rule sets are large.

NLP-based systems parse complete articles and semantically classify extracted relationships (Friedman *et al.*, 2001), apply a combinatory categorial grammar (Park *et al.*, 2001) or a context free grammar (Temkin and Gilder, 2003). Methods can also split complex sentences in simple clausal structures; from these, interactions can be extracted based on Link Grammar (Ahmed *et al.*, 2005). Many systems implement multiple steps of text processing; for example, a system applies a parser based on a syntax-semantic hybrid grammar for relation identification and subsequent semantic constraints as filter (McDonald *et al.*, 2004); another system makes use of a specially developed context-free grammar and lexicon for syntactic parsing to construct a set of alternative semantic sentence structures which are then subjected to a domain-specific filter (Daraselia *et al.*, 2004; Novichkova *et al.*, 2003). NLP-based systems generally do not depend on large sets of manually compiled extraction rules, which reduces human compilation effort. The systems can identify relations which are described in quite diverse ways and thus generally achieve better performance than pattern-based approaches. Yet, in many cases, the computational costs of parsing prohibit application on large collections of texts, and parsers need to be carefully selected or adapted to allow for the peculiarities of biomedical language.

Learning-based relation extraction approaches were introduced some years later than the statistical, pattern-based, and NLP-based approaches. They often use simple representations of texts such as the *bag of words* approach, in which text is represented as a vector where each element represents a word and the value of an element indicates the presence or relevance of the corresponding word in the given text. Support Vector Machines have been applied to locate interaction information (Donaldson *et al.*, 2003). Patterns have been learned from texts by aligning sentences by a dynamic programming algorithm (Huang *et al.*, 2004), and patterns have been improved by a minimum description length (MDL)-based pattern optimization algorithm (Hao *et al.*, 2005). Another approach identified syntax patterns describing interactions by sequence alignments applied to sentences and finite state automata optimized with a genetic algorithm (Hakenberg *et al.*, 2005). Machine learning based on a subsequence kernel has also been combined with a statistical co-occurrence approach (Bunescu *et al.*, 2006).

Machine learning-based systems require large annotated training corpora, which must be compiled specifically for the types of relations to be extracted. The performance of machine learning-based systems depends on the size of the applied training data. For obtaining high-quality training data, manual annotation is inevitable.

Most relation extraction methods in the biomedical domain focus on protein-protein interactions. Yet, other types of relations have also been studied, such as gene locations on chromosomes (Leek, 1997), localizations of proteins within the cell (Craven and Kumlien, 1999), and gene-disease relations (Chun *et al.*, 2006).

All approaches dealing with the analysis of individual abstracts or sentences share an intrinsic limitation: They can only detect relationships that are already reported explicitly in the literature. For generating hypothesis on so far unknown relationships, information derived from distinct abstracts can be combined and thus, indirect links among entities can be revealed (e.g. Swanson (1986, 1988, 1990)).

Even though a large number of relation extraction approaches have been presented, only a small number of corresponding corpora (i. e. annotated texts and extracted relation networks) are available.

iProLINK ([Hu et al., 2004](#)) provides curated data sources in the areas of bibliography mapping, annotation extraction, protein named entity recognition and protein ontology development: It includes mapped citations (PubMed ID to protein entry), annotation-tagged literature corpora (e.g. post-translational modifications (PTMs)), a protein name dictionary, word token dictionaries, and protein name-tagged literature corpora.

[Ramani et al. \(2005\)](#) combined and linked interactions from existing databases and generated a co-citation text mining network from 750 000 abstracts. The resulting network contains 31 609 interactions among 7 748 proteins.

Furthermore, public databases for gene/protein interactions, which generally contain manually curated entries, are useful for the evaluation of relation extraction approaches and interpretation of experimental data. For example, Reactome ([Joshi-Tope et al., 2005](#)) focuses on interactions in core cellular pathways, and HPRD ([Peri et al., 2004](#); [Mishra et al., 2006](#)) is biased towards disease-related genes and is currently the largest available database.

The goal of this work was to develop a relation extraction system that is applicable to large collections of biomedical research publications, that is straightforward, provides pointers into the literature where evidence for a certain relation was found, shows high performance with generally balanced recall and precision values, can cope with diverse types of objects, and can be tuned towards specific types of relations.

To this end, RelEx, a tool for relation extraction, has been developed.

5.2 RelEx - Relation Extraction Utilizing Dependency Parse Trees

RelEx is a modular system for relation extraction. It is based on publicly available NLP tools for text preprocessing; importantly, it makes use of dependency parse trees ([Mel'cuk \(1988\)](#); [Klein and Manning \(2002, 2003\)](#), for details see next section). RelEx applies a small set of straightforward rules on the preprocessed text data and achieves very good performance.

Although the RelEx approach is not restricted to genes and proteins and particular types of interactions, in the following the focus is on physical, genetic, and regulatory relations between genes and proteins.

5.2.1 The RelEx Workflow

The RelEx workflow (Figure 5.1) extracts directed qualified relations starting from free-text sentences. RelEx requires a synonym dictionary containing gene and protein names, and a list of restriction terms that are used to describe relations of interest.

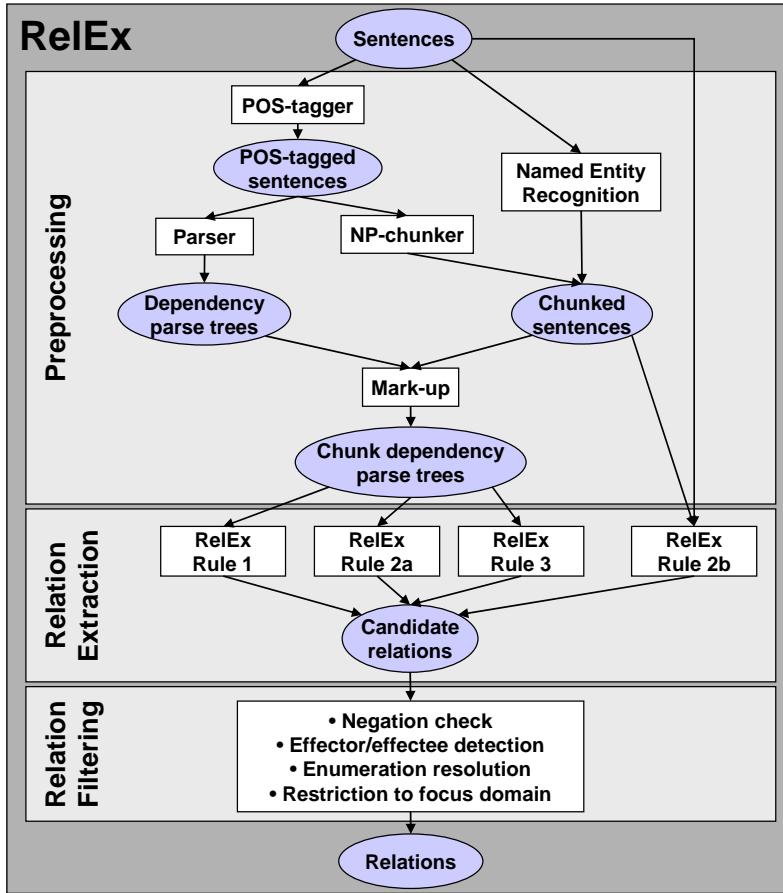


Figure 5.1: The RelEx workflow is subdivided into preprocessing, relation extraction and relation filtering leading from the original free-text sentences to directed, qualified relations. Preprocessing is based on publicly available tools and named entity identification. Candidate relations are extracted according to rules applied on chunk dependency trees and original sentences and subjected to filtering steps.

Text Preprocessing

Sentences are *part-of-speech(POS)* tagged by MedPost¹ (Smith *et al.*, 2004). The POS-tagged sentences are then subjected to parsing and noun-phrase chunking.

Dependency parsing is applied for resolving sentence structures. Dependency grammars represent an alternative to phrase structure grammars. In phrase structure grammar (Example in Figure 5.2, left panel), parse trees contain words only as leaves, and internal nodes are labeled by the class of the word at the respective leaf (e.g. noun, verb, adjective) or phrase class of the respective subtree (e.g. noun phrase, prepositional phrase). The trees reflect nested syntactic structures.

Dependency grammars have been developed by Tesnière (1953). An dependency grammar

¹<ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz>

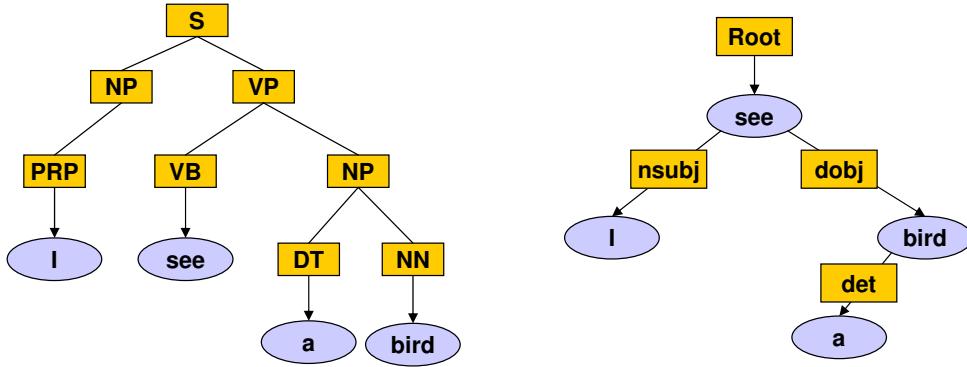


Figure 5.2: Examples of parse trees for the sentence “I see a bird”. Left panel: Phrase structure parse tree showing the syntactic structure of the sentence, the word classes, and phrase classes (S: sentence, NP: noun phrase, VP: verb phrase, PRP: preposition, VB: verb, DT: determiner, NN: noun). Right panel: Dependency parse tree showing words (ellipses), dependencies (edges pointing from the head of a dependency to the dependent word), dependency types (boxes), and the head of the sentence (Root) (nsubj: nominal subject, dobj: direct object, det: determiner).

describes the grammatical dependencies between the words in a sentence construction and represents the resulting structure of a sentence as a directed, labeled tree called *dependency parse tree* (Example in Figure 5.2, right panel). In dependency parse trees, edges connect words directly and words are assigned to leaves and internal nodes; that is, there is a bijection between the nodes of the dependency parse tree and the words of the sentence. In *typed dependency parse trees*, the directed edges between a word and its dependents are labeled with the syntactic role (e.g. *subject*, *object*, *auxiliary*, *modifier*) of the word an edge is pointing to. In a sentence, the verb is seen as the highest level word and thus located at the root of the tree, governing a set of complements, which in turn govern their own complements.

Here, the Stanford Lexicalized Parser² (Version 1.5, Klein and Manning (2002, 2003)) is applied on the POS-tagged sentences to generate a typed dependency parse tree for each sentence. The applied hierarchy of grammatical relations has been described by de Marn-effe et al. (2006). The parser also assigns to each word the word position, which is used for further processing.

Base noun-phrase (NP) chunking takes as input sentences, which might be POS-tagged, and divides them into non-overlapping segments of noun phrases, the base noun phrases. A base noun phrase is a text substring (i.e. a sequence of adjacent word tokens) that contains a noun or pronoun, but no verb, and functions like a single noun or pronoun in a sentence. In RelEx, fnTBL³ (Ngai and Florian, 2001) has been applied for this task. In the following example, the noun-phrase chunks are marked by parentheses:

²<http://nlp.stanford.edu/software/lex-parser.shtml>

³<http://nlp.cs.jhu.edu/~rflorian/fntbl/>

[The sigmaB-dependent promoter] drives [expression] of [yvyD] under [stress conditions] and after [glucose starvation] whereas [a sigmaH-dependent promoter] is responsible for [yvyD transcription].

Named entity recognition is concerned with finding objects in texts. It takes as input sentences and eventually a synonym dictionary and returns a mapping between sentences and found objects together with the sentence positions of the found objects. For extracting relations between genes and proteins, all named entity identification systems described in Chapter 4 are applicable. The results presented in the following were obtained with ProMiner. For detecting objects and terms that are not covered in the available synonym dictionaries, RelEx can extract named entities directly from the noun phrase chunks.

If a noun-phrase chunk contains only part of a multi-word gene or protein name, the chunk is expanded so that it contains the complete name. For each chunk, the corresponding nodes in the dependency tree are combined into a chunk-node returning a simplified *chunk dependency tree* (Figure 5.3).

Relation Extraction

RelEx creates candidate relations by extracting *paths* connecting pairs of objects (e.g. genes/proteins) from dependency parse trees. These paths should contain just the relevant terms describing the relation between the given pair of objects. Currently, three rules are used that reflect the sentence constructs that are most frequently used in English language for describing relations, namely:

- Rule 1: *effector-relation-effectee* (“*A activates B*”)
- Rule 2: *relation-of-effectee-by-effector* (“*Activation of A by B*”)
- Rule 3: *relation-between-effector-and-effectee* (“*Interaction between A and B*”)

Rule 1 (Example in Figure 5.3) extracts paths from the chunk dependency tree that lead from a start point (generally the effector) to an end point (generally the effectee). If the chunk dependency tree contains one or more subject dependencies, the tree is split so that the parent of each subject dependency becomes root of a partial tree; consequently, each resulting partial tree has exactly one subject dependency. The chunks with an incoming edge labeled as subject dependency are marked as *potential start points*. Starting from these, RelEx constructs paths towards the other gene/protein-containing chunks (*potential end-points*). If the dependency tree does not include any subject dependencies all pairs of gene names containing noun-phrase chunks are potential start and end points and thus candidate interaction pairs. For each potential start and end point, the path connecting these two noun phrase chunks is extracted from the chunk dependency tree.

Some of the paths generated by rule 1 are not valid or need to be revised, which is automatically detected and accomplished as follows: A path is invalid if it contains a term occurring after the noun phrase chunk of the end point in the sentence, unless the respective term is

contained in the least common ancestor node of the start and end chunk or is part of an enumeration (see below) with the end chunk. This restriction has been found to reduce the number of false paths, especially for long and complex sentences. It reflects the fact that verb and modifying terms usually occur before the object they refer to.

A path needs to be revised if it contains two nodes tagged as verbs between the least common ancestor and the end node which are directly linked to each other by an *and*, *but*, or *whereas* dependency. In this case the first verb generally is not relevant for the given path but refers to another child node and is therefore removed from the path. For instance, this path revision applies to the sentence “Protein A binds B and inhibits C” where *binds* is not relevant for the interaction between *A* and *C*.

Rule 1 applied on the sentence “This indicates that *the yvyD gene product*, being a member of both the sigmaB and sigmaH regulons, might *negatively regulate the activity of the sigmaL regulon*.” extracts the parts marked in *italics* as candidate relation.

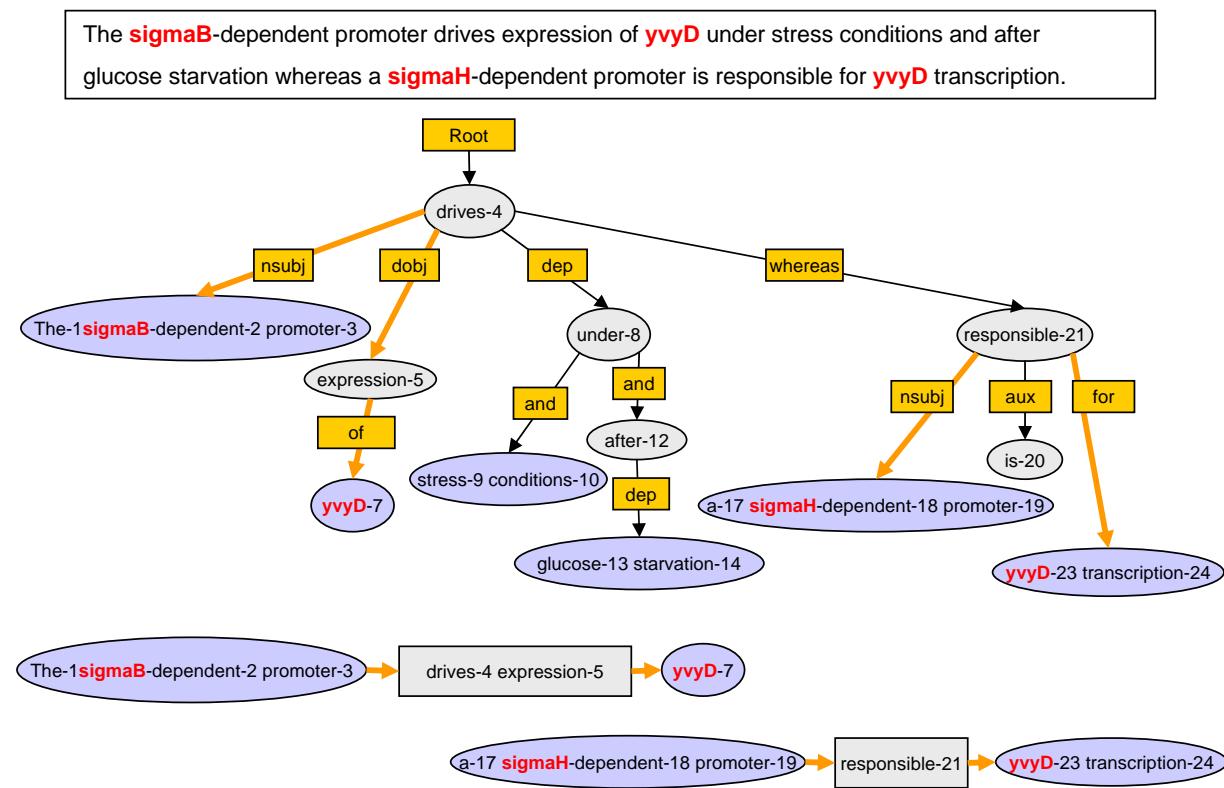


Figure 5.3: Chunk dependency tree which is generated from the dependency parse tree produced by the Stanford Lexicalized Parser; it groups the words in noun phrase chunks (nodes marked in blue). Words marked in bold indicate gene/protein names. Thick yellow edges indicate paths that are extracted by Rule 1. The numbers appended to words indicate word positions in the sentence. The extracted candidate relations are indicated below the chunk dependency tree.

Rule 2a extracts the longest paths through the tree that contain only noun phrase chunks as nodes and dependencies of the types *of*, *by*, *to*, *on*, *for*, *in*, *through*, *with*. The paths containing at least one of these dependencies between two protein name containing chunks are retained as candidate relations (Example in Figure 5.4, left panel).

Rule 2b is similar to Rule 2a, but is applied directly on the chunked sentences. The longest sequences of chunks that are connected by the terms *of*, *by*, *to*, *on*, *for*, *in*, *through*, *with* are extracted. A sequence is retained as candidate relation if it contains at least two of these terms and at least one of these terms between two chunks each containing at least one protein name. Rule 2 extracts relations described like “*Dephosphorylation of SpoIIAA-P by SpoIIE*” or “*sigmaK-dependent transcription of gerE*”.

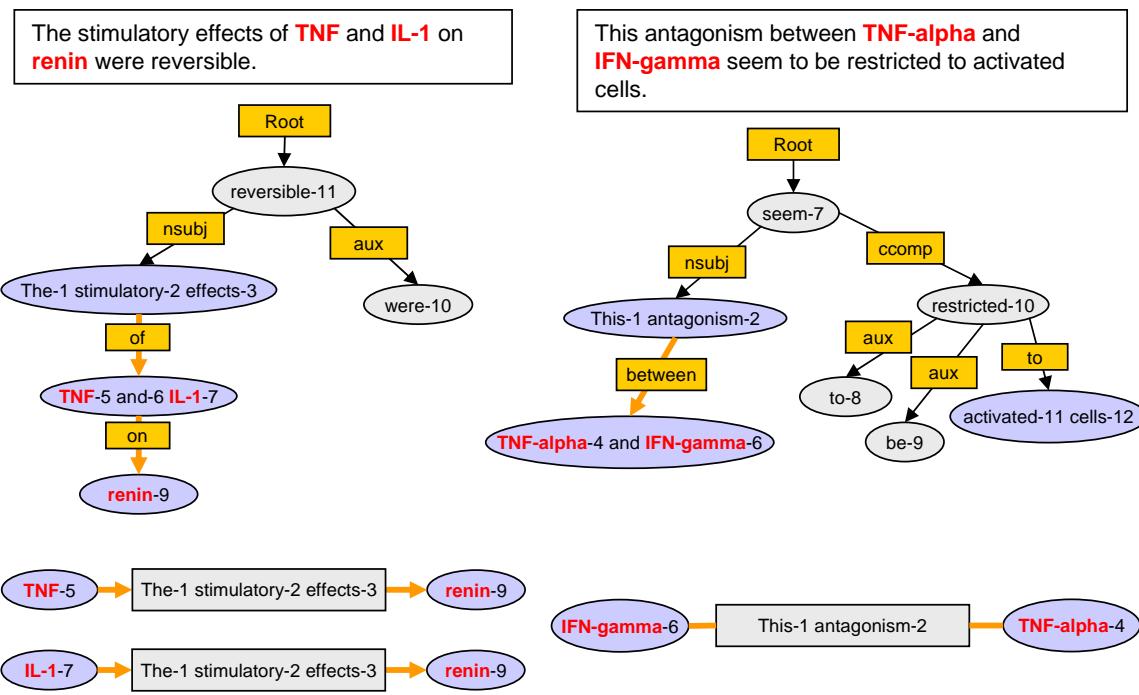


Figure 5.4: Examples of sentences and chunk dependency parse tree representations for which rules 2 (left panel) or 3 (right panel) extract paths marked by thick yellow edges. The extracted candidate relations are indicated below the chunk dependency trees.

Rule 3 extracts two noun phrase chunks connected by a dependency of the type *between* provided that the successor in the tree contains the word *and* (Example in Figure 5.4, right panel) or has a dependent noun phrase chunk which is connected via an *and* dependency. In the latter case, the dependent noun phrase chunk is included in the candidate relation. This rule extracts relations described like “*the physical association between EGFR and p185c-neu*”.

The set of rules can easily be adapted or expanded to extract other types of relations. For example, the *apposition* dependency can be used for searching annotations on genes and proteins as it generally points from an entity to a description of this entity (e.g. Spo0A-P $\xrightarrow{\text{appos}}$ a major transcription factor).

Relation Filtering and Postprocessing Steps

All candidate relations are filtered and post-processed as follows:

Negation check A relation is said to be negated if a node in the candidate relation or one of the respective child nodes contains a negation word (*no*, *not*, *nor*, *neither*, *without*, *lack*, *fail(s,ed)*, *unable(s,d)*, *abrogate(s,d)*, *absen(ce,t)*). Negated relations generally do not carry information that is directly useful for the construction of a set of qualified interactions and for that reason negated relations are removed from further analysis.

Effector-effectee detection Generally, the named entity appearing first in the extracted relation (i.e. with the smaller sentence position) is assumed to be the effector of the relation while the second named entity is assumed to be the effectee. The roles are switched if some form of passive construct is detected; that is, if a predefined expression (Table 5.1) matches the relation and is preceded by a verb, noun, or adjective ending on *-t*, *-d*, *-ion*, *-ing*. For the word *by* the roles are only switched if *by* is not followed by one of the words *time*, *times*, *fold* or by a verb ending on *-ing*.

Single words	by, after, with, if, once, require, requires, when, through
Multi-word expressions	due to, in case, provided that, (effect, result, member) of, in response to, (in, under) control of, depend(s,ed,ent) on

Table 5.1: Effector-effectee detection: Expressions indicating switched roles; i.e., the named entity with the smaller sentence position is defined to be the effectee and the named entity with the larger sentence position is defined to be the effector of the relation.

Enumeration resolution Noun phrase chunks connected to each other by an *and*, *or*, *nn*, *det*, or *dep* dependency form an *enumeration*. If a noun phrase chunk contains more than one protein name, these are considered to describe *alternative agents/targets*. For all candidate relations all gene/protein name containing chunks are analyzed for alternatives from enumerations and chunks containing several protein names. Variants of the candidate relation are generated so that one relation per alternative gene/protein name at each respective position is generated.

Restricting candidate relations to focus domain The words contained in candidate relations are checked against a set of *relation restriction terms*. This list reflects the types of relations that are in the focus of interest. It contains terms that are typically used to

describe a relation of interest, most importantly interaction verbs and derived nouns and adjectives. Here, the focus is on physical, regulatory and genetic interactions; a corresponding list of restriction terms for 151 distinct word stems has been compiled. A candidate relation is retained if it contains at least one relation term.

5.2.2 Evaluation

The evaluation of RelEx on various data sets is described in the following. The used data sets differ in their focus and thus provide complementary information.

Data Sets

Learning Language in Logic (LLL) data set The task of the Learning Language in Logic (LLL) challenge 2005 ([Nédellec, 2005](#)) was to extract genic interactions of the types action, regulon, binding and promoter from sentences dealing with *Bacillus subtilis* transcription. The task required identification of genes/proteins that interact and their roles (i. e. agent or target) together with their position within a sentence. Participating groups focused on machine learning approaches. The organizers provided a synonym dictionary for genes/proteins, a training set (55 sentences, 103 interactions), an evaluation script for the training set, a test set (80 sentences, 54 interactions), and a website for evaluation of the results on the test set.

Manually annotated dataset on human interactions (hprd50) A subset of 50 abstracts referenced by the Human Protein Reference Database (HPRD, [Peri et al. \(2004\)](#); [Mishra et al. \(2006\)](#)) was randomly selected. Direct physical interactions, regulatory relations, as well as modifications (e.g. phosphorylation) were manually annotated by two annotators with biochemical background (R. Küffner (LMU) and K. Fundel). The consensus contains 138 relation instances (i. e. pairs of genes/proteins with abstract and sentence identifier), corresponding to 92 distinct relations in abstracts (i. e. pairs of genes/proteins with abstract identifier). The inter-annotator agreement was 81% (determined as the intersection of annotated relations divided by the total number of relations) which corresponds to a F-measure of 89% (considering one of the annotations as standard of truth and evaluating the other against it).

BioCreAtIvE data set (BioC) [Hakenberg et al. \(2005\)](#) manually annotated 1000 sentences obtained from the first BioCreAtIvE challenge (Task 1A, [Yeh et al. \(2005\)](#)) with 252 interactions in 170 sentences.

Herpes data set A set of manually extracted *Herpes* protein interactions has been compiled from 576 MEDLINE abstracts (personal communication by C. Friedel). 85 distinct interactions have been extracted for five closely related Herpes viruses (EBV, HSV, KSHV, VZV, mCMV). In contrast to the LLL, hprd50 and BioC data sets, interactions were not necessarily described in single sentences and in some cases interactions were described in

detail only in the full text articles. Abstracts were not necessarily fully annotated and directions of interactions were not annotated.

Evaluation Criteria

Results were evaluated in terms of recall, precision, and F-measure (see Definitions 2.1, 2.2, and 2.4).

Definition 5.1 A *relation instance* rel_i , $i \in \{sen, abs, LLL\}$ is defined as:

- rel_{sen} : a tuple (g_1, g_2, i_s) of a pair of genes g_1, g_2 and a sentence identifier i_s
- rel_{abs} : a tuple (g_1, g_2, i_a) of a pair of genes g_1, g_2 and an abstract identifier i_a
- rel_{LLL} : a tuple $(g_1, g_2, i_s, d, p_1, p_2)$ of a pair of genes g_1, g_2 and a sentence identifier i_s , with defined interaction direction d and sentence positions p_1, p_2 of genes g_1, g_2

Each of the above definitions of a relation instance can be used as evaluation criterion. rel_{sen} is the criterion that is most frequently applied in the literature. rel_{abs} is useful for comparing manually annotated or RelEx relations against interactions in public databases (e.g. HPRD) which do not provide sentence information. rel_{abs} is less stringent than rel_{sen} as an interaction might be mentioned in several sentences within an abstract. Here, rel_{LLL} is the most stringent criterion as direction and sentence positions need to be defined. This criterion is applicable for the LLL-challenge data set which is annotated with the required details and only contains directed interactions.

Besides RelEx, co-occurrence search has also been applied. The co-occurrence results ($cooc_{sen}$: all pairs of co-occurring genes/proteins identified within a sentence are assumed to interact) indicate the maximum recall that can be achieved by a relation extraction approach working on individual sentences given the method for gene name identification.

Evaluation with Standard Criteria

Table 5.2 shows the evaluation results with standard criteria (rel_{sen} ; i.e., instances of gene/protein pairs in sentences). For comparison, this table also contains precision and recall that is achieved by co-occurrence extraction. With RelEx, 77–85% of the relations that are found as co-occurrence are extracted as relations. These numbers correspond to inter-annotator agreement for the recognition of gene names and biomedical annotations, which has been determined to be in the range of 69–91% (Colosimo *et al.*, 2005) and 70–80% (Wilbur *et al.*, 2006). Protein relation extraction can be assumed to be more difficult than gene name recognition as the latter forms part of relation extraction and relations can be described in still more ways than gene names/annotations.

RelEx achieves much higher precision and F-measure than co-occurrence search for all data sets. Highest performance was measured on the LLL and BioC data sets, while on the hprd50 set performance is slightly lower. The performance on the Herpes data set is worse than on the other data sets. This is mainly due to the fact that per definition the LLL,

	LLL		hprd50		BioC		Herpes	
nb. sentences	55		88		1000			
nb. abstracts							119	
nb. co-occurrences($cooc_{sen}$)	216		294		1676		1762	
nb. relations(rel_{sen})	97		138		252		85	
	cooc	RelEx	cooc	RelEx	cooc	RelEx	cooc	RelEx
Recall (%)	100	85	100	78	100	83	92	77
Precision (%)	46	79	47	79	60	82	24	48
F-measure (%)	63	82	64	78	75	83	38	59

Table 5.2: Evaluation results of RelEx (rel_{sen} : an instance is a pair of genes/proteins with sentence identifier, $cooc_{sen}$: sentence co-occurrences).

BioC, and hprd50 data sets contain annotated relations in single sentences, and all sentences that contain relations to be extracted are contained in the analyzed set of sentences. The manually extracted Herpes interactions are not restricted to interactions described in single sentences and, in some cases, are described in detail only in the full-text article. Furthermore, abstracts were in many cases not fully annotated. The evaluation of RelEx on the manually curated Herpes data set is a rough estimate of the relation extraction performance for the purpose of network generation compared to a human annotator having access to full text articles.

Some approaches (Hu *et al.*, 2005; Ono *et al.*, 2001; Saric *et al.*, 2005) claim to achieve higher recall and precision. These were evaluated on data sets that were individually created by the authors. These data sets were mostly rather small or focused on a very restricted set of interaction types or descriptions (e.g. phosphorylation events or descriptions with of-by constructs). Often, the used benchmark sets are not available, such that new approaches cannot easily be compared against these methods. The LLL challenge data set can also be considered as rather small, yet, due to the public availability this data set makes it possible to compare methods and, most importantly, it provides for an independent evaluation.

Evaluation with LLL-Challenge Criteria

Evaluation in the LLL challenge scenario (i.e. application on LLL challenge data with criteria rel_{LLL}) (Figure 5.5, F: 75%, R: 83%, P: 68% on the training set; F: 72%, R: 78%, P: 68% for the basic test set) shows that RelEx returns relations with significantly higher recall and precision than the approaches previously applied for the LLL-challenge (F: 51.8%, R: 53.8%, P: 50.0% for the basic and F: 54.3%, R: 53.0%, P: 55.6% for the linguistically enriched test set (Nédellec, 2005)).

The approaches with the best results in the challenge were based on alignment and finite state automata (Hakenberg *et al.*, 2005), and on Markov Logic applied to create a set of weighted clauses which can classify pairs of named entities as genic interactions (Riedel and Klein, 2005). All approaches which were evaluated in the challenge applied machine

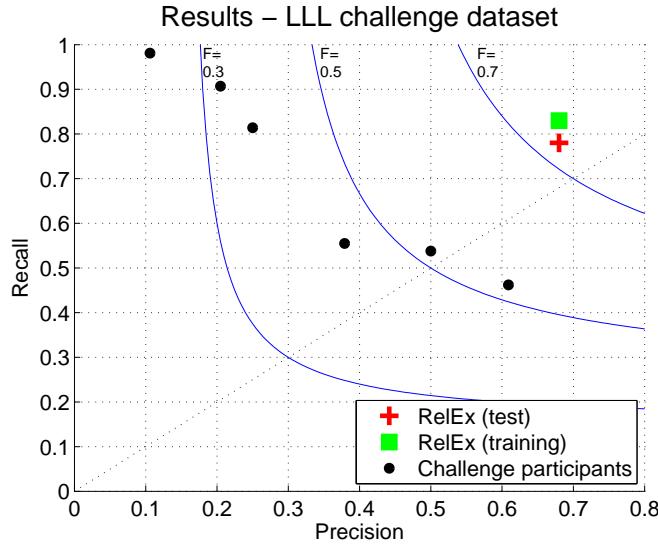


Figure 5.5: Evaluation results on the LLL-challenge data sets obtained with the criteria applied in the challenge (rel_{LLL}).

learning. Approaches based on machine learning generally require large annotated training corpora. Interestingly, the two best performing groups enlarged the training corpus by adding further sentences or clauses. This might indicate that the provided training corpus was not large enough for effective machine learning. The lack of large and consistently annotated corpora in the biomedical domain remains a general problem for developing machine learning-based information extraction approaches as well as for evaluation, even though corpora for specific tasks such as named entity recognition have become available (Hirschman *et al.*, 2005a; Kim *et al.*, 2003).

Analysis of Errors

The detailed analysis of the results on the hprd50 data set indicates the most prominent sources of error: Out of 28 false positive relations, nine relations were generated by the rules not being specific enough or constructs not being correctly resolved, eight describe undesired types of relations (e.g. *homology*, *part of*, *similarity*), six were generated from sentences where a POS-tagging error occurred that lead to erroneous parse trees, and four were generated from sentences where the detected gene/protein name actually does not refer to a gene/protein but forms part of a cell name or description of an experimental technique.

Out of 31 false negative relations, eight are described by a wording that were not covered by the applied rules (e.g. “a and b are receptors that interact”, “a and b form a complex”), eight relations are described in sentences which contained POS-tagging errors, four false negatives were due to anaphora (e.g. *which*, *these proteins*) which RelEx currently does not resolve, four relations were not detected due to erroneous subordinate clause attach-

ment produced by the dependency parser, in two cases the relevant relation terms were not contained on the candidate relation paths, and in another two cases relations were not extracted due to noun phrase chunks erroneously being split up.

The usage of publicly available preprocessing tools clearly causes RelEx to depend on the quality of the output of the applied tools. Thus, the choice of preprocessing tools is important for the overall performance. Biomedical texts are often quite complicated to analyze as they are generally composed of long sentences and contain many non standard English words and complicated sentence structures. Yet, during the last years, a couple of natural language processing tools have emerged that are either adapted to the biomedical domain or work well on this domain even without special adaptation or training. The MedPost part-of-speech tagger has been designed specifically for biomedical texts. The fnTBL noun-phrase chunker is capable of processing pre-POS-tagged sentences. Consequently, the combination of MedPost and the fnTBL chunker is expected to yield high accuracy in detecting biomedical noun-phrase chunks. Yet, in some cases, noun phrase chunks were not identified correctly (e.g. "... either homozygous [or hemizygous] , ... ", "... by [these C] / [EBP-related genes , termed C] / [EBP beta and C] / [EBP delta] , exhibit ... ", where noun phrase chunks are marked by parentheses).

In some cases, the dependency parser seems to return suboptimal results. Nevertheless, for relation extraction dependency parse trees have a significant advantage over syntactic parse trees: dependency parse trees reflect non-local dependencies; that is, dependencies between words that are far apart in a sentence. Sentences of biomedical texts tend to be long and complicated and frequently mention a number of possible effectors and effectees. Dependency parse trees provide a useful structure for the sentences in that nested clauses are more readily recognizable, and subjects and objects are marked. Due to comprehensive preprocessing a small set of rules is sufficient to cover the descriptions of binary, directed relations.

5.3 Large-Scale Network Generation, Analysis, and Applications

RelEx has been designed with a focus on large-scale applicability on biomedical research publications. In the following, the results of large-scale application of RelEx on approximately 1 million abstracts are described. The resulting comprehensive network is analyzed by comparison against HPRD ([Peri et al., 2004](#)), the largest publicly available collection of human protein-protein interactions obtained from manual literature-curation, and the largest currently available human protein-protein interaction data set experimentally determined by yeast two-hybrid technology ([Rual et al., 2005](#)). These comparisons provide insights with respect to the characteristics of the respective networks and their overlaps. Finally, two applications of the generated literature network are highlighted: (1) The RelEx networks are used for expanding manually compiled networks. The different levels of stringency of the RelEx networks ease manual curation. (2) The network schemes approach

integrates the network with automatically compiled context annotations. Thus, queries that specify a certain context can be formulated and searched against the network. Furthermore, frequent context annotations of single relations or subnetworks can be detected automatically, which provides a functional description of the respective relation or subnetwork.

5.3.1 Large-Scale Network Generation

A comprehensive literature-derived network of human gene and protein interactions has been compiled based on the following data: A subset of approximately 1 million MEDLINE abstracts from 1990 or newer were selected for large-scale application. The selection was performed based on MeSH terms and the resulting subset can be assumed to represent a comprehensive set of abstracts dealing with human gene/protein interactions (for details see [Küffner et al. \(2005\)](#)). The applied synonym dictionary has been derived from Entrez Gene ([Maglott et al., 2005](#)), HUGO ([Eyre et al., 2006](#)), and Swiss-Prot ([Bairoch et al., 2005](#)) and contained 338 824 synonyms for 27 141 human genes and proteins (Section 3.2). Several networks of increasing stringency have been obtained from this large-scale application: The *abstract co-occurrence* network offers highest coverage of relations described in the literature (neglecting full-text articles); the *sentence co-occurrence* network represents a baseline for relations described in single sentences; and the *RelEx* network exhibits increased precision and retains high recall, as indicated by the evaluation above (Section 5.2.2). Three different sets of relation restriction terms have been applied in RelEx: if nothing is specified, the full set of relation restriction terms is used (1048 terms with 157 distinct word stems). *red* represents a subset of terms assigned to 24 word stems that were selected for their high frequency in the manually annotated interactions in the *hprd50* data set. *top* refers to a further restricted subset of terms assigned to three word-stems (*interact*, *complex*, *bind*).

	nb. nodes	nb. edges	Clustering coefficient	Characteristic path length	Diameter
Abstract co-occurrences	14 305	677 661	0.53	2.75	≥ 7
Sentence co-occurrences	13 846	359 176	0.45	3.07	≥ 10
RelEx relations	10 821	149 778	0.39	3.25	9
RelEx relations (red)	8 509	76 604	0.38	3.44	11
RelEx relations (top)	6 974	37 732	0.34	3.85	14

Table 5.3: Characteristics of the literature-derived networks of human gene and protein relations (the characteristic path length and diameter for the co-occurrence networks were estimated from subnetwork analyses).

The characteristics of the resulting networks (Table 5.3) show that the RelEx networks have a lower clustering coefficient and a higher characteristic path length than the co-occurrence networks. The clustering coefficient characterizes the overall tendency of nodes

to form clusters or groups. The characteristic path length is the average over the shortest paths between all pairs of nodes and measures the network's overall navigability (Barabasi and Oltvai, 2004). Reducing the set of RelEx relation restriction terms slightly decreases the clustering coefficient and increases the characteristic path length and diameter. This indicates that the cliquishness is reduced and networks get more sparse when the set of relation restriction terms is reduced.

Evaluation of large-scale networks is difficult, especially as the derived networks contain direct physical interactions as well as regulatory dependencies. In the following sections, the networks are analyzed in comparison with networks derived by other means and further characteristics are provided. The overall size of the networks is difficult to judge, but is in the range of published estimates. For example, Rhodes *et al.* (2005) presented a probabilistic model for predicting human protein-protein interactions based on model organism interactome data, protein domain data, genome-wide gene expression data and functional annotation data. They predict nearly 40 000 protein-protein interactions with a false positive rate of 50% and estimate a total of 300 000 protein-protein interactions. Hart *et al.* (2006) estimated the human protein-protein interaction network to contain 154 000–369 000 interactions and state that due to the high false-positive rate, the current maps are expected to be only about 10% complete. Xia *et al.* (2006) predicted PPI networks in human based on genomic, proteomic, and functional annotation; the resulting Integrated Network Database (IntNetDB) contains 180 010 predicted interactions among 9 901 human proteins. The Unified Human Interactome (UniHI), which is based on ten major interaction maps derived by computational and experimental methods, includes more than 150 000 distinct interactions between 17 000 human proteins (Chaurasia *et al.*, 2007). The other characteristics of the RelEx networks (clustering coefficient, characteristic path length, diameter) are similar to the values summarized by Uetz *et al.* (2006) for seven data sets on protein-protein interactions of human, yeast and herpes viruses; the RelEx clustering coefficients are at the upper bound of their values while the characteristic path lengths are at the lower bound of their values.

5.3.2 Comparing RelEx Relations with HPRD Interactions

HPRD (Peri *et al.*, 2004) contains interactions that were manually extracted from MEDLINE full-text articles (For characteristics of the network derived from HPRD see Table 5.5). The comparison of RelEx relations against HPRD interactions provides information with respect to differences and overlaps of the two approaches (Table 5.4). A large fraction of HPRD interactions cannot be retrieved from the abstract sentences. This is demonstrated by the analysis of sentence co-occurrences: only approximately half of the interactions annotated in HPRD can be found in abstract sentences. RelEx extracts a significantly larger number of relations from the abstracts than the number of interactions contained in HPRD as shown by the values of Overlap2.

The manually annotated hprd50 data set (Section 5.2.2) allows us to estimate the performance on the basis of abstracts referenced by HPRD (Table 5.2) and thus to examine the differences between RelEx relations and HPRD interactions. The performance on the

	Co-occurrences	RelEx
Instances ($cooc_{sen}/rel_{sen}$)	3 381 602	731 432
Nb. interacting gene/protein pairs	359 173	149 778
HPRD - Overlap1 (%)	51	40
HPRD - Overlap2 (%)	5	8

Table 5.4: Results of RelEx large-scale application on a comprehensive set of MEDLINE abstracts (approximately 1 million abstracts) and comparison against HPRD. Overlaps were determined for pairs of genes/proteins, restricted to the set of genes/proteins common to HPRD and sentence co-occurrence search (5925 genes/proteins), and irrespective of the individual abstract. Overlap1: Proportion of HPRD-relations found by co-occurrence/RelEx; Overlap2: Proportion of co-occurrences/RelEx-relations available in HPRD.

hprd50 data set is slightly lower than on the LLL-challenge data set (F: 78% vs. 82%). The detailed analysis showed that the hprd50 data set contained several quite long and complicated sentences. Furthermore, in contrast to *Bacillus subtilis* gene names, human gene and protein names are often multi-word terms. In certain cases, these impaired the construction or analysis of the parse trees.

The hprd50 data set also makes it possible to compare the interactions provided by HPRD against our manual annotation (Figure 5.6). We annotated 92 distinct relations for the abstracts, while HPRD annotated 76 interactions for the corresponding articles. Only 26 relations were in common. RelEx identified 22 of these 26 relations.

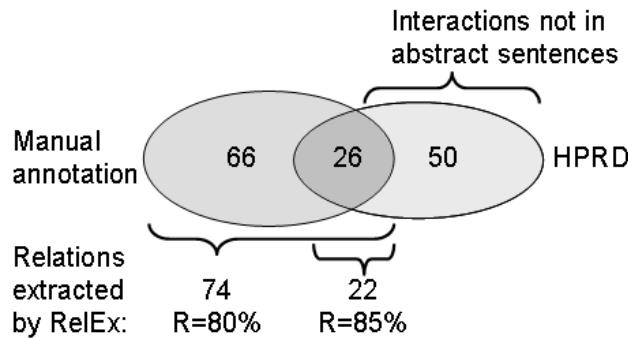


Figure 5.6: Comparison of manually annotated relations, HPRD interactions, and relations extracted by RelEx based on the hprd50 dataset (numbers correspond to relations rel_{abs} , R: Recall).

Only 27 of the 76 HPRD interactions exist as sentence co-occurrences in the abstracts. Manual analysis of the abstracts indicated that the majority of HPRD interactions is not covered in the abstract at all. Evidently, RelEx cannot retrieve HPRD interactions that are not described in abstract sentences. According to our manual annotation, the abstracts contained more interactions than annotated by HPRD for the full text articles.

The moderate overlap between HPRD and relations extracted by RelEx can be explained

by various effects: HPRD interactions are manually extracted and consequently not restricted to single sentences; in many cases the relations are described in detail in the full text only; yet, abstracts and articles are not necessarily completely annotated, and thus only a part of the relations mentioned in an abstract or article may be covered. HPRD does not yet cover the entire gene/protein space as it focuses on disease-related genes. Further differences to our annotation can be explained by the observation that HPRD focuses on direct physical protein-protein interaction data. Gene regulatory relations as well as long-range relations are not covered. An important bias has been observed in the annotated relations; 17 of the 26 HPRD interactions contained in our manually annotated set were described using just two verbs, namely *interact(s/ed/ion)* and *binds/bound*. The remaining relations contain words such as *cross-link*, *coprecipitated*, *adapter*. This indicates that HPRD uses quite stringent annotation guidelines focused on direct physical interactions; most of them are described with a rather limited set of words and expressions.

Our results indicate that HPRD, even though being a very large and valuable source for protein interaction data, currently covers only a small part of the human protein-protein relations from very limited relation categories.

Besides HPRD, numerous other database organize protein interaction data (e.g. BIND ([Bader et al., 2003](#)), DIP ([Salwinski et al., 2004](#)), MIPS ([Pagel et al., 2005](#)), MINT ([Zanzoni et al., 2002](#)), IntAct ([Hermjakob et al., 2004b](#))). [Mathivanan et al. \(2006\)](#) compared eight databases for protein-protein interactions; they found that HPRD contains by far the largest number of interactions and genes, and that the databases vary in terms of annotations, which is due to the use of alternative vocabulary terms. [Gandhi et al. \(2006\)](#) analyzed six PPI databases and found that HPRD covers most of the interactions contained in any of the other databases. Thus, HPRD appears to be the most appropriate data source for large-scale comparisons; the other public data sources can be assumed to contain an even smaller fraction of the RelEx relations.

5.3.3 Comparing RelEx Relations with Yeast Two-Hybrid and Literature-Curated Protein-Protein Interaction Data

Yeast two-hybrid (Y2H) technology can be used to detect direct physical protein interactions. During the last years, a number of large Y2H protein-protein interaction (PPI) maps have been generated. [Rual et al. \(2005\)](#) presented a proteome-scale map of human protein-protein interactions. They analyzed approximately 7200 genes and found 2754 interactions among 1549 proteins (termed *Y2H* data set in the following). Besides, they compiled a binary literature-curated interaction map (LCI) by integrating DIP ([Xenarios et al., 2002](#)), MINT ([Zanzoni et al., 2002](#)), BIND ([Bader et al., 2003](#)), HPRD ([Peri et al., 2004](#)) and MIPS ([Pagel et al., 2005](#)). They generated two subsets of LCI: *LCI-core* contains only interactions that are supported by at least two MEDLINE entries and *LCI-hypercore* contains only interactions that are supported by at least two MEDLINE entries and are contained in at least two of the used databases. Figure 5.7 shows a schematic visualization of the analyzed data sets.

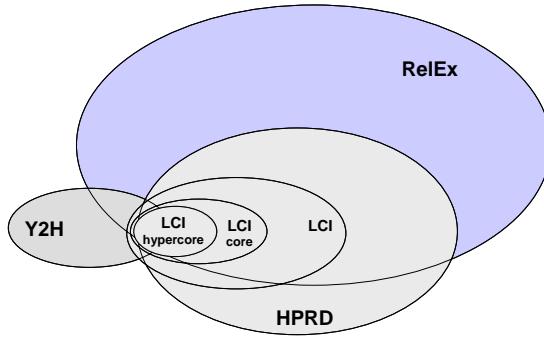


Figure 5.7: Schematic diagram of the analyzed interaction networks. RelEx: text-mining network, Y2H: experimental data (Rual *et al.*, 2005), LCI: literature-curated data (Rual *et al.*, 2005), LCI core and LCI hypercore: subsets of LCI with higher support, HPRD: literature-curated data (Peri *et al.*, 2004).

The topological characteristics of the analyzed public networks (Table 5.5) are very different from those of the co-occurrence and RelEx networks (Table 5.3): The clustering coefficient is higher and the characteristic path length is lower for the text-mining networks than for the public data sets, except for the LCI hypercore data set which has a characteristic path length in the range of the values determined for the RelEx networks. These differences might be explained by the fact that the analyzed public data sets focus on physical protein-protein interactions while co-occurrence search and RelEx extract a broader spectrum of relation types which leads to more dense networks.

	nb. nodes	nb. edges	Clustering coefficient	Characteristic path length	Diameter
Y2H (Rual <i>et al.</i> , 2005)	1533	2729	0.06	4.36	12
HPRD (Peri <i>et al.</i> , 2004)	6147	20638	0.26	4.74	15
LCI (Rual <i>et al.</i> , 2005)	2179	4039	0.14	5.24	14
LCI core	640	623	0.08	7.57	19
LCI hypercore	291	275	0.08	3.31	9

Table 5.5: Characteristics of the Y2H network and literature-curated networks that the co-occurrence and RelEx networks are compared against. Y2H is experimental data while the other data sets are literature curated. LCI: literature-curated interaction map (Rual *et al.*, 2005), LCI core and hypercore: subsets of LCI with higher support. For details see text.

Next, overlaps of co-occurrence and RelEx networks with Y2H and public literature-curated PPI data sets were analyzed:

Definition 5.2 Given a RelEx/co-occurrence set of interactions I_{prop} and public interaction data set I_{public} , the **interaction overlap** O_i , $i \in \{public, prop, union\}$ is defined

as:

$$O_i = \frac{I_{prop} \cap I_{public}}{I_x}$$

with

- O_{prop} : quotient of intersection and RelEx/co-occurrence: $I_x = I_{prop}$
- O_{public} : quotient of intersection and public data set: $I_x = I_{public}$
- O_{union} : quotient of intersection and union of RelEx/co-occurrence and public data set: $I_x = I_{prop} \cup I_{public}$

where an interaction is a pair of genes/proteins, irrespective of the individual abstract, and restricted to the set of genes/proteins in common to the co-occurrence search and the respective public data source.

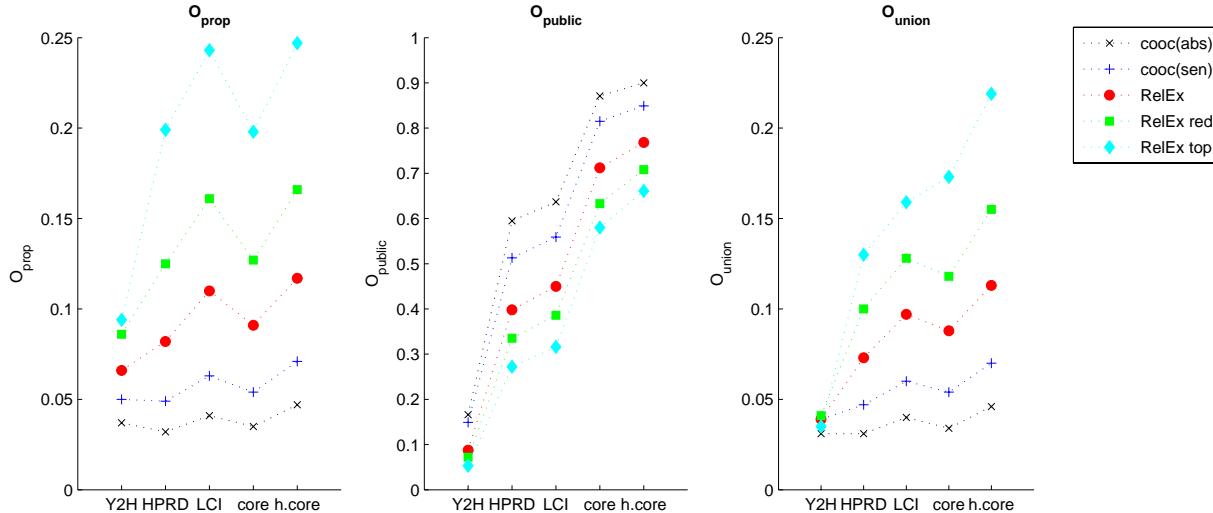


Figure 5.8: Overlap of co-occurrence and RelEx networks with public Y2H and literature-curated protein-protein interaction maps. Overlaps: quotient of intersection and data in text mining (O_{prop}), public data set (O_{public}), union text mining and public data set (O_{union}).

The results of the overlap analysis (Figures 5.8 and 5.9) show that the co-occurrence and RelEx interaction maps contain an important part of most investigated public data sources, but also contain significantly more relations. Using RelEx instead of co-occurrence search leads to increased overlaps O_{prop} and O_{union} ; that is, the fraction of relations only contained in the text-mining interaction map is reduced. Reduced sets of relation restriction terms (*red* and *top*) lead to a further increase in O_{prop} and O_{union} ; O_{prop} and O_{union} are maximized for the comparison RelEx top against LCI hypercore, which indicates convergence of RelEx relations and manually curated interactions derived from public databases.

The overlaps of co-occurrence/RelEx relations with the experimentally determined Y2H interactions are lower than the overlaps with the public literature-derived interaction maps.

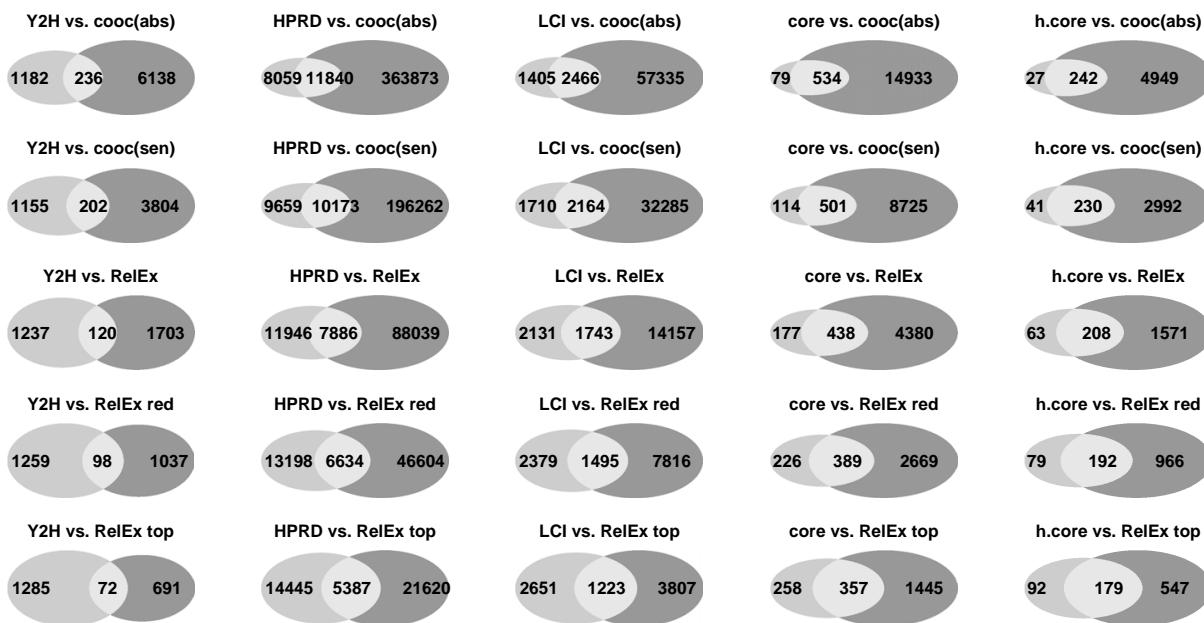


Figure 5.9: Venn diagrams showing the overlaps of co-occurrence and RelEx networks with public Y2H and literature-curated protein-protein interaction maps. Numbers in ellipses: interactions; left ellipses: public datasets, right ellipses: co-occurrence/RelEx-relations.

Restricting the RelEx text-mining networks by reduced sets of relation terms increases O_{prop} from 3.7 to 9.4%, but O_{union} varies only slightly (3.1–4.1%).

Low overlap between Y2H data and literature derived interactions was found in a number of studies; in the following, these will be briefly summarized:

Rual *et al.* (2005) found that the set of experimentally determined Y2H PPIs contains 2.3% of the interactions contained in LCI, 4.6% of LCI-core, and 8.4% of LCI-hypercore. Calculating O_{union} for their comparisons, one obtains 1.4% (Y2H vs. LCI), 0.9% (Y2H vs. LCI-core), and 0.7% (Y2H vs. LCI-hypercore). In a second large-scale human Y2H study (Stelzl *et al.*, 2005), a human protein-protein interaction network of 3186 interactions among 1705 proteins has been compiled; only 16 HPRD interactions were contained in this Y2H network (3%). For a *Drosophila* protein interaction map, only 2.3% of the Y2H interactions were contained in a literature derived set (Formstecher *et al.*, 2005); for an earlier data set (Giot *et al.*, 2003), a corresponding value of 1.8% has been determined. Reguly *et al.* (2006) compiled a literature-curated network for yeast containing 33 311 interactions. According to their study, high-throughput protein-interaction datasets achieve only around 14% coverage of the protein interactions in the literature. They found that only 2–3% of the interactions reported in Y2H screens have been confirmed by other means which is, presumably, due to the high false positive rate of two-hybrid methodology. For the genetic interactions derived from literature curation and HTP methods, they found less than 5% of either data set confirmed in the other data set.

Protein interaction datasets obtained from high-throughput experiments are generally known to be prone to a high rate of false negative and false positive results. A number of approaches for filtering false positives observed in high-throughput experiments have been presented. Bader and Hogue (2002) recommend to integrate several datasets obtained from different methods. Methods for assessing the reliability of interactions have been presented based on expression data or information on paralogs (Deane *et al.*, 2002), screening statistics and network topology (Bader *et al.*, 2004), or gene expression data and functional prediction (Deng *et al.*, 2003). Deane *et al.* (2002) found that only approximately 30–50% of the observed interactions are biologically relevant. Suthram *et al.* (2006) compared seven confidence assignment schemes and found the score of Deng *et al.* (2003) to yield the overall best performance.

5.3.4 Using RelEx for Network Expansion

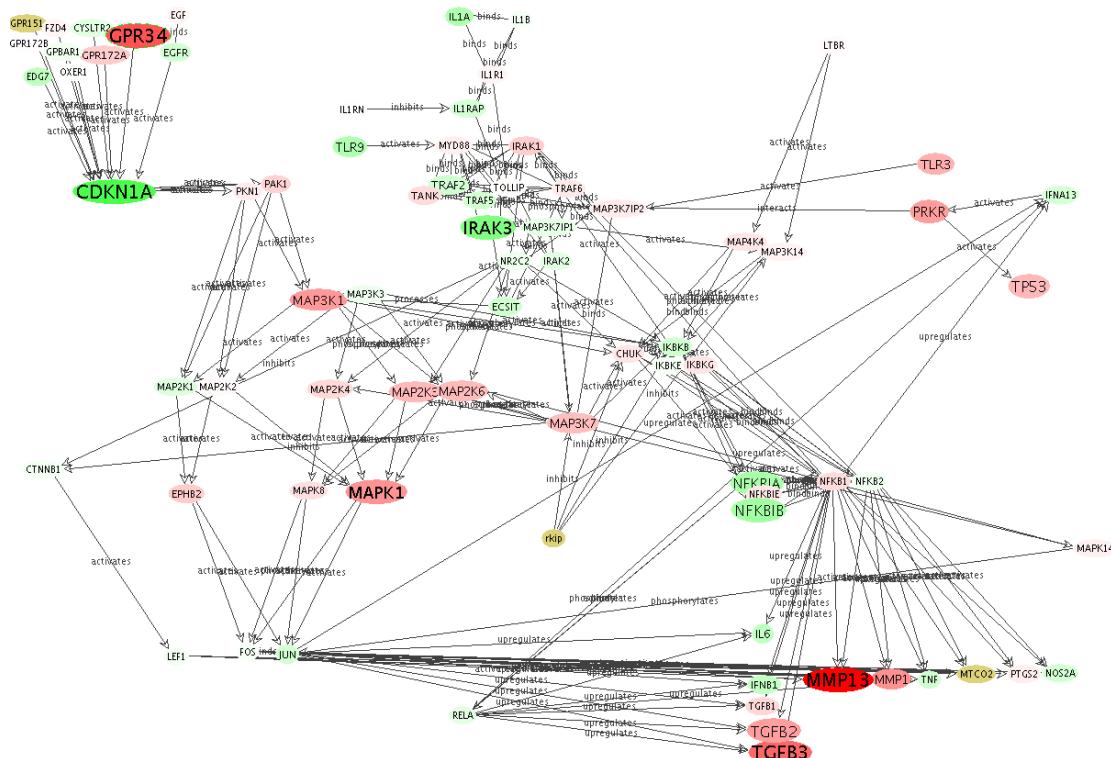


Figure 5.10: Manually compiled regulation network describing the IL1 pathway as obtained from combining BioCarta data and expert information kindly provided by Thomas Aigner (University of Leipzig, personal communication 2004). Nodes are annotated with the Affymetrix two-class osteoarthritis data set (Section 7.2), red: up-regulation, green: down-regulation in osteoarthritis, size: significance of p-values of differential expression, yellow: value not available. The network was visualized with ToPNet (Hanisch *et al.*, 2004).

Often, manually compiled regulation networks are used for providing an overview of regulatory cascades. Such networks help in interpreting experimental data and understanding biochemical phenomena. Manual compilation of networks is labor intensive; consequently, the respective networks are generally rather small. Figure 5.10 shows an example of a network that has been generated by combining data from BioCarta⁴ with expert information. The figure shows the IL1 pathway which integrates the regulatory reactions that take place when IL1 binds to its receptor: The signal is transferred via the MAP-kinase cascade and several transcription factors and leads to regulation of transcription of a number of genes. Color and size of the network nodes in Figure 5.10 indicate differential expression of the respective genes in an experiment investigating the difference between normal and osteoarthritic cartilage cells (see Sections 6.4 and 8.1 for some background on osteoarthritis and biological interpretation of an osteoarthritis gene expression data set). Only few genes in the network appear clearly regulated; one of them is MMP13 which is one of the target genes of the IL1 network. The manually compiled network lacks information on potential interaction partners and down-stream targets of MMP13. Given that MMP13 is known to play an important role in osteoarthritis, it appears useful to expand the network so that it includes genes and proteins with which MMP13 interacts.

	nb. genes/proteins	nb. abstracts	nb. sentences
Abstract co-occurrences	440	1964	-
Sentence co-occurrences	238	1110	1589
RelEx relations	122	280	351
RelEx relations (red)	50	98	116
RelEx relations (top)	14	16	17

Table 5.6: Characteristics of the hull of MMP13 in large-scale text-mining networks; that is, the network of genes that are directly linked to MMP13 in the text-mining networks. nb. abstracts/nb. sentences is the number of abstracts/sentences from which a co-occurrence/relation path for one of the relations in the respective MMP13 hull has been extracted.

Automatically generated networks can be used to expand manually compiled networks. Thus, hypotheses on relations beyond the manually compiled network can be generated. The networks can be exploited by automatic approaches or by manual analysis. The literature-derived networks provide links into the source articles that make it possible to directly check the correctness of the respective relation and to extract additional information.

Accordingly, the networks obtained from large-scale text mining (Table 5.3) can be used to identify literature-derived relations beyond those in the manually compiled IL1 network. Importantly, the different levels of stringency speed up the manual curation process as the most stringent network already contains the most relevant interaction partners, but contains a significantly reduced number of links into the literature (Table 5.6). For example, inspection of only 16 abstracts derived from the high-confidence RelEx network makes it

⁴<http://www.biocarta.com>

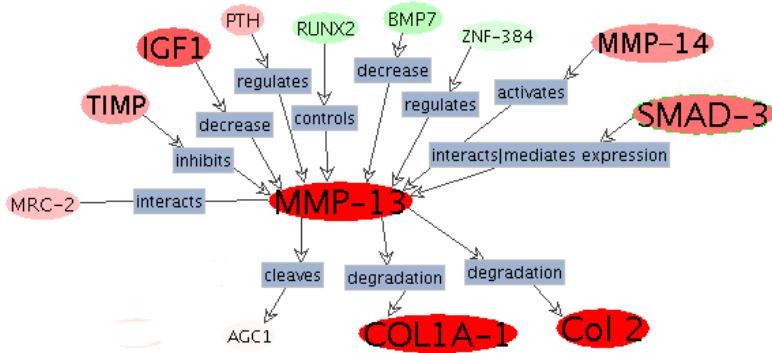


Figure 5.11: MMP13 interaction network showing genes that are directly linked to MMP13 in the text-mining network, but not contained in the IL1 pathway shown above, as obtained from manual curation. Nodes are annotated with the Affymetrix two-class osteoarthritis data set (Section 7.2): red: up-regulation, green: down-regulation in osteoarthritis, larger nodes indicate more significant p-values of differential expression. The network was visualized with ToPNet (Hanisch *et al.*, 2004).

possible to generate the high-quality network of MMP13 interaction partners shown in Figure 5.11. This network shows that MMP13 is under control of far more genes than indicated in the manually compiled IL1 network, and it indicates some targets of MMP13.

5.3.5 Network Schemes: A Means for Exploiting Context

Network schemes represent a means for specifying text-mining contexts (Figure 5.12). Network schemes are defined as Petri nets, which are bipartite graphs containing one set of vertices called places and the other set of vertices called transitions. Places are defined by biological entities, such as genes/proteins, cell-types, tissues, diseases, or organisms. Places can be occupied by specific entities, by alternatives between entities or by more general expressions (e.g. any tissue, three or more organisms). Transitions can be defined as co-occurrences in sentences or abstracts or as RelEx relations. RelEx relations can further be specified by specific types of relations, which are defined by respective subsets of relation restriction terms (e.g. type *down-regulation* contains the words: *inhibit*, *down-regulate*, *block*, *inactivate*, etc.).

Network schemes take as input data (1) a network which provides, for each relation, links into the respective articles where evidence for the relation was found, and (2) context annotations for the literature articles. The network has been generated by RelEx (Section 5.2); each relation in the RelEx network is labeled with the abstracts from which it was extracted. Context annotations have been compiled automatically by exact matching of non-gene and non-protein dictionaries (Section 3.5.1) against title, abstracts, and MeSH annotations of all publications in MEDLINE. The matching returned, for each abstract, a set of identifiers for the objects which were found in the respective abstract.

Network schemes have two main applications: First, schemes can be used for *user input*

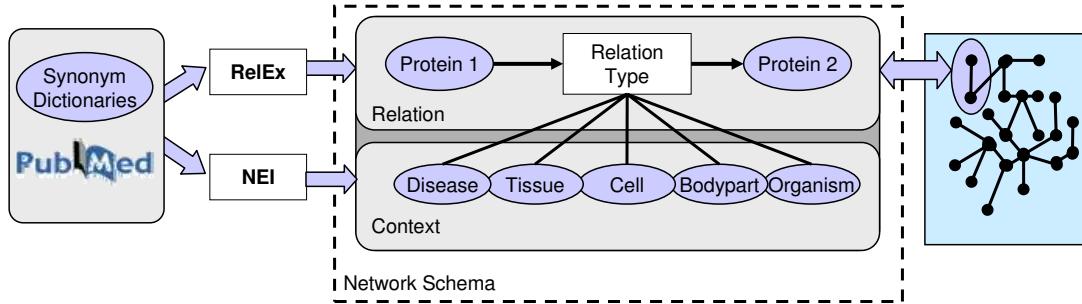


Figure 5.12: Network schema used for exploiting context in text-mining networks: Relations are extracted from texts by co-occurrence search or RelEx. Context annotations from the respective texts are extracted by named entity identification with appropriate synonym dictionaries. Thus, relations can be restricted to those with specific contexts for network generation, or single relations or subnetworks can be analyzed for common and statistically overrepresented contexts.

specification. This represents a flexible means for defining contexts of interest. Instances of the specified schema can then be searched by text mining and reported to the user.

Second, network schemes can be used as *templates* to be filled with contexts as derived automatically from texts. This provides a means detect common contexts and identify frequent patterns in the respective biological contexts. Statistical analysis is then applied for ranking and reporting relevant common contexts from the expanded network schemes. The statistical significance of a context annotation c can be determined by Fisher's exact test (Fisher, 1932), which makes use of the hypergeometric distribution:

$$p_c(x \geq r; N, n, k) = 1 - \sum_{i=0}^{r-1} \frac{\binom{k}{i} \binom{N-k}{n-i}}{\binom{N}{n}}$$

where: N is the number of abstracts in the entire set used for analysis (i. e. the abstracts subjected to RelEx analysis); n is the number of abstracts in the selected subset (i. e. containing a relation which matches the applied network schema); k is the number of abstracts in the entire set containing annotation c ; $p_c(x \geq r)$ is the probability of observing r or more abstracts with annotation c by chance.

This second approach can be used for detecting contexts, such as diseases or tissues, for which a specific relation has been described. The approach is based on the same statistical principles as gene ontology overrepresentation analysis (Section 9.1). Yet, the network schema approach is more generally applicable as it directly exploits the literature and thus does not require manual annotations. Similarly, it is more flexible, as the non-gene and non-protein dictionaries used for compiling context annotations contain more entries and are more fine-grained than gene ontology. Furthermore, it does not rely on information for single genes but focuses on articles discussing specific relations or interaction events. The approach can be applied for single relations/interactions or larger network modules.

In the following, the application of network schemes for the detection of common contexts is described by two examples, namely the IL1 pathway and the MMP13 interaction network as extracted from the RelEx network (see Section 5.3.4). These networks are small subnetworks of the entire RelEx network. By use of network schemes, these subnetworks can be functionally categorized.

IL1 Pathway:	
disease:	inflammation, rheumatoid arthritis, arthritis, osteoarthritis, acute phase reaction, septic shock, inflammatory response, synovitis, endotoxemia, thymoma, infection, fever
tissue:	articular cartilage, vascular endothelium, cartilage, epidermis, Media, smooth vascular muscle, respiratory mucosa
cell:	cultured cells, macrophages, cell line, monocytes, tumor cells cultured, fibroblasts, hela cells, macrophages peritoneal, granulocyte, chondrocytes, alveolar macrophages, keratinocytes, osteoblasts, jurkat cells, leukocytes mononuclear, epithelial cells, 3t3 cells, endothelial cells, osteoclasts, microglia, synoviocytes, neutrophils, kupffer cells, mesangial cells, astrocytes, cell line transformed, t cell, smooth muscle cells, bone marrow cells, u937 cells
body part:	synovial membrane, glomerular mesangium, veins, gingiva, pulmonary alveoli, joint
organism:	mice, mice inbred c3h, mice inbred balb c, mice knockout
MMP13 interaction network (red):	
disease:	osteoarthritis knee, osteoarthritis, rheumatoid arthritis, chondrosarcoma, periodontitis, osteosarcoma, carcinoma squamous cell, odontogenic cysts, arthritis
tissue:	cartilage, tissue, cartilage articular, extracellular matrix, Bone
cell:	chondrocytes, osteoblasts, fibroblasts, cells cultured, neutrophils, squamous cell
body part:	synovial membrane, bone, parathyroid, skull, tibia

Table 5.7: Results of application of network schemes for the detection of common contexts: Manually generated network of the IL1 pathway (presented in Section 5.3.4) and MMP13 interaction network as extracted from the RelEx network restricted with term set *red*). The table only shows highest ranked annotations ($p\text{-value} \leq 10^{-7}$ for the IL1 Pathway, $p\text{-value} \leq 10^{-5}$ for the MMP13 interaction network).

The results are shown in Table 5.7. BioCarta⁵, which served as one data source for generating the manually curated IL1 network, provides a description for each pathway. The first sentences of the description for the IL1 pathway are given in the following:

Interleukin-1 (IL-1) is a pro-inflammatory cytokine that signals primarily through the type 1 IL-1 receptor (IL-1R1). The activities of IL-1 include induction of *fever*, expression of vascular adhesion molecules, and roles in *arthritis* and *septic shock*. The *inflammatory activities* of IL-1 are partially derived by transcriptionally inducing expression of cytokines such as TNF-alpha and interferons, as well as inducing the expression of other inflammation-related genes.

⁵<http://www.biocarta.com>

The terms found by the network-schema context analysis (Table 5.7) fit well to the Bio-Carta description. The terms marked in italics in the above description are contained in the disease section of overrepresented annotations. The section on cell types contains several cell types that play a role in the immune system. The overrepresentation of mice in the organism annotation indicates that phenomena induced by IL1 are often studied in this model organism.

The second example investigates context overrepresentation for interactions of MMP13. The top-ranked annotations indicate that these interactions are frequently discussed in the literature in context of osteoarthritis and cartilage. It is known that MMP13 is involved in the destabilization of the joint cartilage collagen network and is used as a marker for hyper-catabolism in cartilage degradation (Sections 6.4 and 6.4). Thus, it makes sense that MMP13 interactions are closely related to this context.

The network scheme approach thus makes it possible to detect common contexts for interactions. By this approach, individual interactions or network modules can be functionally characterized.

5.4 Characterization of Gene/Protein Interactions

Biochemical networks are often considered as undirected, unlabeled graphs where each node represents a gene/protein and each edge represents an interaction. Public databases generally contain no additional information for physically interacting protein pairs. Yet, most biochemical relations and interactions exhibit characteristics which are useful to know for data analysis and for understanding biochemical phenomena; for example, an interaction can be directed, activating, and affect gene-expression.

RelEx (Section 5.2) extracts pairs of interacting genes/proteins from free text and specifies interaction directions. Furthermore, RelEx returns sentence *paths* (i. e. sequences of words that describe an interaction) and returns the sentences in which evidence for an interaction was found. Thus, it provides for each interaction a *context*, which can be used to further specify the interaction.

In the following, an approach to confirm and characterize interactions derived from text mining is described (Küffner *et al.*, 2006). The approach comprises a curation protocol, manual text annotation, and application of machine learning to predict interaction characteristics for a given pair of genes/proteins. It has mainly been developed by Robert Küffner and Timo Duchrow. The approach makes use of the synonym dictionaries (Section 3.2), named entity identification (Chapter 4), and relation extraction by RelEx (Section 5.2).

5.4.1 Data Preparation and Classification Approach

Data Preparation

Several databases (DIP (Salwinski *et al.*, 2004), BIND (Alfarano *et al.*, 2005), HPRD (Mishra *et al.*, 2006)) were screened for MEDLINE references. From the respective ab-

stracts, sentences that contained at least two protein names, at least one ReEx path, and an interaction keyword (from a list of about 300 keywords such as *activate*, *phosphorylation*) were extracted. From each of the selected abstracts one sentence was randomly selected to avoid bias from abstracts referring to a particular interaction several times.

10736564.2.6	sent	Immobilized-1 FGFR4-2 also-3 bound-4 FGF-8-5 besides FGF-1-7 and FGF-2-9				
10736564.2.6	mark	[Immobilized-1] pm0119069-2 [also-3 bound-4 pm0118903-5] besides [pm0124654-7 and pm0106166-9]				
10736564.2.6	annot	2	5	interacting	5	4
10736564.2.6	annot	2	7	interacting	5	4
10736564.2.6	annot	2	9	interacting	5	4
10736564.2.6	annot	5	7	interacting	1	
10736564.2.6	annot	5	9	interacting	1	
10736564.2.6	annot	7	9	interacting	1	

Table 5.8: Annotation schema applied for characterization of gene/protein interactions. In the example, ProMiner identified four distinct proteins at sentence positions 2, 5, 7, and 9. The annotation section contains one entry for each pair of proteins. The protein at position 2 interacts with the proteins mentioned at the positions 5, 7, and 9. An interaction between the latter proteins is not described. The slots after the attribute label may be used for hints, which are defined by the sentence position of the respective word.

For each pair of gene/protein names in a given sentence, five attributes (Table 5.9) are annotated with one of five values (Example in Table 5.8). Together, the attribute labels describe the relation type of the respective gene/protein pair. For each attribute, the curator can add hints; that is, words in a sentence that lead the curator to his interpretation. In total, 269 sentences containing 1 090 co-occurrences have been annotated. The distribution of label frequency (Table 5.9) shows that interactions described in texts are biased. For example, activation events are more frequently described than inhibition events, and direct interactions are more frequent than long-range interactions.

Attribute	Meaning of label 1	label 1	label 2	label 3	label 4	label 5	Meaning of label 5
interacting	no	661	0	0	37	392	yes
directed	undirected	186	4	3	6	240	directed
activating	inhibiting	36	0	280	10	113	activating
immediate	indirect	101	13	33	64	228	direct
expression	protein-protein	258	32	44	9	96	protein-gene

Table 5.9: Attributes, their labels, and number of corresponding entries in the data set. Labels 1 and 5 indicate annotation with strong confidence, labels 2 and 4 moderate confidence cases, and label 3 indicates that an attribute cannot be specified from the given sentence.

Classification

Classification requires the individual instances (here: sentence identifier, sentence position protein 1, sentence position protein 2) to be represented as feature vectors. Two feature sources were used:

- Bag-of-words (BOW): all stemmed words in a sentence are used as features.
- Bag-of-words+path (BOW+path): the RelEx paths for the respective sentence that contains the two proteins and at least one of the indicated hints (if appropriate) are filtered; the stemmed words contained in these paths are used as features in addition to the BOW features.

All co-occurrence instances were classified with respect to the five attributes given in Table 5.9 by support vector machines (SVM). For learning and prediction, labels 1 and 2 were combined as well as 4 and 5. The prediction of the first attribute (interaction) is a two class problem. The prediction of the other attributes constitute three class problems; these are modeled by two two-class problems. For example, inhibiting (1+2) vs. activating (4+5) vs. not specified (3) is modeled by 1+2 vs. 3+4+5 and 4+5 vs. 1+2+3. For these attributes, *not specified* is predicted if a sample is located on the side of the negative training samples with respect to the decision hyperplane for both classifiers. Otherwise, the class corresponding to the maximum value of the SVM decision functions of the two classifiers is selected. Thus, a total of nine classifiers were applied for prediction of the five attributes. For all SVMs linear kernels were applied and the cost ratio for training errors on positive samples was set to the ratio of the negative to the positive class sizes.

5.4.2 Evaluation

The evaluation of the classification approach by repeated stratified ten-fold cross-validation indicated very good performance (Table 5.10).

Protocol	Precision	Recall	F-measure
bag-of-words (BOW)	35.5	68.3	46.7
BOW + hints	36.1	69.0	47.4
BOW + path	78.2	82.3	79.4
BOW + path + hints	82.7	87.2	84.9

Table 5.10: Evaluation of interaction characterization and the used feature sources. The use of RelEx paths (path) as feature source lead to an important increase in classification performance.

The slightly lower performance for attributes occurring at lower frequency indicates that accuracy could be increased by enlarging the training data set. The comparison of the different feature sources indicates that features based on RelEx paths entail significantly increased performance (second versus third row in Table 5.10). Best performance was achieved by

using RelEx paths together with hints for feature generation. These results indicate that the RelEx paths, indeed, contain useful information on the individual relations.

5.5 Conclusions

RelEx has been developed for compiling a comprehensive set of causal and physical protein/gene interactions from free text. RelEx is based on several publicly available preprocessing tools and a small set of rules. It is able to cope with different organism domains. The system can be tuned towards different kinds of relations by use of appropriate relation restriction terms and/or entity synonyms. Compared to other approaches it is fairly straightforward to implement but still achieves competitive performance.

RelEx has been validated on publicly available data sets for prokaryote and human interactions. When RelEx is compared with the rather stringent criteria of the LLL challenge data set (Nédellec, 2005), performance is significantly higher than previously reported results. In this scenario, the ability to specifically extract directed relations from particular sentences was analyzed. Most of the published approaches are evaluated with respect to the extraction of relations from abstracts, which is considerably relaxed compared to the former criteria. Here, the RelEx performance is in the range of existing approaches (e.g. Hu *et al.* (2005); Ono *et al.* (2001); Saric *et al.* (2005)), but RelEx can extract a significantly broader spectrum of relation types.

In contrast to many other approaches, RelEx can be applied to large corpora. It has been applied to approximately 1 million abstracts, which represent a comprehensive subset of MEDLINE enriched in human protein-protein interactions. The resulting network contains about 150 000 relations between approximately 11 000 genes/proteins and about 731 000 text fragments describing these interactions with an expected recall of 78% and precision of 79%. The performance estimates have been obtained from evaluation on several hand-curated benchmark sets. Of course, the performance estimate for the whole MEDLINE is affected by uncertainties as the benchmark data sets are very small subsets of the entire MEDLINE.

Importantly, RelEx returns additional information besides pairs of objects identified to interact. First, by including gene name identification (Chapter 4), it assigns public database identifiers to the detected objects. Thus, other data sources such as experimental data can be mapped to the objects which enables network-based analysis methods (e.g. Hanisch *et al.* (2002, 2004); Sohler *et al.* (2004)). Second, RelEx provides, for each extracted relation, references to the abstracts where the relation was found. This makes it possible to detect common contexts for the relations. By application of network schemes, contexts can be analyzed systematically, and overrepresented context annotations can be detected; these provide functional descriptions for the analyzed relations. Third, RelEx returns, for each detected relation, a sentence path which contains the subset of terms from a sentence describing a given relation. The paths have proven useful for the classification of relations as activating/inhibitory, physical/indirect, protein-gene/gene-gene. Typed relations will help in analyzing pathways and provide a first step in inferring regulatory cascades.

5.6 Chapter Summary

The discovery of regulatory pathways, signal cascades, metabolic processes or disease models requires knowledge on individual relations such as physical or regulatory interactions between genes and proteins. Most interactions mentioned in the free text of biomedical publications are not yet contained in structured databases.

In this chapter, RelEx, an approach for relation extraction from free text with large-scale applicability, has been presented ([Fundel et al., 2007](#)). It expands on natural language preprocessing by applying a small number of rules to achieve competitive recall and precision. RelEx has been applied on a comprehensive set of one million MEDLINE abstracts dealing with gene and protein relations. Thus a comprehensive human gene and protein relation network consisting of approximately 150 000 relations between 11 000 genes/proteins has been compiled.

The analysis of this network showed that the overlap with physical protein interactions detected by yeast two-hybrid technology is low. This finding is in accordance with other studies that compared literature and experimentally derived interaction networks. Importantly, the generated network contains significantly more interactions than any of the analyzed publicly available interaction networks obtained from manual literature curation or experimental yeast two-hybrid measurements. Yet, the order of magnitude of the network is in the range of published estimates. Furthermore, the network is not restricted to direct physical interactions, but also contains regulatory and other relations.

The RelEx network has several applications: It can be used (1) for methods integrating network and experimental data (e.g. [Hanisch et al. \(2002, 2004\)](#); [Sohler et al. \(2004\)](#)), (2) for fast manual network curation, (3) for searching descriptions of interactions observed in specific contexts by means of network schemes, or (4) for analyzing single relations or subnetworks for common contexts that functionally characterize the respective relations via overrepresentation analysis on automatically compiled context annotations.

Gene and protein relations can be characterized by a classification approach ([Küffner et al., 2006](#)). Therefore, sentences were manually annotated with several attributes describing the respective relation type (e.g. activating, direct physical, gene regulatory). A classifier is then trained to predict interaction types. Here, using the RelEx paths as features entails an increase in performances such that interaction types can be predicted with high accuracy.

Part II

Gene Expression Data Analysis

Chapter 6

Background: Gene Expression Data Analysis

Microarrays represent a cutting-edge technology for analyzing biological processes and phenomena at the molecular level. They enable researchers to monitor expression patterns of thousands of genes simultaneously. Osteoarthritis is an important degenerative joint disease that is in the focus of the gene expression data sets investigated here.

This chapter gives a general introduction to the biological background that is relevant for microarrays (Section 6.1), presents principles of microarray technology (Section 6.2), and describes the necessary and frequently applied data processing steps (Section 6.3). Finally, a short introduction to Osteoarthritis is given (Section 6.4).

6.1 Microarrays – Biological Background

The minimal unit of life is a *cell*. A cell is a highly complex system composed of a wide range of molecules, the most complex being macromolecules (DNA, proteins, and polysaccharides) that govern most of the activities of life. The *central dogma of molecular biology* (Figure 6.1) summarizes the transfer of genetic information in living organisms.

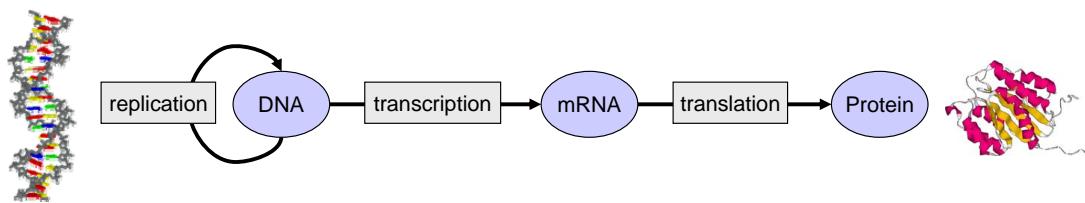


Figure 6.1: The central dogma of molecular biology summarizes how information is transferred from DNA via mRNA to proteins.

Deoxyribonucleic acid (DNA) molecules contain the genetic information of a cell; they store information which is necessary for the synthesis of other macromolecules. DNA is copied and passed on to a cell's progeny by *replication*. When applying the genetic information

stored in DNA, a molecule of *ribonucleic acid (RNA)*, the *transcript*, is synthesized as copy of the DNA template by *transcription*. Primary RNA is processed (e. g. removal of introns by splicing, attachment of poly(A) tail at the 3' end) to end up in mature mRNA. The mature mRNA is transferred from the nucleus to cytoplasm where it serves as template for *translation* into a protein molecule. *Proteins* fulfill manifold tasks in a cell; for example, they catalyze chemical reactions and serve as building blocks for cellular structures.

A *gene* is a segment of DNA that encodes a functional RNA. In the case of a protein coding gene, it contains a coding region (i. e. a sequence of nucleotides that corresponds to the sequence of amino acids in the protein) and regulatory elements. Eucaryotic genes typically consist of interrupted coding regions containing protein-encoding fragments (*exons*) and interspersed non-coding fragments (*introns*). The *genome* describes the total of DNA found in a cell. The *genotype* describes the total of genes and genetic construction of an organism or cell. The *phenotype* describes the characteristics displayed by an organism given a particular set of environmental factors; that is, the outward appearance of an organism. The phenotype of an organism may or may not directly reflect its genotype.

The DNA sequence encodes the genetic information and consists of nucleotides. Each nucleotide contains a phosphate group, a deoxyribose sugar, and one of the four bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The nucleotides are linked by phosphodiester bonds between the 5' hydroxyl phosphate group of one pentose ring of the deoxyribose sugar and the 3'OH group of the next pentose ring; this gives the resulting chain a polarity. Two anti-parallel DNA strands bind together in a right-handed double helix by noncovalent hydrogen bonds between the *complementary pairs* of nucleic acids: Adenine (A) pairs with Thymine (T) and Cytosine (C) pairs with Guanine (G). The process of association of two complementary sequences of DNA is called *hybridization*. Two DNA strands can separate from each other (*denaturation*) and re-form a double helix (*renaturation*) under physiologic conditions. A fragment of DNA is typically written as:



The *genetic code* describes the relation between the DNA sequence and the encoded protein sequence, in that *codons* (i. e. sequences of three nucleotides) correspond to one amino acid. The genetic code is degenerated; that is, several codons code for the same amino acid. Special codons mark the beginning and end of a coding sequence. The sequence between an initiation and a termination codon is called an *open reading frame (ORF)*.

A *mutation* is a sequence modification of a DNA molecule with respect to the parent or original nucleotide sequence by insertion or deletion of one or more nucleotides or by replacement of one nucleotide by another. Mutations can be caused by errors during replication or by spontaneous sequence changes as can be induced by irradiation.

RNA resembles DNA but contains ribose instead of deoxyribose and Uracil (U) instead of Thymine. RNA is typically found as single-stranded nucleotide chain. RNA molecules are subcategorized by their function: *transfer RNA (tRNA)* and *ribosomal RNA (rRNA)* take functionally part in the synthesis of proteins, and *messenger RNA (mRNA)* serve as template for protein synthesis. *Complementary DNA (cDNA)* has the same chemical

constitution as DNA. A cDNA sequence is complementary to the corresponding mature mRNA; that is, it contains only the coding regions. In nature, cDNA is found in viruses. In laboratory, cDNA is synthesized from a mRNA template by the enzyme *reverse transcriptase*. cDNA synthesis is used for amplification and for incorporation of dyes, for example. *Polymerase chain reaction (PCR)* is a technique that *amplifies* or replicates DNA fragments by creating copies. DNA replication is performed by a DNA polymerase enzyme isolated from the bacterium *Thermus aquaticus (Taq)*, or a recombinant version thereof, at high temperatures with fast reaction rates and high fidelity. *Reverse Transcription PCR (RT-PCR)* is a variant thereof starting from mRNA molecules and resulting in numerous sequence copies in cDNA molecules.

Plasmids are independently replicating small extrachromosomal DNA molecules. A *cDNA library* is a set of plasmid vectors with inserted cDNA segments obtained from mRNA by reverse transcription, usually harbored in bacterial clones.

Gene expression describes the process of genes becoming active in that mRNA and protein molecules are synthesized from the template DNA of a gene. Some protein-encoding genes are expressed at an approximately constant rate; these genes generally perform basic reactions within the cell and are called *housekeeping genes*. The expression of most genes is influenced by signals arising from external conditions such as stimulation with chemical agents, nutrition, temperature, disease stage. *Microarray technology* can be used to measure mRNA abundances and thus to analyze the effect of external conditions on the molecular level of gene expression.

6.2 Microarray Technology

Southern and Northern blots have long been used to identify and quantify DNA and RNA molecules, respectively. With these techniques, single genes or transcripts can be identified within nucleic acid samples: Samples are separated by gel electrophoresis, then transferred to a solid support, typically a membrane, and subsequently hybridized with labeled nucleic acid molecules.

Microarrays have evolved from these blotting techniques. The most important advances that lead from blotting techniques to modern microarrays were (1) the use of a solid support that facilitates production, handling and analysis, (2) the development of precise spotting devices for attaching oligonucleotides and cDNA with high density, and (3) the improvement of fluorescent labeling of nucleic acids, fluorescence detection, and image processing.

A *microarray* consists of a solid support on which fragments of nucleic acids are immobilized at precise locations at high density; the fragments are called *probes*. The nucleic acid sample to be analyzed is labeled and hybridized to the microarray. A nucleic acid molecule that is intended to hybridize to a probe on the array is called *target*.

The different technological microarray platforms are described in the following: In *spotted cDNA arrays*, full-length cDNA clones or expressed sequence tags (EST) libraries are robotically spotted on the support. With this technique, custom-designed arrays are easy

to create, the required technical equipment is relatively affordable, and the sequences of spotted cDNA fragments do not need to be known. *Spotted oligonucleotide arrays* contain spotted synthetic oligonucleotides (typically 20–70 nucleotides) as probes. The oligonucleotides can be selected so as to improve specificity by avoiding cross-hybridization. *In-situ oligonucleotide arrays* contain short oligonucleotides (typically 20–25 nucleotides) which are synthesized directly on the array by photolithography; by this technique probes can be synthesized at very high density (about 450 000 probes per 1.28 square centimeter, representing approximately 12 000 target transcripts). Several carefully designed probes per target are synthesized on the array to yield good sensitivity and specificity. This technique is predominately applied by commercial vendors (e.g. Affymetrix GeneChip arrays) for premanufactured arrays.

Affymetrix probes are 20 nucleotides long, and every probe sequence comes as a *probe pair* of a *perfect-match (PM)* and a *mismatch (MM)* probe. The perfect-match probe is exactly complementary to the target sequence. The mismatch probe equals the perfect-match probe except for the nucleotide at the central position, which is replaced so that it forms a mismatch. The mismatch probes serve as controls for cross-hybridization. On Affymetrix arrays, a *probe set* consisting of 11–20 probe pairs is used to detect one transcript.

Probe selection is an important issue for all types of microarrays. In case of cDNA arrays splice variants have to be considered. For oligonucleotide-arrays, the probes should be specific in order to avoid cross-hybridization, they should have similar hybridization properties and exclude palindromic sequences to avoid self-hybridization.

For hybridization detection, the samples need to be labeled. This is typically done by reverse transcription with incorporation of modified nucleotides. The labeling method affects experiment design and data analysis. With Radio-labeling and Biotin-labeling, only one sample per array can be analyzed (*one-channel measurement*). In order to compare two samples, such as healthy versus diseased tissue, one has to hybridize two arrays. With Fluorescence-labeling, two samples can be labeled each one with a different fluorescent dye (e.g. Cyanine 3 (Cy3, green) and Cyanine 5 (Cy5, red)) and then hybridized to a same array (*two-channel measurement*). The two dyes can then be read out by fluorescence detection of the two wavelengths. Experiments with two-channel measurements require only half the number of arrays compared to the one-channel methods. Furthermore, the direct comparison of paired samples on one chip avoids effects caused by differences in handling. The labeled cDNA sample in a hybridization buffer is then added onto the microarray and hybridization is performed under stringent control of hybridization conditions (temperature, pH, ionic strength of the buffer, etc.). Afterwards, the microarray is washed, dried, and scanned. Cy3/Cy5-labeled arrays yield microarray images with red, green, and yellow dots which reflect abundant hybridization of the sample labeled with Cy3, Cy5, or equal hybridization with both samples, respectively, and black dots indicate probes that hybridize with none of the samples.

Limitations of Microarray Technology and Alternatives

Despite their high experimental value and ubiquitous usage microarrays suffer from numerous limitations:

- The activity of a gene product is not directly correlated with the amount of mRNA. The number of protein molecules synthesized from a mRNA molecule varies, so do post-translational modification and degradation rates.
- Microarray measurements only provide relative expression levels between two samples or two experiments
- Dye incorporation and hybridization efficiency depends on sequence composition
- Fluorescence is linear only over a limited range of intensity
- The efficiency of reverse transcription varies between different mRNA molecules.
- cDNA probes might remain in double-stranded form which decreases the number of probes available for hybridization with target sequences
- fluorescence light from sample molecules hybridized to the glass slide and not washed off causes measurement noise.
- Dust, fibers and organic particles can shade fluorescence light and thus cause mis-readings.

Other techniques for measuring gene expression levels are available for the analysis of comparatively small sets of genes. These are therefore typically used as verification methods. *Serial analysis of gene expression (SAGE)* involves the sequencing of very short, carefully selected unique sequence *tags* (i. e. nucleotide sequences of length 9–11 nucleotides) from a sample. The abundance of a tag is interpreted to represent the level of gene expression in the sample. *Northern Blot* is a membrane-based technique: single-stranded mRNA molecules are size-separated by gel electrophoresis, then transferred to a membrane support by capillary transfer, immobilized and finally probed with radioactively labeled single-stranded cDNA complementary to the gene to be detected. Semi-quantitative determination of expression levels is possible by radioactivity measurement. *Quantitative Real-time PCR (QRT-PCR)* monitors fluorescence in every cycle of PCR amplification of a nucleotide sequence. The number of cycles required for obtaining a significant fluorescence signal together with standard curves are then used to determine the original amount of nucleotide molecules.

6.3 Microarray Expression Data Analysis Overview

Acquisition of microarray gene expression data is a multi-step process. After image capturing, the image is segmented. Spots are identified according to the grid of spotted or synthesized probes, and background intensity is estimated. The analysis of location and shape of each spot returns an intensity value for each spot. Low quality spots need to be

marked and discarded from further analysis. Finally, the intensities of spots representing a same transcript need to be combined in order to yield one value per gene and sample. Microarray gene expression data analysis includes a number of sequentially performed data processing steps (for a review see [Allison et al. \(2006\)](#)). Numerous alternatives exist for each step. Typical gene expression data analysis steps and tasks are shortly described in the following.

Normalization is applied for compensating for systematic variations within gene expression data which can be caused by varying efficiency of dye incorporation, by sample and array preparation, background effects, etc. A wide and constantly evolving range of normalization techniques exist. They can be classified by their applicability on one-channel or two-channel expression data, or by their scope (global or local).

Transformation is performed to obtain approximately normally distributed data. Generally, gene expression data is inherently not normally distributed. Yet, parametric tests such as Student's t-test assume normally distributed data. Several transformation methods exist, a frequently used one is the log-transformation.

Detection of Differentially Expressed Genes describes the task of detecting genes that are differentially expressed between classes of samples, such as different tissues, disease stages, or experimental conditions. Differentially expressed genes can give hints on biochemical reactions or pathways responsible for observable differences between sample classes; they can be used as disease marker, or even as disease treatment targets. Differentially expressed genes are detected by determination of *fold change* (i. e. the ratio of expression values for a gene between two classes) and *p-value*, which indicates the statistical significance and thus the level of confidence in the designation of a gene as being differentially expressed.

Unsupervised Data Analysis reveals inherent structure in gene expression data without imposing information on the the desired outcome. This can be useful to identify sample or gene subgroups, such as disease subtypes or genes involved in a same pathway. For example, *Cluster Analysis* partitions data in subsets and thus reveals the global organization of data. Examples of clustering methods include hierarchical agglomerative clustering, k-means clustering, and self-organizing maps. *Principal Component Analysis* is a global linear method for dimension reduction that projects the original data to a number of principal components.

Supervised Learning Methods start from a set of input objects (e.g. samples) and desired outputs (the training set) and generate a model that maps the input data to the predefined outputs. The learned model can then be applied on unlabeled samples for sample classification or prediction of a continuous output value. *Supervised Classification* or *Discriminant Analysis* analyzes a number of classified samples and constructs a discriminant or prediction rule for assigning a class label to unseen samples. Corresponding methods are: *Support Vector Machines*, *Neural Networks*, *Classification Trees*, *Nearest-Neighbor Rules*. The latter two and some other of these supervised classification methods can also indicate the *features* (e.g. genes) that are most relevant for the prediction.

Feature selection is concerned with reducing the number of genes to be analyzed with the

aim of selecting the most relevant genes with respect to a given criterion. Feature selection can be useful for increasing speed and performance of unsupervised as well as supervised data analysis methods.

Microarray gene expression data is typically represented as a matrix of expression levels. An experiment monitoring N genes for M samples is described by a matrix $N \times M$, where each column contains the expression levels of the N genes for one sample (the *expression signature*) and each row contains the expression levels of one gene for all M samples (the *expression profile*).

Several public microarray repositories have been set up to share gene expression data: ArrayExpress ([Parkinson et al., 2007](#)), GEO (Gene Expression Omnibus, [Barrett et al. \(2007\)](#)), and SMD (Stanford Microarray Database, [Demeter et al. \(2007\)](#)). Considerable international effort is underway to standardize the way information is stored so as to ensure that all information required for data interpretation is also publicly available and data from different experiments can be combined. MIAME (Minimum Information About a Microarray Experiment) ([Brazma et al., 2001](#)) specifies the minimum information that should be reported on a gene expression experiment for being entered in a public database.

Basically, gene expression data represents a snapshot of gene expression activity for a given set of samples and defined experimental conditions. Single gene expression measurements do not contain sufficient information to infer regulatory contexts; only more comprehensive experiments, such as time-course experiments, allow to derive more detailed or causal models. During the last years, several directions for integrated gene expression data analysis have been tackled. Bayesian networks have been used to model complex stochastic processes and thus to describe interactions between genes ([Friedman et al., 2000](#)). The Microarray Experiment Functional Integration Technology (MEFIT) is a scalable Bayesian framework that can be applied on large compendia of microarray datasets and predicts functional relationships within the context of specific biological processes ([Huttenhower et al., 2006](#)). Finer structured interactions between genes, such as causality, mediation, activation, and inhibition can also be discovered from bayesian networks ([Pe'er et al., 2001](#)). Integration of a large number of gene expression experiments can be used for detection of modules; that is, gene sets that participate in specific biological processes, their regulators and the conditions under which regulation occurs ([Segal et al., 2003a,b, 2004](#)). Second order expression analysis ([Zhou et al., 2005](#)) integrates data from different platforms by extracting expression patterns from each data set and then analyzing patterns across multiple data sets. Thus, genes of the same function yet without coexpression patterns can be identified, and transcription factor activities can be quantified.

Furthermore, analysis gene expression data in conjunction with data derived from other sources appears promising.

6.4 Osteoarthritis

Osteoarthritis is an important degenerative joint disease (for reviews see Aigner *et al.* (2002, 2003, 2004, 2006b)); it is the most common disabling condition in the Western world and thus represents a very significant social and economic burden. The risk of developing osteoarthritis increases with age. Osteoarthritis is a multifactorial disease that affects in particular articular cartilage (mainly knees, hips, finger joints) and is characterized by sequential loss of cartilage. Due to the aging society, the number of patients, whose quality of life is severely affected by the disease, increases. Risk factors include unphysiological loading, genetic predisposition, endogenic factors, and injuries. Osteoarthritis implies diverse underlying patho-physiological mechanisms and is thus often referred to as disease group.

The most frequent symptoms of OA are pain, especially start-up pain, and reduced range of motion of the concerned joint. The early disease stage is characterized by roughening and degradation of joint cartilage and resynthesis of cartilage matrix. In the late stage, the subchondral bones are thickened and show outgrowths, cartilage is increasingly lost, and smaller joints can even fuse (Figure 6.2). The disease process affects, besides joint cartilage, the entire joint structure including the synovial membrane, subchondral bone, ligaments, and periarthritis muscles. The synovial membrane is often affected by inflammation (Synovitis), which sometimes can reach the extent observed in (mild) rheumatoid arthritis, and is reflected in many of the signs and symptoms of OA, including joint swelling, stiffness, and sometimes redness of the overlying skin.

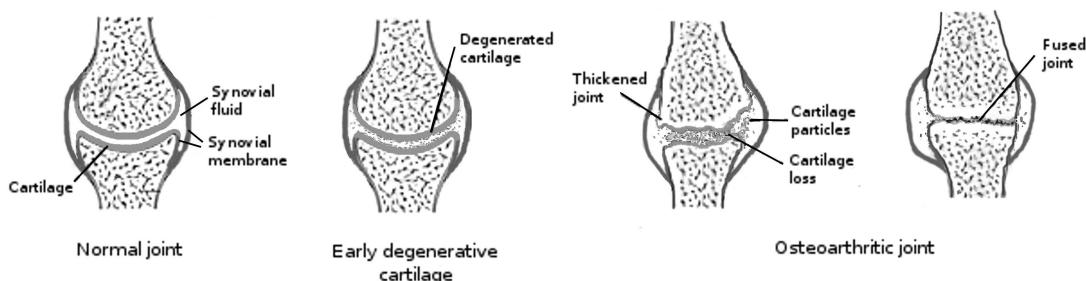


Figure 6.2: Articular capsule of a normal joint, a joint with early degenerative cartilage showing a roughened surface, and a joint affected by osteoarthritis, which may lead to joint fusion.

So far, therapy is limited to pain medicamentation (e. g. by non-steroidal anti-inflammatory drugs (NSAID) or cyclooxygenase (COX)-2-specific inhibitors), physiotherapy (e. g. exercise, temperature therapy), and finally the replacement of the affected joint by an endoprosthetic surgery. None of the drugs on the market is disease modifying and thus slows, halts, or reverses the disease progress. Consequently, there is a high need for the development of disease modifying agents in order to improve quality of life as well as to reduce the enormous socio-economic burdens of the disease.

Joint Physiology

Joints are highly specialized organs that allow largely frictionless movements. Joints consist of the joint capsule, cartilage, ligaments, subchondral bone, and different types of connective tissues (Figure 6.2). Articular cartilage covers the joint surfaces and is responsible for sustaining high loadings and allowing pain-free movements. More than 95% of the volume of articular cartilage consists of extracellular matrix.

The synovial membrane plays a crucial role in nourishing the chondrocytes as well as removing metabolites and matrix degradation products from the synovial space. The synoviocytes (cells of the synovial membrane) maintain the basic metabolic homeostasis of the joints and produce large amounts of glycoproteins and hyaluronic acid, which provides the joint surfaces with its gliding capacity. Substances such as nutrients, metabolites, and oxygen are transferred between the synoviocytes and the articular cartilage by diffusion via the synovial fluid.

The synovial lining cells are capable of secreting matrix-degrading proteases (MMPs) and catabolic cytokines (IL-1, TNF-alpha). Cartilage matrix catabolism might be in part induced or promoted by these catabolic mediators. Inflammatory cytokines such as IL-1 β and TNF- α are top candidates for therapeutic intervention because both are able to down-regulate matrix anabolism in articular chondrocytes and to induce expression and secretion of matrix degrading proteases.

Articular Cartilage

Adult articular cartilage is avascular, alymphatic, and aneural; that is, cartilage does not contain blood vessels, lymphatic vessels, or nerves. Cartilage contains only one type of cells, the *chondrocytes*. These are sparsely distributed in the cartilage and responsible for matrix turnover; that is, controlled degradation and resynthesis. Major constituents of the extracellular matrix are collagens, proteoglycans, and a heterogeneous group of other proteins. The major part of the cartilage matrix consists of a fibrillar and an extrafibrillar matrix. The main constituents of the fibrillar matrix are collagen type II, IX, XI, XVI. The collagen network is responsible for the tensile strength; it hinders expansion of the aggrecan component and provides stiffness to the tissue. The aggrecan-hyaluron aggregates are highly hydrophilic and bind intercellular water, which is responsible for the elasticity of the tissue. When compressed, the cartilage matrix is compliant; when unloaded, water is drawn into the matrix and thus it regains elasticity. Besides collagen fibrils and aggrecan aggregates, a large number of other components, such as small non-aggregating proteoglycans (e.g. decorin, biglycan, and fibromodulin), are important for matrix cohesion and cross-linking the collagen network with the proteoglycans inbetween.

A balanced turnover is required for proper functioning of the matrix. In normal adult articular cartilage, the half life of aggrecan ranges from days to months; some components have been shown to persist for years. The normal proteolytic turnover is highly regulated and is most probably implemented by Matrix Metalloproteases (MMPs), particularly MMP-3.

The collagen type II network is extremely stable; destabilization can only be brought about by cleavage by collagenases, especially MMP-1 and MMP-13.

Primary osteoarthritis generally results from an imbalance between mechanical stress and the physico-chemical ability of the articular cartilage to resist the stress. The disease is primarily characterized by cartilage destruction (Figure 6.2), even though degradation processes within the surrounding tissues also play an important role. The destruction of articular cartilage is largely due to the destruction and loss of the cartilage matrix, which results from an imbalance between degradation and de novo synthesis of matrix components. Loss of aggrecan is characteristic of the early stages of cartilage degeneration. The overall collagen content remains rather constant throughout the disease progress, but the collagen network is loosened. Degradation processes appear to be specifically prominent in the surface zone and around the chondrocytes in osteoarthritic cartilage. Enhanced levels of metalloproteinases (MMP-2, -7, -8, -9, -13, -14, ADAMTS-4, -5, ADAM-10, -15) have been reported to accompany the increased matrix degradation, but it is not yet clear which proteases are crucial for the degradation. Arthritic cartilage also shows expression of molecules that are not present in normal cartilage (e.g. tenascin, collagen types IIA and III). The composition of the proteoglycans changes. Collagen type X becomes a prominent component in the calcified cartilage components and might be involved in the calcification process that is characteristic for osteoarthritic cartilage degeneration. Collagen type VI is synthesized and degraded at an increased rate and its distribution within the cartilage changes, which might be responsible for the changes in physical properties.

Chondrocytes

Cartilage contains only one type of cells, the chondrocytes. For microarray analysis, this represents an important advantage, as the general issue of diverse cell populations in a tissue and thus unclear origin of the signal does not apply. Chondrocytes are responsible for the controlled turnover of the extracellular matrix (Figure 6.3).

In normal articular cartilage, proteoglycan shows turnover, but collagen type II turnover is very low to non-existent. During the osteoarthritic disease process, chondrocytes are heterogeneous and show, depending on the region or zone of the cartilage and degradation stage, different reaction patterns: cell death (apoptosis/necrosis), proliferation to compensate for cell loss, and increase of synthetic activity. Chondrocytes can undergo phenotypic modulation by modifying the overall gene expression profile; for example, synthetic activity can be modified by regulation of anabolic gene expression.

In the early-stage osteoarthritic cartilage chondrocytes neoexpress genes, such as enzymes that degrade the matrix, cytokines, and growth factors relevant for modifying the catabolic processes. The degradation phenotype of chondrocytes is initially observed in the superficial zone of the cartilage and later propagates to the intermediate and deep zones and may involve epigenetic mechanisms like the loss of DNA methylation of the promoters of MMP-3, -9, -13, and ADAMTS-4.

In OA, chondrocytes show enhanced synthesis of extracellular matrix components; thus, they attempt to repair the damaged matrix. Collagen type II expression is more up-

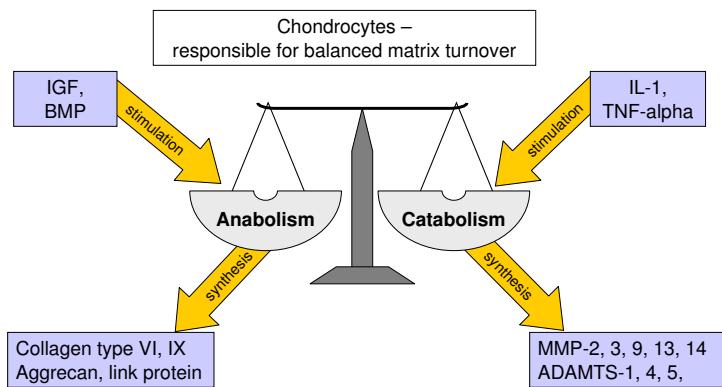


Figure 6.3: Chondrocytes are responsible for maintaining balanced turnover of extracellular matrix. Substances such as cytokines can stimulate anabolism or catabolism of matrix components.

regulated than aggrecan, which resembles expression in fetal cartilage. In osteoarthritic cartilage, a net loss of proteoglycan has been shown.

Cytokines and growth factors are responsible for regulating anabolic and catabolic activity as well as phenotypic alterations of chondrocytes (Figure 6.3). Insulin-like growth factor (IGF) 1, bone morphogenic proteins (e. g. BMP-2, BMP-7), and transforming growth factor beta (TGF- β) promote anabolic activity; i. e., they up-regulate expression of collagens and proteoglycans. Pro-inflammatory cytokines, primarily tumor necrosis factor alpha (TNF- α) and interleukin 1 beta (IL-1 β) stimulate expression of degradative enzymes and thus promote catabolic activity. Besides these factors, other molecules such as chemokines, nitric oxide, and oxygen radicals are also involved in the maintenance of homeostasis and controlled matrix turnover. Most of the anabolic and catabolic factors are synthesized by the chondrocytes in an auto- and paracrine manner. Thus, functional genomics of chondrocytes in terms of understanding the cellular gene expression patterns appears to be an important clue in understanding the disease.

In osteoarthritis research, cartilage degradation is assumed to be tightly related to three underlying main phenomena (Table 6.1), all of which are important for understanding the degeneration process and the disease. Especially, the imbalance between anabolic and catabolic events in osteoarthritis represents one highly interesting target for therapy.

Chondrocyte behavior	Functional category of genes	Marker Genes	
		up-regulated	down-regulated
hypo-anabolism	Suppression of anabolic activity		Col2, Aggrecan
hyper-catabolism	Increase of catabolic activity	MMP-2, MMP-13	
phenotypic alterations	Modulation of phenotype	Col3, Col10	

Table 6.1: Marker genes for different categories of functional processes relevant for cartilage degradation.

Models for Understanding Osteoarthritis

Osteoarthritis is difficult to analyze on the molecular level as samples from human joint cartilage are not readily available. Samples of normal and slightly affected joint cartilage can only be obtained from autopsies; these samples might be affected by post-mortem processes. Samples of diseased joints are available from replacement surgery. The respective patients are generally severely affected by the disease, and thus the amount of cartilage that can be obtained from a joint is small. Besides the study of human samples, other specific approaches and models are applied for understanding the disease mechanisms and functional validation of the role of individual genes:

The '**evo-devo**' approach is based on the observation that reaction patterns of cells and tissues in the adult often resemble processes occurring during development. Processes such as matrix anabolism and catabolism as well as cellular differentiation are central in osteoarthritis and are also observed in the fetal growth plate. Development of fetal growth plate cartilage has thus been used for studying chondrocyte behavior. The validity and limitations of this approach remain to be analyzed in detail.

***In vivo* models** are the most accurate possibility for studying a disease, especially in terms of studying how to modify the disease processes. Induced animal systems for specific aspects of osteoarthritis exist, but no animal model for the disease as a whole.

***In vitro* models** (i. e. cell cultures) represent a useful means for studying phenomena for which *in vivo* models are not readily available or their usage should be reduced in response to social and ethic responsibility. Isolation and cell culture can lead to destabilized phenotypes or altered gene expression patterns. Thus, these systems need to be compared to their *in vivo* counterparts.

Chapter 7

Gene Expression Data Analysis

Microarrays make it possible to capture the expression level of thousands of genes simultaneously. Integrated data analysis systems require appropriately processed data. Most approaches require as input a set of genes where genes are assigned with numerical values that describe the magnitude (fold change) and statistical significance (p-value) of differential expression between two sample classes. Besides, many approaches require as input the set of all genes that have been analyzed, and a subset thereof that should contain just the differentially expressed genes. Numerous methods exist for the determination of fold changes and p-values for differential expression.

This chapter first describes the osteoarthritis-related data sets (Section 7.2) that are used for the data processing approaches presented in this chapter.

An extensive study points out the effects of primary data processing with a focus on normalization and gene p-value determination. Guidelines for the selection of methods are described and a method for gene p-value determination is presented (Section 7.3, Fundel *et al.* (2005b)).

In Section 7.4, a model-selection approach is presented that selects an appropriate combination of feature subset selection method and classifier and derives subsets of genes which lead to good classification performance and are stable; that is, they are consistently selected for many sample subsets (Davis *et al.*, 2006). This approach is joint work with Caroline Friedel, Robert Küffner, Chad Davis, Fabian Gerick, and Volker Hintermair.

In the next chapter (Chapter 8), the presented data sets are analyzed from a more biological point of view.

7.1 Introduction and Literature Review

Microarrays make it possible to measure the expression level of a large number of genes simultaneously; thus, expression profiles can be investigated under different conditions.

Today, numerous methods and tools exist for analyzing gene expression data. New normalization techniques are presented (e. g. Edwards (2003); Futschik and Crompton (2004);

[Wilson et al. \(2003\)](#); [Zhao et al. \(2005\)](#)), so are methods for detecting differentially expressed genes (e.g. [Comander et al. \(2004\)](#); [Cui and Churchill \(2003\)](#); [Cui et al. \(2005\)](#); [Tusher et al. \(2001\)](#); [Yan et al. \(2005\)](#)). A number of tools analyze microarray data in a largely automated way (e.g. [Chung et al. \(2004\)](#); [Hanisch et al. \(2002\)](#); [Herrero et al. \(2004\)](#); [Knudsen et al. \(2003\)](#); [Pandey et al. \(2004\)](#)), many of them even integrate gene expression data with further information obtained from ontologies, pathway databases, or text mining.

Yet, comparisons between different normalization methods focused mainly on Affymetrix and two-channel cDNA microarrays (e.g. [Bolstad et al. \(2003\)](#); [Park et al. \(2003\)](#)), and do not consider sample groups. Generally, existing literature offers little guidance on how to decide which method to use, how to compare different methods and their outcomes, and how to check the correspondence of possible outcomes to biological expectation and downstream interpretation; especially for one channel cDNA data.

It is crucial to determine an appropriate combination of individual processing steps for a given dataset in order to ensure the validity and reliability of expression data analysis.

Today, gene expression data analysis tends towards large scale integration, while specific methods (e.g. for normalization, detection of marker genes) are also continuously being developed. Importantly, large scale data integration and advanced data analysis methods strongly depend on the results of preprocessing methods.

7.2 Data Sets

Several microarray gene expression data sets are analyzed in the following sections. All deal with osteoarthritis (OA, see Section 6.4 and [Aigner et al. \(2002, 2003, 2004, 2006b\)](#)).

- **GPC¹ four-class data set:** 83 samples of human articular cartilage classified into four osteoarthritis related groups were analyzed with custom design radio-labeled cDNA microarrays. The 78 samples remaining after outlier analysis are classified as follows: 18 normal (*n*), 20 early degenerative cartilage (*e*), 21 peripheral OA (*p*), and 19 central OA (*c*).

This data set contains the largest number of samples among the analyzed data sets; given the difficulty of obtaining human joint samples, this represents a large data set in the domain. This data set is used for analyzing normalization and data processing effects (Section 7.3), for applying the classification procedure described in Section 7.4.1, and for detailed biological analysis of gene expression patterns (Section 8.1).

- **Affymetrix² two-class data set:** 26 samples of human articular cartilage classified into healthy and osteoarthritic tissue were analyzed with Affymetrix Human Genome U133 Plus 2.0 arrays. After outlier removal, the data set contains 13 samples of normal articular cartilage and 12 samples of osteoarthritic cartilage.

¹<http://www.gpc-biotech.com/>

²<http://www.affymetrix.com/>

- **Zeptosens³ two-class data set:** 10 samples of osteoarthritic human articular cartilage were compared against pooled normal samples from 9 donors with custom-designed two-color Zeptosens SensiChip microarrays containing 865 distinct probes each.
- **In vivo–in vitro data set:** 12 samples of human articular cartilage, split into healthy and osteoarthritic, directly after extraction (in vivo) and three days after transfer into cell culture (in vitro), were analyzed with Affymetrix Human Genome U133 Plus 2.0 arrays (i.e. three replicates, four classes). This data set is used to analyze the validity of chondrocytes in cell culture as a model for chondrocytes in the living organism (Section 8.3).
- **Time series data set:** A time series of chondrocytes stimulated with IL-1 was obtained from three samples measured at 5/6 time points with Affymetrix Human Genome U133 Plus 2.0 arrays. IL-1 stimulates catabolic processes in chondrocytes which play an important role in cartilage degradation (see Section 8.4).

Experiments comparing normal and osteoarthritic cartilage were conducted with different microarray platforms (GPC, Affymetrix, Zeptosens). These measurements allow us to compare the platforms against each other (Section 8.2).

7.2.1 GPC Four-Class Data Set

The GPC four-class data set has been used for the analysis of several aspects and is thus referred to in various sections (Sections 7.3, 7.4.1, and 8.1). Therefore, it is described in more detail in the following.

Sample Preparation

Cartilage from human femoral condyles was extracted for gene expression analysis. 83 samples of human cartilage were analyzed. Normal articular cartilage (termed *n*, 18 samples, 45 to 88 years) and early degenerated cartilage (*e*, 20 samples, 43 to 91 years) was obtained from autopsies, within 48 hours after death. Osteoarthritic cartilage was obtained from total knee replacements: 21 samples were assigned to low grade (termed *p* for *peripheral*, 61 to 84 years); and 19 samples classified as moderate/high grade (termed *c* for *central*, 61 to 84 years). Classification into the disease groups was performed by domain experts. Cases of rheumatoid arthritis were excluded from the study. Only primary degenerated and no regenerative cartilage (osteophyte tissue) was used. For details on RNA extraction see [McKenna *et al.* \(2000\)](#).

Array Production

Custom designed cDNA microarrays were produced and measured by GPC-Biotech AG (Martinsried, Germany). Two cDNA libraries were constructed from mRNA pools using

³<http://www.zeptosens.com/>

chondrocytes obtained from normal and osteoarthritic joints and transfected into *E. coli*. Bacterial colonies were plated; clones were picked and subjected to oligonucleotide finger-printing. The inserts were PCR amplified and spotted onto nylon filters. Short oligonucleotides were radioactively labeled and hybridized to the filters. Clones with sufficient hybridization information were subjected to clustering which resulted in 8821 different clusters. A part of the sequences had been preselected for OA-relevant genes.

More than 700 identical cDNA arrays were produced; cDNAs were spotted with a needle printer. Each microarray contains 7467 spots; 5517 spots represent 3648 genes. There are 1 to 74 spots per gene on the array, and 1062 genes are represented by more than one spot. cDNA synthesis for expression analysis was primed using random hexamers and ^{33}P for labeling. For each sample, four identical arrays were hybridized. After washing, filters were read out by a phosphor imager and then scanned. Proprietary software was used for determining spot intensities and local background correction. Duplicate spots on an array that showed too much difference were eliminated from further processing. Replicate data was used to estimate the probability of significant expression and to verify correlation between measurements.

For expression level verification, real-time PCR was performed for ten selected genes (Agrecan, btg2, collagen types I, II, and III, GAPDH, GPX3, MMP-3, SOD2, SOX9, tob1).

7.3 Analyzing the Effects of Primary Data Processing

Integrated data analysis requires appropriately preprocessed data. Most approaches require fold changes and p-values that quantify the differential expression of genes. Fold changes and p-values depend on the methods applied for calculating the respective values, and on the prior data processing steps.

In the following, a study is presented that investigates how the *higher-level* outcome of a microarray experiment, namely a list of differentially regulated genes, is related to the *low-level* details of data processing. The focus is on normalization and methods for the identification of differentially expressed genes.

Therefore, different normalization techniques are applied to the GPC four-class data set and the differences in the final result are evaluated. Furthermore, different methods for combining spot p-values into gene p-values are analyzed. The proposed method (Stouffer's method) can be considered as an alternative to standard procedures that shows advantages for data that exhibits large inter-spot expression value differences.

Finally, the large number of samples makes it possible to perform a stability analysis on the significantly regulated genes. It has been shown ([Michiels et al., 2005](#)) that in numerous published large studies on differential gene expression differentially expressed genes are highly unstable for subsets of the analyzed samples. Thus, a procedure for estimating the error via a robustness analysis is proposed.

Background Knowledge

The GPC four-class microarray experiment was conducted to identify differentially expressed genes for the group pairs $n-e$, $n-p$, $n-c$, $e-p$, $e-c$, $p-c$, $n-l$, $e-l$ (where l is the combined set of p and c). For the given data, the following background knowledge is available and corresponding expectations apply for data processing:

1. It has experimentally been confirmed that mRNA content was the same for all sample preparations, and thus expression intensities are expected to be similar for all measurements.
2. The number of up- and downregulated genes is expected to be balanced for each comparison. This general biological expectation should always hold, provided that the experiment does not investigate specific activation events.
3. From previous experiments, it is known that the degree of similarity varies substantially between the sample classes. Specifically, n and e as well as p and c are very similar, whereas n is very different from p and c and, consequently, also from l . Previous cluster analysis showed a good separation between the class pairs ne and pc , whereas the groups n and e as well as p and c were not separated from each other. This leads to the expectation that more genes are significantly regulated in the comparisons $n-p$, $n-c$, and $n-l$ than in the comparisons $n-e$ and $p-c$. Interestingly, also in terms of clinical staging, n and e and p and c resemble each other, whereas the two group pairs ne and pc are clearly distinct.

Several of the effects described in the following sections are linked to the data distribution of the analyzed data set. The expression value distribution (Figure 7.1) shows most data concentrated in a very small range (75% of the values are <0.13 , 99% are <5.91) and some values are significantly larger (overall maximum at 539.9). Due to the technique (cDNA spots of different sequences, radioactive detection), the expression values for different spots representing the same gene can vary significantly. The distribution differs significantly from a log-normal distribution (Figure 7.1, right panel).

7.3.1 Normalization

The analyzed data has already been subjected to within-array normalization. Between-array normalization is applied to remove systematic variances between arrays. In the following, several methods for between-array normalization are applied to the data and the results are analyzed. The group-level plot is presented as an appropriate tool for visualizing differences in expression levels between groups.

Normalization Methods

Centralization ([Zien et al., 2001](#)) estimates for each pair of arrays the quotient of the constants of proportionality and subsequently computes an optimally consistent scaling for the samples based on the matrix of pairwise quotients. Centralization needs two parameters

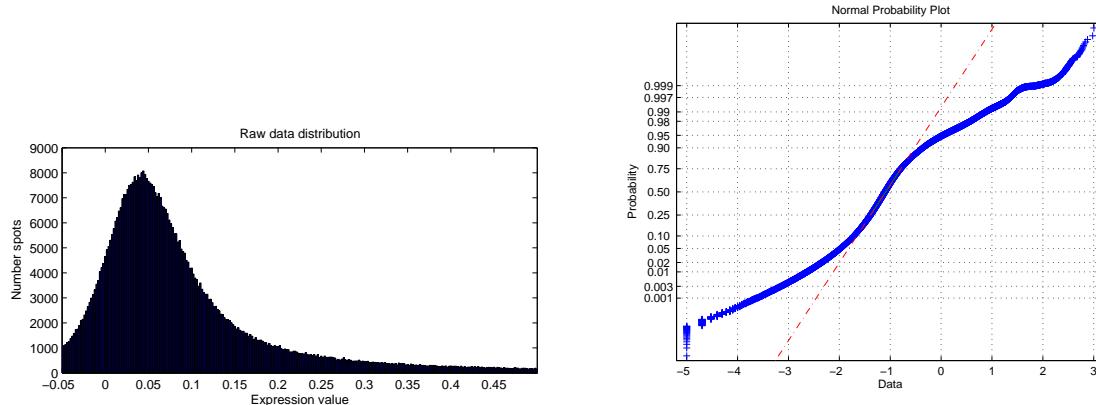


Figure 7.1: Raw data distribution (left panel) and normal probability plot (right panel) of the GPC four-class osteoarthritis data set.

describing the range of reliable measurements to be used (here: 0.03–1).

Percentile Normalization adjusts a certain percentile to the same level for all samples by applying a multiplicative factor to each sample. Here, the generally applied 50% (eq. median) and 75% percentiles have been used.

MAD Scale Normalization adjusts the median and median absolute deviation (MAD), which are robust measures for the location and spread of a distribution, to a common level (typically 0 and 1, respectively). Here, a variant is applied that transforms the data spread and location to the original scales. For each sample k and spot s the normalized value x'_{ks} from the original value x_{ks} is determined by:

$$x'_{ks} = \frac{x_{ks} - \text{median}(x_k)}{\text{MAD}(x_k)} \cdot \text{MAD}(X) + \text{median}(X)$$

with : $\text{MAD}(x_k) = \text{median}(|x_k - \text{median}(x_{k\cdot})|)$

where: MAD : median absolute deviation; X : entire dataset.

Variance Stabilization (Huber *et al.*, 2002) incorporates data calibration, an intensity-dependent error model and data transformation; it is intended to lead to an intensity independent measure of differential expression.

LOESS (Cleveland, 1979) (Local Regression) fits simple models to data segments defined by measured intensity; thus it does not require to specify a global function of any form to fit a model to the data.

Quantile Normalization normalizes the distributions of the expression values (i. e. each quantile) for each array.

Flooring The raw expression intensities contain negative values due to background correction of the original data performed by GPC-Biotech. Negative data as well as expression values close to zero are not appropriate for computing fold changes and p-values. The floor value was estimated by analysis of p-values versus spot expression values: p-values smaller

than 10^{-3} were based almost exclusively on expression values above 0.01. Thus 0.01 was used as floor value.

Analysis of Normalized Data

Generally, normalization effects are visually inspected in boxplots (Figure 7.2, left panel) which show for each sample (on the x-axis) the 25% percentile and 75% percentile of the data as box and the median as horizontal line within the box; whiskers are of a length proportional to the interquartile range, and all data points lying outside these whiskers are displayed individually as outliers. Boxplots of data with a distribution like shown in Figure 7.1 show small boxes and whiskers but numerous outlier values.

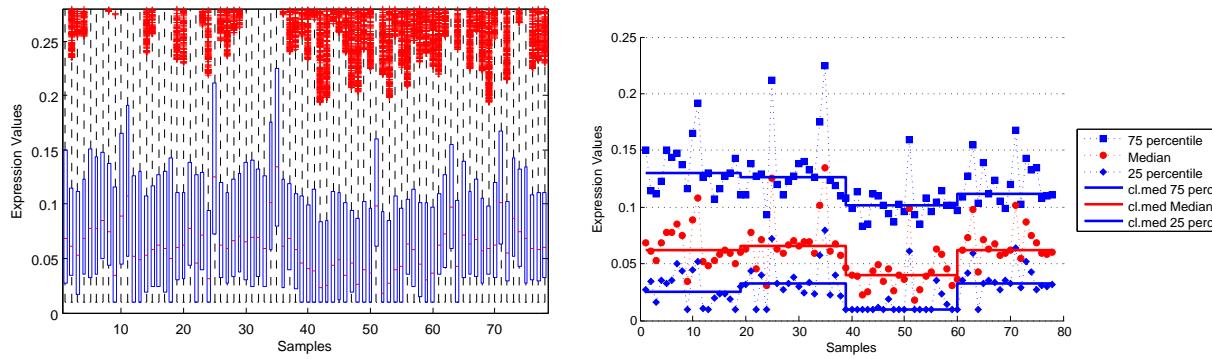


Figure 7.2: Boxplot (left panel) and group-level plot (right panel) for the GPC four-class OA data set ($n:1\text{--}18$, $e:19\text{--}38$, $p:39\text{--}59$, $c:60\text{--}78$). The group-level plot shows the 25%, 50% and 75% percentile for each sample (as does the boxplot) and additionally shows the median over these values for each sample group representing different disease stages (cl.med: class median).

The **group-level plot** (Figure 7.2, right panel) depicts group-specific variations. Differing expression levels between sample groups affect the p-value and fold change calculation and can result in an artificially high number and/or biased direction of differentially regulated genes. The group-level plot shows, in addition to the 25%, 50%, and 75% percentiles of the individual samples the group-levels; that is, the median of the 25%, 50%, and 75% percentiles over all samples belonging to the same group.

Normalization can significantly alter group-levels and thus also p-value and fold change distributions (Figure 7.3). For the analyzed data without between-array normalization, more genes appear upregulated than downregulated from p to c due to the differences in group level. This must be due to a systematic error as the total mRNA content was the same for all samples, and thus approximately the same number of genes are expected to be up- and down-regulated. After 50% percentile normalization, more genes appear downregulated. Only after 75% percentile normalization, MAD scale normalization, Variance Stabilization, LOESS, or Quantile Normalization up- and downregulation appears balanced. LOESS makes use of a parameter, the degree of smoothing, that determines how much of the data is used to fit each local polynomial. The normalization result depends

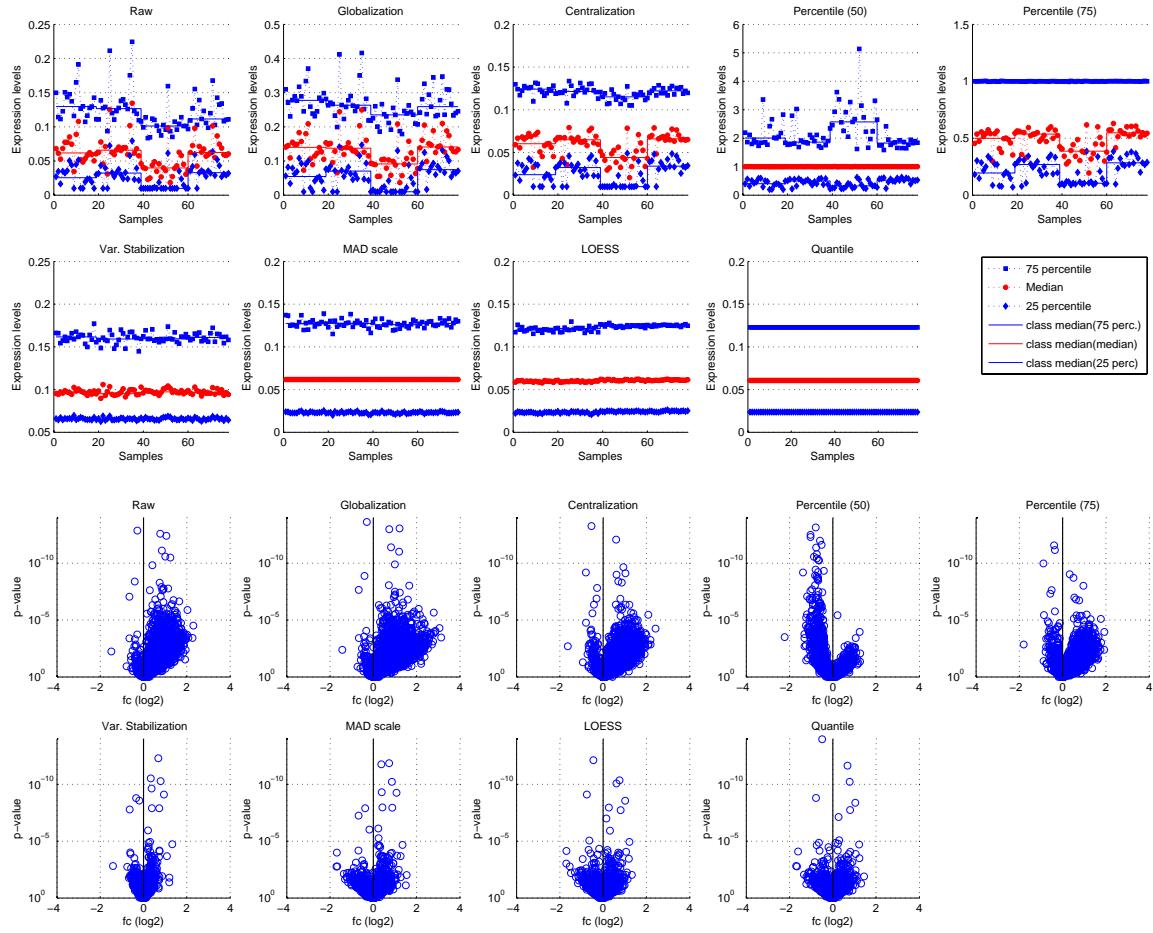


Figure 7.3: Effects of normalization for group comparison p versus c for raw and normalized data: group-level plots (upper panel, $p:39-59$, $c:60-78$) and volcano plots (lower panel).

significantly on this parameter; for example, if the parameter is set to 40% of the data size, the sample dendrogram is significantly altered compared to the other normalization methods.

Between-array normalization methods differ in how rigorously they modify the original data. Some methods apply a multiplicative factor (e.g. globalization, percentile normalization, centralization), or a multiplicative and an additive factor (e.g. scale normalization). Other methods fit the location and the shape of the original distribution (e.g. quantile normalization), or adjust data in a intensity-dependent way (e.g. variance stabilization, LOESS); these modify the original data most.

Far more normalization techniques than the ones analyzed here exist (e.g. [Yang et al. \(2002\)](#); [Smyth and Speed \(2003\)](#); [Edwards \(2003\)](#); [Ballman et al. \(2004\)](#)); most of the newer normalization techniques are non-linear. Some of them focus on two-channel or

Affymetrix-type data and can not easily be applied to other kind of data; others can directly or after slight adaptation be applied to, say, one-channel cDNA microarray data. Generally, normalization is desired to modify the underlying data as slightly as possible, but as much as necessary to remove systematic biases, yet, to conserve biological relevant information. Thus, the normalization method should be selected in accordance to the data set under investigation with the general guideline *enough* normalization with only slight data modification. Clearly, this entails several issues: A data processing pipeline cannot be fully automated as user input is required for certain decisions; the result depends on the available knowledge about the analyzed data, and, to some extent, the result depends on the judgement of the individual researcher.

7.3.2 Differential Expression

Differential expression is quantified by the statistical significance of the change in expression level (p- or q-value) and by the difference in expression level between the two investigated sample groups (fold change).

Why Combining p-Values?

Generally, genes are represented on a microarray by a varying number of spots (or probe sets). The spots for a given gene often contain different nucleotide sequences; they can cover distinct regions within the gene sequence, vary in binding affinity and specificity, represent splice variants, and annotation can be of varying reliability. These differences can result in high variability of measured expression intensities. Most advanced microarray data analysis methods as well as direct biological interpretation of gene expression experiments require for determination of differentially expressed genes. Thus, data obtained for individual spots need to be combined into data for genes.

Often, gene expression levels for individual spots are simply averaged, frequently after log-transformation, to yield a gene expression level. This approach is not appropriate if spot data for a gene is spread over several orders of magnitude (such as in the given data set), as illustrated by an example of the effects for artificial data (Figure 7.4): A gene is measured by two spots in a comparison of two sample groups (1–20, 21–40). One spot yields high (green), the other low intensities (red). In the left plot only the low-intensity spot shows significant regulation while in the right plot only the high-intensity spot shows significant regulation between the groups. The average in linear space (blue) is more similar to the signal of the high-intensity spot, while the average in logarithmic space (pink) is more similar to the low-intensity spot, in terms of expression values as well as in terms of p-values for differential expression.

P-value combination represents an alternative to the above approach. Importantly, its overall result does not depend on the intensity of the individual regulated spot.

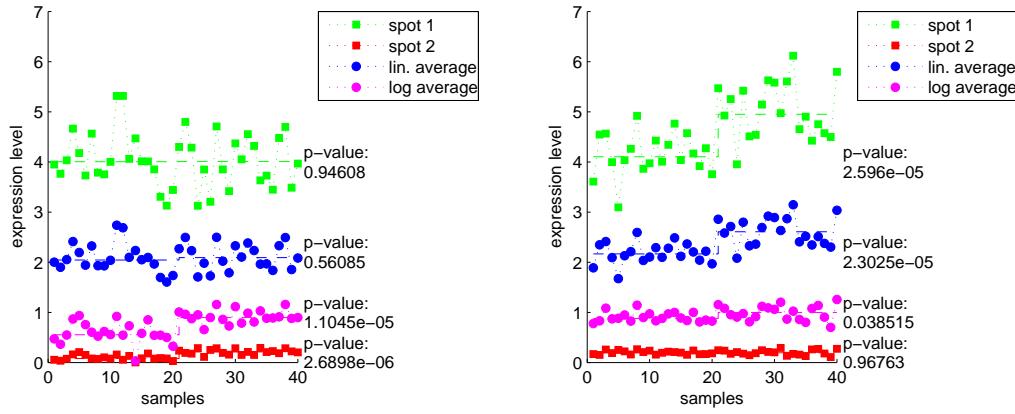


Figure 7.4: Example of possibilities for combining spot expression values into gene expression values and corresponding p-values. Each plot shows artificial expression values for a gene measured by two spots (red, green) in a comparison of two sample groups of 20 samples each. For details see text (Section 7.3.2).

P-Value Combination

Gene p-values that quantify the statistical significance of differential expression are calculated based on spot p-values (determined by the two-sided Wilcoxon ranksum test) by one of the following methods:

- *Fisher's inverse chi-square method* (Fisher, 1932). This method uses the fact that given a uniform distribution U , $-2 \cdot \log(U)$ has a chi-square distribution with two degrees of freedom, and the sum of two independent chi-square variables is again chi-square distributed (with four degrees of freedom). The combined p-value $p_{chi}(g)$ for a gene g can be computed as:

$$p_{chi}(g) = 1 - \chi^2_{2d}(\sum_s -2 \cdot \log(p_s))$$

where p_s is the p-value for spots s representing gene g , d is the number of spots s representing gene g , and $\chi^2_d(x)$ is the cumulative distribution function of the chi-square distribution with d degrees of freedom.

- *A variant of Fisher's inverse chi-square method* that considers the directions associated to individual spot p-values:

$$p_{dirchi}(g) = \min_{dir}(1 - \chi^2_{2d}(\sum_s -2 \cdot \log(p_s^{dir})))$$

where p_s^{dir} are the one-sided spot p-values (Wilcoxon ranksum test) for all spots s representing gene g ; these one-sided spot p-values are determined for both regulation directions; the overall combined gene p-value is set to the smaller of the two combined p-values, each of them corresponding to one test direction.

- *Stouffer's method* (Rosenthal, 1984). This method transforms p-values to Z-scores assuming a normal distribution ($p_s \rightarrow Z_s$), which is a straightforward calculation as the one-sided p-value $p_s^{\text{one-sided}}$ corresponds to the area under the normal cumulative distribution function between $-\infty$ and $-|Z_s|$:

$$p_s^{\text{one-sided}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-|Z_s|} e^{\frac{-t^2}{2}} dt$$

Each Z_s gets the sign corresponding to the regulation direction of the corresponding spot, the Z-scores of spots representing one gene are summed, and the sum is scaled:

$$Z_{\text{overall}} = \sum_s Z_s / \sqrt{k}$$

where k is the number of tests (i. e. the number of spots to be combined). Finally the Z-scores are transformed back to p-values ($Z_{\text{overall}} \rightarrow p_{\text{overall}}$) by

$$p_{\text{overall}}^{\text{one-sided}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-|Z_{\text{overall}}|} e^{\frac{-t^2}{2}} dt$$

Gene p-values obtained without p-value combination Mean expression values for each sample and gene were calculated by averaging spot expression values over all spots representing a gene, and Wilcoxon ranksum p-values were determined (*mean expr. value*) from these mean expression values. Furthermore, for each gene the most (*min. p-value*) and least (*max. p-value*) significant corresponding spot p-value was used as gene p-value.

Fold Change

The fold change for a gene g between two sample groups $C_1, C_2 \in \{n, e, p, c, l\}$, $C_1 \neq C_2$ has been estimated as follows: A spot s for the gene g is taken into account if at least one expression value in the groups under investigation is above the floor value (0.01). For each spot, fold changes ($\text{sfc}_{S_g}^{C_1, C_2}$) are computed for all pairs of samples derived from the two groups to be compared. The median (or the trimmed mean) of these spot fold changes is used as overall estimate for the gene-fold change ($\text{fc}(g)^{C_1, C_2}$).

$$S'_g := \{s \text{ spot} | s \text{ represents gene } g\}$$

$$s \in S_g := \{s \in S'_g | \exists k \in \{C_1 \cup C_2\} : x_{ks} > 0.01\}$$

$$\text{sfc}_{S_g}^{C_1, C_2} := \{\log_2(x_{is}/x_{js}) | i \in C_1 \wedge j \in C_2, s \in S_g\}$$

$$\text{fc}(g)^{C_1, C_2} = 2^{\text{median}(\text{sfc}_{S_g}^{C_1, C_2})}$$

where: x_{ks} is the expression value of spot s in sample k .

The independent determination of gene-fold change and directed gene p-value makes it possible that the direction of gene p-value and fold change differ. This effect was found to occur only for few genes and the respective fold changes were very close to 1; therefore, biological interpretation is not affected.

Evaluation of Gene P-Values

A number of heuristic yet intuitive criteria are used for evaluating gene p-values:

- if all spots are regulated in the same direction, then the gene p-value should be at least as significant as the least significant spot p-value
- if spots show inconsistent direction of regulation, then the gene p-value should be of lower significance than the most significant spot p-value
- if spots show inconsistent direction of regulation and of approximately equal significance, the gene p-value should tend towards 1

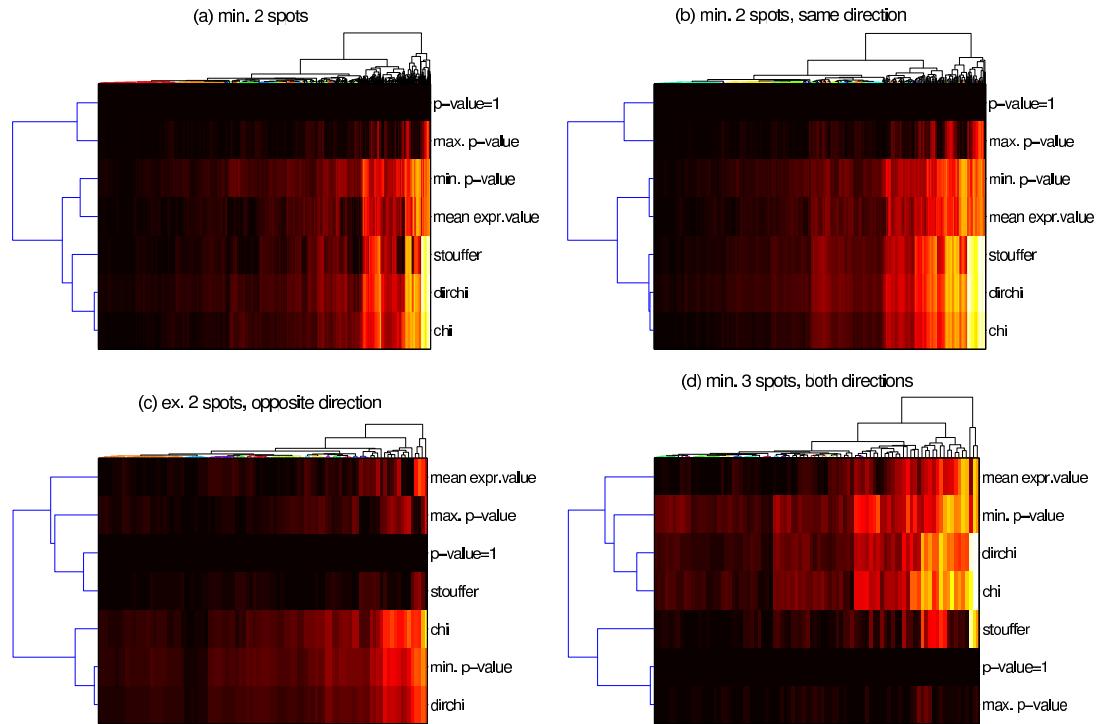


Figure 7.5: Comparison of different methods for gene p-value determination. Hierarchical clustering (Euclidean distance, average linkage) of p-values (\log_{10} transformed, black: least significant p-value, white: most significant p-value) for genes (columns) obtained from various combination methods (rows), and the baseline ($p\text{-value}=1$). The plots show all genes represented with (a) min. 2 spots; (b) min. 2 spots with consistent regulation direction; (c) exactly 2 spots with inconsistent direction; (d) min. 3 spots with inconsistent direction.

Figure 7.5 gives an overview of the obtained p-values for the group comparison normal versus late ($n-l$). The dendrograms on the y-axes reflect the similarity of results. For all genes represented by at least two spots (subplot a), Stouffer's method yields results that are most similar to the chi-square method and the variant thereof while p-values based on

mean expression values are rather similar to the minimum of the underlying spot p-values. For consistently regulated spots (b), these similarities become even more pronounced as indicated by the dendrogram branch lengths. For genes represented by exactly two spots which show opposite regulation (c), p-values from Stouffer's method are most similar to the baseline ($p\text{-value}=1$). For genes represented by at least three inconsistently regulated spots (d), Stouffer's p-values are generally most similar to the baseline and the maximum spot p-value. These results indicate that Stouffer's method returns conservative p-values for doubtful spot information, yet it is as sensitive as the chi-square method for consistent spot information.

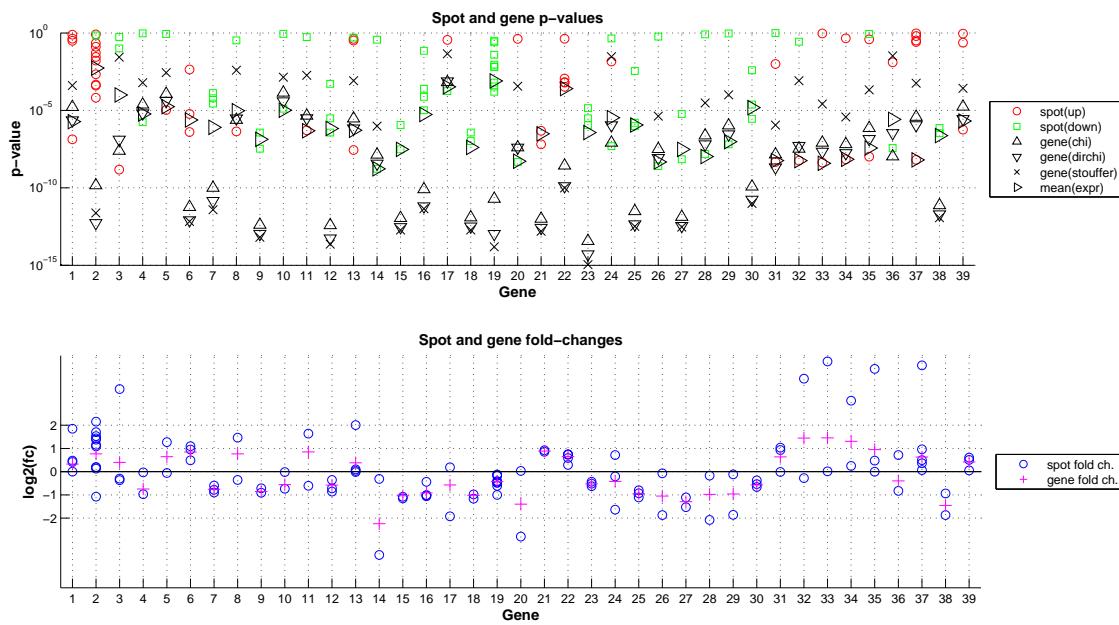


Figure 7.6: Different methods for combining spot p-values to gene p-values: Exemplary results of the GPC four-class data set; upper panel: spot and gene p-values; lower panel: spot and gene fold changes.

The detailed results for exemplary genes (Figure 7.6) show how Stouffer's method penalizes inconsistent regulation direction (e.g. genes 3, 8, 36), while it returns very significant gene p-values for (predominantly) consistently regulated spots (e.g. genes 2, 19, 22). The Affymetrix two-class data set (Figure 7.7) also shows inconsistently regulated probe sets in a number of cases. In approximately half of these cases, Stouffer p-values are close to those determined from mean expression values, in the other cases, Stouffer's p-values are more significant. Thus, for this data set, p-value determination on mean expression levels penalizes inconsistent regulation similarly to Stouffer's method, which is due to the underlying expression value distribution that does not show an as important spread as the GPC data set.

Overall, among the analyzed methods, Fisher's inverse chi-square method is most predom-

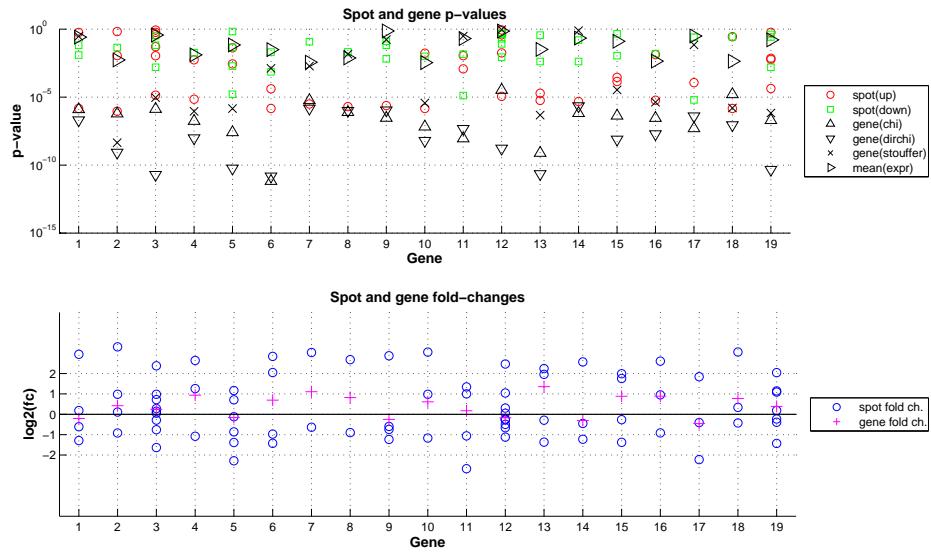


Figure 7.7: p-value combination - results for exemplarily selected genes of the Affymetrix two-class data set. Upper panel: spot and gene p-values; lower panel: spot and gene fold changes.

inantly used for combining p-values. A clear drawback of this method is that it ignores the direction of the underlying spot p-values. The presented variant of Fisher's inverse chi-square method considers regulation directions: Gene p-values for inconsistently regulated spots are less significant than obtained from the original Fisher's method. Yet, p-values cannot cancel out each other. Stouffer's method reflects p-values of opposite direction in a decrease in significance of the resulting overall p-value. Here, two spots of opposite directions and approximately equally significant p-values can cancel out each other. The comparison against combination of expression values (Figures 7.5 and 7.6) indicates that Stouffer's method generally returns very significant p-values for consistently regulated spots and conservative p-values for inconsistent spot information which indicates high discriminative power. P-values based on the mean expression values do not depend on the number of spots representing a same gene, yet they are biased towards the spot p-value of the underlying spot with highest expression intensity.

Stouffer's method has so far predominately been used in studies that integrate various types or sets of previous results; a process referred to as *meta-analysis* (e.g. integration of medical studies (Hall *et al.*, 1996; Cardozo *et al.*, 1998; Standish *et al.*, 2004) and studies in social sciences (Storm and Ertel, 2001)). We are not aware of its application for combining spot p-values into gene p-values. Besides the widely used methods analyzed here, a number of other more rarely used methods for combining p-values exist (e.g. Zaykin *et al.* (2002); Whitlock (2005)).

The analyzed methods for p-value combination assume statistical independence of the in-

put data, which is not necessarily given for spot p-values. Gene p-values are predominantly used for ranking genes and for giving a rough estimate of statistical significance; e.g. for selecting genes for experimental validation. The p-values derived from Stouffer's method are perfectly suited for these tasks, especially as they reflect consistent and inconsistent regulation in a more prominent way than other methods.

It is important to keep in mind that p-value combination makes more significant p-values possible; that is, genes represented by more spots/probe sets can achieve more significant p-values than those represented by fewer spots/probe sets. Thus, the number of spots for a given gene needs to be reported together with the p-value and fold change. The argument that genes represented by few spots/probe sets are per se discriminated can be brought forward, yet on the other hand it appears reasonable that the detection of differential expression via multiple spots/probe sets increases overall confidence.

In the literature, the issue of how to deal with within-array replicates has been addressed in terms of removing outlier spots ([Tseng et al., 2001](#); [Konig et al., 2004](#)), assessment of spot-quality ([Beissbarth et al., 2000](#)), or specific normalization methods [Fan et al. \(2004\)](#). [Smyth et al. \(2005\)](#) estimated the strength of correlation between within-array replicate spots; their method improves the precision of estimated genewise variances and thus identification of differentially expressed genes. Other approaches focus on redefinition of probe set mappings (e.g. [Stalteri and Harrison \(2007\)](#); [Sandberg and Larsson \(2007\)](#)), which implies detailed sequence analysis.

Unfortunately, large scale gold standards for evaluating fold change and significance of differential expression for all genes represented by more than one spot/probe set on an array are not available. Some spike-in experiments are publicly available (e.g. [Cope et al. \(2004\)](#); [Choe et al. \(2005\)](#)). These are useful for evaluating methods for the determination of expression values, but they make no assertions on differentially expressed genes represented by more than one spot/probe set. Confirmation of differential expression can be obtained from alternative techniques such as quantitative PCR. Ten genes have been verified by quantitative PCR ([Aigner et al., 2006a](#)) and the results showed good agreement with Stouffer's method.

7.3.3 Number of Regulated Genes

Based on gene p-values, q-values can be determined by use of the R-library *qvalue* ([Storey and Tibshirani, 2003](#)). The q-value quantifies the false discovery rate: A q-value of 0.01 indicates that when selecting the subset of all genes having a q-value ≤ 0.01 as significant, 1% of the selected genes have to be expected to be false positives. The q-value computation implies the estimation of π_0 , which is the number of non-regulated genes, by analyzing the distribution of p-values: The uniform distribution underlying the given p-value distribution is estimated and the area under this uniform distribution estimates π_0 . The number of regulated genes can thus be estimated by $1 - \pi_0$.

According to the background knowledge for the analyzed experiment (Section 7.3), fewer genes are expected to be regulated in the comparisons $n - e$ and $p - c$ than in the other comparisons.

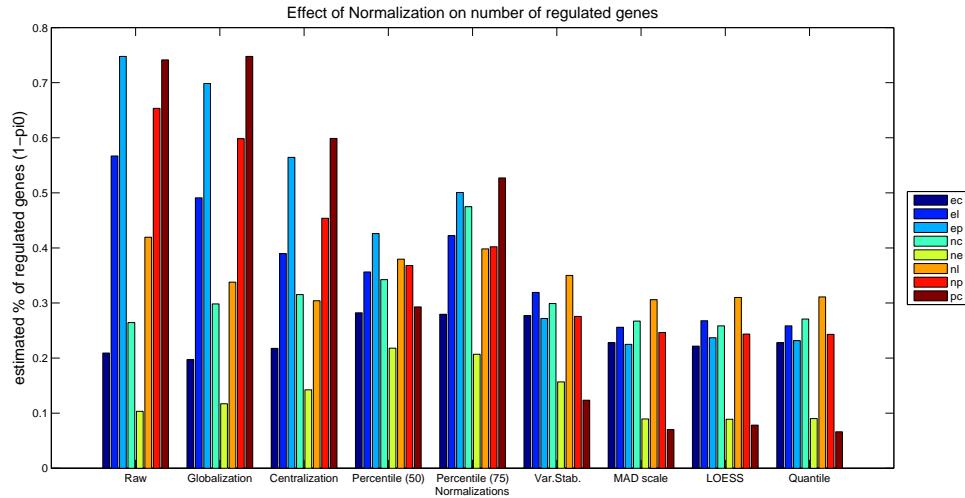


Figure 7.8: Effect of normalization on the number of significantly regulated genes.

The effect of normalization on the estimated number of significantly regulated genes (Figure 7.8) is most pronounced in the comparison p - c ; depending on the normalization between 7% and 74% of the genes appear regulated. Only after percentile normalization to the median, Variance Stabilization, MAD scale normalization, LOESS and quantile normalization, fewer genes appear regulated in the comparisons n - e and p - c than in the other comparisons. The latter four yield significantly smaller numbers of regulated genes in all comparisons than other methods.

The estimation of the number of differentially expressed genes via the R-library *qvalue* can be used for testing the appropriateness of normalization methods, especially for multi-group comparisons. As this approach returns a single number per group comparison and normalization, the overall number of regulated genes and the specific pattern of various group comparisons is easy to check against background knowledge.

7.3.4 Robustness Analysis

The large number of samples allows us to assess the robustness of differentially expressed genes between two sample groups by leave-one-out and subset sampling analysis.

Leave-One-Out Analysis

One sample is disregarded at a time and p-values are calculated for the remaining samples. The p-values obtained from the full dataset are considered as standard of truth. A series of cutoff p-values ($p_{cut} = 10^{-7}, \dots, 10^{-1}$) is applied and the fraction of genes from the full dataset that are significant at the cutoff p-value in the leave-one-out data sets is determined. *Robust* differentially expressed genes are selected according to two criteria:

- *exact*: The fraction of genes that are significant at p_{cut} in all leave-one-out data sets
- *relaxed*: The fraction of genes that are significant with $p\text{-value} \leq 2 \cdot p_{cut}$ in all leave-one-out data sets.

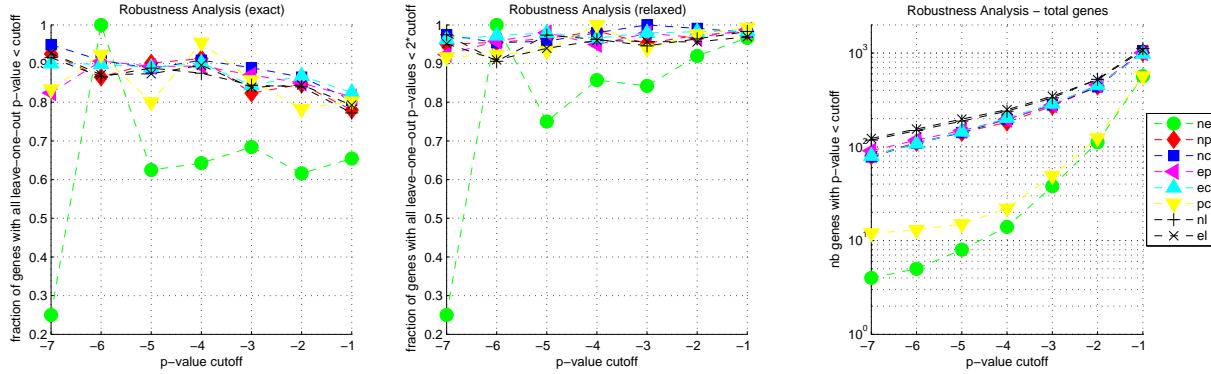


Figure 7.9: Leave-one-out robustness analysis. Fraction of genes significant at a certain p-value level in the entire data set that are also significant in the leave-one-out data sets according to *exact* (left panel) and *relaxed* (middle) evaluation. Right panel: number of genes significant at a certain p-value level (MAD scale normalized data, p-values determined by Stouffer's method).

The results (Figure 7.9) indicate that the p-values are generally very robust. At a cutoff p-value of 10^{-3} , the agreement of most group comparisons is above 82% in the strict analysis and above 93% in the relaxed analysis. P-values between normal and early degenerative cartilage are least robust, which reflects the small number of significantly regulated genes in this comparison (only 8 genes have a p-value $\leq 10^{-5}$, 38 genes have a p-value $\leq 10^{-3}$) and the similarity of normal and early degenerative cartilage samples.

The results indicate that an error of about 10% of the significantly differentially regulated genes has to be expected.

Subset Sampling

Random sample subsets (50 subsets of size $m=10$ to 18 each) for each group pair are generated, and p-values are calculated for the subsets. The t top p-value genes obtained from the entire sample set are used as standard of truth and the fraction of these top candidates that are also among the t top candidates of at least $s\%$ of the subset p-value sets is determined. For t , the values 50, 75, 100 were used; for s , the values 100, 80, 50 were used.

The result for $t=50$ (Figure 7.10, $t=75$ and $t=100$ yielded very similar results) show that an increase in subset size (m) leads to an important increase in the fraction of stable genes. The genes for the group comparisons *n-e* and *p-c* are less robust than for the other comparisons. For the other comparisons, at a subset sample size of 10, about 50% of the

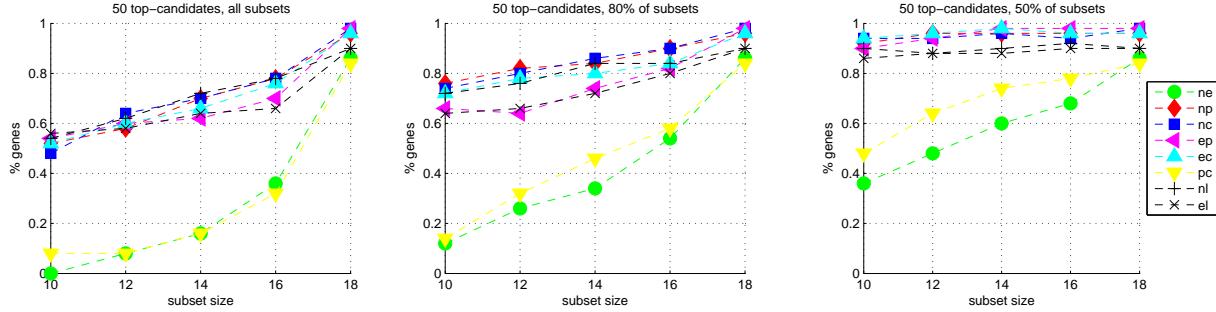


Figure 7.10: Subset sampling robustness analysis. Fraction of the 50 top p-value genes from the entire data set that are also among the 50 top candidates in at least $s\%$ of the subset-based p-values (left panel: $s = 100$, middle panel: $s = 80$, right panel: $s = 50$; MAD scale normalized data, p-values determined by Stouffer's method).

top-candidates are obtained from all subsets and about 90% of the top-candidates are obtained from half of the subsets.

The subset sampling analysis reflects the differences between the individual group comparisons in a more prominent way than the leave-one-out analysis. Especially, the comparison $p-c$ here yields similar results as the comparison $n-e$, which was not the case in leave-one-out analysis. Subset sampling represents a more sensitive means for robustness analysis than the leave-one-out analysis; it provides an overview of the error to be expected for a group comparison and the reported differentially expressed genes.

7.4 Deriving Reliable Gene Signatures for Microarray Classification

For many tasks it is required to select, based on gene expression data, a subset of the analyzed genes which is then subjected to further analysis. For example, this is the case for biological interpretation, where only a manageable number of genes can be analyzed, but also for many integrated analysis methods (e.g. gene ontology overrepresentation analysis). Therefore, the selected genes should not only be differentially expressed, but also reflect the majority of the analyzed samples, and the measurements should be consistent for different sample subsets.

The previous section (Section 7.3) showed that, for the GPC four-class osteoarthritis data set, the individual samples vary significantly and many genes are not stable (see subset sampling analysis in Section 7.3.4). Furthermore, a cluster analysis (Section 8.1.3) indicated that sample classification is rather difficult for this data set.

Deriving a gene subset from this data that separates well between the sample groups and is consistent for different sample subgroups is thus a challenging task.

In the following, an approach for model selection is presented that not only evaluates classification performance but also requires the selected genes to be stable, that is, frequently

selected in various sampling steps (Davis *et al.*, 2006); thereby, the approach selects reliable subsets of genes which are suited for further analysis. It has been implemented by Chad Davis, Fabian Gerick, and Volker Hintermair during a practical course on gene expression data analysis which I co-supervised together with Caroline Friedel and Robert Küffner.

Sample classification

Sample classification describes the assignment of samples to predefined classes. Gene expression data is frequently used for sample classification, based on the assumption that samples which belong to the same class exhibit similar gene expression patterns. Sample classification has two important aspects: (1) The classification of new samples into one of the given classes, and (2) the identification of genes which are important for the distinction between the classes. The latter is especially important for the biomedical interpretation of gene expression data. For this aspect, it is necessary that the selected genes represent the majority of the samples and are not sensitive to individual, possibly outlier, samples. Generally, sample classification implies several components:

- *feature subset selection (fss)*: Selection of a subset of features (genes) which reflect the differences between sample classes and thus are useful for classification.
- *supervised classification*: application of a method that makes use of classified training samples to predict the class of each presented test sample.
- *model*: combination of feature subset selection method, classification method, and the respective set of parameters.

A wide variety of methods exist for feature subset selection and classification. Classification accuracy depends not only on the difficulty of the classification task but also on the applied methods for feature subset selection and classification and the respective parameters.

7.4.1 StabPerf: Stable Model Selection by Optimizing Reliable Classification Performance

StabPerf (Figure 7.11) generates random subsets of arrays (sampling) and applies all models from a model library. The standard library of feature subset selection methods contains:

- F-test (F_t): All features with an F-test statistic above a threshold t are selected.
- Pearson correlation (PC_t , $PC_{[n]}$): All features with an absolute Pearson correlation above a threshold t or the top n genes with highest absolute Pearson correlation are selected.
- P-value combined with fold change (PV_tFC_f): Genes with a t-test p-value of less than t and $|\log_2(foldchange)|$ greater than f are selected.

- Decision tree (DT_t): A decision tree is trained and genes are selected starting from the root and moving down the tree, as long as they exceed a significance threshold t .
- Support Vector Machine (SVM_t): A SVM with linear kernel is trained and genes with a normed weight above t are selected.

Overrepresentation analysis (ORA) of Gene Ontology (GO) terms was used as optional postfilter for the above fss methods: Only genes being annotated with a GO-category that is overrepresented (determined by a threshold p-value p_t) in the subset of genes are retained.

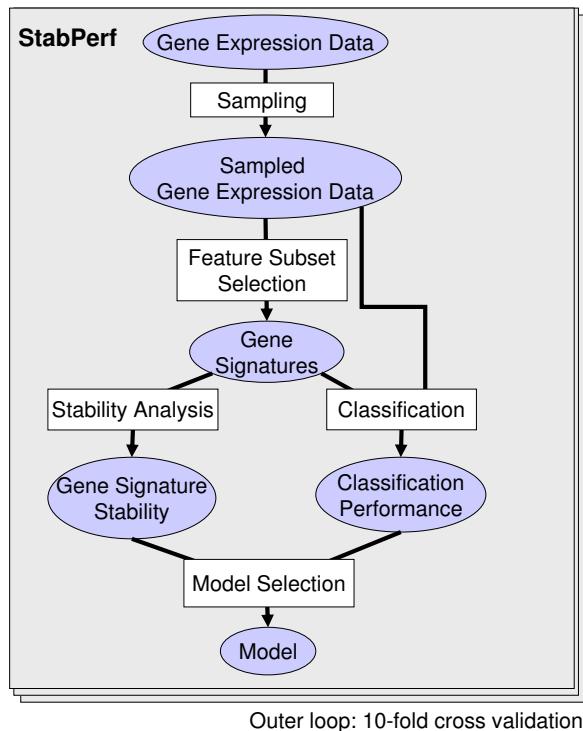


Figure 7.11: Overview of StabPerf (**S**table model selection by optimizing reliable classification **P**erformance). Random sampling from the gene expression dataset produces multiple training subsets; gene signatures are derived for each of these and for various feature subset selection methods. Signature stability and classification performance are used to select a best model.

The standard library of classification methods contains:

- Nearest shrunken centroid (NSC): a sample is assigned to the class with the nearest centroid, where the centroid of each class is shrunk to the overall centroid.
- k-nearest neighbors (kNN): a sample is classified based on the classes of the k nearest neighbors by a voting scheme.
- Support Vector Machines (SVM): a separating hyperplane with maximal margin is determined and samples are classified by their position with respect to the hyper-

plane (here: third order polynomial kernels (SVM-P) and radial basis function kernels (SVM-R) were used).

- Decision Trees (DT): a classifier in form of a tree where the root and each internal node describes a test on a feature and has one child node for each possible value or range of values; each leaf node is associated with a class label. A sample is classified by sequential testing of the features as indicated by the nodes from the root to a leaf.

Sampling allows the measurement of the non-adjusted stability $Stab_{NA}$ of a fss method S :

$$Stab_{NA}(S) = \frac{\sum_{f \in F} freq(f)/n}{|F_1|}$$

where $freq(f)$ is the number of sampling steps in which a feature $f \in F_1$ has been selected and F_1 is the set of all features which have been selected in at least one of the n sampling steps. The length adjusted stability $Stab$ penalizes long signatures and is defined as:

$$Stab(S) = \max \left[0, Stab_{NA}(S) - \alpha \cdot \frac{\mu}{|F|} \right]$$

where α is a penalty factor (here: $\alpha = 10$), μ is the median number of selected features, and F is the total number of features (i.e. genes) per array. The classification performance $Perf(M)$ of a model $M = (S, C)$ consisting of a fss method S and a classification method C is a measure of model reliability and is defined as:

$$Perf(M) = \max [0, Acc(M) - \beta \cdot MAD(M)]$$

where $Acc(M)$ is the total model accuracy, $MAD(M)$ is the median absolute deviation of accuracy over all sampling steps, and β is a weighting factor (here: $\beta = 0.5$).

The model score $ModScore(M)$ is based on the gene signature stability and the classification performance of the model:

$$ModScore(M) = \gamma \cdot Stab(S) + (1 - \gamma) \cdot Perf(M)$$

where γ determines the relative importance of stability and classification performance (here: $\gamma=0.5$).

Feature frequencies over all gene signatures provide a relevance ranking for the selected features, and the selection of genes consistently found to be significant can be assumed to return more biologically relevant genes. Consensus gene signatures are constructed by extracting those genes that occur in more than a given fraction τ of all signatures.

The classification performance of the entire model selection approach is estimated via stratified 10-fold cross validation.

7.4.2 Application on Osteoarthritis Data

The StabPerf approach has been applied on the four-class osteoarthritis data set described in Section 7.2.1. The 78 samples allow sampling and the different group similarities represent a challenge for classification.

It was found that the classification performance significantly depends on the number of sampling steps. Without repeated sampling, over-estimation of the classification performance of approximately 20 percentage points (pp) and under-estimation of approximately 10 pp were observed. By increasing the number of sampling steps performance estimates became more stable. For further analysis, the number of sampling steps was set to 400.

FSS	Classifier	$Stab_{NA}$	μ	Acc	MAD	Modscore
$PV_{0.001}FC_{2.5}$	NSC	66.6%	165	70.3%	13.5%	0.536
	5NN			71.2%	13.5%	0.543
	SVM-P			55.4%	8.1%	0.428
	SVM-R			70.2%	13.5%	0.535
	DT			63.0%	9.4%	0.486
$PV_{0.01}FC_{2.0} + ORA$	NSC	63.9%	139	73.4%	12.4%	0.563
	5NN			77.6%	14.8%	0.592
	SVM-P			59.0%	9.9%	0.453
	SVM-R			77.2%	12.4%	0.594
	DT			66.7%	12.4%	0.509

Table 7.1: Exemplary results of StabPerf applied on osteoarthritis data. For the individual criteria, the best values obtained of all analyzed models are marked in bold.

The analysis of all pairwise combinations of fss and classification methods (Exemplary results are shown in Table 7.1) showed that no method is consistently superior with respect to the criteria accuracy (*Acc*), *MAD* of accuracy, stability $Stab_{NA}$, and model score *ModScore*. Overall, the values of the individual criteria were highly variable for the different models. The model that was ranked best (fss: $PV_{0.01}FC_{2.0} + ORA$, classifier: *SVM-R*) according to the *ModScore* was not ranked best according to any of the underlying criteria. The signatures derived by this best method were analyzed in more detail. The signatures contained 96 genes in total, 33 of these were present in 75% of the signatures and retained as consensus signature. Interestingly, 66 of the 96 genes, and 32 of the 33 genes in the consensus signature have already been attributed to the context of OA in the literature. The fss method $PV_{0.01}FC_{2.0} + ORA$ tends to select functionally coherent sets of genes due to the enclosed overrepresentation analysis. Nevertheless, the consensus signature contains a significantly increased fraction of clearly relevant genes. Thus, sampling not only allows reliable performance estimation, but also for the generation of a consensus signature which shows high relevance for the experimental context and thus alleviates biological interpretability. Furthermore, the feature frequency over signatures can be used to rank genes, for example, for selecting candidates for follow-up studies.

The overall classification accuracy of StabPerf was 73.4% for the four-way classification of the OA data set and 97.5% for the two-way classification between the sample groups normal-early and peripheral-central as estimated by 10-fold stratified cross validation.

7.5 Chapter Summary

This chapter presents an exemplary study on how microarray data normalization and processing affects the final outcome, especially the identification of differentially expressed genes ([Fundel et al., 2005b](#)). It introduces the group-level plot as a helpful means for visual inspection of normalization effects on data from classified samples. Furthermore, different methods for combining spot p-values into gene p-values have been compared, which represents an alternative to standard gene p-value determination methods that is especially suited for data that exhibits large inter-spot expression value differences. Stouffer's method has been found to work best. This study shows on exemplary data that it is of vital importance to check every individual step of gene expression data analysis for its appropriateness. For most individual processing steps numerous alternatives exist. It is important to test different possibilities and analyze the effects of the decision with appropriate tools. Especially, when a data set does not contain experimental quality controls such as spike-in data, one has to rely on biological background knowledge for selecting the most appropriate combination of methods. Clearly, this introduces a bias as no standard procedure can be defined and thus the result depends, to some extent, on the judgement of the individual researcher. The use of global robustness and quality measures for analyzing results can help in estimating the reliability of final microarray study outcomes.

The StabPerf approach ([Davis et al., 2006](#)) addresses two important objectives in microarray classification: Achieving high classification performance and determining feature subsets and classification results that are stable, that is, resilient against variations between different subsets of expression arrays. StabPerf evaluates all possible combinations of a wide spectrum of feature subset selection and classification methods and applies repeated random sampling for each of these combinations. The approach selects the best combination among the given methods for a given data set. Repeated sampling entails important benefits: (1) realistic performance estimates, (2) information on the distribution of accuracies, and (3) a consensus gene signature can be built from the most stable genes. The application of StabPerf on an osteoarthritis data set showed that, indeed, StabPerf achieved good classification performance and returned a stable and biologically meaningful consensus signature.

Chapter 8

Gene Expression in Osteoarthritis

Osteoarthritis is an important degenerative joint disease for which, so far, no disease-modifying drug is on the market (for background see Section 6.4). Osteoarthritis is considered to be caused by an imbalance between matrix anabolism and catabolism; yet, many details are unknown. Therefore, intense research effort is devoted to the exploration of the underlying molecular mechanisms.

In this chapter several gene expression data sets investigating osteoarthritis related samples (Section 7.2) are analyzed from a more biological point of view. First, the detailed biological interpretation of a data set comparing normal and osteoarthritic cartilage samples is summarized (Section 8.1). This part is joint work with Thomas Aigner ([Aigner et al., 2006a](#)).

As human joint cartilage samples are difficult to obtain, cell cultures represent an appealing alternative. Different microarray platforms are compared against each other (Section 8.2), and the similarity between cell culture samples and in-vivo samples is analyzed (Section 8.3).

Interleukin 1 promotes catabolic processes and is able to downregulate anabolic processes. A time-series experiment investigating the genes and processes implicated in Interleukin 1 stimulation is analyzed in Section 8.4.

8.1 Analysis of Gene Expression in Osteoarthritis

This section deals with the GPC four-class gene expression data set (Section 7.2.1). The data analysis methods applied for this data set are described in Section 7.3. Here, the focus is on the biological data interpretation and the analysis of the disease phenomena implicated in osteoarthritis on the molecular level; results are described in terms of general observations and detailed biological analyses. Most of this work has been prepared together with Thomas Aigner ([Aigner et al., 2006a](#)).

Overall, the gene expression levels of many genes showed high variability between different

donors for all four sample groups. This might be expected for the diseased sample groups but seems surprising for normal cartilage. Furthermore, this implicates that a wide range of expression levels are compatible with similar functioning of the tissue.

When interpreting gene expression data from the different sample groups, one has to bear in mind that peripheral and central OA samples are obtained from total knee replacements (i. e. from living patients) while normal and early degenerative cartilage samples are obtained from autopsies (i. e. after death). Post-mortem effects might alter mRNA levels. In the present experiment, post-mortem time was kept short and no significant correlation between expression levels and post-mortem time could be found. Gene groups known to be related to post-mortem processes were analyzed in detail; for example, hypoxia-induced genes are expected to be up-regulated when oxygen supply is suppressed due to death. No significant regulation could be detected for these genes.

8.1.1 Expression Levels of Genes Relevant for Anabolism

Chondrocytes are responsible for a balanced turnover of the extracellular matrix (Figure 6.3). A number of genes are known to be implicated in matrix anabolism and catabolism. For quality control of the analyzed data set, it is important to check the expression levels of these well-known genes. Here, the focus is on genes implicated in matrix anabolism. Collagen type II, III, VI, IX, and XI are expected to be up-regulated, which was confirmed in the analyzed data. Other collagens so far not known to be expressed in adult articular chondrocytes (COL5A1, COL15A1) were also found to be expressed. The non-collagenous matrix proteins fibromodulin, CILP, fibronectin, tenascin, and osteonectin/SPARC showed significantly higher expression levels in osteoarthritic cartilage. Other non-collagenous matrix proteins (e. g. aggrecan, COMP, decorin, biglycan) showed similar expression levels in normal and osteoarthritic cartilage. In normal cartilage, most non-collagenous proteins showed higher expression than the collagens. The turnover of non-collagenous matrix proteins thus appears to have a higher rate than turnover of collagenous proteins.

These observations correspond well to previous observations and thus confirm validity of the analyzed data set. The results of the genes analyzed by quantitative PCR also confirmed the microarray measurements.

8.1.2 Differential Expression between Sample Groups

Normal versus early degenerative cartilage: Only very few genes appeared regulated between normal and early degenerative cartilage (15 genes with $q < 0.05$). This confirms that chondrocytes in these two groups are very similar. One of the up-regulated genes was fibronectin which is implicated in cartilage degeneration.

Normal versus late-stage osteoarthritic cartilage: Significantly more genes were regulated between normal and late-stage osteoarthritic cartilage. The gene ontology (GO) overrepresentation analysis (Figure 8.1, for background see Section 9.1) showed that a significant number of genes implicated in extracellular matrix formation were up-regulated, which is expected from previous knowledge.

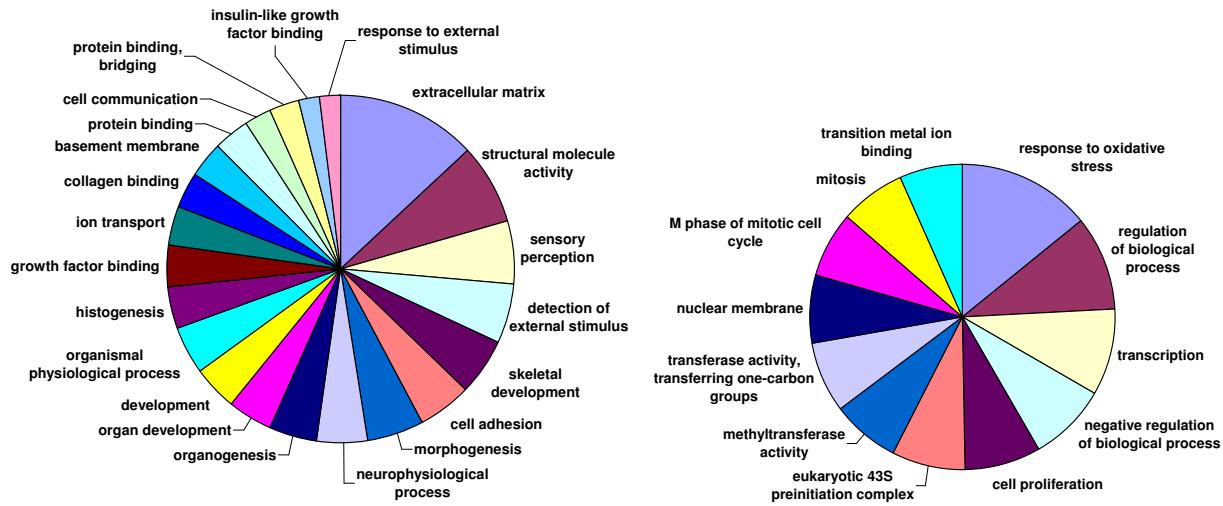


Figure 8.1: Overrepresentation analysis of genes differentially expressed in the comparison normal versus late-stage osteoarthritis cartilage. Left panel: up-regulated genes ($p \leq 0.01$, $\log_2(\text{fold change}) > 1$), right panel: down-regulated genes ($p \leq 0.01$, $\log_2(\text{fold change}) < -1$). The area assigned to a functional category corresponds to the statistical significance of overrepresentation determined from the hypergeometric distribution.

Many genes involved in oxidative defense appear down-regulated (e.g. glutathion peroxidase 3 (GPX3), superoxide dismutase 2 and 3 (SOD2, SOD3), thioredoxin interacting protein (TXNIP)). This might be one reason for the increased accumulation of reactive oxygen species (ROS) within the cells and the increased oxidative stress in osteoarthritic chondrocytes, which in turn might enhance matrix break-down. Genes involved in oxidative defense have so far not been reported to be directly relevant for osteoarthritis; this finding indicates a new target for disease treatment.

A group of cytokines and genes involved in cytokine signaling were regulated. Interestingly, many genes related to the IL-1 pathway were down-regulated (IL-1 β) or showed no significant expression (IL-6, IL-8, LIF), which contrasts to previous expectation. Yet, as measured expression levels were low for these genes the observations should be checked by further measurements.

Low-grade late-stage versus high-grade late-stage osteoarthritic cartilage: Only few genes appear differentially regulated between low grade and high grade late stage (14 genes with $q < 0.05$).

8.1.3 Clustering Analysis

Clustering has been performed for analyzing the similarity between the individual samples and sample groups. Genes have been neutrally preselected for clustering (i. e. without using knowledge on sample classification) by significant expression levels and significant variance between samples.

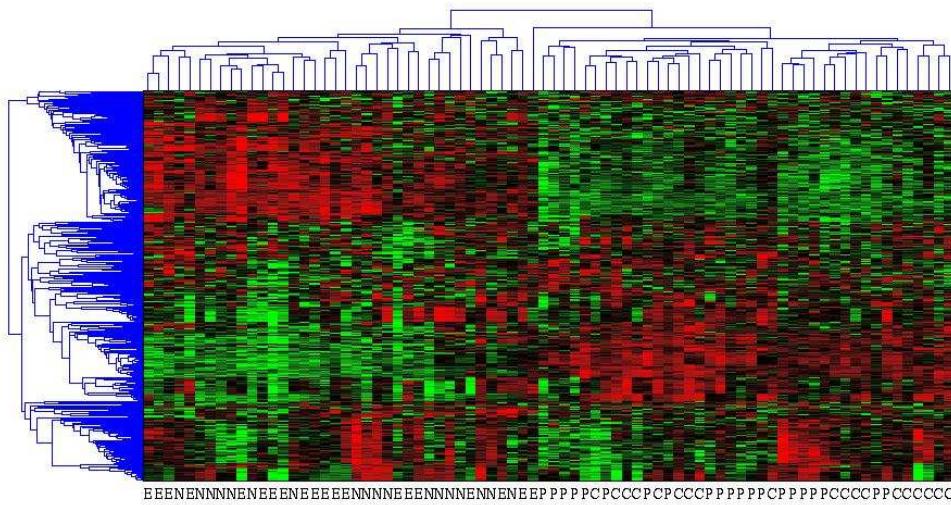


Figure 8.2: Heatmap and dendrograms obtained from clustering gene expression data of osteoarthritis related samples (Spearman correlation, average linkage).

The resulting dendrogram (Figure 8.2) shows a clear separation between normal and early degenerative cartilage versus peripheral and central OA samples. The normal samples are not separated from the early degenerative cartilage samples, and peripheral and central OA samples are also not separated. This confirms that the differences between normal and early degenerative cartilage as well as between peripheral and central OA stages are relatively minor while the two combined groups are clearly different.

In macroscopic analysis, the peripheral late stage cartilage samples are rather similar to normal or early degenerative cartilage samples. Yet, the analysis of gene expression profiles shows that the less damaged cartilage from late stage osteoarthritic joints is clearly distinct from normal and early degenerative cartilage. Thus, the late-stage peripheral cartilage samples do not represent a good model for early OA on the molecular level. The similarity of peripheral and central cartilage samples might be due, at least in part, to the fact that both areas are exposed to the same synovial factors, e.g. cytokines and growth factors. Furthermore, the difference between early and peripheral samples might in part be due to the difference in sample origin with early degenerative cartilage samples being derived from dead donors and peripheral samples being derived from living patients. Given the fact that no clear dependency of gene expression on post-mortem time could be shown, this presumably plays a minor role.

8.2 Comparison of Microarray Platforms

The four data sets dealing with normal and osteoarthritic joint cartilage can be used to compare the different platforms and investigate reproducibility of results. Overall, 158 genes have been measured in all four data sets. As the different techniques return significantly varying expression levels, the normalized expression value distributions of this gene subset were assimilated by quantile normalization.

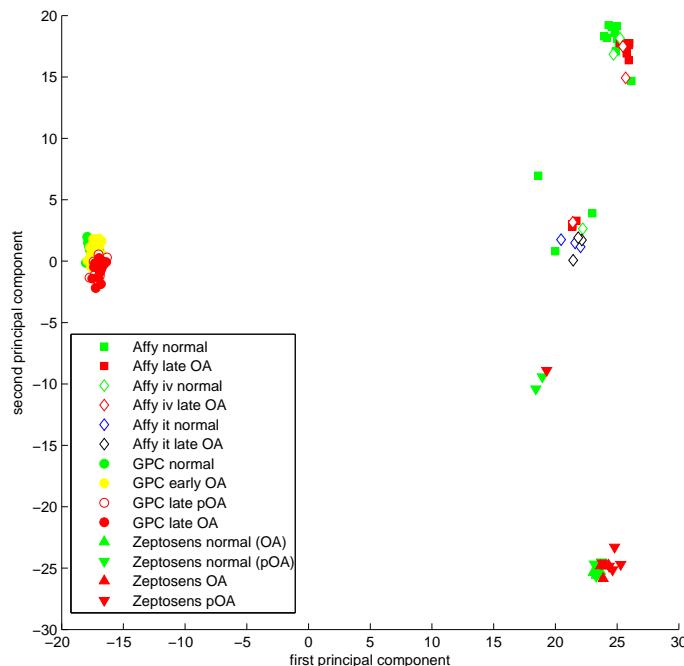


Figure 8.3: Comparison of microarray platforms and experiments. Principal Component Analysis (PCA) based on expression values of 158 genes measured in all four experiments investigating normal and osteoarthritic joint cartilage and showing expression values above noise level (iv: in vitro, it: in vitro).

Principal component analysis of the expression values (Figure 8.3) indicates that the outcomes of the different microarray platforms differ from each other significantly. The GPC data set is most different from the others. Interestingly, the two Affymetrix data sets as well as the Zeptosens data set are split up in the visualization of the first two principal components.

Hierarchical clustering of the normalized expression values (Figure 8.4) also separates the platforms. The dendrogram clearly reflects the structure of the investigated samples: In the GPC data set, the normal and early samples are separated from the peripheral and central OA samples with few exceptions. In the Affymetrix data sets, the in-vitro samples are

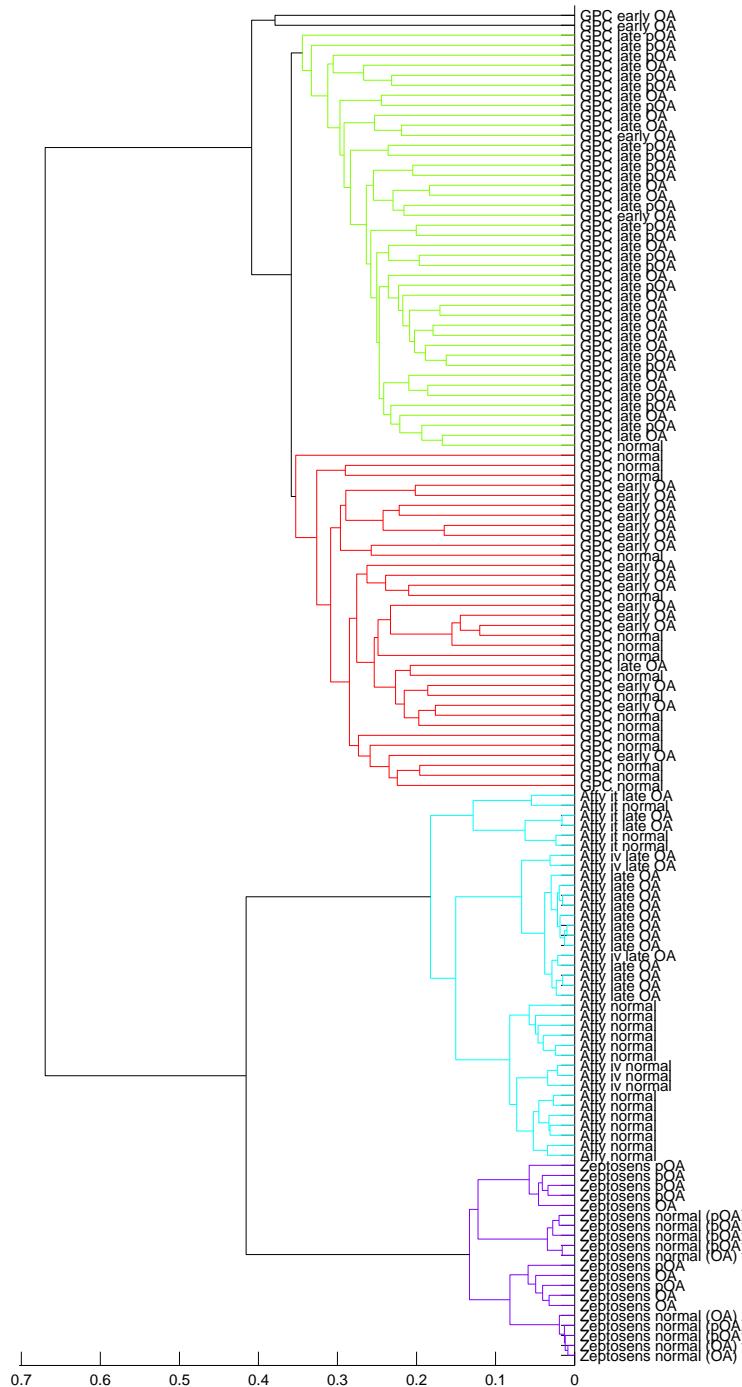


Figure 8.4: Comparison of microarray platforms and experiments. Hierarchical clustering (Spearman correlation, average linkage) is based on expression values of 158 genes measured in all four experiments investigating normal and osteoarthritic joint cartilage and showing expression values above noise level (iv: in vitro, it: in vitro).

separated from the in-vivo and cartilage samples. The in-vivo OA samples are mixed with the cartilage OA samples and the in-vivo normal samples are mixed with the normal cartilage samples. This indicates that the in-vitro samples do not represent a good model for cartilage samples (Section 8.3). Zeptosens SensiChip arrays apply two-color measurement technology. With these chips, OA samples were compared against pooled normal samples. In the dendrogram, two subgroups can be observed; each subgroup contains five pairs of normal and OA samples. The difference between these subgroups is more important than the differences between the disease stages. This observation is not an effect of the analysis of only the small set of common genes, the analysis of all measured genes returns a similar result. The detailed biological analysis of the Zeptosens data set has been described by [Gebauer et al. \(2005\)](#).

In several previous studies generally good inter- and intra-platform reproducibility of gene expression measurements has been claimed ([Barnes et al., 2005](#); [Jarvinen et al., 2004](#); [Shi et al., 2006](#)). Yet, in a number of studies poor concordance has also been observed ([Shi et al., 2005](#); [Woo et al., 2004](#); [Yauk et al., 2004](#)). For a review on reliability and reproducibility issues in microarray measurements see [Draghici et al. \(2006\)](#). In review of different studies on reproducibility and correlation of data produced across different approaches, [Yauk and Berndt \(2007\)](#) found that initial investigations in the years before 2004 found discrepancies while more recent studies showed much higher levels of correlation.

8.3 Analysis of Osteoarthritis Models

Cell cultures make it possible to keep cells in a controlled environment, and importantly, to increase the amount of biological material. This is of special interest for human chondrocytes as donors are not easily available and the amount of biological material that can be obtained from a donor is generally small. The in-vivo/in-vitro experiment was set up to investigate whether chondrocytes transferred into cell culture represent a good model for in-vivo samples.

IL1 stimulation enhances extracellular matrix catabolism and is therefore sometimes used as model for osteoarthritis. The time series data set investigating IL1 stimulation was obtained from cultured chondrocytes, similar to the in-vitro control samples.

The Affymetrix two-class, in-vivo/in-vitro, and the IL1-stimulation time series data sets were obtained with the same array type, and thus these can be compared to each other via the complete set of measured genes.

Figures 8.5 and 8.6 show that the in-vitro samples are rather dissimilar to the in-vivo samples and the samples of the Affymetrix two-class experiment. Normal and OA samples are clearly separated from each other for the two-class experiment and the in-vivo group, but not for the in-vitro group. This indicates that the differences between normal and osteoarthritic cells get lost when cells are transferred into cell culture. Two of the in-vitro samples are in fact rather similar to the control samples of the time-series experiment. This might be explained by the fact that these sample classes were both derived from cell culture. These results indicate that chondrocytes in cell culture represent a model of rather

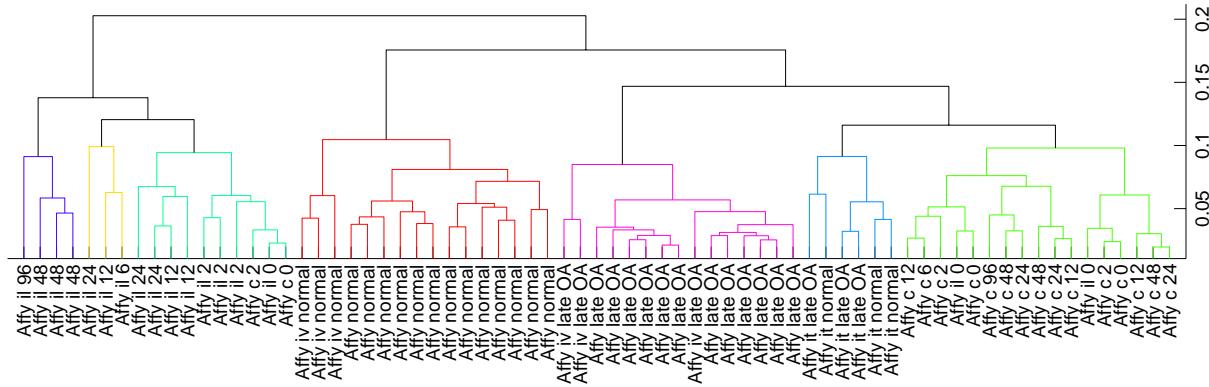


Figure 8.5: Analysis of osteoarthritic models by clustering of three Affymetrix data sets. Hierarchical clustering (Spearman correlation, complete linkage) is based on expression values (iv: in vitro, it: in vitro, c: controls of the time series experiment, il: IL1 stimulation).

poor quality for in-vivo cartilage samples. Furthermore, the effects on gene expression level induced by IL1-stimulation differ from those induced by osteoarthritis.

8.4 IL1-Stimulation Time Series Analysis

The Short Time-series Expression Miner (STEM, [Ernst *et al.* \(2005\)](#); [Ernst and Bar-Joseph \(2006\)](#)) is a tool for clustering, comparing, and visualizing short time series gene expression data from microarray experiments. The cluster algorithm has been specifically designed for short time series (about 8 time points or fewer), and is thus ideally suited for the IL1-stimulation time-series data set.

All genes with a significant fold change ($\text{abs}(\log_2(\text{fc})) > 1$) for at least one time point, which corresponds to 9237 genes, were analyzed with STEM. The analysis results in eight significant clusters, three of them containing more than one profile (Figure 8.7). The gene ontology overrepresentation analysis for the clusters and profiles provides a description for the underlying processes.

Cluster 1 (profiles with red background; profiles 43, 29, 49) contains genes that are upregulated after 2–24 hours with subsequent decay (total $434 + 221 + 288 = 943$ genes). The most significant profile in this cluster (profile 43) shows significant overrepresentation of several processes related to catabolism. This corresponds to the chondrocyte model in which IL1 stimulates catabolic processes (Figure 6.3).

Cluster 2 (green background; profiles 41, 14, 18) groups profiles that show up-regulation (eventually after initial down-regulation) and remain on an elevated expression level (total $337 + 270 + 186 = 793$ genes). For this cluster, annotations with secretion and transport processes are overrepresented.

Cluster 3 (blue background; profiles 19, 8) groups genes that are downregulated with the most significant change occurring after 12–24 hours, and then stabilize at the lower expres-

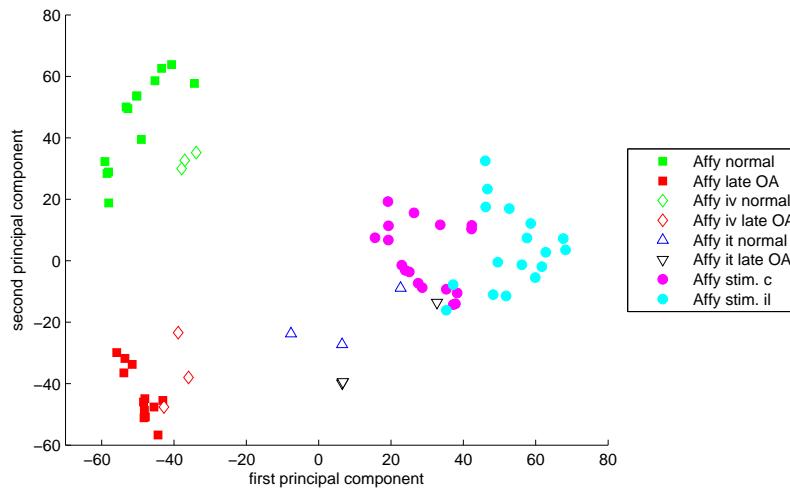


Figure 8.6: Analysis of osteoarthritic models by Principal Component Analysis of three Affymetrix data sets. Principal Component Analysis (PCA) has been performed with expression values of all samples (iv: in vitro, it: in vitro, c: controls of the time series experiment, il: IL1 stimulation).

sion level (total $253+224=477$ genes). For these profiles, GO-terms such as cell adhesion, cell growth, development, metabolism are overrepresented.

Two of the significant single profiles appear interesting: Profile 38 groups genes that show rapid up-regulation and then remain on a constant level (217 genes). This profile is clearly related to inflammatory processes: respective overrepresented GO-terms are: defense response, response to wounding, inflammatory response.

Profile 9 groups genes that are continually down-regulated (215 genes). This profile is clearly related to extracellular matrix, cell adhesion and morphogenesis and the respective overrepresentation p-values are very significant ($p < 10^{-6}$).

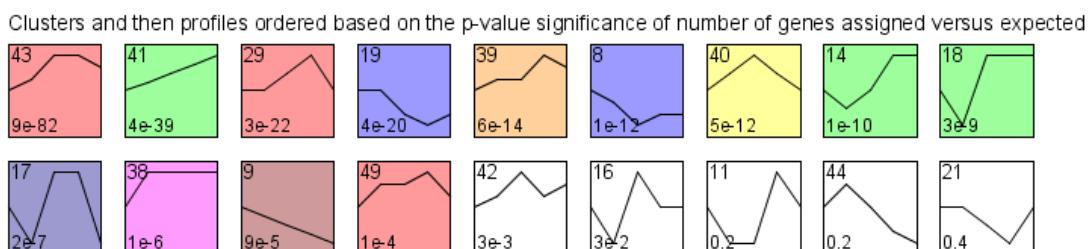


Figure 8.7: IL1-stimulation time-series experiment analyzed with STEM. Each expression profile is shown in a separate box and assigned with a profile number (upper left corner); profiles are sorted by significance (lower left corner). Significant profiles are colored. The same color is used for profiles within a cluster, where a cluster groups profiles with similar shape.

In summary, the obtained profiles and clusters together with the respective GO overrepresentation analyses provide an overview of the main phenomena implicated in IL1-stimulation of chondrocytes: IL1 is known as a stimulatory agent and accordingly more genes are up-regulated than down-regulated by IL1. Inflammatory processes show earliest activation upon stimulation and thus represent early response processes. Then, catabolism is stimulated temporarily. Finally, secretion and transport phenomena are activated as late response. Processes concerning the extracellular matrix and cellular adhesion are continually downregulated.

8.5 Chapter Summary

Osteoarthritis, a degenerative joint disease, is the most common disabling condition in the Western world. This chapter describes the analysis of several microarray data sets investigating osteoarthritis under various aspects. The detailed biological analysis of genes which are differentially expressed between normal and osteoarthritic cartilage indicated that, for example, genes implicated in extracellular matrix turnover appear predominantly upregulated while genes implicated in oxidative stress response appear predominantly downregulated ([Aigner *et al.*, 2006a](#)). The overrepresentation analysis provided an overview of the functional processes implicated in osteoarthritis. In the clustering analysis, the normal and early degenerated cartilage samples were mixed as well as the peripheral and central cartilage samples, while these two pairs of sample classes were clearly separated from each other. This reflects the degree of similarity of the sample classes.

The comparison of expression data sets obtained with three microarray platforms indicate important differences between the data sets; this might reflect differences in the techniques. The analysis of in-vivo and in-vitro samples showed that in-vitro cultures of chondrocytes represent a poor model for cartilage obtained directly from human donors.

Finally, the analysis of the IL1 stimulation time-series data set revealed regulation patterns with associated GO-terms which together provide a description of the processes implicated in IL1-stimulation. Especially, processes implicated in catabolism, secretion, transport processes, defense and inflammatory response are assigned to upregulated profiles. Processes implicated in cell adhesion, cell growth, development, metabolism, and extracellular matrix turnover are assigned to downregulated profiles.

Together, these results provide a description of the most important processes implicated in osteoarthritis and demonstrate the relevance of experimental setup for biological interpretation. Different microarray platforms can return significantly differing results, and in-vitro models can, even though being frequently used, differ importantly from its in-vivo counterpart.

Part III

Integrated Data Analysis and Conclusions

Chapter 9

Background: Integrated Gene Expression Data Analysis

Microarray data represents a snapshot of gene expression activity. For explaining biological phenomena such as diseases or cell differentiation based on gene expression data, extensive background knowledge is required. Knowledge on biological mechanisms such as signal transduction cascades, metabolic pathways, or regulatory events can be included in data analysis by utilizing text mining or other kinds of additional data. The combination of gene expression data with data from other sources can alleviate data interpretation. Most integrated data analysis methods focus on manually compiled data such as ontology annotations (Section 9.1), large-scale networks (Section 9.2), or specific text-mining applications (Section 9.3); a review on these approaches is given in the following.

9.1 Integration with Manually Compiled Data

Various types of manually compiled data have been used in combination with gene expression data for advanced analysis of biological phenomena. One of the most frequently used approaches is *Overrepresentation analysis (ORA)*, which is based on the assignment of genes to predefined categories. Generally, given a gene expression data set, a subset of genes needs to be defined which should contain just the differentially expressed genes. ORA evaluates for each category whether the fraction of genes annotated with the respective category in the set of differentially expressed genes is significantly larger than expected from the total set of analyzed genes. ORA is frequently performed with Gene Ontology (GO) annotations (e.g. Onto-Express and Onto-Tools ([Khatri et al., 2002, 2006](#)), MAPPFinder ([Doniger et al., 2003](#)), GoMiner ([Zeeberg et al., 2003](#)), GeneMerge ([Castillo-Davis and Hartl, 2003](#)), FatiGO ([Al-Shahrour et al., 2004](#)), goCluster ([Wrobel et al., 2005](#))). ORA has also been applied for the identification of statistically significantly enriched pathways (e.g. KOBAS ([Mao et al., 2005](#))). [Curtis et al. \(2005\)](#) reviewed several methods of pathway analysis and compare the performance of three methods.

Generally, a cutoff needs to be chosen to select the genes considered to be significantly

regulated for ORA; the overall result usually depends on this cutoff. Methods that do not require a cutoff include the work by Al-Shahrour *et al.* (2005); they proposed a method to scan ordered list of genes for over-represented annotations, and the Gene Set Enrichment Analysis (Subramanian *et al.*, 2005), which ranks genes and calculates an enrichment score that reflects overrepresentation at the extremes of the entire list. ORA reveals processes that are relevant for the experiment under investigation, yet the predefined categories limit the level of detail of the result. Datta and Datta (2006) made use of functional annotations to systematically judge the results of an unsupervised clustering of genes; to this end, they introduced two performance measures that evaluate biological homogeneity and stability. Manually curated networks and pathway models contain information on causal relationships and thus provide knowledge beyond single genes. Several public databases contain such networks and pathway models (e.g. Transpath (Krull *et al.*, 2006), KEGG (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2002), BioCarta¹, GenMAPP (Dahlquist *et al.*, 2002)). Yet, due to the important effort required for manual curation, their content is generally limited and mostly focused on well-known phenomena. Gene expression data can be visualized in context of networks, and thus dependencies of regulatory events can be detected. This can help in understanding experimental observations. GenMAPP (Dahlquist *et al.*, 2002) is a stand-alone tool for viewing and analyzing gene expression data in the context of biological pathways; it integrates color-coding of genes and access to annotation for genes. Pathway Miner (Pandey *et al.*, 2004) maps genes onto pathways and extracts gene product association networks for genes that co-occur in pathways. Thus, gene expression profiles can be visualized in the context of biological pathways, and pathways can be ranked based on statistical tests.

Overrepresentation analysis as well as analysis of hand-curated pathways and networks rely on manually compiled data (gene annotations or networks) which requires significant human effort and thus can never be comprehensive. Text mining can overcome the rate-limiting step of manual annotation/curation and thus represents a useful alternative for the generation of context for gene expression data analysis.

9.2 Integration with Large-Scale Networks

Large-scale networks can, for example, be obtained from experimental measurements (e.g. by yeast two-hybrid technology) or text mining. Numerous methods make use of large-scale networks for the interpretation of gene expression data. Co-clustering (Hanisch *et al.*, 2002) is an approach for joint clustering of genes and vertices of a network. Co-clustering makes use of a combined distance function integrating correlation of gene expression measurements and graph distance on networks. The approach identifies processes that are relevant for the condition under which the gene expression data was obtained.

Significant area search (Sohler *et al.*, 2004) is an approach that maps gene expression data on networks and extracts subnetworks that show a significant number of regulated genes.

¹<http://www.biocarta.com>

These subnetworks often provide hints on the biological processes which are affected in the measured conditions. The algorithm is based on the selection of a set of seed genes according to a specified threshold and subsequent greedy expansion by including the most significant neighboring molecules. The selected subnetworks are evaluated by combining individual p-values by Fisher's inverse χ^2 method (Fisher, 1932), where the individual p-values are adjusted based on local graph topology.

Pathway Queries (Sohler *et al.*, 2004; Sohler and Zimmer, 2005) represent a means to specify hypotheses as queries against a network, given gene expression data and functional annotation of individual genes. The functional annotations can be generated automatically (e.g. transcription factor binding sites identified by sequence analysis), or derived from manual annotations (e.g. gene ontology). Pathway queries make use of an XML-based language in which pathway templates are formulated as graph-like structures. The user can specify queries and define required properties of genes or proteins and pose restrictions on edges. The algorithm then extracts instances matching a given query from the network. A scoring function can be applied to identify active transcription factors or kinases. Pathway queries thus provide a step towards explaining gene expression data by upstream events. The latter two algorithms are implemented in ToPNet (Hanisch *et al.*, 2004), a tool for combined visualization and exploration of gene networks and expression data. ToPNet makes use of networks in form of Petri Nets; that is, bipartite graphs in which the places represent molecules, and transitions represent relationships between these molecules. ToPNet also implements basic graph algorithms like computing hulls around genes for exploring the neighborhood of genes or computing all shortest paths among selected molecules. Data maps contain annotations to places and can be used for gradient color coding or defining the size of visualized places representing molecules. Data maps also provide a means to annotate places with links to external databases, or with Gene Ontology terms that can be used for filtering networks or specifying queries.

9.3 Integration with Text Mining

Text mining can be used to generate networks. A number of text mining techniques have been applied to generate networks (see Section 5.1). The resulting networks are significantly more comprehensive than those obtained from manual curation and are thus suited for application of the approaches described in the previous section.

Besides, literature on genes can be used for detecting groups of similar genes. In contrast to manual annotation of individual genes, text mining makes use of information on genes in a more comprehensive way and increases the speed of data analysis. Furthermore, the usage of predefined annotation categories can be avoided with text mining.

A number of approaches and tools have been presented that directly integrate gene expression data analysis and text mining. In the following, several of these approaches are shortly described: MedMiner (Tanabe *et al.*, 1999) is an Internet-based program which filters and organizes large amounts of textual and structured information obtained from public search engines such as GeneCards or PubMed and that can be used for gene ex-

pression data analysis. PubGene ([Jenssen et al., 2001](#)) contains a database of gene-gene co-citations annotated with MeSH and GO-terms and web tools for gene expression analysis. [Chaussabel and Sher \(2002\)](#) derived literature profiles containing term frequencies from MEDLINE and performed a cluster analysis; they found that the resulting clusters gave a coherent picture of the functional relationships among lists of genes. LACK ([Kim et al., 2003](#)) determines the statistical significance of apparent lexical bias in microarray datasets by assessing the frequency of a user-specified list of search terms in a set of differentially expressed genes in comparison to randomly generated datasets. The microGENIE system ([Korotkiy et al., 2004](#)) is a tool for semi-automated querying of PubMed that combines information from UniGene and Swiss-Prot and thus obtains information on the biological relevance of differentially expressed genes. [Aubry et al. \(2006\)](#) presented a method for functional annotation of gene sets based on a combination of GO annotations and gene-term associations detected by literature mining. The LMMA approach ([Li et al., 2006](#)) refines a literature-based co-occurrence network by a multivariate selection procedure based on microarray data and thus generates more reliable networks.

An example of a system that integrates information from diverse databases, text-mining, and graph-based analysis and visualization is the ONDEX system ([Kohler et al., 2006](#)). Most of the approaches presented so far primarily focus on one of the analyzed data types, gene expression data or text mining, and then add, in a second step, the other type of data. For example, several approaches select a subset of regulated genes or cluster gene expression data and annotate the resulting clusters by use of text mining (e.g. [Masys et al. \(2001\); Masys \(2001\)](#)). Other approaches first apply text mining and then map expression data to the text data to provide functional annotations (e.g. [Chaussabel and Sher \(2002\); Chaussabel \(2004\)](#)). In the following, an approach is presented that differs significantly from the above approaches as it integrates text and gene expression data simultaneously; thus, both types of data can contribute equally.

Chapter 10

Text Mining applied for the Interpretation of Gene Expression Data

The biological interpretation of gene expression data generally requires detailed knowledge on the investigated systems which is not always readily available. Large repositories of biomedical publications such as MEDLINE represent an enormous and valuable information resource. Text mining can be applied to automatically exploit such large text resources and extract relevant information and thus, it can compensate for missing knowledge or complement available knowledge. Thus, text mining represents a useful means for extracting information that can complement experimental data for combined analysis and interpretation.

In this chapter an approach is presented that integrates gene expression data and text mining and extracts clusters of genes that are predominantly regulated and have a coherent literature background ([Küffner et al., 2005](#)). The clusters and associated concepts provide interpretations and biological hypotheses for the expression experiment. The approach has been developed by Robert Küffner.

It relies on gene and protein name identification ([Chapter 4](#)) and on appropriately processed gene expression data ([Chapter 7](#)). The ConceptMaker approach represents one example of how text mining and gene expression data analysis can be integrated to provide relevant gene groups together with literature contexts.

10.1 ConceptMaker

ConceptMaker combines gene expression measurements with text mining and derives interpretations together with biological hypotheses ([Figure 10.1](#)). In contrast to previous approaches which combine the data types sequentially (e.g. [Masys et al. \(2001\)](#); [Chaussabel and Sher \(2002\)](#)), ConceptMaker integrates text and expression data simultaneously. It identifies gene clusters that exhibit significantly regulated genes and a coherent literature profile; the so-called *concepts* that describe active functional contexts. The approach does

not rely on controlled vocabularies, manual annotations or pathway resources.

The ConceptMaker approach is based on two essential features: (1) a method to derive a literature topic or hypothesis given a set of genes with interesting gene expression patterns and (2) a method to select genes that belong to a given literature topic and expression pattern. Here, a topic is defined as a consistent and coherent set of literature features ([Shatkay et al., 2000](#)).

Data Representation

Each object (documents, terms, genes, topics) is represented as a vector of term weights. Words derived from MEDLINE abstracts and MeSH headings, bigrams within a window of three consecutive words, and protein identifiers are filtered based on statistical criteria and used as terms t . Documents are represented via the occurrence of terms in a vector space model. Thus, a set of documents d is represented by a matrix A containing a row for each document and a column for each term. The term weights are determined according to a variant of $tfidf$ (term frequency – inverse document frequency, [Kim et al. \(2001\)](#)). Latent semantic indexing (LSI) is then applied: A is subjected to singular value decomposition (SVD): $A = D \sum T^*$, where T^* is the conjugate transpose of T , and D and T are matrices of document and term singular vectors, respectively, and \sum is a diagonal matrix with the singular values on the main diagonal. The singular vectors with the smallest singular values are then discarded to reduce the noise in term usage in the corpus. The number of remaining dimensions k is chosen significantly smaller than the number of documents $|d|$ and terms $|t|$.

Principles of the ConceptMaker Algorithm

Human gene and protein name identification in MEDLINE abstracts from 1990 or later returned 1.7 million abstracts containing 16 100 different protein objects and 4.3 million abstract–protein links. Abstracts were split up in words and bigrams. Very frequent and rare words were pruned as well as documents containing no or only very frequent proteins resulting in 516 000 documents with 128 000 terms including 43 000 bigrams and 38 000 MeSH/qualifier terms.

ConceptMaker uses the following definitions:

A *Cluster* is a set of significantly regulated genes that share a coherent literature context.

A *Topic* refers to the literature features that are shared by the members of a cluster.

A *Concept* represents associated clusters and topics.

A set of terms q or a cluster of genes C can be projected into the LSI space. This returns a *term/gene group profile* (*tgp*): $tgp(A_k, C) = tgp(A_k, q) = q^T K_k \sum^{-1}$. Thus, objects or sets of objects can be compared in SVD space by the cosine between the respective profiles.

Starting from a cluster of all genes, the ConceptMaker algorithm constructs a sequence of topic matrices A_k^i and finds the best topic with respect to the matrix and expression data. ConceptMaker iteratively reduces the current cluster to a coherent topic (analogous

to the gene shaving method presented by [Hastie *et al.* \(2000\)](#)) by selecting the best genes according to a score that measures compatibility of genes to the gene cluster, its regulatory pattern and literature profile. Then, the cluster size is kept constant and members of the cluster are replaced by alternative ones so that the topic may shift. Thus a cluster of a particular size is determined that is most compatible with a topic. Finally, the best cluster is selected from the set of nested clusters and returned as the best topic for the current topic matrix.

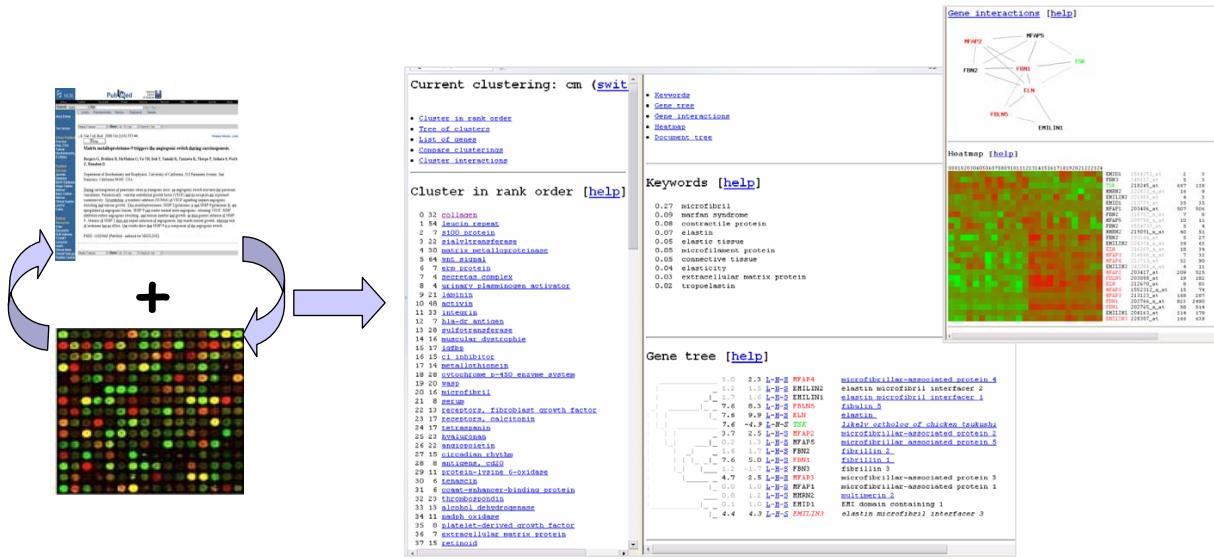


Figure 10.1: The ConceptMaker approach: ConceptMaker integrates analysis of literature and gene expression data and identifies genes that are functionaly related and significantly regulated; these define the so called active functional contexts. The results are visualized by a web page.

Once a cluster is selected, the matrix is orthogonalized with respect to the found cluster; that is, the singular vectors in the SVD are replaced by vectors which are orthogonal to the group profile of the selected cluster, and thus a new matrix is constructed. Thereby, the contribution of the identified cluster and its literature profile is removed from the topic matrix, and thus the final clusters will cover distinct topics.

For analyzing and visualizing the resulting clusters, a summary is derived for each cluster by analyzing the similarity between the individual term profiles and the respective cluster profile; the terms which achieve highest similarity describe the respective cluster. Furthermore, the term and gene group profiles make it possible to determine the similarity of terms/genes within a cluster. Analogously, the similarity between publications can be determined. Thus, substructures within gene clusters can be revealed by hierarchical clustering. ConceptMaker takes advantage of this possibility by visualizing the genes and documents of a cluster in a dendrogram of genes and documents, respectively. The dendro-

grams enable the user to quickly screen the literature as similar articles are grouped and to get an overview of the relevant genes.

Results

ConceptMaker has been applied on the Affymetrix two-class data set that compares osteoarthritic to normal human joint cartilage (Section 7.2). The right part of Figure 10.1 shows a small part of the web page that summarizes the results. The web page indicates the identified clusters. For each cluster, the page lists a set of keywords. Furthermore, it shows a gene tree that depicts a dendrogram of the genes in the cluster, a visualization of gene interactions in form of a network, a heatmap that indicates the expression levels of the genes, and a document tree that contains the documents which are relevant for the respective cluster. The gene and document tree can be generated based on the information obtained from text mining; they present the data so that similar objects appear close to each other.

For the analyzed data set, 59 clusters were identified. Among the top-ranked clusters, there are clusters termed *collagen*, *matrix metalloproteinase*, *wnt signal*, *secretase complex* and *integrin*. As a literature analysis showed, for all of these clusters, evidence can be found that the genes that are part of these clusters presumably play a role in the context of osteoarthritis.

The ConceptMaker web page is extensively cross-linked within the web page and to other data sources; for example, each gene is linked to Entrez Gene ([Maglott et al., 2007](#)), HUGO ([Eyre et al., 2006](#)), and Swiss-Prot ([Bairoch et al., 2005](#)), and literature references are linked to PubMed. Thus, ConceptMaker supports browsing of the results and enables the researcher to get an overview of the most relevant gene groups and related contexts for the given experiment.

Conclusions

ConceptMaker identifies the relevant gene groups and corresponding literature topics for a given gene expression experiment. It thus supports the researcher in the generation of hypotheses concerning the relevant processes.

Importantly, ConceptMaker overcomes several limitations of previous approaches. One of the most widely used approaches for integrated analysis of gene expression data and context information is overrepresentation analysis (ORA), which is usually performed based on gene ontology annotations to the individual genes (Section 9.1). ORA makes use of pre-defined ontologies, requires genes to be annotated to contribute information, and generally, requires a cutoff to be set for selecting differentially expressed genes.

ConceptMaker is more flexible: It does not depend on predefined ontologies, genes do not need to be annotated, and it does not require the selection of a subset of differentially expressed genes. By using gene and protein name identification (Chapter 4), ConceptMaker can directly make use of literature data. This entails important benefits as biomedical liter-

ature represents a much larger, more comprehensive, and more detailed information source than any ontology currently available. ConceptMaker can thus make use of information for genes that are not yet annotated with ontology terms. Furthermore, it can generate specific concepts for any domain, irrespectively of whether it is covered in an ontology or not, by extracting terms that describe clusters directly from text.

In terms of usage of gene expression data, it is important to bear in mind that ConceptMaker does not require a cutoff for selecting significantly regulated genes. Instead, all measured genes contribute to the overall result according to their rank. Genes can be included in a cluster even if only moderately regulated, provided that they fit well to the cluster topic. Thus, clusters are not restricted to the most strongly regulated genes as important conclusions can also be drawn from non-regulated genes in a predominantly regulated cluster. Similarly, sets of genes, where the single genes are inconspicuous but the entire set is relevant, can be detected.

Chapter 11

Conclusions

Techniques that can be used for large-scale experiments are nowadays used routinely. For example, microarrays can monitor the expression levels of all genes of a genome simultaneously. The amount of data which must be processed and evaluated necessitates the adoption of automatic analysis methods. For biological data interpretation, detailed background knowledge on the investigated systems is required. This knowledge can be retrieved from scientific publications, which are nowadays publicly available via large repositories such as MEDLINE. For large scale experiments, such as microarray gene expression studies or yeast two-hybrid protein interaction studies, the amount of relevant literature is too large to be manually reviewed. Here, text mining provides a means to support biomedical research.

Adequate processing of gene expression data is not only required for direct biological interpretation, but also for integrating expression measurements with other types of data in advanced analysis methods.

11.1 Contributions of this thesis

This thesis addresses two major points which are of high importance for an combined interpretation of text information and gene expression data.

First, several aspects in text mining which are relevant for biomedical data interpretation are addressed: Methods for deriving high-quality synonym dictionaries for genes and proteins, as well as for other biological entities have been developed. The resulting dictionaries are useful for named entity identification; that is, for searching occurrences of objects in texts and mapping these occurrences to specific objects. The synonym dictionaries make it possible to map each mention of an object to its identifier irrespectively of the individual name being used in a text. Dictionaries thus make it possible to retrieve texts dealing with specific objects and to integrate information. The compiled dictionaries are publicly available via several tools.

New methods and systems for named entity identification have been developed and an existing system has been expanded. The systems rely on synonym dictionaries and imple-

ment a modular concept which supports straightforward customization. A system can thus easily be tuned to allow for the characteristics of individual organism nomenclatures. This is highly important as nomenclatures of different organisms have been shown to vary significantly, from very strict and easy to recognize to highly variable and bearing a high degree of ambiguity. The developed methods together with the synonym dictionaries achieve very high recall and precision in named entity identification, as shown by the results of the BioCreAtIvE challenge evaluations ([Fundel et al., 2005a](#); [Hanisch et al., 2005](#); [Fundel and Zimmer, 2007](#)).

All dictionaries for gene and non-gene objects have been searched against the largest collection of biomedical research publications; this provides a comprehensive annotation of the individual articles.

Based on the aforementioned methods, an approach for relation extraction has been developed. This approach makes use of publicly available preprocessing tools for natural language processing. By application of a small number of rules it detects and extracts relations with high recall and precision ([Fundel et al., 2007](#)). The type of relations to be extracted can be tuned to fit individual requirements, and the extracted relations can be restricted to a reduced set of high-confidence relations. The approach has been applied on a comprehensive subset of the largest collection of biomedical research publications, and thus a network containing approximately 150 000 relations between 11 000 genes and proteins has been generated. This network contains significantly more relations than any of the available public databases for experimental or literature-curated interactions, and exhibits high recall and precision estimates. Thus, it represents a valuable data source for various applications.

The network can be used for the expansion of manually compiled networks, which can be done by automatic means or manual curation. The relation extraction approach alleviates manual curation as it provides links to the literature and the restricted sets of relations already contain the most relevant information. Together with the context annotations compiled by searching the non-gene dictionaries against the texts, the network can be screened for relations which were described in certain contexts, or overrepresented contexts can be automatically detected for individual relations or subnetworks, which provides a functional description of the respective relation or subnetwork. A classification approach makes it possible to characterize the extracted relations further; thus, each relation can be assigned to predefined categories.

The second major point in this thesis concerns gene expression data processing. This part focused on the analysis of several gene expression data sets that have been generated to support the analysis of molecular processes implicated in osteoarthritis. Most of the data sets were generated in a large research project on osteoarthritis (“Leitprojekt Diagnose und Therapie der Osteoarthrose”).

One of the data sets was not suited for analysis with standard methods for p-value determination as it showed large inter-spot variances for genes represented by several spots on a chip; yet, the data set was very interesting from a biological point of view, as it represented the largest data set in the domain. Therefore, the effects of primary gene ex-

pression data processing on the higher level data analysis have been analyzed in detail for this data set, and a new method for the identification of differentially expressed genes has been proposed. This method can be used as an alternative to existing methods; it offers important advantages for highly skewed data showing large inter-spot variances for genes represented by several spots on a chip. The appropriate processing of the data together with the detailed biological interpretation provided new insights in the molecular events implicated in osteoarthritis ([Aigner et al., 2006a](#)).

Similar cartilage samples were measured with diverse types of microarrays. The resulting data sets have been used for comparing the platforms against each other. Furthermore, a data set obtained from cell lines was compared against the data obtained from *in vivo* samples; that is, samples that were directly derived from human joint cartilage. This comparison indicated that the investigated cell lines are a poor model for osteoarthritis. The analysis of an IL1 stimulation time series experiment revealed the involved processes, which supports biological interpretation.

The third part of this work provides the connection between the first two parts. This part deals with combined data analysis methods. Many combined analysis approaches rely on named entity identification, relation extraction, and gene expression data analysis as a prerequisite. The methods that have been developed in this thesis showed very good performance and the generated data are of very high quality. The methods and data directly contribute to the development of integrated approaches. An example for an integrated approach is given by the ConceptMaker algorithm ([Küffner et al., 2005](#)), which selects genes that are differentially expressed and at the same time have a coherent literature background. This approach builds on the tools and data derived in this work. Its application on an osteoarthritis-related data set resulted in several gene clusters that are clearly relevant for osteoarthritis.

11.2 Perspectives for Future Research

The methods developed in this thesis showed good performance and large-scale applicability. In the respective sections, possibilities for further development and extension of the individual methods were proposed and discussed. Yet, as the performance values are already very good, it seems desirable to shift the focus from method performance improvement towards the enlargement of scope of the methodologies or towards new areas of application. The increase in biological knowledge as well as recent technical advances suggest a wide range of future trends and research directions related to the topics addressed in this thesis. In the following, several future directions are suggested; the main directions include integration of further data sources and data types, integration issues in text mining, and the development of combined analysis methods.

Integration of further data sources and data types The presented approaches for text mining as well as for gene expression data analysis focus on individual genes and

proteins, assuming that these are well defined objects. Yet, a protein is not necessarily a one-to-one reproduction of a gene and, generally, it is not static and clearly defined by sequence alone. It is known that synthesis of an important fraction of proteins involves *alternative splicing*, a process by which, distinct subsequences of the mRNA molecule are excised based on a single species of original mRNA and used as template for protein synthesis. Thus, several distinct proteins can be derived from a single gene. Furthermore, post-translational modifications of proteins (e.g. glycosylation, phosphorylation) can alter functional properties and e.g. regulate enzymatic activity. These modifications can either be static and depending on the environment (e.g. tissue) or continuously be changed according to a cell's state, influence of stimulatory agents, etc. So far, most databases, structured terminologies, and scientific articles do not reflect this level of detail, even though specific databases for the respective information have been set up (e.g. FAST DB ([de la Grange et al., 2005](#)), which defines the exon content of all known transcripts produced by human genes). Information on protein variants and modifications should be considered not only in the analysis and interpretation of experimental data, but also in text mining.

With the progression of experimental techniques, increasing amounts of data on more and more detailed level becomes available. *Mass spectrometry (MS)* can be used to detect post-translational modifications and protein expression. *Exon arrays* can be used to detect alternatively spliced mRNAs. *Tiling arrays* even make it possible to analyze entire genomes for expressed sequence stretches at very high resolution. Thus, expression events in genomic regions which have, so far, not been characterized as genic regions, can be detected. These array types recently became available for off-the-shelf use. Data analysis methods for these techniques are currently being developed by the scientific community. The integration of the derived data with other data sources clearly represents a future challenge, in terms of algorithms as well as in terms of appropriate nomenclature. Therefore, a common vocabulary for the respective splicing and modification events and resulting subspecies of proteins is required.

Furthermore, genes vary between individuals. Most of these variations originate from *single nucleotide polymorphisms (SNPs)*; that is, DNA sequence variations of a single nucleotide at specific positions in the genome. These genetic variances can cause inter-individual differences in susceptibility to pathogens and response to treatment by specific drugs. SNPs thus influence biochemical phenomena and accordingly represent a further dimension in the gene and protein universe. Pharmacogenomics is concerned with analyzing the influence of genetic variations on drug response, a question of immense interest to the pharmaceutical industry. Microarrays make it possible to experimentally determine an individual's genotype by SNPs analyses. So far, SNPs data has mainly been analyzed by itself or together with disease or phenotype data. The systematic integration of SNPs analyses with other types of data represents a future challenge. Here again, common nomenclature becomes essential.

Besides more details on genes and proteins, data types other than numeric or text should be exploited and integrated. An important amount of information on biological knowledge is being published as images or figures. Thus, the integration of image data represents an important, yet challenging step towards more extensive usage of public information.

Automated image analysis is a difficult task and a research domain on its own; it has not yet been applied for large-scale biomedical information extraction. Only first steps in this direction have been proposed, most of them being based on the analysis of figure legends rather than figures. For example, Liu *et al.* (2004) presented a method for indexing and classification of figures based on the figure legends, and Shatkay *et al.* (2006) presented a method for text categorization based on features derived from image data alone and in combination with text data. The detailed exploitation of images from biomedical research publications appears very interesting; yet, this task also appears rather difficult as image data is very heterogeneous and generally additional text information and background knowledge is required to correctly interpret images.

Furthermore, information beyond the molecular level should be integrated in a comprehensive and consistent way. In fact, researchers are often interested in dependencies between effects on the molecular level and observations at the *phenotype* level. The systematic integration of phenotypic data again requires a common vocabulary, and the large flexibility that natural language offers for describing phenotypes suggests specific adjustments of text-mining systems.

Integration issues in text mining For text mining, the different data types mentioned above represent an important challenge. Similarly to current gene and protein nomenclature approaches, common vocabularies that define unique identifiers and designations for the respective data are required to render data integration possible.

Text mining approaches dealing with individual aspects of the above data types have already been presented. For example, Shah *et al.* (2005) presented an approach for extracting information on transcript diversity from MEDLINE. Based on this work, Shah and Bork (2006) described a classification approach that inductively learns to identify sentences talking about physiological transcript diversity from MEDLINE; the sentences were subjected to semantic role labeling for identifying semantic categories, and the obtained information is made publicly available in the LSAT database. Bonis *et al.* (2006) presented OSIRIS, a tool for retrieving literature about sequence variation (SNPs) of a gene.

So far, the presented approaches focus on specific types of data; the systematic integration of the diverse data types represents a future challenge.

Especially, due to technical advances, the *level of detail* of scientific articles generally increases; that is, newer articles contain more detailed information than older articles. None of the text-mining approaches presented so far distinguishes between older and newer literature for data integration and proposes an integration procedure that considers the respective standard of knowledge. Therefore, the information derived from older literature would either have to be automatically broken down to a more fine-grained level, or different levels representing the standard of knowledge would be required for merging older and newer literature data.

For large-scale data integration, it is desirable to represent data in a consistent, uniform format. Several standards for the *representation of pathway and network information* have been presented (e. g. SBML (Hucka *et al.*, 2003), PSI MI (Hermjakob *et al.*, 2004a), BioPAX

(Luciano, 2005), for a comparison see Stromback and Lambrix (2005)); these make it possible to add detailed context annotations to networks, which is useful for integrated data analysis approaches. The conversion of the generated data into one of these standard formats can be considered as a short-term goal.

Development of combined analysis methods Several directions of integrated data analysis methods would be of interest. One research area that is currently very active is the domain of *systems biology*. Systems biology includes approaches that model an entire system (e.g. pathway, cell) by differential equations which describe processes in a quantitative way. Systems biology thus requires detailed information on reactions and kinetic parameters. These are time-consuming and sometimes difficult to determine experimentally. Text mining could provide a useful means for extracting relevant information from the literature; first approaches in this directions have been presented (e.g. Hakenberg *et al.* (2004)).

So far, biomedical text mining is mainly used for extracting static observations. A new dimension could be introduced by considering time-dependent effects and thus *dynamic observations*. This could be considered as a counterpart to dynamic experimental data, such as time-series gene expression measurements. So far, pathway visualizations generally only provide a static view of regulatory dependencies. Altogether, text-mining, pathway data, and time-series measurements should provide an insight in the dynamics of cellular events.

The systematic integration of gene expression data and yeast two-hybrid data could offer further possibilities. So far, these data types have mainly been combined to screen data for differentially expressed subnetworks (Ideker *et al.*, 2002), to group functional modules (Tornow and Mewes, 2003; Segal *et al.*, 2003a), and to estimate the level of confidence for protein interactions (e.g. Deng *et al.* (2003)). Ma *et al.* (2007) presented a method for prioritizing genes associated with a certain phenotype by combining gene expression and protein interaction data. The combination of these data types might offer more information, for example, on varying complex composition in specific contexts, such as different tissues or cell types.

Few publications present data analysis methods that deal with more than two data types. For example, Herrgard *et al.* (2006) presented an approach that integrates analysis of gene expression and protein–DNA interaction data sets, growth phenotype experiments, a transcriptional regulatory and metabolic network model. By an iterative procedure they could identify regulatory cascades and new interactions; thus, they could predict growth phenotypes of transcription factor knockout strains. Aerts *et al.* (2006) presented an approach for gene prioritization based literature data, functional annotation, microarray expression, EST expression, protein domains, protein–protein interactions, pathway membership, transcriptional motifs, sequence similarity, and a set of training genes. Clearly, the integration of multiple data types appears promising as it covers more aspects of biomedical phenomena than single data sets. Issues concern how to best integrate the diverse data sets, and eventually how to weight the relevance of individual data sets to reflect the importance

and reliability.

Finally, the integration of diverse data types from multiple *organisms* is an important challenge that seems promising. An important fraction of current research is devoted to a small number of organisms, such as human, mouse, drosophila, or yeast, whereas for most other organisms, little data and information is available. When studying an organism for which little information is available, it appears appealing to transfer as much knowledge as possible from well-investigated organisms to the one of interest; yet, inter-species differences must be accounted for. On genome and protein sequence level, data from diverse organisms is routinely compared and integrated, for example, for determining the degree of similarity between species or for protein structure prediction by homology modeling. For example, [Lu et al. \(2006\)](#) integrated gene expression data with sequence similarity information for the identification of cycling genes; their method represents genes from multiple species as nodes in a graph and can use a high quality dataset from one species to overcome noise problems in another. On the higher levels (e.g. pathways), no methods are readily available for transferring the information from one organism to another. Such methods would be very useful as they would speed up research progress. Eventually, the comparison of pathways between organisms could reveal that not single genes or proteins, but their interactions make the difference between organisms, which would be very important to know.

The methods and data generated in this work provide an important step towards several of the future challenges described above. The developed text mining methods are flexible and can thus easily be applied on new data types. The text mining methods depend on data sources that are appropriate for the construction of dictionaries. As the construction of common vocabularies and ontologies is a very active field of research, dictionaries for a wide range of objects and terms are readily available. Thus, network models can be compiled from texts together with context annotations.

The large-scale gene and protein text-mining network that has been generated in this work together with the extracted context annotations is immediately usable for integrated data analysis. Thus, the development of methods that exploit networks together with the extracted contexts and experimental data is a promising task that can directly be tackled. Immediate next steps could aim at providing a more detailed characterization of individual network components (for example, certain protein interactions or protein–gene relations might play a crucial role in a certain disease), at deriving descriptions of biomedical phenomena and consequently also at deriving disease hypotheses. Towards this goal, combination with more ambitious systems biology approaches will be the next big challenge.

Abbreviations

F	:	F-measure
GO	:	Gene Ontology
HMM	:	Hidden Markov Model
IE	:	Information Extraction
IR	:	Information Retrieval
LSI	:	Latent Semantic indexing
MGI	:	Mouse Genome Informatics
MeSH	:	Medical Subject Headings
NE	:	Named Entity
NER	:	Named Entity Recognition
NEI	:	Named Entity Identification
NLP	:	Natural Language Processing
ORA	:	Overrepresentation analysis
P	:	Precision
PCA	:	Principal Component Analysis
POS	:	Part of Speech
pp	:	percentage points
PPI	:	Protein-protein interaction
R	:	Recall
RGD	:	Rat Genome Database
SGD	:	Saccharomyces Genome Database
SNP	:	Single Nucleotide Polymorphism
SVD	:	Single Value Decomposition
SVM	:	Support Vector Machine
UMLS	:	Unified Medical Language System
Y2H	:	Yeast Two-Hybrid

Bibliography

- Adar, E. 2004. SaRAD: a Simple and Robust Abbreviation Dictionary. *Bioinformatics*, 20(4):527–33. [24](#), [47](#)
- Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. 2006. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–44. [184](#)
- Ahmed, S., Chidambaram, D., Davulcu, H., and Baral, C. 2005. IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 54–61. [89](#)
- Aigner, T., Bartnik, E., Sohler, F., and Zimmer, R. 2004. Functional genomics of osteoarthritis: on the way to evaluate disease hypotheses. *Clin Orthop Relat Res*, (427 Suppl):S138–43. [128](#), [134](#)
- Aigner, T., Bartnik, E., Zien, A., and Zimmer, R. 2002. Functional genomics of osteoarthritis. *Pharmacogenomics*, 3(5):635–50. [128](#), [134](#)
- Aigner, T., Fundel, K., Saas, J., Gebhard, P. M., Haag, J., Weiss, T., Zien, A., Obermayr, F., Zimmer, R., and Bartnik, E. 2006a. Large-scale gene expression profiling reveals major pathogenetic pathways of cartilage degeneration in osteoarthritis. *Arthritis Rheum*, 54(11):3533–44. [xiv](#), [xvi](#), [147](#), [157](#), [166](#), [181](#)
- Aigner, T., Sachse, A., Gebhard, P. M., and Roach, H. I. 2006b. Osteoarthritis: pathobiology-targets and ways for therapeutic intervention. *Adv Drug Deliv Rev*, 58(2):128–49. [128](#), [134](#)
- Aigner, T., Zien, A., Hanisch, D., and Zimmer, R. 2003. Gene expression in chondrocytes assessed with use of microarrays. *J Bone Joint Surg Am*, 85-A Suppl 2:117–23. [128](#), [134](#)
- Al-Shahrour, F., Diaz-Uriarte, R., and Dopazo, J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–80. [169](#)

- Al-Shahrour, F., Diaz-Uriarte, R., and Dopazo, J. 2005. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, 21(13):2988–93. 170
- Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F., and Hogue, C. W. 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*, 33(Database issue):D418–24. 6, 114
- Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65. 126
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402. 56
- Andrade, M. A. and Valencia, A. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–7. 15
- Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):D226–9. 6
- Aronson, A. R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, pages 17–21. 7
- Ashburner, M. and Lewis, S. 2002. On ontologies for biologists: the Gene Ontology—untangling the web. *Novartis Found Symp*, 247:66–80; discussion 80–3, 84–90, 244–52. 7, 64
- Aubry, M., Monnier, A., Chicault, C., de Tayrac, M., Galibert, M. D., Burgun, A., and Mosser, J. 2006. Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of gene sets. *BMC Bioinformatics*, 7:241. 172
- Bader, G. D., Betel, D., and Hogue, C. W. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–50. 105

- Bader, G. D. and Hogue, C. W. 2002. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–7. [109](#)
- Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. 2004. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1):78–85. [109](#)
- Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 33 Database Issue:D154–9. [6](#), [23](#), [27](#), [29](#), [44](#), [72](#), [102](#), [176](#)
- Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R., Oughtred, R., Skrzypek, M., Theesfeld, C. L., Binkley, G., Dong, Q., Lane, C., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J. M. 2005. Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the *Saccharomyces* Genome Database (SGD). *Nucleic Acids Res*, 33 Database Issue:D374–7. [6](#), [7](#), [27](#)
- Ballman, K. V., Grill, D. E., Oberg, A. L., and Therneau, T. M. 2004. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, 20(16):2778–86. [140](#)
- Barabasi, A. L. and Oltvai, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–13. [103](#)
- Barnes, M., Freudenberg, J., Thompson, S., Aronow, B., and Pavlidis, P. 2005. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res*, 33(18):5914–23. [163](#)
- Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R. 2007. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res*, 35(Database issue):D760–5. [127](#)
- Beissbarth, T., Fellenberg, K., Brors, B., Arribas-Prat, R., Boer, J., Hauser, N. C., Scheideler, M., Hoheisel, J. D., Schutz, G., Poustka, A., and Vingron, M. 2000. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16(11):1014–22. [147](#)
- Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, 35(Database issue):D301–3. [6](#)
- Blake, J., Richardson, J., Bult, C., Kadin, J., Eppig, J., and the members of the Mouse Genome Database Group. 2003. MGD: The Mouse Genome Database. *Nucleic Acids Res*, 31:193–195. [6](#), [7](#), [71](#)

- Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, pages 60–7. **16**, 88
- Blaschke, C., Hirschman, L., and Valencia, A. 2002. Information extraction in molecular biology. *Brief Bioinform*, 3(2):154–65. **88**
- Blaschke, C. and Valencia, A. 2001. The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform Ser Workshop Genome Inform*, 12:123–34. **88**
- Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32 Database issue:D267–70. **7**, 31, 46, 64
- Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193. **134**
- Bonis, J., Furlong, L. I., and Sanz, F. 2006. OSIRIS: a tool for retrieving literature about sequence variants. *Bioinformatics*, 22(20):2567–9. **183**
- Box, D., Ehnebuske, D., Kakivaya, G., Layman, A., Mendelsohn, N., Nielsen, H., Thatte, S., and Winer, D. 2000. Simple Object Access Protocol (SOAP) 1.1. W3C Note 08 May 2000. URL: <http://www.w3.org/TR/SOAP>. **50**
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4):365–71. **127**
- Brill, E. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy*. **31**, 59
- Bunescu, R., Ge, R., Kate, R., Mooney, R., Wong, Y., Marcotte, E., and Ramani, A. 2003. Learning to Extract Proteins and their Interactions from Medline Abstracts. In *Proceedings of ICML-2003 Workshop on Machine Learning in Bioinformatics*, pages 46–53. **56**
- Bunescu, R., Mooney, R., Ramani, A., and Marcotte, E. 2006. Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline. In *HLT-NAACL Workshop on Linking Natural Language Processing and Biology: Towards deeper biological literature analysis (BioNLP-2006)*, pages 49–56, New York City, NY. Association for Computational Linguistics. **89**

- Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, 32(Database issue):D262–6. 7
- Cantor, M., Sarkar, I., Bodenreider, O., and Lussier, Y. 2005. Genestrace: Phenomic knowledge discovery via structured terminology. *Pac Symp Biocomput*, 103:14. 7
- Cardozo, L., Bachmann, G., McClish, D., Fonda, D., and Birgerson, L. 1998. Meta-analysis of estrogen therapy in the management of urogenital atrophy in postmenopausal women: second report of the Hormones and Urogenital Therapy Committee. *Obstet Gynecol*, 92(4 Pt 2):722–7. 146
- Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummeacker, M., Paley, S., Pick, J., Rhee, S. Y., Tissier, C., Zhang, P., and Karp, P. D. 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34(Database issue):D511–6. 6
- Castillo-Davis, C. I. and Hartl, D. L. 2003. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–2. 169
- Chang, C.-C. and Lin, C.-J. 2001. LIBSVM: a library for support vector machines. 16, 59
- Chang, J. T., Schutze, H., and Altman, R. B. 2002. Creating an online dictionary of abbreviations from MEDLINE. *J Am Med Inform Assoc*, 9(6):612–20. 6, 24, 47, 48, 64
- Chang, J. T., Schutze, H., and Altman, R. B. 2004. GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, 20(2):216–25. 14
- Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E. E., and Futschik, M. E. 2007. UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, 35(Database issue):D590–4. 103
- Chaussabel, D. 2004. Biomedical literature mining: challenges and solutions in the 'omics' era. *Am J Pharmacogenomics*, 4(6):383–93. 172
- Chaussabel, D. and Sher, A. 2002. Mining microarray expression data by literature profiling. *Genome Biol*, 3(10):RESEARCH0055. 172, 173
- Chen, L., Liu, H., and Friedman, C. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248–56. 24, 30, 34, 35, 36
- Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M., and Halfon, M. S. 2005. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol*, 6(2):R16. 147

- Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., Hong, E. L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C. L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., and Cherry, J. M. 2004. 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):D311–4. [71](#)
- Chun, H. W., Tsuruoka, Y., Kim, J. D., Shiba, R., Nagata, N., Hishiki, T., and Tsujii, J. 2006. Extraction of gene-disease relations from MedLine using domain dictionaries and machine learning. *Pac Symp Biocomput*, pages 4–15. [89](#)
- Chung, H. J., Kim, M., Park, C. H., Kim, J., and Kim, J. H. 2004. ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res*, 32(Web Server issue):W460–4. [134](#)
- Cleveland, W. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836. [138](#)
- Cohen, A. M., Hersh, W. R., Dubay, C., and Spackman, K. 2005a. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics*, 6:103. [23](#)
- Cohen, K. B., Fox, L., Ogren, P. V., and Hunter, L. 2005b. Empirical data on corpus design and usage in biomedical natural language processing. *AMIA Annu Symp Proc*, pages 156–60. [18](#)
- Colosimo, M. E., Morgan, A. A., Yeh, A. S., Colombe, J. B., and Hirschman, L. 2005. Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, 6 Suppl 1:S12. [83](#), [98](#)
- Comander, J., Natarajan, S., Gimbrone, M. A., J., and Garcia-Cardena, G. 2004. Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics*, 5(1):17. [134](#)
- Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. 2004. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20(3):323–31. [147](#)
- Craven, M. and Kumlien, J. 1999. Constructing biological knowledge bases by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol*, pages 77–86. [89](#)
- Crosby, M. A., Goodman, J. L., Strelets, V. B., Zhang, P., and Gelbart, W. M. 2007. FlyBase: genomes by the dozen. *Nucleic Acids Res*, 35(Database issue):D486–91. [27](#)
- Cui, X. and Churchill, G. A. 2003. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210. [134](#)

- Cui, X., Hwang, J. T., Qiu, J., Blades, N. J., and Churchill, G. A. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6(1):59–75. [134](#)
- Curtis, R. K., Oresic, M., and Vidal-Puig, A. 2005. Pathways to the analysis of microarray data. *Trends Biotechnol.*, 23(8):429–35. [169](#)
- Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., and Conklin, B. R. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet*, 31(1):19–20. [170](#)
- Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. 2004. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–11. [89](#)
- Datta, S. and Datta, S. 2006. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, 7:397. [170](#)
- Davis, C. A., Gerick, F., Hintermair, V., Friedel, C. C., Fundel, K., Küffner, R., and Zimmer, R. 2006. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–63. [xiv, xvi, 133, 151, 155](#)
- de la Cruz, N., Bromberg, S., Pasko, D., Shimoyama, M., Twigger, S., Chen, J., Chen, C. F., Fan, C., Foote, C., Gopinath, G. R., Harris, G., Hughes, A., Ji, Y., Jin, W., Li, D., Mathis, J., Nenasheva, N., Nie, J., Nigam, R., Petri, V., Reilly, D., Wang, W., Wu, W., Zuniga-Meyer, A., Zhao, L., Kwitek, A., Tonellato, P., and Jacob, H. 2005. The Rat Genome Database (RGD): developments towards a phenome database. *Nucleic Acids Res*, 33 Database Issue:D485–91. [6, 27](#)
- de la Grange, P., Dutertre, M., Martin, N., and Auboeuf, D. 2005. FAST DB: a website resource for the study of the expression regulation of human gene products. *Nucleic Acids Res*, 33(13):4276–84. [182](#)
- de Marneffe, M.-C., MacCartney, B., and Manning, C. D. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Conference on Language Resources and Evaluation (LREC) 2006*. [92](#)
- Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5):349–56. [109](#)
- Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese, J. C., Nitzberg, M., Wymore, F., Zachariah, Z. K., Brown, P. O., Sherlock, G., and Ball, C. A. 2007. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res*, 35(Database issue):D766–70. [127](#)

- Deng, M., Sun, F., and Chen, T. 2003. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput*, pages 140–51. 109, 184
- Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. 2002. Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, pages 326–37. 87
- Dingare, S., Nissim, M., Finkel, J., Manning, C., and Grover, C. 2005. A system for identifying named entities in biomedical text: how results from two evaluations reflect on both the system and the evaluations. *Comparative and Functional Genomics*, 6(1-2):77–85. 14
- Dobrokhotov, P. B., Goutte, C., Veuthey, A. L., and Gaussier, E. 2003. Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. *Bioinformatics*, 19 Suppl 1:i91–4. 16
- Domedel-Puig, N. and Wernisch, L. 2005. Applying GIFT, a Gene Interactions Finder in Text, to fly literature. *Bioinformatics*, 21(17):3582–3. 88
- Donaldson, I., Martin, J., de Brujin, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., Pawson, T., and Hogue, C. W. 2003. Pre-BIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4:11. 89
- Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7. 169
- Donner, C. 2003. *Erkennung von Gennamen in wissenschaftlichen Texten - automatische Generierung von Synonymlisten und Präprozessierung der Texte*. Bachelor thesis, LMU München. 21, 42, 54
- Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*, 22(2):101–9. 163
- Drysdale, R. A., Crosby, M. A., Gelbart, W., Campbell, K., Emmert, D., Matthews, B., Russo, S., Schroeder, A., Smutniak, F., Zhang, P., Zhou, P., Zytkovicz, M., Ashburner, M., de Grey, A., Foulger, R., Millburn, G., Sutherland, D., Yamada, C., Kaufman, T., Matthews, K., DeAngelo, A., Cook, R. K., Gilbert, D., Goodman, J., Grumbling, G., Sheth, H., Strelets, V., Rubin, G., Gibson, M., Harris, N., Lewis, S., Misra, S., and Shu, S. Q. 2005. FlyBase: genes and gene models. *Nucleic Acids Res*, 33 Database Issue:D390–5. 6, 7, 27, 64, 71
- Edwards, D. 2003. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, 19(7):825–33. 133, 140

- Egorov, S., Yuryev, A., and Daraselia, N. 2004. A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc*, 11(3):174–8. [56](#)
- Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., and Richardson, J. E. 2007. The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res*, 35(Database issue):D630–7. [27](#)
- Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., Blake, J. A., Anagnostopoulos, A., Baldarelli, R. M., Baya, M., Beal, J. S., Bello, S. M., Boddy, W. J., Bradt, D. W., Burkart, D. L., Butler, N. E., Campbell, J., Cassell, M. A., Corbani, L. E., Cousins, S. L., Dahmen, D. J., Dene, H., Diehl, A. D., Drabkin, H. J., Frazer, K. S., Frost, P., Glass, L. H., Goldsmith, C. W., Grant, P. L., Lennon-Pierce, M., Lewis, J., Lu, I., Maltais, L. J., McAndrews-Hill, M., McClellan, L., Miers, D. B., Miller, L. A., Ni, L., Ormsby, J. E., Qi, D., Reddy, T. B., Reed, D. J., Richards-Smith, B., Shaw, D. R., Sinclair, R., Smith, C. L., Szauter, P., Walker, M. B., Walton, D. O., Washburn, L. L., Witham, I. T., and Zhu, Y. 2005. The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res*, 33 Database Issue:D471–5. [6](#), [7](#), [27](#)
- Eriksson, G., Franzen, K., Olsson, F., Asker, L., and Liden, P. 2002. Exploiting syntax when detecting protein names in text. In *Proceedings of Workshop on Natural Language Processing in Biomedical Applications*. [14](#)
- Ernst, J. and Bar-Joseph, Z. 2006. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7:191. [164](#)
- Ernst, J., Nau, G. J., and Bar-Joseph, Z. 2005. Clustering short time series gene expression data. *Bioinformatics*, 21 Suppl 1:i159–68. [164](#)
- Eyre, T. A., Ducluzeau, F., Sneddon, T. P., Povey, S., Bruford, E. A., and Lush, M. J. 2006. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res*, 34(Database issue):D319–21. [6](#), [27](#), [29](#), [43](#), [72](#), [102](#), [176](#)
- Fan, J., Tam, P., Woude, G. V., and Ren, Y. 2004. Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc Natl Acad Sci U S A*, 101(5):1135–40. [147](#)
- Fisher, R. 1932. *Statistical methods for research workers*. Oliver and Boyd, London, 4th edition edition. [112](#), [142](#), [171](#)
- Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehan, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B., Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault, J.-A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M.-P., Mirouse,

- V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rosse, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J., and Daviet, L. 2005. Protein interaction mapping: A *Drosophila* case study. [108](#)
- Frantzi, K., Ananiadou, S., and Tsujii, J. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *Proceedings of the ECDL*, page 585–604. Springer. [14](#)
- Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P., and Coster, J. 2002. Protein names and how to find them. *Int J Med Inform*, 67(1-3):49–61. [14](#), [18](#)
- Friedel, C. C. 2003. *Automatische Generierung und Verbesserung von Synonymlisten und Entwicklung eines geeigneten Benchmarksystems*. Bachelor thesis, LMU München. [21](#), [42](#), [54](#)
- Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(1):S74–82. [89](#)
- Friedman, N., Linial, M., Nachman, I., and Pe'er, D. 2000. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620. [127](#)
- Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. 1998. Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput*, 707:18. [14](#)
- Fundel, K., Güttler, D., Zimmer, R., and Apostolakis, J. 2005a. A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics*, 6 Suppl 1:S15. [xiii](#), [xv](#), [55](#), [58](#), [65](#), [83](#), [180](#)
- Fundel, K., Küffner, R., Aigner, T., and Zimmer, R. 2005b. Data Processing Effects on the Interpretation of Microarray Gene Expression Experiments. *German Conference on Bioinformatics (GCB) 2005: GI-Edition Lecture Notes in Informatics (LNI)*. [xiv](#), [xvi](#), [133](#), [155](#)
- Fundel, K., Küffner, R., and Zimmer, R. 2007. RelEx—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–71. [xiii](#), [xvi](#), [87](#), [118](#), [180](#)
- Fundel, K. and Zimmer, R. 2006. Gene and protein nomenclature in public databases. *BMC Bioinformatics*, 7:372. [xiii](#), [xv](#), [21](#), [53](#)
- Fundel, K. and Zimmer, R. 2007. Human Gene Normalization by an Integrated Approach including Abbreviation Resolution and Disambiguation. In *Second BioCreAtIvE Challenge Workshop – Critical Assessment of Information Extraction in Molecular Biology*, Madrid, Spain. [xiii](#), [xvi](#), [55](#), [65](#), [84](#), [180](#)
- Futschik, M. and Crompton, T. 2004. Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol*, 5(8):R60. [133](#)

- Galperin, M. Y. 2007. The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Res*, 35(Database issue):D3–4. 5
- Gandhi, T. K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. 2006. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–93. 105
- Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. 2005. Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21(18):3658–3664. 48
- Gebauer, M., Saas, J., Haag, J., Dietz, U., Takigawa, M., Bartnik, E., and Aigner, T. 2005. Repression of anti-proliferative factor Tob1 in osteoarthritic cartilage. *Arthritis Res Ther*, 7(2):R274–84. 163
- Gilbert, D. 2005. Biomolecular interaction network database. *Brief Bioinform*, 6(2):194–8. 6
- Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., J., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. 2003. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–36. 108
- Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M., and Orengo, C. A. 2007. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res*, 35(Database issue):D291–7. 6
- Gruber, T. 1993. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human Computer Studies*, 43(5-6):907–928. 6
- Güttler, D. 2006. *LiMB - Ein interaktiver Browser für die Analyse biomedizinischer Texte mit Hilfe von Textmining*. Diploma, LMU München. 21, 49, 54
- Hakenberg, J., Plake, C., Leser, U., Kirsch, H., and Rebholz-Schuhmann, D. 2005. LLL'05 Challenge: Genic Interaction Extraction - Identification of Language Patterns Based on Alignment and Finite State Automata. In *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*. 14, 89, 97, 99

- Hakenberg, J., Schmeier, S., Kowald, A., Klipp, E., and Leser, U. 2004. Finding kinetic parameters using text mining. *Omics*, 8(2):131–52. [184](#)
- Hall, J. A., Roter, D. L., Milburn, M. A., and Daltroy, L. H. 1996. Patients' health as a predictor of physician and patient behavior in medical visits. A synthesis of four studies. *Med Care*, 34(12):1205–18. [146](#)
- Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7. [6](#)
- Hanisch, D., Fluck, J., Mevissen, H. T., and Zimmer, R. 2003. Playing biology's name game: identifying protein names in scientific text. In *Pac Symp Biocomput*, volume 8, pages 403–14. [xiii](#), [xv](#), [29](#), [40](#), [44](#), [55](#), [56](#), [60](#)
- Hanisch, D., Fundel, K., Mevissen, H. T., Zimmer, R., and Fluck, J. 2005. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S14. [xiii](#), [xv](#), [40](#), [44](#), [55](#), [65](#), [80](#), [83](#), [180](#)
- Hanisch, D., Sohler, F., and Zimmer, R. 2004. ToPNet—an application for interactive analysis of expression data and biological networks. *Bioinformatics*, 20(9):1470–1. [109](#), [111](#), [117](#), [118](#), [171](#)
- Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. 2002. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 Suppl 1:S145–54. [3](#), [117](#), [118](#), [134](#), [170](#)
- Hao, Y., Zhu, X., Huang, M., and Li, M. 2005. Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*, 21(15):3294–300. [89](#)
- Hart, G. T., Ramani, A. K., and Marcotte, E. M. 2006. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120. [103](#)
- Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, 1(2):RESEARCH0003. [175](#)
- Hatzivassiloglou, V., Duboue, P. A., and Rzhetsky, A. 2001. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, 17 Suppl 1:S97–106. [19](#), [56](#)
- Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe,

- B., Hogue, C., and Apweiler, R. 2004a. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–83. [183](#)
- Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. 2004b. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):D452–5. [105](#)
- Herrero, J., Vaquerizas, J. M., Al-Shahrour, F., Conde, L., Mateos, A., Diaz-Uriarte, J. S., and Dopazo, J. 2004. New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res*, 32(Web Server issue):W485–91. [134](#)
- Herrgard, M. J., Lee, B. S., Portnoy, V., and Palsson, B. O. 2006. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res*, 16(5):627–35. [3](#), [184](#)
- Hirschman, L. 1998. The Evolution of evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12(4):281–305. [13](#)
- Hirschman, L., Colosimo, M., Morgan, A., and Yeh, A. 2005a. Overview of BioCreAtIVe task 1B: normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1:S11. [19](#), [24](#), [56](#), [100](#)
- Hirschman, L., Morgan, A. A., and Yeh, A. S. 2002a. Rutabaga by any other name: extracting biological names. *Journal of Biomedical Informatics*, 35(4):247–259. [17](#), [24](#), [30](#), [56](#)
- Hirschman, L., Park, J. C., Tsujii, J., Wong, L., and Wu, C. H. 2002b. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–61. [13](#)
- Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. 2005b. Overview of BioCreAtIVe: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1. [18](#), [56](#)
- Hoffmann, R. and Valencia, A. 2003a. Life cycles of successful genes. *Trends Genet*, 19(2):79–81. [17](#)
- Hoffmann, R. and Valencia, A. 2003b. Protein interaction: same network, different hubs. *Trends Genet*, 19(12):681–3. [17](#)
- Hofmann, O. and Schomburg, D. 2005. Concept-based annotation of enzyme classes. *Bioinformatics*, 21(9):2059–66. [15](#)
- Hoglund, A., Blum, T., Brady, S., Donnes, P., Miguel, J. S., Rocheford, M., Kohlbacher, O., and Shatkay, H. 2006. Significantly improved prediction of subcellular localization by integrating text and protein sequence data. *Pac Symp Biocomput*, pages 16–27. [15](#)

- Hu, Z. Z., Mani, I., Hermoso, V., Liu, H., and Wu, C. H. 2004. iProLINK: an integrated protein resource for literature mining. *Comput Biol Chem*, 28(5-6):409–16. 90
- Hu, Z. Z., Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K., and Wu, C. H. 2005. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, 21(11):2759–65. 88, 99, 117
- Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K., and Li, M. 2004. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–12. 89
- Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104. 138
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novere, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–31. 183
- Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. 2006. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22(23):2890–7. 127
- Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–40. 184
- Jarvinen, A. K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O. P., and Monni, O. 2004. Are data from different gene expression microarray platforms comparable? *Genomics*, 83(6):1164–8. 163
- Jelier, R., Jenster, G., Dorssers, L. C., van der Eijk, C. C., van Mulligen, E. M., Mons, B., and Kors, J. A. 2005. Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics*, 21(9):2049–58. 87
- Jensen, L. J., Saric, J., and Bork, P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7(2):119–29. 15
- Jenssen, T. K., Laegreid, A., Komorowski, J., and Hovig, E. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1):21–8. 87, 88, 172

- Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database issue):D428–32. [90](#)
- Joyce, A. R. and Palsson, B. O. 2006. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210. [4](#)
- Kanehisa, M. and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30. [6](#), [170](#)
- Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. 2002. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30(1):42–6. [6](#), [170](#)
- Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res*, 28(1):56–9. [6](#)
- Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. 2002. Tuning Support Vector Machines for Biomedical Named Entity Recognition. In *Workshop on Natural Language Processing in the Biomedical Domain (at ACL'2002)*, pages 1–8. [14](#)
- Küffner, R., Duchrow, T., Fundel, K., and Zimmer, R. 2006. Characterization of Protein Interactions. *German Conference on Bioinformatics (GCB) 2006: GI-Edition Lecture Notes in Informatics (LNI)*. [xiv](#), [xvi](#), [87](#), [114](#), [118](#)
- Küffner, R., Fundel, K., and Zimmer, R. 2005. Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics*, 21 Suppl 2:ii259–ii267. [xiv](#), [xvi](#), [102](#), [173](#), [181](#)
- Khatri, P., Desai, V., Tarca, A. L., Sellamuthu, S., Wildman, D. E., Romero, R., and Draghici, S. 2006. New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate. *Nucleic Acids Res*, 34(Web Server issue):W626–31. [169](#)
- Khatri, P., Draghici, S., Ostermeier, G. C., and Krawetz, S. A. 2002. Profiling gene expression using onto-express. *Genomics*, 79(2):266–70. [169](#)
- Kim, J. D., Ohta, T., Tateisi, Y., and Tsujii, J. 2003. GENIA corpus—semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–2. [18](#), [47](#), [100](#), [172](#)
- Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. In *International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*. [14](#), [18](#), [47](#)
- Kim, W., Aronson, A., and Wilbur, W. 2001. Automatic MeSH term assignment and quality assessment. *Proc AMIA Symp*, 319:23. [174](#)

- Klein, D. and Manning, C. D. 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems*, 15(NIPS 2002). 90, 92
- Klein, D. and Manning, C. D. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*. 90, 92
- Knudsen, S., Workman, C., Sicheritz-Ponten, T., and Friis, C. 2003. GenePublisher: Automated analysis of DNA microarray data. *Nucleic Acids Res*, 31(13):3471–6. 134
- Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., and Philippi, S. 2006. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–90. 172
- Koike, A. and Takagi, T. 2004. Gene/Protein/Family Name Recognition in Biomedical Literature. In *Proceedings of BioLink 2004 Workshop: Linking Biological Literature, Ontologies and Databases: Tools for Users*. 42, 56
- Konig, R., Baldessari, D., Pollet, N., Niehrs, C., and Eils, R. 2004. Reliability of gene expression ratios for cDNA microarrays in multiconditional experiments with a reference design. *Nucleic Acids Res*, 32(3):e29. 147
- Korotkiy, M., Middelburg, R., Dekker, H., van Harmelen, F., and Lankelma, J. 2004. A tool for gene expression based PubMed search through combining data sources. *Bioinformatics*, 20(12):1980–2. 172
- Krauthammer, M., Kra, P., Iossifov, I., Gomez, S. M., Hripcsak, G., Hatzivassiloglou, V., Friedman, C., and Rzhetsky, A. 2002. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics*, 18 Suppl 1:S249–57. 16
- Krauthammer, M. and Nenadic, G. 2004. Term identification in the biomedical literature. *J Biomed Inform*, 37(6):512–26. 15
- Krauthammer, M., Rzhetsky, A., Morozov, P., and Friedman, C. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1-2):245–52. 19, 56
- Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kronenberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O., and Wingender, E. 2006. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res*, 34(Database issue):D546–51. 6, 170
- Leek, T. 1997. *Information Extraction Using Hidden Markov Models*. Masters thesis, University of California San Diego. 89
- Leroy, G. and Chen, H. 2002. Filling preposition-based templates to capture information from medical abstracts. *Pac Symp Biocomput*, pages 350–61. 88

- Leroy, G., Chen, H., and Martinez, J. D. 2003. A shallow parser based on closed-class words to capture relations in biomedical text. *J Biomed Inform*, 36(3):145–58. 88
- Li, S., Wu, L., and Zhang, Z. 2006. Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinformatics*, 22(17):2143–50. 172
- Ling, X., Jiang, J., He, X., Mei, Q., Zhai, C., and Schatz, B. 2006. Automatically generating gene summaries from biomedical literature. *Pac Symp Biocomput*, pages 40–51. 15
- Liu, F., Jenssen, T. K., Nygaard, V., Sack, J., and Hovig, E. 2004. FigSearch: a figure legend indexing and classification system. *Bioinformatics*, 20(16):2880–2. 183
- Liu, H., Aronson, A. R., and Friedman, C. 2002a. A study of abbreviations in MEDLINE abstracts. *Proc AMIA Symp*, pages 464–8. 24, 47
- Liu, H., Johnson, S. B., and Friedman, C. 2002b. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assoc*, 9(6):621–36. 56
- Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. 2003a. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83. 7
- Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. 2003b. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, pages 601–12. 7
- Lu, Y., Rosenfeld, R., and Bar-Joseph, Z. 2006. Identifying cycling genes by combining sequence homology and expression data. *Bioinformatics*, 22(14):e314–22. 185
- Luciano, J. S. 2005. PAX of mind for pathway researchers. *Drug Discov Today*, 10(13):937–42. 184
- Ma, X., Lee, H., Wang, L., and Sun, F. 2007. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, 23(2):215–21. 184
- MacCallum, R. M., Kelley, L. A., and Sternberg, M. J. 2000. SAWTED: structure assignment with text description-enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics*, 16(2):125–9. 16
- Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 33 Database Issue:D54–8. 6, 23, 27, 72, 102
- Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 35(Database issue):D26–31. 27, 176

- Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. 2005. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 21(19):3787–93. [169](#)
- Masys, D. R. 2001. Linking microarray data to the literature. *Nat Genet*, 28(1):9–10. [16](#), [172](#)
- Masys, D. R., Welsh, J. B., Lynn Fink, J., Gribskov, M., Klacansky, I., and Corbeil, J. 2001. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319–26. [16](#), [172](#), [173](#)
- Mathivanan, S., Periaswamy, B., Gandhi, T., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y., and Pandey, A. 2006. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7 Suppl 5:S19. [105](#)
- McDonald, D. M., Chen, H., Su, H., and Marshall, B. B. 2004. Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics*, 20(18):3370–8. [89](#)
- McKenna, L. A., Gehrsitz, A., Soder, S., Eger, W., Kirchner, T., and Aigner, T. 2000. Effective isolation of high-quality total RNA from human adult articular cartilage. *Anal Biochem*, 286(1):80–5. [135](#)
- Mel'cuk, I. 1988. *Dependency Syntax: Theory and Practice*. State University Press of New York, New York. [90](#)
- Michiels, S., Koscielny, S., and Hill, C. 2005. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–92. [136](#)
- Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakant, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H. G., Nagini, M., Kumar, G. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S., and Pandey, A. 2006. Human protein reference database—2006 update. *Nucleic Acids Res*, 34(Database issue):D411–4. [90](#), [97](#), [114](#)
- Morgan, A., Wellner, B., Colombe, J., Arens, R., Colosimo, M., and Hirschman, L. 2007. Evaluating the Automatic Mapping of Human Gene and Protein Mentions to Unique Identifiers. In *Pac Symp Biocomput*, volume 12, pages 281–291. [73](#)
- Morgan, A. A., Hirschman, L., Colosimo, M., Yeh, A. S., and Colombe, J. B. 2004. Gene name identification and normalization using a model organism database. *J Biomed Inform*, 37(6):396–410. [56](#)

- Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J. D., Sigrist, C. J., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. 2007. New developments in the InterPro database. *Nucleic Acids Res*, 35(Database issue):D224–8. [42](#), [46](#)
- Narayanaswamy, M., Ravikumar, K. E., and Vijay-Shanker, K. 2005. Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics*, 21 Suppl 1:i319–27. [88](#)
- Nash, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hong, E. L., Livstone, M. S., Oughtred, R., Park, J., Skrzypek, M., Theesfeld, C. L., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Dolinski, K., Botstein, D., and Cherry, J. M. 2007. Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res*, 35(Database issue):D468–71. [27](#)
- Naur, P. 1960. Revised Report on the Algorithmic Language ALGOL 60. *Communications of the ACM*, 3(5):299–314. [25](#)
- Nédellec, C. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge. In *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*. [18](#), [97](#), [99](#), [117](#)
- Nenadic, G., Mima, H., Spasic, I., Ananiadou, S., and Tsujii, J. 2002. Terminology-driven literature mining and knowledge acquisition in biomedicine. *Int J Med Inform*, 67(1-3):33–48. [47](#)
- Ngai, G. and Florian, R. 2001. Transformation-Based Learning in the Fast Lane. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 NAACL '01*, pages 40–47. [68](#), [92](#)
- Nobata, C., Collier, N., and Tsujii, J. 1999. Automatic term identification and classification in biology texts. *Proc. of the 5th NLPRS*, pages 369–374. [14](#)
- Novichkova, S., Egorov, S., and Daraselia, N. 2003. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, 19(13):1699–706. [89](#)
- Okazaki, N. and Ananiadou, S. 2006. Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, 22(24):3089–95. [48](#)

- Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. 2001. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–61. 56, 88, 99, 117
- Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H. W., Ruepp, A., and Frishman, D. 2005. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–4. 105
- Pandey, R., Guru, R. K., and Mount, D. W. 2004. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, 20(13):2156–8. 134, 170
- Park, J., Kim, H., and Kim, J. 2001. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Pac Symp Biocomput*, 6:396–407. 89
- Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S., and Simon, R. 2003. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4(1):33. 134
- Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. 2007. ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 35(Database issue):D747–50. 127
- Pe'er, D., Regev, A., Elidan, G., and Friedman, N. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17 Suppl 1:S215–24. 127
- Pehkonen, P., Wong, G., and Toronen, P. 2005. Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, 6:162. 16
- Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T. K., Chandrika, K. N., Deshpande, N., Suresh, S., Rashmi, B. P., Shanker, K., Padma, N., Niranjan, V., Harsha, H. C., Talreja, N., Vrushabendra, B. M., Ramya, M. A., Yatish, A. J., Joy, M., Shivashankar, H. N., Kavitha, M. P., Menezes, M., Choudhury, D. R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C. K., Prasad, C. K., Kumar-Sinha, C., Deshpande, K. S., and Pandey, A. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32(Database issue):D497–501. 90, 97, 101, 103, 105, 106
- Pillet, V., Zehnder, M., Seewald, A. K., Veuthey, A. L., and Petrak, J. 2005. GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics*, 21(8):1743–4. 6
- Porter, M. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137. 64, 67

- Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H. 2001. The HUGO Gene Nomenclature Committee (HGNC). *Hum Genet*, 109(6):678–80. 6, 29
- Proux, D., Rechenmann, F., Julliard, L., Pillet, V. V., and Jacq, B. 1998. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Inform Ser Workshop Genome Inform*, 9:72–80. 14
- Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., and Morrell, M. 2001. Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo*, 10(Pt 1):371–5. 18, 48
- Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. 2005. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, 6(5):R40. 90
- Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–64. 6
- Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O. G., Ideker, T., Dolinski, K., Batada, N. N., and Tyers, M. 2006. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol*, 5(4):11. 108
- Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. 2005. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 23(8):951–9. 103
- Riedel, S. and Klein, E. 2005. Genic Interaction Extraction with Semantic and Syntactic Chains. In *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05), Bonn, Germany, 2005*. 99
- Rimer, M. and O'Connell, M. 1998. BioABACUS: a database of abbreviations and acronyms in biotechnology and computer science. *Bioinformatics*, 14(10):888–9. 48
- Rosenthal, R. 1984. *Meta-analytic procedures for social sciences*. Beverly Hills, CA: Sage Publications. 143
- Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhoute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8. 101, 105, 106, 108

- Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–51. [6](#), [105](#), [114](#)
- Sandberg, R. and Larsson, O. 2007. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinformatics*, 8:48. [147](#)
- Sandler, T., Schein, A. I., and Ungar, L. H. 2006. Automatic term list generation for entity tagging. *Bioinformatics*, 22(6):651–7. [23](#)
- Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P. 2005. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*. [99](#), [117](#)
- Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P. 2006. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, 22(6):645–50. [88](#)
- Sarkar, I. N. and Rindflesch, T. C. 2002. Discovering protein similarity using natural language processing. *Proc AMIA Symp*, pages 677–81. [16](#)
- Schacherer, F., Choi, C., Gotze, U., Krull, M., Pistor, S., and Wingender, E. 2001. The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, 17(11):1053–7. [6](#)
- Schuemie, M. J., Weeber, M., Schijvenaars, B. J., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B., and Kors, J. A. 2004. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–604. [17](#)
- Schwartz, A. S. and Hearst, M. A. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput*, pages 451–62. [24](#), [47](#)
- Segal, E., Friedman, N., Koller, D., and Regev, A. 2004. A module map showing conditional activity of expression modules in cancer. *Nat Genet*, 36(10):1090–8. [127](#)
- Segal, E., Wang, H., and Koller, D. 2003a. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19 Suppl 1:i264–71. [127](#), [184](#)
- Segal, E., Yelensky, R., and Koller, D. 2003b. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19 Suppl 1:i273–82. [127](#)
- Sehgal, A. K. and Srinivasan, P. 2006. Retrieval with gene queries. *BMC Bioinformatics*, 7:220. [16](#)
- Seki, K. and Mostafa, J. 2003. A probabilistic model for identifying protein names and their name boundaries. *Proc IEEE Comput Soc Bioinform Conf*, 2:251–8. [14](#)
- Settles, B. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–2. [14](#)

- Shah, P., Perez-Iratxeta, C., Bork, P., and Andrade, M. 2003. Information extraction from full text scientific articles: where are the keywords. *BMC Bioinformatics*, 4(1):20. 17
- Shah, P. K. and Bork, P. 2006. LSAT: learning about alternative transcripts in MEDLINE. *Bioinformatics*, 22(7):857–65. 183
- Shah, P. K., Jensen, L. J., Boue, S., and Bork, P. 2005. Extraction of transcript diversity from scientific literature. *PLoS Comput Biol*, 1(1):e10. 183
- Shatkay, H., Chen, N., and Blostein, D. 2006. Integrating image data into biomedical text categorization. *Bioinformatics*, 22(14):e446–53. 183
- Shatkay, H., Edwards, S., Wilbur, W. J., and Boguski, M. 2000. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc Int Conf Intell Syst Mol Biol*, 8:317–28. 174
- Shatkay, H. and Feldman, R. 2003. Mining the biomedical literature in the genomic era: an overview. *J Comput Biol*, 10(6):821–55. 15
- Shi, L. and Campagne, F. 2005. Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics*, 6:88. 23
- Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. M., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Slikker, W., J., Shi, L., and Reid, L. H. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, 24(9):1151–61. 163
- Shi, L., Tong, W., Fang, H., Scherf, U., Han, J., Puri, R. K., Frueh, F. W., Goodsaid, F. M., Guo, L., Su, Z., Han, T., Fuscoe, J. C., Xu, Z. A., Patterson, T. A., Hong, H., Xie, Q., Perkins, R. G., Chen, J. J., and Casciano, D. A. 2005. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, 6 Suppl 2:S12. 163
- Smalheiser, N. R. and Swanson, D. R. 1998. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*, 57(3):149–53. 16
- Smith, L., Rindflesch, T., and Wilbur, W. J. 2004. MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, 20(14):2320–1. 68, 91
- Smyth, G. K., Michaud, J., and Scott, H. S. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–75. 147

- Smyth, G. K. and Speed, T. 2003. Normalization of cDNA microarray data. *Methods*, 31(4):265–73. 140
- Sohler, F., Hanisch, D., and Zimmer, R. 2004. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–21. 3, 117, 118, 170, 171
- Sohler, F. and Zimmer, R. 2005. Identifying active transcription factors and kinases from expression data using pathway queries. *Bioinformatics*, 21(Suppl 2):ii115–ii122. 3, 171
- Srinivasan, P. and Libbus, B. 2004. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20 Suppl 1:I290–I296. 16
- Stalteri, M. A. and Harrison, A. P. 2007. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics*, 8:13. 147
- Standish, L. J., Kozak, L., Johnson, L. C., and Richards, T. 2004. Electroencephalographic evidence of correlated event-related signals between the brains of spatially and sensory isolated human subjects. *J Altern Complement Med*, 10(2):307–14. 146
- Stapley, B. J., Kelley, L. A., and Sternberg, M. J. 2002. Predicting the sub-cellular location of proteins from text using support vector machines. *Pac Symp Biocomput*, pages 374–85. 15
- Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobisch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68. 108
- Storey, J. D. and Tibshirani, R. 2003. Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol Biol*, 224:149–57. 147
- Storm, L. and Ertel, S. 2001. Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychol Bull*, 127(3):424–33; discussion 434–8. 146
- Stromback, L. and Lambrix, P. 2005. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21(24):4401–7. 184
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50. 170
- Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., and Ideker, T. 2006. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360. 109

- Swanson, D. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, 30(1):7–18. 16, 89
- Swanson, D. 1988. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med*, 31(4):526–57. 89
- Swanson, D. 1990. Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspect Biol Med*, 33(2):157–86. 89
- Szugat, M., Gütter, D., Fundel, K., Sohler, F., and Zimmer, R. 2005. Web servicing the biological office. *Bioinformatics*, 21 Suppl 2:ii268–ii269. xiii, xv, 21, 49, 50, 51, 54
- Takeuchi, K. and Collier, N. 2003. Bio-Medical Entity Extraction using Support Vector Machines. *Proc ACL 2003 Workshop on NLP in Biomedicine*. 14
- Tamames, J. and Valencia, A. 2006. The success (or not) of HUGO nomenclature. *Genome Biol*, 7(5):402. 22
- Tan, K., Shlomi, T., Feizi, H., Ideker, T., and Sharan, R. 2007. Transcriptional regulation of protein complexes within and across species. *Proc Natl Acad Sci U S A*, 104(4):1283–8. 3
- Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L., and Weinstein, J. N. 1999. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27(6):1210–4, 1216–7. 171
- Tanabe, L. and Wilbur, W. J. 2002. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–32. 14, 56
- Tanabe, L., Xie, N., Thom, L. H., Matten, W., and Wilbur, W. J. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S3. 18
- Temkin, J. M. and Gilder, M. R. 2003. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19(16):2046–53. 89
- Tesnière, L. 1953. *Esquisse d'une syntaxe structurale (Sketch of a Structural Syntax)*. Klincksieck, Paris. 91
- Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. 2000. Automatic extraction of protein interactions from scientific abstracts. *Pac Symp Biocomput*, pages 541–52. 88
- Tornow, S. and Mewes, H. W. 2003. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*, 31(21):6283–9. 184

- Tsai, R. T., Wu, S. H., Chou, W. C., Lin, Y. C., He, D., Hsiang, J., Sung, T. Y., and Hsu, W. L. 2006. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7:92. 14
- Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*, 29(12):2549–57. 147
- Tsuruoka, Y. and Tsujii, J. 2004. Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform*, 37(6):461–70. 56
- Tuason, O., Chen, L., Liu, H., Blake, J. A., and Friedman, C. 2004. Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pac Symp Biocomput*, pages 238–49. 24, 30, 34, 36
- Tusher, V. G., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21. 134
- Twigger, S. N., Shimoyama, M., Bromberg, S., Kwitek, A. E., and Jacob, H. J. 2007. The Rat Genome Database, update 2007—Easing the path from disease to data and back again. *Nucleic Acids Res*, 35(Database issue):D658–62. 27
- Uetz, P., Dong, Y. A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S. V., Roupelieva, M., Rose, D., Fossum, E., and Haas, J. 2006. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239–42. 103
- Weeber, M., Klein, H., Berg, L., and Vos, R. 2001. Using Concepts in Literature-Based Discovery: Simulating Swanson’s Raynaud-Fish Oil and Migraine-Magnesium Discoveries. *Journal of the american society for information science and technology*, 52(7):548–557. 16
- Weeber, M., Schijvenaars, B. J., Van Mulligen, E. M., Mons, B., Jelier, R., Van Der Eijk, C. C., and Kors, J. A. 2003. Ambiguity of human gene symbols in LocusLink and MEDLINE: creating an inventory and a disambiguation test collection. *AMIA Annu Symp Proc*, pages 704–8. 24, 30
- Whitlock, M. C. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach. *J Evol Biol*, 18(5):1368–73. 146
- Wilbur, W. J., Rzhetsky, A., and Shatkay, H. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7(1):356. 18, 98
- Wilson, D. L., Buckley, M. J., Helliwell, C. A., and Wilson, I. W. 2003. New normalization methods for cDNA microarray data. *Bioinformatics*, 19(11):1325–32. 134

- Woo, Y., Affourtit, J., Daigle, S., Viale, A., Johnson, K., Naggert, J., and Churchill, G. 2004. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J Biomol Tech*, 15(4):276–84. 163
- Wren, J. D., Chang, J. T., Pustejovsky, J., Adar, E., Garner, H. R., and Altman, R. B. 2005a. Biomedical term mapping databases. *Nucleic Acids Res*, 33 Database Issue:D289–93. 24
- Wren, J. D. and Garner, H. R. 2002. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods Inf Med*, 41(5):426–34. 6, 24
- Wren, J. D., Hildebrand, W. H., Chandrasekaran, S., and Melcher, U. 2005b. Markov model recognition and classification of DNA/protein sequences within large text databases. *Bioinformatics*, 21(21):4046–53. 14, 16
- Wrobel, G., Chalmel, F., and Primig, M. 2005. goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics*, 21(17):3575–7. 169
- Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J., Thompson, M. J., Marcotte, E. M., and Eisenberg, D. 2001. DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res*, 29(1):239–41. 6
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–5. 105
- Xia, K., Dong, D., and Han, J. D. 2006. IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, 7:508. 103
- Yan, X., Deng, M., Fung, W. K., and Qian, M. 2005. Detecting differentially expressed genes by relative entropy. *J Theor Biol*, 234(3):395–402. 134
- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15. 140
- Yauk, C. L. and Berndt, M. L. 2007. Review of the literature examining the correlation among DNA microarray technologies. *Environ Mol Mutagen*. 163
- Yauk, C. L., Berndt, M. L., Williams, A., and Douglas, G. R. 2004. Comprehensive comparison of six microarray technologies. *Nucleic Acids Res*, 32(15):e124. 163
- Yeh, A., Morgan, A., Colosimo, M., and Hirschman, L. 2005. BioCreAtIVe task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6 Suppl 1:S2. 14, 97

- Yoshida, M., Fukuda, K., and Takagi, T. 2000. PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, 16(2):169–175. 47
- Yu, H., Hripcsak, G., and Friedman, C. 2002. Mapping abbreviations to full forms in biomedical articles. *J Am Med Inform Assoc*, 9(3):262–72. 23, 24, 47
- Yu, Y. J., Chen, C., Yu, Y. M., and Lin, H. Z. 2003. Web multimedia information retrieval using improved Bayesian algorithm. *J Zhejiang Univ Sci*, 4(4):415–20. 23, 47
- Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. 2002. MINT: a Molecular INTeraction database. *FEBS Lett*, 513(1):135–40. 6, 105
- Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. 2002. Truncated product method for combining P-values. *Genet Epidemiol*, 22(2):170–85. 146
- Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28. 169
- Zhao, Y., Li, M. C., and Simon, R. 2005. An adaptive method for cDNA microarray normalization. *BMC Bioinformatics*, 6(1):28. 134
- Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–90. 14
- Zhou, W., Torvik, V. I., and Smalheiser, N. R. 2006. ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, 22(22):2813–8. 48
- Zhou, X. J., Kao, M. C., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O. M., Finch, C. E., Morgan, T. E., and Wong, W. H. 2005. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23(2):238–43. 127
- Zien, A., Aigner, T., Zimmer, R., and Lengauer, T. 2001. Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, 17 Suppl 1:S323–31. 137

Acknowledgments

I am very grateful to all people who enabled this thesis and helped and encouraged me while I was undertaking the work described here.

First of all I would like to thank my advisor Ralf Zimmer for giving me the opportunity to write my PhD thesis in the exciting field of bioinformatics, for inspiring me to work on challenging topics in text mining and gene expression data analysis and for supporting my work that lead to this thesis.

I owe a debt of gratitude to the members of the bioinformatics group at the LMU where I could learn a lot. Especially, I would like to thank the people who directly contributed to parts of this work (in more or less chronological order): Daniel Ziemek and Theo Mevissen for their support with respect to ProMiner, Caroline Friedel and Cornelia Donner for writing their bachelor theses on hierarchical synonym dictionaries, benchmarks, and relation extraction, Joannis Apostolakis and Daniel Güttler for the joint work that lead to the exact matching approach for named entity identification, Robert Küffner for the joint text-mining projects, Martin Szugat for setting up the ProThesaurus web service and Wiki, Thomas Aigner for the information on osteoarthritis and the respective data sets as well as for the many discussions and ideas, and last but not least Frank Steiner for keeping the computers running.

Thanks to Peter Zbinden for giving me, after my studies of biotechnology, the opportunity to work in software engineering and computational data analysis; this experience motivated me to go further in this direction and is very valuable to me.

I am grateful to Oliver Kohlbacher for reviewing this thesis, and to Hans-Peter Kriegel and Volker Heun for being part of my dissertation committee.

Special thanks go to my parents for always being there for me and supporting me, my brother Rainer and sister Sibylle for all kinds of encouragement. Sincere thanks to Gerald for everything. Thanks to my friends for keeping in touch with me during these busy years.

Curriculum vitae

Personal Information

Name Katrin Fundel
Birth 28.05.1974 in Friedrichshafen, Germany
Permanent Address Königsbergerstr. 10
D-88045 Friedrichshafen

Education and Professional Experience

- | | |
|-----------|---|
| 2002–2007 | Scientific assistant and PhD-student
Ludwig-Maximilians-Universität München
Teaching and Research Unit for Practical Informatics and Bioinformatics, chair: Prof. Ralf Zimmer |
| 1999–2002 | Software Engineer
Discovery Partners International AG, Allschwil, Switzerland |
| 1995–1998 | Ecole Supérieure de Biotechnologie de Strasbourg
Universities of Strasbourg, Basel, Freiburg and Karlsruhe
Major: Biotechnology |
| 1993–1995 | University of Konstanz
Major: Physics |
| 1984–1993 | Secondary School
Graf-Zeppelin-Gymnasium Friedrichshafen |

Publications

2007

- Human Gene Normalization by an Integrated Approach including Abbreviation Resolution and Disambiguation.
Second BioCreative Challenge Evaluation Workshop, Madrid, Spain, 2007.
Katrin Fundel, Ralf Zimmer
- RelEx - Relation extraction using dependency parse trees
Bioinformatics. 2007 Feb 1; 23(3):365-371. Epub 2006 Dec 1.
Katrin Fundel, Robert Küffner, Ralf Zimmer

2006

- Large-Scale Gene Expression Profiling Reveals Major Pathogenetic Pathways of Osteoarthritis Cartilage Degeneration
Arthritis and Rheumatism, 2006 Nov;54(11):3533-44.
Thomas Aigner, **Katrin Fundel**, Joachim Saas, Pia M. Gebhard, Jochen Haag, Tilo Weiss, Alexander Zien, Franz Obermayr, Ralf Zimmer, Eckart Bartnik
- Gene and Protein Nomenclature in public databases
BMC Bioinformatics 2006 Aug 9;7:372
Katrin Fundel and Ralf Zimmer
- Reliable gene signatures for microarray classification: assessment of stability and performance
Bioinformatics, 2006 Oct 1;22(19):2356-63. Epub 2006 Jul 31.
Chad A. Davis, Fabian Gerick, Volker Hintermair, Caroline C. Friedel, **Katrin Fundel**, Robert Küffner, Ralf Zimmer
- Characterization of Protein Interactions
D. Huson, O. Kohlbacher, A. Lupas, K. Nieselt, A. Zell (eds.): German Conference on Bioinformatics (GCB) 2006: GI-Edition Lecture Notes in Informatics (LNI); P-83: 64-73
Robert Küffner, Timo Duchrow, **Katrin Fundel**, and Ralf Zimmer

2005

- ProMiner: rule-based protein and gene entity recognition
BMC Bioinformatics 2005, 6(Suppl 1):S14 (24 May 2005)
Daniel Hanisch, **Katrin Fundel**, Heinz-Theodor Mevissen, Ralf Zimmer, Juliane Fluck
- A simple approach for protein name identification: prospects and limits
BMC Bioinformatics 2005, 6(Suppl 1):S15 (24 May 2005)
Katrin Fundel, Daniel Güttler, Ralf Zimmer, Joannis Apostolakis
- Data Processing Effects on the Interpretation of Microarray Gene Expression Experiments
A. Torda, S. Kurtz, Matthias Rarey (eds.): *German Conference on Bioinformatics (GCB) 2005. GI Lecture Notes in Informatics; P-71*: 77-91
Katrin Fundel, Robert Küffner, Thomas Aigner, Ralf Zimmer
- Web Servicing the Biological Office
Bioinformatics. 2005; 21(Suppl. 2):ii268-ii269.
Martin Szugat, Daniel Güttler, **Katrin Fundel**, Florian Sohler, Ralf Zimmer
- Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts.
Bioinformatics. 2005; 21(Suppl.2):ii259-ii267
Robert Küffner, **Katrin Fundel**, Ralf Zimmer

2004

- ProMiner: Organism-specific protein name detection using approximate string matching.
BioCreative Challenge Evaluation Workshop, Granada, Spain, 2004.
Daniel Hanisch, **Katrin Fundel**, Heinz-Theodor Mevissen, Ralf Zimmer, Juliane Fluck
- Exact versus approximate string matching for protein name identification.
BioCreative Challenge Evaluation Workshop, Granada, Spain, 2004.
Katrin Fundel, Daniel Güttler, Ralf Zimmer, Joannis Apostolakis