Text mining applied to molecular biology

Rob Jelier

The financial contributions of the WikiProfessional Initiative, the BAZIS foundation and SUWO, the Urological Research Foundation, for the publication of this thesis is gratefully acknowledged.

Jelier R.
Text-mining applied to molecular biology.
PhD Thesis Erasmus University Rotterdam — with summary in Dutch.
Cover design by the author, inspired by the work of Tord Boontje and Piet Mondriaan.

ISBN: 978-90-8559-335-5

This thesis was typeset by the author with LaTeX 2_ε.

Text mining applied to molecular biology

Text mining toegepast voor de moleculaire biologie

Proefschrift

ter verkrijging van de graad van doctor

aan de Erasmus Universiteit Rotterdam

op gezag van de rector magnificus

Prof. dr. S.W.J. Lamberts

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 10 januari 2008 om 16.00 uur

door

Rob Jelier

geboren te Dirksland

Promotiecommissie

Promotor

Prof. dr. J. van der Lei

Copromotoren

Dr. ir. J.A. Kors
Dr. ir. G.W. Jenster

Overige leden

Prof. dr. C.M. van Duijn
Prof. dr. G.J.B. van Ommen
Prof. dr. P.J. van der Spek

The studies described in this thesis were performed when the author was a member of the Biosemantics group, department of Medical Informatics, Erasmus MC, the Netherlands.

Acknowledgments

Over the past few years I've frequently been asked, in one way or another: "Why do you forgo the spoils of capitalism for a mostly solitary life behind a computer?". Well, because science is fun. I feel it's a privilege to have the freedom to not constrain my curiosity, to play around a bit, to see how far I can push myself. It's been great to start work at eleven in the morning. But perhaps most importantly, I enjoyed, in Feynman's words, the pleasure of finding things out.

It's no fun playing alone, and indeed, a lot of people contributed to this booklet. First, and foremost, I would like to thank my supervisors Jan Kors and Guido Jenster. Jan, you are a devoted scientist and mentor, and I've greatly appreciated our many discussions and your throrough criticism on my writings. Guido, your enthousiasm is legendary; thank you for sharing your energy and wit. A special thank you goes to Lambert Dorssers. Lambert, thanks for your support and many useful comments over the past years. I'm most grateful to professor Van der Lei for being my promotor and to professors Gert-Jan van Ommen, Peter van der Spek and Cock van Duijn for participating in the reading committee.

The atmosphere in the Biosemantics group has always been excellent. I'll fondly remember the team work. Many thanks to Kristina, Antoine, Peter-Jan, Christiaan, Marc and Erik. Martijn, I've greatly appreciated our heated discussions and productive cooperation. Barend, you have amazing dialectical skills and a gift to make people enthousiastic, thank you for your input and support. Many thanks to Peter-Bram 't Hoen for a most pleasant and fruitful cooperation. Thanks to my colleagues at the department of medical informatics, Renske (both), Cobus, Rashindra, Michiel, Marissa and others for pleasant distractions and social endeavours.

I'm very grateful to the beautiful people I've had the privilege to befriend the past few years. I would like to single out a few. First I would like to mention Josef, my flatmate for more than four years, who endured the full brunt of my bouts of depression, my singing, my attempts at piano playing and the turbulences of my love life. Christa, Bas, dear friends, thank you for acting as "paranimf" during the graduation ceremony.

Finally, a heartfelt thank you to my parents. Lieve pa en ma, jullie zijn een rustpunt in mijn leven. Dank voor jullie onvoorwaardelijke liefde en steun. I dedicate this thesis to you.

Rob Jelier, October 2007

Contents

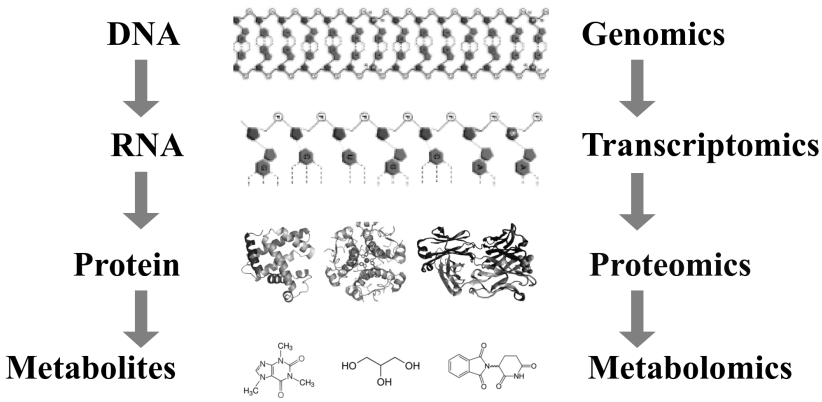
1 Introduction

The recent years have seen a revolution in the life sciences. We have gained the understanding and the technology to intelligently manipulate the molecular building blocks that constitute living organisms. Currently, research groups around the world are rallying to completely engineer and synthesize a living cell for the first time [1, 2], though at first still closely following nature's example. This would be a tremendous achievement given the high complexity involved in modeling all the relevant biochemical and physical processes. Still, the first artificial cell will be equipped with only the bare essentials for survival, an evolutionary first timer, and will be of much lower complexity than higher organisms. With the growing understanding of how the molecular building blocks of life cooperate to form creatures like ourselves, we start to see the delineation of their bewildering complexity. We now know that the human genome contains approximately 3.2 billion bases (about 1.5 gigabyte of sequence information) and codes for only about 22680 genes[1] (`www.ensembl.org`). The regulated transcription of these genes can result in a virtually infinite number of transcriptome states [4] (2^{22680} for mere bit-wise regulation), quite sufficient for each of the approximately 10^{14} human cells to have a distinct transcriptome. A final explosion of complexity occurs when we consider the interactions between all the actors in the cell, the DNA molecules, proteins, RNA molecules and metabolites and other small molecules. Say we would like to understand an average liver cell based on its proteins. According to a recent estimate liver cells contain 13000 non-redundant proteins [5]. Assume that on average every protein in our network is involved in 5 interactions [4], then 32500 interactions have to be taken into account. This is an underestimate of the number of interactions that would have to be modelled, as it ignores the small molecules and metabolites[2] with which these proteins interact. The interactions include, for instance, protein-binding to stabilize or inactivate protein complexes, molecular transport, as well as the enzymatic reactions involved in metabolic networks. Many of the interactions involve more than 2 factors, are non-linear over different concentrations, and can constitute

[1] The total number of unique transcripts in the transcriptome, including non-coding RNA, is expected to range between 10^6 and 10^7[3]. The expected number of unique proteins is several times the number of protein coding gene due to alternative splicing and post-translational modifications.

[2] A recent estimate is that there could be as much as 2000 metabolites in humans [6].

Figure 1.1: Overview of the different -omics levels.



complex feedback loops. Currently though, for many proteins their roles in biological processes are not completely understood, or completely unknown, and the modeling of human cells is in its infancy.

1.1 High-throughput technologies

The available information on protein activity is limited, but tremendous technological progress has been made on the automation and up-scaling of biochemical assays, and other approaches, to gain this information. Progress has been especially large in the area of determining the presence and abundance of the actors in the biochemical network. With high-throughput technologies it has become possible to quantitatively measure large numbers of transcripts, proteins and metabolites simultaneously. We can now retrieve snapshots of the transcriptome, proteome and metabolome (the -omics age, see Figure 1.1) and follow changes over time and with varying conditions. mRNA molecules are chemically very homogeneous, contrary to proteins and metabolites, and are therefore most suitable for comprehensive scale measurements. The first reports date back to 1995 [7, 8]. The most popular method to date to measure mRNA molecules has been DNA microarray technology. A DNA microarray is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface. The spots bind specifically to complementary DNA or RNA molecules in a solution. To measure how many molecules bind to the spots, the RNA/DNA molecules in the solution have been tagged with a fluorochrome, which can be detected by laser excitation. DNA microarrays have been extensively used and encouraging results have been achieved, for instance in the study of cancer. The technology has allowed for more differentiation between cancer patients, which will eventually lead to more accurate prognoses and better treatment [9, 10]. Also, the technology facilitated a better understanding of the changes in biological processes associated with the progression of cancer, e.g. from low-grade and localized prostate

Table 1.1: A biologist's wish list of information about genes when interpretating DNA microarray data sets. The list is transcribed from the research proposal that preceded this thesis.

Information needs
Gene function
Part of regulatory pathway
Tissue/cell type specific expression
Protein localization
Chromosomal gene location
Role in cancer
Known co-expressed genes

cancer to hormone insensitive metastases [11, 12]. Still, without workable models of biological processes to help explain the observations, interpretation of the DNA microarray datasets is often difficult.

1.2 Analysis of DNA microarray data

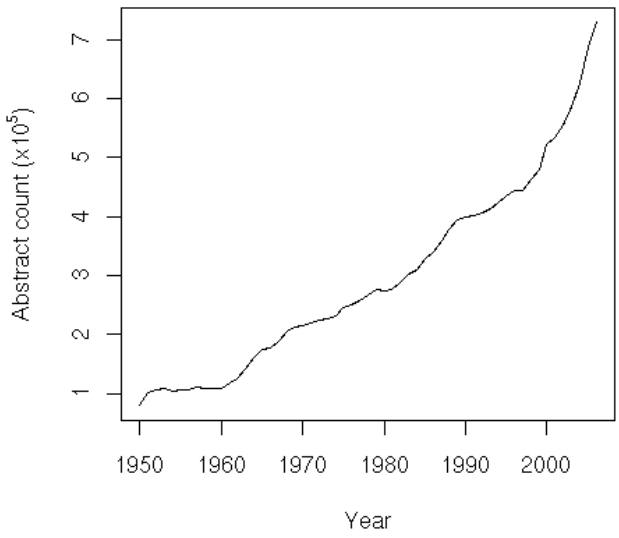
In a typical DNA microarray experiment, a perturbed state of a biological system is compared to a normal state, e.g. prostate cancer cells are compared to normal prostatic cells. This can result in the differential expression of potentially several hundreds of genes. It is then up to the researcher to characterize the observations, and, if possible, come up with a mechanistic explanation of the phenotype. One approach is to compare the measured profile of differential expression to a compendium of reference profiles of well-defined perturbations, such as the response to toxins whose mechanism of toxicity is well understood, or gene knockouts (e.g. through RNAi) [13, 14]. Though promising, the usability of the approach is limited by the size of the compendium, which should eventually reflect the number of pathways that could be perturbed. In addition, it remains to be seen how useful the approach is for comparisons across biological model systems and microarray platforms; both factors introduce considerable variation.

The alternative approach is to systematically evaluate the expression profile with the available information on the genes in the hope of identifying insightful patterns. To perform this analysis researchers are confronted with unprecedented information needs (see e.g. Table 1.1). In the early days of DNA microarray data analysis hardly any of this information could be retrieved from structured databases.

1.3 Databases in molecular biology

The advent of high-throughput technology has spawned great interest in systematically storing data, information and knowledge, and the resulting databases are often freely available through the internet [15]. Nowadays, several of the information needs can therefore be resolved: with the sequencing of the human genome as well as several other model organisms, the genomic location of genes can readily be retrieved and has been integrated with DNA microarray data [16]. Information on co-expression as well as tissue-specific

Figure 1.2: The number of abstracts published in Medline per year from 1950 to 2006.



expression can be retrieved by systematically analyzing data from online databases that store large numbers of DNA microarray data such as the Gene Expression Omnibus [17] and the Stanford Microarray Database [18]. Unfortunately, information needs concerning higher level information, such as gene function, still cannot be retrieved from available data. This information is typically published primarily in the unstructured free-text form of scientific publications and cannot be used directly in computational systems. Therefore retrieving this information is problematic, especially given the size of the biomedical scientific literature, which comprises millions of scientific papers, and the fact that thousands of new papers are added every day [19] (see Figure 1.2). In the biomedical field the most important bibliographical database is Medline which is maintained by the U.S. National Library of Medicine's (NLM) and contains over 17 million references to journal articles.

1.4 Information overload

It is not only in the context of the interpretation of high-throughput datasets that biomedical researchers struggle to keep abreast of the available knowledge. It has become impossible for researchers to read all publications in their field of interest, which forces them to make a stringent selection of relevant articles to read. The need to better manage this information overload has spawned a lot of activity to retrieve and structure our current knowledge. Ontologies play an important role in this [20]. In this context, an ontology

defines the concepts that are used in a certain field, as well as the relationships between the concepts. Ontologies are important to facilitate the exchange of information, both between people and automated systems. An important example in the life sciences is the Gene Ontology (GO) [21] which provides over 23000 concepts to describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Another example is the Unified Medical Language System (UMLS), an initiative to combine terminology systems, including GO, from the whole breadth of the biomedical field into a single comprehensive ontology [22].

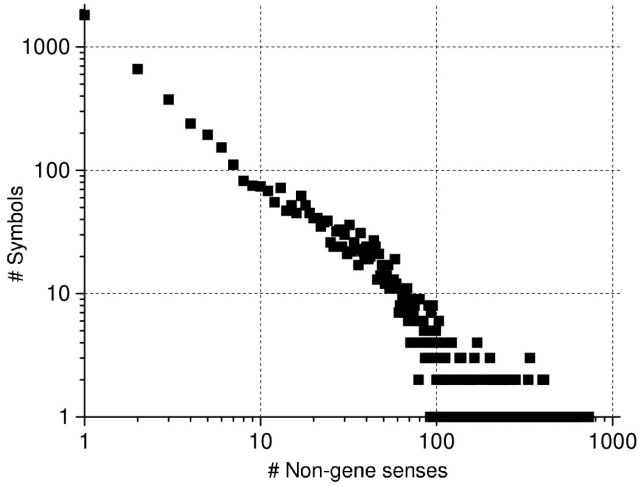
Several databases have become available that offer structured information on genes and proteins. There are several public databases, e.g. the databases offered by the Gene Ontology Annotation project [23], which assigns GO concepts to genes, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) project [24] which mainly focuses on mapping metabolic networks in several organisms. Several commercial projects have also generated databases, e.g. as offered by GeneGO (`www.genego.com`) and Ingenuity (`www.ingenuity.com`). For a large part, these databases are filled with manually encoded information in triplet form, such as "[subject] Rab27a - [relation] has - [object] GTPase activity", and are generated by experts reading the scientific literature. Manual encoding should be reliable and accurate[3], but is limited in scope and flexibility due to its labor-intensive nature. Complementary to manual encoding, currently much research effort is spent for the development of computerized algorithms to extract information from the scientific literature [19, 26]. Automated methods have the advantage of speed and adaptability, though it is more difficult to achieve high precision and recall.

1.5 Natural language processing

Understanding human language through computers is exceedingly difficult. The main reason is that language is intended for people who are knowledgeable both of the language and the world around us. People use their background knowledge to verify if what they read or hear is consistent and logical, and by doing so they can cope with the ambiguity that pervades human language. Erhardt et al. [27] reviewed and illustrated the different ambiguities in natural language: 1. Ambiguity at the part of speech level, like "complex" which can be an adjective: "complex timing relationships", a verb: "the capacity to complex metals", and a noun: "a complex of six proteins". 2. Ambiguity at the syntactical level, where sentences can be parsed in multiple ways. Consider the following examples: "AFB1 binds preferentially to DNA with an alternating G-C sequence" and "GMPPCP binds to tubulin with a low affinity relative to GTP". In the first example, the prepositional phrase introduced by "with" should be attached to the preceding noun phrase, "DNA". In the second sentence, the prepositional phrase should be attached to the noun phrase before the verb "bind" ("CMPPCP"), and not to "tubulin". 3. Semantic ambiguities, such as in these examples: "to seed coffee", "to roast coffee", "to drink coffee". "Coffee" refers to the coffee plant, to the coffee beans and then to the drink made from the roasted coffee beans. Handling these ambiguities automatically with a computer is and will be difficult, as they lack most of the knowledge every human has, at least for the foreseeable future

[3] It should be noted that the agreement between encoders can be low (see e.g. [25]), which suggests that it is necessary to have extensive checks and balances in place for manual encoding to be truly accurate and consistent.

Figure 1.3: Number of non-gene meanings for gene symbols. Dots indicate the number of human gene symbols (on the vertical axis) and, for each of these symbols, the estimated number of non-gene meanings (horizontal axis). Reproduced from Schijvenaars et al. [28].

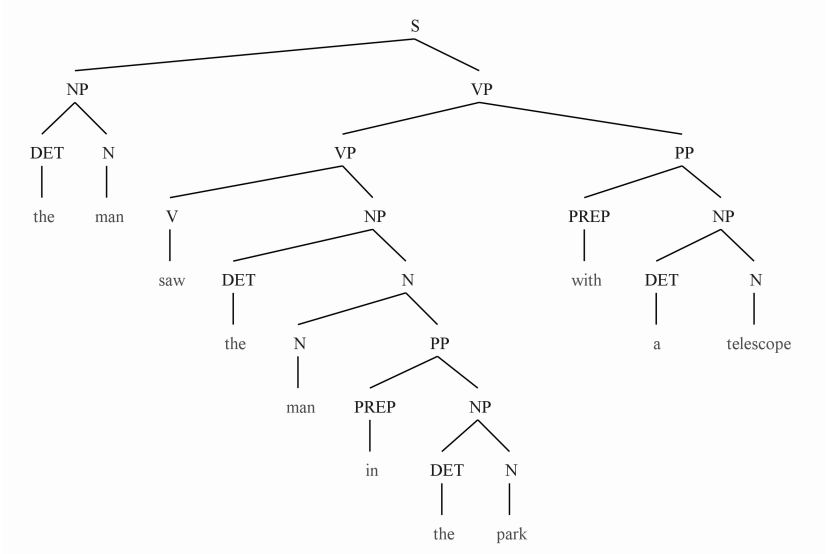


[27]. It is therefore with the necessary approximations and restrictions that computerized systems are applied to unlock the knowledge stored in natural language.

1.6 Named entity recognition

For many natural language processing applications in biomedicine it is necessary to identify the concepts to which words, terms or phrases in a text refer. This is the goal of named entity recognition, and it is not a trivial task, especially for genes and proteins. For genes, the first problem is that there can be many different names for the same gene, for human genes on average about 5.2 synonyms per gene [27]. The second more intricate problem is the ambiguity of many gene names and symbols (see e.g. Figure 1.3). Schijvenaars et al. [28] found that approximately 33% of the human genes has one or more homonymous symbols, and 13% of all gene symbols is ambiguous. In some cases genes symbols are aspecific abbreviations also used for other concepts (PSA, PC), or are common English names (hairy, fruitless), or are specific but used for multiple genes (p22, SCP1). In general, humans are very efficient at resolving word sense ambiguity, and require only very limited context to do so (see [29] for an overview of word sense disambiguation in the biomedical domain). To disambiguate gene symbols automatically a wide variety of approaches have been proposed, e.g. scanning the text for non-ambiguous synonyms (mostly a full gene name to disambiguate a gene symbol), or looking for keywords derived from full gene names [30], e.g. prostate as a keyword for PSA, the acronym for prostate specific antigen. Also machine learning approaches have been proposed to compare the textual context of an ambiguous symbol to reference contexts [28]. Resolving gene name ambiguity in texts has received quite some attention, for instance through challenges like BioCreAtIvE ([31], `biocreative.sourceforge.net`) which provide a forum as well as test sets to evaluate

Figure 1.4: A syntax tree (`http://ironcreek.net/`). Note the ambiguity, multiple parse trees are possible. The prepositional phrase "with a telescope" can be attached to the first man, as it is now, or to the second man, who would then have a telescope in the park. S stands for sentence, NP is noun phrase, DET determiner, N noun, VP verb phrase, V verb, PREP preposition.



and compare different approaches.

1.7 Information extraction

The field of information extraction in biomedicine focuses on the extraction of relations between genes and other biomedical concepts. In a typical information extraction excercise, the first step is the tokenization of text, retrieving the boundaries of sentences and words, followed by part -of-speech tagging, the assignment of tags like "noun", "verb" or "adjective" to every word. Subsequently a syntax or parse tree can be constructed (Figure 1.4), which represents the syntactic structure in a sentence according to some formal grammar. As the next step, semantic tags, like "gene", can be assigned through dictionary approaches or pattern matching. Finally, relationships are extracted from the annotated sentence.

The techniques used to extract relations vary in complexity, starting from very simple matching, without syntactic parsing, of patterns such as "protein A - action X - protein B" [32, 33]. More advanced methods make use of predicate-argument structures, a normalized form of syntactic relations, which combine shallow parsing techniques with semantic patterns to extract relations [34, 35]. This method is more generic than using explicit syntactic patterns as not all possible patterns have to be made explicit. The most complex methods, and in theory the most powerful, use full parsing to analyze the structure of the whole sentence [36, 37]. Due to the great flexibility of natural language, its ambiguity, and domain-specific peculiarities and vocabulary, a lot of manual tweaking and coding of

patterns is required in order to achieve acceptable recall and precision. Therefore applications in this area are typically strictly focused to efficiently retrieve a limited number of relationship types. For a further introduction into the field of information extraction see [38], and for a review on its applications in biomedicine see [39].

1.8 Co-occurrence based text-mining

One of the emerging approaches is text-mining, which infers associations between biomedical entities by combining information from multiple papers. Text-mining typically uses the occurrence and co-occurrence statistics of concepts or lexical features, such as words or bigrams. The approach avoids the necessity to "understand" natural language and is therefore more suitable for applications with a broad scope. Text-mining can be used to retrieve associations between concepts that are explicitly mentioned in the same document, but also implicit associations can be found: associations never mentioned together in a single document, but inferred from other associations.

The classic application for text-mining is literature-based knowledge discovery, which attempts to link disjunct sets of literature, in order to derive promising new hypotheses [40–44]. Swanson was a pioneer in this field and was able to publish several new hypotheses derived with the help of text-mining (see e.g. [45]). His well-known first discovery was the hypothesis that Raynaud's disease could be treated with fish oil [46], which was later corroborated experimentally [47]. For literature-based knowledge discovery, both an open and a closed so-called A-B-C discovery model are used [48]. An open discovery process is characterized by the generation of a hypothesis. For Swanson's discovery, the task was to find a new treatment for Raynaud's disease (the A concept). First, interesting associated concepts are retrieved, here typically physiological processes affected by the disease (the B concepts). Next, drugs or substances are retrieved that act on the selected intermediate terms, the relevant physiological processes (the C concepts). In the discovery process, it is likely that many Bs and Cs will be found and the most important challenge for discovery support tools is to prioritize the most likely suggestions. In the closed discovery model a hypothesis is elaborated and tested on the basis of the literature. For example, given a disease A and substance C which do not have a published relation, the researcher tries to find intermediate B-terms associated to both A and C, such as cellular or physiological functions, which provide clues to confirm the hypothesis.

Text-mining has also been applied for other applications in the biomedical domain. In 2001, Jenssen et al. [49] introduced PubGene, a gene network consisting of 13,712 genes based on co-occurences derived from Medline. The weights of the connections were weighted by the number of times the genes co-occur in abstracts. The authors demonstrated that the network could be helpful in the analysis of DNA microarray data. Another approach was followed by Shatkay et al. [50]. They identified for each gene a *kernel* document describing the gene's function. Subsequently, similar documents were retrieved through a standard information retrieval approach to yield document sets per gene. Next, associations between genes were inferred by measuring the overlap between document sets. Stapley et al. [51] used weighted word counts in combination with a classification approach, support vector machines, to predict the sub-cellular location of proteins. For certain tasks it can be beneficial to combine text-derived data with other data sources. For instance, Perez-Iratxeta et al. [52] derived a scoring system to prioritize genes that may have a functional relationship with genetically inherited diseases. In their

approach they used indexed keywords in Medline to first retrieve associations between pathological conditions and chemical concepts. Then chemical concepts were associated with GO terms through documents attached to the GO terms. Finally, both association types were combined, and, together with information on the chromosomal region linked to the disease, used to suggest disease genes.

1.9 Thesis overview

The aim of the research presented in this thesis is the application of co-occurrence based text-mining to assist biomedical researchers in data interpretation, with a special focus on the analysis of data generated by high-throughput technology, and to facilitate literature-based discovery. In this thesis we present and evaluate two text-mining approaches, the associative concept space (ACS), and concept profiling. The ACS is a Euclidean space in which thesaurus concepts are positioned and where the distances between concepts indicate their relatedness. In chapter 2, the ACS is evaluated on a controlled test set for the retrieval of functionally associated genes. For the same task we introduce concept profiles in chapter 3. The idea behind concept profiles is to relate concepts to each other based on their associated sets of texts. A concept profile characterizes a set of texts associated to a concept. The profile consists of a list of concepts and each concept in the profile has a weight to signify its importance. The ACS and concept profile technology are compared on the controlled test set introduced in chapter 2. Subsequently, concept profiles are applied to two case studies involving actual DNA microarray datasets. In chapter 4, we further explore the properties of concept profiles and the influence of weighting schemes by means of a large test set for the task of assigning GO concepts to genes. Next, in chapter 5, concept profiles are applied to a current challenge in molecular biology: the comparative analysis of DNA microarray studies. Such analysis can be used to confirm findings from individual studies as well as identify interesting parallels between studies. However, such analyses are hampered by the large influences of design, technical and statistical factors on the found differentially expressed genes. Comparisons based on perturbed biological processes could be more robust as different genes may hint at the same process. The use of concept profiles for this purpose is evaluated on a large dataset comprising 102 DNA microarray experiments from the field of muscle disease and development. In chapter 6 Anni 2.0 is introduced, a tool designed to aid biomedical researchers with a broad range of information needs by means of concept profiles. Anni is evaluated with two user cases: the analysis of a set of genes differentially expressed between localized and metastatic prostate cancer, and the reproduction of a published literature-based knowledge discovery.

2

Retrieving associated genes with ACS

Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes

R. Jelier[1], G. Jenster[2], L.C.J. Dorssers[3], C.C. van der Eijk[1], E.M. van Mulligen[1], B. Mons[1] and J.A. Kors[1]

Departments of [1]Medical Informatics, [2]Urology and [3]Pathology
Erasmus MC, Rotterdam

Bioinformatics 2005, 21:2049-58

Abstract

The advent of high-throughput experiments in molecular biology creates a need for methods to efficiently extract and use information for large numbers of genes. Recently, the associative concept space (ACS) has been developed for the representation of information extracted from biomedical literature. The ACS is a Euclidean space in which thesaurus concepts are positioned and the distances between concepts indicates their relatedness. The ACS uses co-occurrence of concepts as a source of information. In this paper we evaluate how well the system can retrieve functionally related genes and we compare its performance with a simple gene co-occurrence method. To assess the performance of the ACS we composed a test set of five groups of functionally related genes. With the ACS good scores were obtained for four of the five groups. When compared to the gene co-occurrence method, the ACS is capable of revealing more functional biological relations and can achieve results with less literature available per gene. Hierarchical clustering was performed on the ACS output, as a potential aid to users, and was found to provide useful clusters. Our results suggest that the algorithm can be of value for researchers studying large numbers of genes.

2.1 Background

The availability of whole genome sequences and the advent of high-throughput technology for molecular biology have dramatically changed the nature of biomedical research. Thousands of genes or proteins can now be studied in a single experiment. With this development arose the challenge to efficiently handle the huge amounts of data produced by these experiments. An important issue in the interpretation of data produced by DNA microarrays is the identification of the biological processes that underlie the observed differences in gene expression. Information needed for this task is for the larger part available in millions of free-text scientific publications, with thousands of new publications being added every day. When many genes are studied, the number of relevant publications will frequently be prohibitively large. This renders the traditional approach of manually searching bibliographic databases for every gene and reading scientific articles inadequate. It is therefore an important challenge at this time to make the available information both accessible and interpretable for molecular biologists.

An interesting current development is the use of annotations of genes with gene ontology (GO) terms [21, 23] for the analysis of the results of microarray experiments [53, 54]. The most reliable annotations are based on manually assigning GO codes to genes based on scientific literature. GO provides a structured description of biological information which is very amenable for use in bioinformatics. These methods are useful, though limited in flexibility by the focus of the ontology. GO annotations are for instance not very useful if one is interested in gene-disease relations. Additionally, the most reliable annotations are obtained by a difficult, slow and labor intensive manual process. Clearly there is much more information stored in the whole body of literature than captured in current GO annotations. Therefore mining texts directly for relevant information on genes would be more flexible and could make an important addition to the molecular biologist's toolbox for microarray data analysis.

The recent years have seen new methods to efficiently use the large amounts of literature for biomedical research. In an early effort, Masys et al. [55] made keyword profiles for genes based on the manual annotations of articles with the controlled vocabulary Medical Subject Headings (MeSH) in the National Library of Medicine's MEDLINE database . For a group of selected genes, these profiles are combined and every keyword is given a value indicating its specificity for the group. An important developing field is the automatic extraction of relevant information from scientific texts (for a review see [56]). The most important distinctions between current text-mining methods are the amount of linguistic information that is used and the number of documents that can be handled efficiently. One approach is to extract detailed information from documents by using natural language-processing techniques [36, 57]. Many approaches though extract information about genes from scientific texts using only information about the co-occurrence of terms in a sentence or abstract [42, 49, 51, 58–60]. The use of simple co-occurrence is popular, because it allows for easy implementation and the efficient processing of huge amounts of texts. Also, the co-occurrence of gene names in an abstract frequently reflects an actual biological relationship between the two genes, as was shown by Jenssen et al. [49], for example, and Stapley and Benoit [51].

Recently, we developed a new co-occurrence based text meta-analysis tool, the associative concept space (ACS) [61]. To construct the ACS, thesaurus concepts are automatically identified in texts. The use of a thesaurus allows that synonyms are mapped to the

same concept, which reduces noise caused by natural language variation. Additional advantages are the possibilities to include multi-word terms and to use thesaurus hierarchies. The thesaurus we use contains genes but also many other biomedical concepts.

The ACS algorithm is a Hebbian-type of learning algorithm that in an iterative process positions the thesaurus concepts in a multidimensional Euclidean space. In this space the dimensions do not take a specific meaning, but just allow the positioning of the concepts relative to each other. The position of a concept follows from the mapping of co-occurrence relations (paths) between concepts to distances. A distance between two concepts will not only reflect the co-occurrence of the two concepts, a one-step relation, but also indirect, multi-step relations between the two concepts. The idea behind the algorithm is that concepts that are placed close to each other will be more likely to share an actual semantic relationship. An important feature of the ACS is that the multidimensional space can be visualized using standard dimension reduction techniques. The visualized ACS allows for easy and intuitively appealing browsing for relations between concepts that are derived from the underlying literature. The ACS can thus be used as a kind of portal to the literature, but it can also be used as a knowledge discovery tool. When in the ACS two concepts are placed close to each other while they do not have a co-occurrence, this would suggest that a relationship is not explicit in the literature set, but is likely to exist.

The ACS can be used in a similar way to how other authors have used co-occurrence as a basis for a knowledge discovery system. Swanson and Smalheiser [62] discovered valuable knowledge hidden in medical literature. They searched for paths between two sets of related terms allowing for one intermediary term to connect terms from the two sets. Several other authors have built on their work using similar models [40–42].

Compared to previously published algorithms, the ACS has the potential to stand out on several points. The ACS could improve on the performance of using only direct co-occurrence of genes by improving recall. When only direct gene-gene co-occurrences are used some relations will be missed, for example the relation between two genes that are involved in the same cellular process would be missed when their roles happen to be described in separate papers. The ACS can reveal relations between genes based on their contexts, i.e. the other concepts with which they are mentioned, and does not require the genes to be mentioned in the same article. The method introduced by Chaussabel and Sher [58] also uses other co-occurring terms, and can pick up relations between concepts that do not necessarily co-occur in the same article. Our approach differs in that we use a thesaurus for identifying concepts in texts, which, as mentioned earlier, has several advantages. Additionally the ACS differs as it implicitly uses more information in that concept relations that involve more than two steps play a role. Raychaudhuri and Altman [63] developed a method that assesses whether a group of genes is related by measuring the similarity of literature attributed to group members. Wren and Garner [64] use a thesaurus-based approach like we do. They use a probabilistic approach to identify whether a group of genes is functionally cohesive according to the literature and identify which terms connect the genes. Different from the previous two methods, the ACS does not assess functional coherence of groups. Instead, distances between concepts in the ACS reflect relatedness. Groups can be identified by clustering, as we shall illustrate, but relations between concepts can also be visualized. In this way a molecular biologist can quickly and intuitively inspect, based on a set of literature about a group of genes, relationships between these genes and other concepts associated with these genes.

In this paper we will assess whether the ACS is useful for molecular biologists. We

will do this by evaluating how well the positioning of genes in the ACS reflects actual functional biological relationships between genes. This is the first systematic evaluation of the ACS on real data. A test set is constructed based on groups of genes that are known to be functionally related. We measure how well the method reproduces these groupings based on the literature about the genes. The performance of the ACS will be compared to a simple approach that only uses co-occurrences of genes. The results of the quantitative analysis are thoroughly reviewed in an attempt to understand the underlying phenomena. Additionally, we demonstrate how the ACS may assist molecular biologists in the interpretation of DNA microarray data.

2.2 Methods

Selection of gene groups We chose five groups of genes, each defined by a different aspect of gene biology, being function, organelle, biological process, metabolic pathway or association with a disease. Only human genes were taken into consideration. Three groups were derived from the functional annotation by the Gene Ontology (GO) annotation project [21, 23] as stated in the Locuslink database of June 19th 2003. As evidence, the following annotation tags were accepted as being trustworthy: IDA, TAS, IGI, IMP, IPI, ISS (see `http://www.geneontology.org/doc/GO.evidence.html`). Note that the most prevalent annotation, inferred from electronic annotation (IEA), was not accepted as sufficient proof. The other two groups were acquired by alternative approaches. For genes associated with a disease, a review on breast carcinomas was used to identify 8 genes regularly associated with this type of cancer [65]. For a metabolic pathway we used the KEGG database [24] to identify the 10 genes involved in glycolysis in man. The selected groups are:

1. Spermatogenesis; GO code 0007283, 41 genes, a biological process;
2. Lysosome; GO code 0005764, 25 genes, an organelle;
3. Chaperone activity; GO code 0003754, 23 genes, a biological function;
4. Breast cancer; review, 8 genes, genes related to a disease;
5. Glycolysis; KEGG database, 10 genes, a metabolic pathway;

None of the selected genes occurred in more than one group. From these genes, only those for which at least 10 abstracts could be retrieved by a PubMed query were added to the set used for the evaluation.

Selection of literature

Literature was selected by a PubMed query performed for each gene [66]. The query was composed of gene symbols, including aliases, and full names that were derived from Locuslink [66]. To avoid the use of ambiguous terms in the query, we only used full gene names or gene symbols with a number in it. Gene symbols or full gene names that refer to more than one LocusLink gene were rejected as well. The accepted gene names and gene symbols were combined by 'OR' and were required to be found as text

words. Additional requirements were the presence of the MeSH annotation 'Human' and an electronic publication date (EDAT) between 1-1-1965 and 31-12-2002.

Some genes within our set were found in many more abstracts than others. To assess how the number of abstracts per gene affects the outcome, three sets of literature were produced. For the first set, each gene contributed exactly 10 abstracts to the set randomly selected from the set of all abstracts available for that gene. In the second set, each gene contributed a maximum of 100 abstracts, though some contributed less. Similarly, we constructed a 1,000 abstracts set. For each set, three versions were made to account for sampling effects. To assess the sensitivity to changes in the literature set we also experimented by adding 10,000 Medline abstracts randomly selected from those published in the years from 1997 to 2002.

Indexing

In this context, indexing means the identification of thesaurus concepts in text. The thesaurus we used for indexing was composed of three parts: the freely available thesauri/ontologies MeSH and GO, and a LocusLink derived human gene thesaurus. For each gene in the thesaurus, we considered all fields from LocusLink describing gene symbols, gene names, aliases and product names as synonymous. To match a common spelling variation, for every symbol that ends with a number, we also added to the thesaurus the same symbol with the number separated by a hyphen or a space and vice versa. For every word in the thesaurus we included the uninflected form produced by the normalizer of the lexical variant generator [67].

MEDLINE titles, MeSH headings, and abstracts, if available, were indexed using Collexis software (`http://www.collexis.com` and [68]). Each concept found was assigned a concept weight to represent the importance of the concept for a particular citation. A document was thus represented by an M-dimensional vector $W = (w_1, w_2, \ldots, w_M)$, where M is the number of distinct concepts in the thesaurus, and$w_i = 0$ if t_i is not in the document. A concept's weight is defined as term frequency (TF) times inverse document frequency (IDF):

$$w_i = TF \times IDF = f_i \times ({}^2\log \frac{N}{N_i} + 1)$$

$TF$is the number of occurrences f_i of a concept t_i in a given document. IDF is a correction factor for the number of documents N_i containing t_i in a given set of N documents. Frequently occurring concepts, or general concepts, are thus given a lower weight. To calculate IDF we used 10 years of Medline. For each concept fingerprint the weights were normalized, i.e. divided by the largest value.

ACS and gene co-occurrence

For the gene co-occurrence method two genes co-occur if they are both found in the abstract, title or MeSH headings of one document. The gene co-occurrence method is based on a co-occurrence matrix. The matrix contains the number of times genes from the set co-occur.

The ACS is a multi-dimensional Euclidean space, in which concepts are positioned. For the ACS per document only co-occurrences of concepts with a weight above a threshold

are used, to diminish the impact of general terms. Concepts are positioned based on their co-occurrences, one-step relations, and multi-step relations. For example, a two-step relation exists between concepts X and Z, if they both co-occur with a concept Y. Concepts that are connected by many co-occurrence paths, either one step or multi-step, are expected to have a small distance in the ACS, while concepts with few or no paths between them should be far apart. The algorithm starts by randomly positioning the concepts in the ACS. Subsequently, for each fingerprint, co-occurring concepts are moved towards each other. After all fingerprints are processed in this manner, the concept cloud is expanded and all concepts are moved away from each other. These attraction and relaxation steps are repeated until the relative position of the concepts is stable. In this way single and multi-step co-occurrence relations are mapped to a Euclidean space. The idea behind the algorithm is that in the ACS distance between concepts takes the meaning of a semantic relatedness. The dimensions in this space do not have a meaning; they only accommodate the placing of concepts relative to each other. For a more detailed description of the algorithm see [61].

For the learning of the ACS standard settings were used. The ACS algorithm iterated 150 times and every ACS had 10 dimensions. Because the ACS algorithm has a random initialization, the final positioning of concepts can be different each time a new ACS is build, even with the same literature set. To take this factor into account, we built and evaluated an ACS three times for each literature set. The results of the evaluation were averaged.

Evaluation

Both the ACS and the gene co-occurrence method were employed to produce a ranking of the set of genes relative to one so-called seed gene. All genes in turn served as a seed, producing a ranking for each of the 53 genes in our set. For the gene co-occurrence method, the genes from the set are rank-ordered according to their number of co-occurrences with the seed. Ties are ordered randomly. For the ACS, genes from the set are rank-ordered according to their Euclidean distances to the seed gene.

For each gene a receiver operating characteristics (ROC) curve was then constructed. ROC curves are commonly used to evaluate classifiers [69]. They are two-dimensional graphs in which the true-positive (TP) rate is plotted against the false-positive (FP) rate. The TP rate is defined as correctly classified positives divided by all positives. The FP rate is defined as incorrectly classified negatives divided by all negatives. For each seed the set of genes was divided in two classes: members from the same functional group as the seed (positives) and non-group members (negatives). As input for the ROC curve served the set of genes ranked relative to the seed. The TP and FP rates were calculated for every rank. The area under the curve (AUC) was used as a performance measure [70]. This value varies between 0 and 1. An AUC of 1 represents perfect ordering, i.e. all positives are at the top of the list with no negatives in between. The AUC has the useful property that a value of 0.5 represents random ordering [70]. This property provides us, in a way, with built-in negative control.

To determine whether the AUC scores differed significantly between the two methods, we used the non-parametric Wilcoxon signed ranks test. The test requires the AUC scores of the genes to be independent. Because this is not true in this case, we had to apply bootstrapping [71] to estimate the distribution of the Wilcoxon test statistic. we

generated 100 new sets of genes by sampling genes from the original set with replacement. The sampling was stratified over the 5 gene groups to obtain groups of equal size as in the original set. In the subsequent selection of literature every gene appearing more than once in the set was given the same set of literature, but with different IDs. This is important for the ACS as we have observed that the size of the literature set can have an influence. During indexing duplicate genes are treated as synonyms. AUCs are calculated for both simple gene co-occurrence and ACS, and the Wilcoxon signed ranks test is applied to measure the difference between the two methods per gene group. These 100 results are used to determine if the two methods differ in performance at the 0,05 level.

It is possible that relations exist between genes in different gene groups. In order to evaluate whether this is the case, we manually checked 108 of all possible 1081 inter-group gene pairs for functional biological relationships. Information sources used were GO annotations, KEGG, Online Mendelian Inheritance in Man (OMIM, `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM`), abstracts in which a co-occurrence was observed, and Swiss-Prot (`http://www.expasy.org/sprot/`). Relationships were acknowledged if they were of the following types: same or similar biological process, biological function, specific organelle, metabolic pathway, protein family, direct interaction, or association with the same disease.

For visual inspection of the multi-dimensional ACS we utilized Sammon mapping [72], which reduces the dimensionality to two, and hierarchical clustering as introduced for microarray analysis in [73]. To apply the latter, the ACS coordinates of the set of genes were translated so that the centre of the set was at the origin. The resultant coordinates were used as input for the clustering program. We used average linkage clustering with correlation (uncentered) as similarity metric.

2.3 Results

We selected five groups of genes, with a total of 53 genes (Table 2.1). Genes in each group share a distinct functional biological characteristic: role in spermatogenesis, breast cancer, glycolysis, lysosome, or chaperone activity. MEDLINE abstracts were selected by PubMed queries for every gene. For the PubMed query we used gene names and symbols extracted from Locuslink, excluding most of the ambiguous terms. We only included a gene in our study if at least 10 abstracts could be retrieved. The median number of retrieved articles for the breast cancer genes (median 2674) and glycolysis genes (median 787) is considerably higher than for the chaperones (median 61), lysosome genes (median 127) and spermatogenesis genes (median 21.5). The same tendency holds for the number of co-occurrences between a gene and other genes from the set (see Table 2.1, medians in same order: 15, 12, 3, 6, 2). Twenty-nine genes co-occur with five or less genes and seven do not co-occur with any. To evaluate the quality of the grouping we estimated the amount of accidental inter-group relationships. From all 1081 possible pairs of genes from different groups, we manually assessed 108 (10%) randomly picked pairs for functional biological relationships. Seven gene pairs (6.5%) were found to have a relevant relationship (see Table 2.2).

The ACS and simple gene co-occurrence were employed to produce for each gene, termed the seed, a ranking of the other 52 genes. A perfect ranking is when all genes that have a functional biological relationship with the seed, rank highest. To produce these rankings we used for the ACS the distances between genes and for simple co-occurrence

Table 2.1: Selected genes from five functional groups. Given are the Entrez Gene identification number (ID), preferred symbol, gene name, number of abstracts retrieved by the PubMed query (A) number of genes from the set with which the gene co-occurs, taking all abstracts into account (C), and the functional group to which the gene belongs (G): a. chaperone activitiy; b. lysosome; c. glycolysis; d. spermatogenesis; e. breast cancer.

ID	Gene symbol	Gene name	A	C	G
325	APCS	Serum amyloid P component	325	7	a
3998	LMAN1	Mannose-binding lectin 1	61	6	a
6102	RP2	Retinitis pigmentosa 2	96	4	a
6687	SPG7	Spastic paraplegia 7	17	0	a
6950	TCP1	t-complex 1	35	3	a
7249	TSC2	Tuberous sclerosis 2	279	3	a
11140	CDC37	Cell division cycle 37 homolog (S.cerevisiae)	17	0	a
154	ADRB2	Beta-2-adrenergic receptor	1309	9	b
410	ARSA	Arylsulfatase A	434	10	b
411	ARSB	Arylsulfatase B	101	5	b
412	STS	Steroid sulfatase	286	12	b
1200	CLN2	Neuronal ceroid-lipofuscinosis 2	112	3	b
2548	GAA	Acid alpha-glucosidase	312	34	b
2581	GALC	Galactosylceramidase	201	4	b
3916	LAMP1	Lysosomal-associated membrane protein 1	36	6	b
4036	LRP2	Low density lipoprotein-related protein 2	127	2	b
4353	MPO	Myeloperoxidase	3837	17	b
4758	NEU1	Sialidase 1	753	14	b
7103	TM4SF3	Transmembrane 4 superfamily member 3	16	0	b
8692	HYAL2	Hyaluronoglucosaminidase 2	27	7	b
10266	RAMP2	Receptor (calcitonin) activity modifying protein 2	47	3	b
10268	RAMP3	Receptor (calcitonin) activity modifying protein 3	22	2	b
226	ALDOA	Fructose-bisphosphate aldolase A	1853	10	c
2023	ENO1	Enolase 1	2550	14	c
2597	GAPD	Glyceraldehyde-3-phosphate dehydrogenase	26	18	c
2821	GPI	Glucose phosphate isomerase	1015	15	c
5230	PGK1	Phosphoglycerate kinase 1	110	7	c
5236	PGM1	Phosphoglucomutase 1	558	8	c
2302	FOXJ1	Forkhead box J1	13	4	d
2492	FSHR	Follicle stimulating hormone receptor	310	6	d
2649	NR6A1	Nuclear receptor subfamily 6, group A, member 1	13	1	d
3010	HIST1H1T	Histone 1, H1t	26	0	d
3206	HOXA10	Homeo box A10	33	3	d
3640	INSL3	Insulin-like 3	36	2	d
5619	PRM1	Protamine 1	74	3	d
5620	PRM2	Protamine 2	65	3	d
6046	BRD2	Bromodomain containing 2	26	2	d
6847	SYCP1	Synaptonemal complex protein 1	10	0	d
8287	USP9Y	Ubiquitin specific protease 9, Y chromosome	11	2	d
8607	RUVBL1	RuvB-like 1 (E.coli)	11	0	d
8900	CCNA1	Cyclin A1	17	2	d
9191	DEDD	Death effector domain containing	14	0	d
9240	PNMA1	Paraneoplastic antigen MA1	30	4	d
23626	SPO11	Sporulation protein 11 homolog (S.cerevisiae)	11	1	d
672	BRCA1	Breast cancer 1, early onset	2674	12	e
675	BRCA2	Breast cancer 2, early onset	1530	8	e
1956	EGFR	Epidermal growth factor receptor	7502	22	e
2064	ERBB2	Erythroblastic leukemia viral oncogene homolog 2	2791	16	e
2066	ERBB4	Erythroblastic leukemia viral oncogene homolog 4	227	7	e
2099	ESR1	Estrogen receptor 1	36	15	e
5241	PGR	Progesterone receptor	3656	22	e
5915	RARB	Retinoic acid receptor, beta	15	5	e
7157	TP53	Tumor protein p53	19919	29	e

Table 2.2: Found inter-group functional biological relations. The last column gives examples of PubMed identification numbers of articles that support the identified relationship.

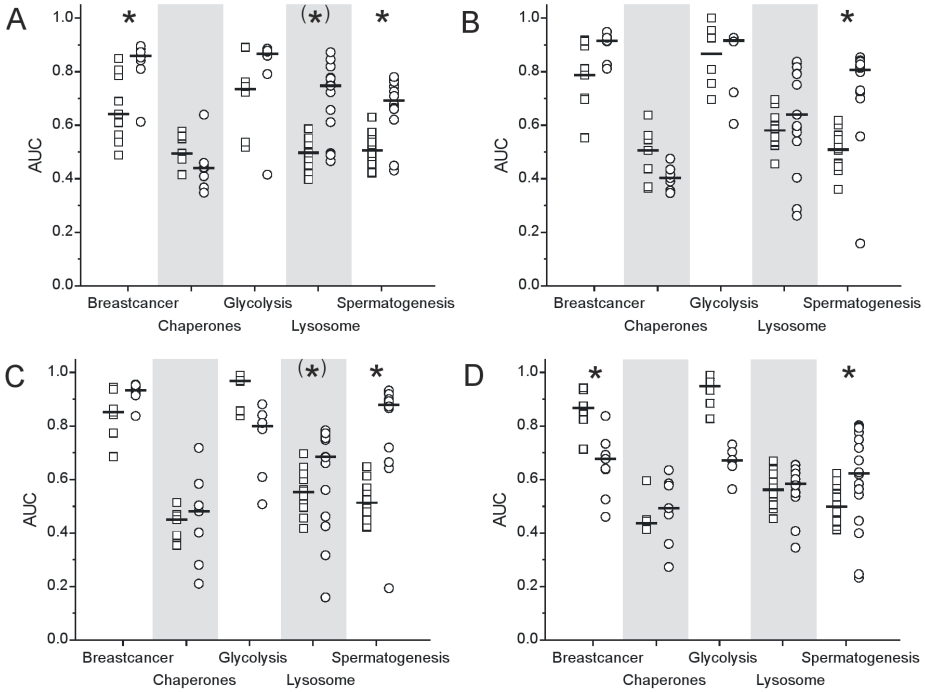
Description of relation	Gene pair	PMID
Cancer	RARB–PNMA1	10050892
Cancer	PGR–PNMA1	10050892
Cancer	RAMP2–TP53	11420706
Alzheimer	MPO–APCS	12052532, 12015594
Cryptorchidism	STS–INSL3	6135610, 10319319
Female reproductive cycle	FSHR–ESR1	11089565, 10342864
Epilepsy	BRD2–TSC2	12830434

the count of gene co-occurrences. To assess the quality of the rankings we determined Receiver Operating Characteristics (ROC) curves and used the area under the ROC curve (AUC) as an outcome measure [70, 74]. The AUC has a value of 1 for perfect ordering, a value of 0.5 for random ordering, and for the worst possible ordering (genes related to the seed have the lowest ranks) the AUC is 0. We varied the maximum number of abstracts a gene could contribute to the set of literature used in the analysis, to take into account that some genes were mentioned in thousands of abstracts whereas others are only mentioned in ten.

Figure 2.1 shows the performance of both the gene co-occurrence method and the ACS for the five gene groups. For the gene co-occurrence method performance for the chaperone, lysosome and spermatogenesis groups is not much better than random ordering, with AUC scores close to 0.5. These low scores are explained by a lack of co-occurrence between its group members (cf. Table 2.1). For the breast cancer and glycolysis groups, performance is moderate for 10 abstracts per gene and improves when more abstracts are available. For the literature set of max 1000 abstracts per gene the score is good for the breast cancer genes (median 0.85) and excellent for the glycolysis genes (median 0.97). The addition of 10,000 randomly selected abstracts to the literature set does not affect the scores much. We found that only very few gene co-occurrences were extracted from these additional abstracts. Following from the AUC scores and also from a manual evaluation of co-occurrences of genes from different groups showed that almost all correctly found co-occurrences did represent actual biological relationships. Some wrong associations were found as a consequence of incorrect indexing due to ambiguity in the gene names. Pairs of genes with a general, though not a functional, biological relationship, were also found several times, such as the localization of two genes on the same chromosome, e.g. MPO and ERBB2.

The results for the ACS show that the ranking of genes scores better than random arrangement for all groups, except for the group of chaperones. The breast cancer genes have very high scores for the first three literature sets (median up to 0.93 for maximally 1000 abstracts per gene). The glycolysis group also has a very high score for the first two sets (0.92 for maximally 100 abstracts per gene) but decreases (median 0.8) for the set of 1000 abstracts per gene. The spermatogenesis group scores best for the set of 1000 abstracts per gene (median 0.88). The lysosome group scores best for the smallest literature set (median 0.75). The addition of 10,000 random abstracts results in a substantial decrease

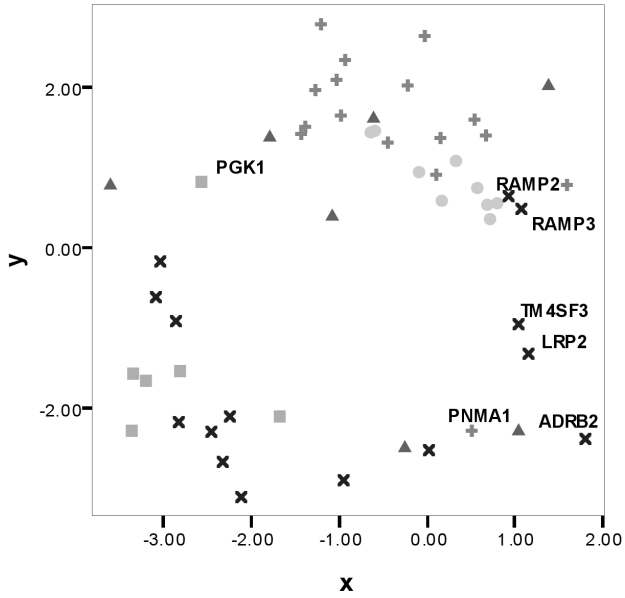
Figure 2.1: AUC scores for individual genes per group for the gene co-occurrence method (open boxes) and the ACS (open circles). The different graphs represent results for the different literature sets: (A) 10 abstracts per gene, (B) maximum 100 abstracts, (C) maximum 1000 abstracts, (D) maximum 1000 per gene + 10 000 randomly selected abstracts. An asterisk above a group indicates a statistically significant difference between the two methods(at the 0.05 level). An asterisk in parentheses indicates a significant difference when wrongly annotated genes are removed (see Results section).



of the AUC scores for most groups. Using the Wilcoxon signed ranks test in combination with bootstrapping the ACS performs significantly better than the gene co-occurrence method for all literature sets when all gene groups are combined, but not when random literature is added. Results for the same test on a per group level are shown in the figure. For only 10 abstracts per gene ACS performance is better for the breastcancer group. ACS performs better in all literature sets for the spermatogenesis group. We observe that when no randomly selected abstracts are added and not considering the chaperone group (see below), the ACS tends to score higher for all groups, with the sole exception for the glycolysis group in the literature set of 1000 abstracts per gene. We should note that due to the small size of the gene groups, statistical power is limited. The gene co-occurrence method only performs better when the randomly selected abstracts are added and only for the breastcancer group. The wrongly annotated genes and their effect on performance will be discussed below.

There are large differences in scores between different groups as well as between individual genes from the same group. The group of chaperones was not retrieved by both methods. In table 2.1 it is shown that its group members have relatively few publications

Figure 2.2: Two-dimensional projection of the ACS with Sammon mapping. The different groups are marked; triangles, Chaperone activity; circles, Breast cancer; squares, Glycolysis; cross symbols, Lysosome; plus symbols, Spermatogenesis. Some genes with aberrant behavior are labeled.



per gene and only a small number of gene co-occurrences. Upon closer inspection, it appeared that typical terms for chaperone activity were very scarce (except for TCP1). Chaperone activity was almost never the topic of the abstracts for these genes. Not surprisingly, in most cases, these genes were mentioned in the context of a disease or syndrome.

For the ACS, some of the genes have scores far below 0.5, which indicates that they were placed away from their group members. Especially the genes from the lysosome group have a large range of scores. Analysis of the function of some of these genes showed interesting results. To visually inspect the ACS, we made a 2-dimensional projection of a typical ACS for the literature set of a maximum of 100 abstracts per gene (Figure 2.2). Six genes of the lysosome group had relatively low AUC scores ($\leq$0.6). It turned out that the gene products of TM4SF3, LRP2, RAMP2, RAMP3, and ADRB2 are not active in the lysosome. As can be seen in Figure 2.2, they are positioned dispersed and away from the majority of the lysosome genes. TM4SF3 is a membrane protein and was assigned the GO annotation via an apparently incorrect "traceable author statement" [75]. For the other genes their products are either a receptor or part of a receptor at the cell surface. LRP2 is a multi-ligand endocytic receptor, which binds molecules and facilitates their internalization by endocytosis [76]. After internalization these endosomes can become lysosomes [77]. RAMP2, RAMP3 and ADRB2 are involved with the activity of receptors whose activity is regulated, upon agonist activation, via internalization and degradation in a lysosome [78, 79]. If the lysosome group would have been better defined without these genes, the score for the lysosome group would have been improved, up to a

median of 0.87 for the set of 100 abstracts per gene. Using the statistical test mentioned earlier the ACS would in this case perform significantly better than using simple gene co-occurrence for the set of 10 abstracts per gene and max 1000 abstracts per gene. Interestingly, RAMP2 and RAMP3 are placed near breast cancer genes (see Figure 2.2). In the abstracts retrieved for these genes, they were not directly implicated in cancer nor directly linked to the breast cancer genes. The only exceptions are two co-occurrences between TP53 and RAMP2 in non-cancer related abstracts. These proteins are involved with the adrenomedullin receptor. Adrenomedullin is an angiogenic factor, and has been linked to a response in (breast) cancer cells in solid tumors that protects against hypoxic cell death [80–82]. Because of the role of adrenomedullin in cancer, the genes that are associated with its receptor are also relevant for the study of cancer [83].

The position of some genes in the ACS is not easily explained in terms of functional relations. The spermatogenesis gene PNMA1, for instance, is placed apart from the spermatogenesis group. Its position was found to be caused by an ambiguity problem. The term MA1 is a synonym for the gene and was used for the PubMed query, but is unfortunately also used for numerous other concepts, such as monoclonal antibody 1. The glycolysis gene PGK1 had very different AUC scores for different builds of the ACS for the same literature set. Apparently it did not find a stable position in the ACS. Study of the abstracts in which PGK1 was mentioned showed that it was referred to in a large number of very different contexts and only occasionally in the context of its role in glycolysis. Given the different contexts in which the gene is mentioned it is hard to imagine how it could be placed in the ACS so that its surroundings correctly reflect all contexts.

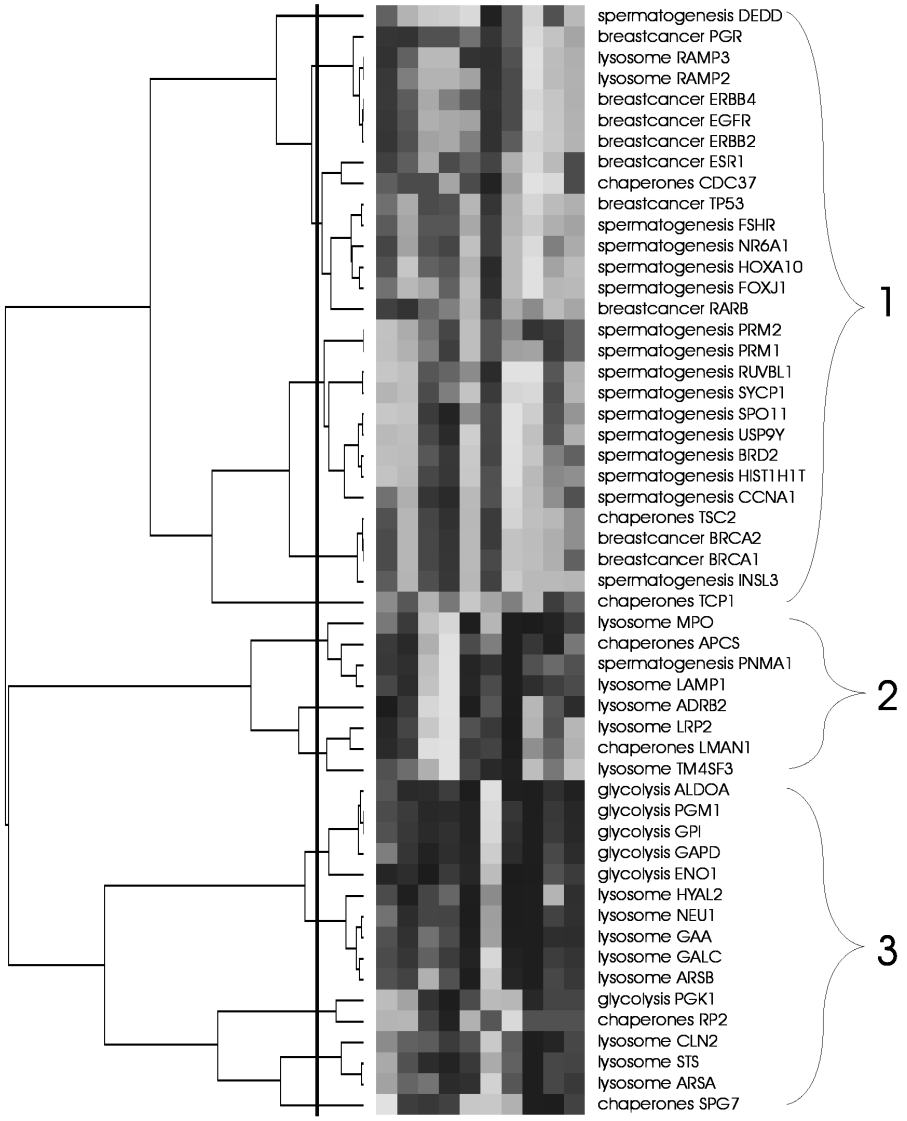
In practical applications of the ACS where the labeling of genes is unknown, clustering algorithms can be used to provide a grouping of the genes of interest. Figure 2.3 shows an example of how a standard clustering technique can be applied. The result for 3 clusters gives a group of genes with roles in cell-cycle control, regulation of gene expression and other forms of DNA-protein interactions (cluster 1), a group mostly containing genes with ambiguity problems or deviating annotations (cluster 2), and a group of enzymes (cluster 3, except for the chaperones). If we allow 14 clusters, 4 clusters contain only one gene and half of the remaining 10 clusters contain only genes which share a functional biological relationship.

2.4 Discussion

Our experiments show that the positioning of genes in the ACS reflects functional biological relationships. Four of the five functional groups that were tested were clustered very well (median AUC > 0.85), if we exclude the aberrant annotations from the lysosome group. Genes with aberrant annotations were correctly placed away from their supposed group members. Interestingly, the ACS placed two of these genes, RAMP2 and RAMP3, in the breast cancer cluster, while there are hardly any co-occurrences between these genes and the breast cancer genes. Study of the literature revealed that a relation to breast cancer is supported by, among others, the role of these genes in angiogenesis. Although not the focus of this study, this is an example of how the ACS could be useful as a knowledge discovery tool.

Both the gene co-occurrence method and the ACS show large differences in performance for different groups. The genes from the breast cancer group have very good scores for both methods. The genes from the glycolysis group score close to perfect for the gene

Figure 2.3: Analysis of the structure of the ACS by hierarchical clustering. The rows represent the different genes and the columns represent the ten axes of the 10D ACS. The marked clusters indicate by approximation: 1, genes with roles in cell-cycle control, regulation of gene expression and other forms of DNA–protein interactions; 2, genes with ambiguity problems or deviating annotation and 3, enzymes. The vertical line in the clustering tree indicates 14 clusters.



co-occurrence method and good for the ACS. For both groups, the number of abstracts retrieved per gene was high when compared to the other groups. The group of chaperones could not be reconstructed by both methods. The poor performance for this group is explained by the scarce reference to their chaperone activity. Clearly, the relationships that the text meta-analysis tools can extract are limited to those described in the literature set, and biomedical abstracts are better represented in Medline than those about basic biology. For the spermatogenesis and lysosome groups it appears that the genes are frequently referred to in the expected context. The ACS method can reproduce both the lysosome and spermatogenesis groups quite well. The gene co-occurrence method on the other hand, scored very poorly for these groups. Most gene co-occurrences between genes do reflect actual biological relations, see also [49]. The low scores for these groups were caused by a lack of actual gene co-occurrences. As these groups could be retrieved with the ACS, this is a clear indication that additional information from the abstract should be used.

The ACS can produce good results with a limited amount of literature, such as only 10 abstracts per gene. This is an important feature as for a large number of genes limited literature is available. Contrary to the gene co-occurrence method though, performance of the ACS was affected by the addition of large amounts of randomly selected abstracts. We hypothesize that the effect of the addition of these abstracts is caused by the appearance of new relationships between concepts. In order to reflect these changes, these concepts will move away from their original positions in the ACS, which apparently disrupts a meaningful clustering. With more relations added the ACS has more problems with accurately representing them in its Euclidean space. This finding makes it necessary to focus the selected literature set on the studied genes, e.g. similar to what we did for the automated selection of literature in this paper. Though this is not a large limitation, it is important to take into account. We are currently working on improving the robustness of the ACS algorithm.

The use of homonymous gene names is wide-spread and has a large impact on text mining applications. The amount of different meanings for one symbol can be quite startling, especially for gene symbols that are two or three letter acronyms, such as ER or GAA [84]. We therefore adapted our literature selection exclude ambiguous gene symbols. For some genes this will have reduced sensitivity, as their preferred terms (according to LocusLink) are ambiguous acronyms, e.g. GAPD, MPO, and PGR. While the selection step did reduce the amount of ambiguity in our literature set, the manual analysis still revealed some problems with indexing, such as GAA which is also a DNA sequence or also, as in the case of PGK1, when the promoter is intended instead of the gene. Clearly methods for disambiguation are needed. Word sense disambiguation has been studied for years (e.g. [85, 86], but only recently a disambiguation tool has been developed specifically for the disambiguation of gene names [87]. A tool for the disambiguation of gene names will be built in to our indexing engine as soon as possible. Another common case of ambiguity is that a gene symbol can refer to the gene itself, its associated mRNA, or relevant proteins. In this article we chose not to distinguish between genes, mRNA or proteins. Such a distinction will sometimes be artificial, is difficult to achieve [88], and is not relevant for our purposes. Currently, a problem with biomedical literature mining tools is the lack of gold standards and established evaluation procedures [56]. The evaluation method we used handles this problem by depending on external and high-quality functional annotation of genes. The use of a genuine list of differentially expressed

genes derived from microarray experiments for a quantitative analysis of performance is difficult and requires a substantial investment, as the annotation process would require extensive reading of scientific literature and extensive expert knowledge. While such annotated datasets are available for several organisms, to our knowledge no exhaustive annotation has been performed on a microarray dataset for human genes. The evaluation set we used was limited in size with its five functional groups and 53 genes, and this allows for the detailed and useful analysis we performed. The five gene groups were drawn from the broad category of functional biological relationships between genes to reflect the broad types of relations in biology. The generalizations concerning ACS performance that we can make based on our results apply only to this category. The manual annotation of genes with GO codes gives a gold standard and has been used by several authors [63, 64]. It is far from perfect though, as our analysis showed that almost half of the genes from the lysosome group had only a remote connection to the lysosome. These cases did have an impact on performance, as the ACS correctly positioned them away from the lysosome group.

The outcome of a DNA microarray experiment can be a sizable set of genes (>100) that are differentially expressed. A tool to quickly identify the genes that according to literature have a functional biological relationship, would facilitate the identification of biological processes underlying the gene expression profile and assist in selecting genes for further analysis. Since distances between genes in the ACS reflect functional biological relatedness, the ACS offers an intuitively appealing presentation that can be of value for molecular biologists. We are currently developing a user-friendly and interactive interface to allow for better browsing of the ACS for genes, related concepts and their relations and to give easy access to descriptions of concepts, database entries for genes and the underlying literature.

In conclusion, the positioning of genes in the ACS reflects functional biological relationships. When the literature set is focused on the studied genes, performance of the ACS is good to excellent. A focused literature set is important, as it was shown that when large amounts of randomly selected abstracts are added, performance decreases. When compared to a simple gene co-occurrence method, the ACS is capable of revealing more functional biological relations and can achieve results with less literature available per gene. The ACS can be of value for researchers studying large numbers of genes, for example in DNA microarray analyses.

2.5 Acknowledgements

We would like to thank Theo Stijnen for his advice on the statistical test.

3

Concept profiles to annotate DNA microarray data

Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation.

R. Jelier[1], G. Jenster[2], L.C.J. Dorssers[3], B.J. Wouters[4], Peter J.M. Hendriksen[2], B. Mons[1], R. Delwel[4] and J.A. Kors[1]

Departments of [1]Medical Informatics, [2]Urology, [3]Pathology and [4] Hematology
Erasmus MC, Rotterdam

BMC Bioinformatics 2007, 8:14

Abstract

High-throughput experiments, such as with DNA microarrays, typically result in hundreds of genes potentially relevant to the process under study, rendering the interpretation of these experiments problematic. Here, we propose and evaluate an approach to find functional associations between large numbers of genes and other biomedical concepts from free-text literature. For each gene, a profile of related concepts is constructed that summarizes the context in which the gene is mentioned in literature. We assign a weight to each concept in the profile based on a likelihood ratio measure. Gene concept profiles can then be clustered to find related genes and other concepts. The experimental validation was done in two steps. We first applied our method on a controlled test set. After this proved to be successful the datasets from two DNA microarray experiments were analyzed in the same way and the results were evaluated by domain experts. The first dataset was a gene-expression profile that characterizes the cancer cells of a group of acute myeloid leukemia patients. For this group of patients the biological background of the cancer cells is largely unknown. Using our methodology we found an association of these cells to monocytes, which agreed with other experimental evidence. The second data set consisted of differentially expressed genes following androgen receptor stimulation in a prostate cancer cell line. Based on the analysis we put forward a hypothesis about the biological processes induced in these studied cells: secretory lysosomes are involved in the production of prostatic fluid and their development and/or secretion are androgen-regulated processes. Our method can be used to analyze DNA microarray datasets based on information explicitly and implicitly available in the literature. We provide a publicly available tool, dubbed Anni, for this purpose.

3.1 Background

The outcome of high-throughput experiments, such as DNA microarray experiments, is typically a list of hundreds of genes that could be relevant to the studied phenomenon. Further analysis is required to relate the genes to relevant biological processes and to identify potentially interesting relationships between the genes. In the early days of DNA microarray data analysis, extracting the required information about genes depended solely on researchers retrieving information from the huge corpus of scientific literature. Nowadays, the need for computational support in the interpretation of high-throughput experiments has become widely recognized.

However, much of the knowledge on genes and proteins is locked in unstructured free text and cannot be used directly in computational systems. To make this knowledge more accessible, several databases have become available that offer structured information on genes and proteins. These databases are either public, e.g. the databases offered by the Gene Ontology Annotation (GOA) project [23] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) project [24], or corporate, e.g. as delivered by GeneGO (`www.genego.com`) and Ingenuity (`www.ingenuity.com`). For a large part, these databases are filled with manually encoded information generated by experts reading scientific literature. Manual encoding is generally considered a reliable method for extracting information from literature, but due to its labor-intensive nature it is limited in scope and flexibility. Complementary to manual encoding, research effort is currently spent on text-mining: the development of computerized algorithms for extracting information from scientific literature [26]. Automated methods have the advantage of speed and adaptability, with the challenging obligation to achieve both high precision and recall.

In text-mining, broadly two approaches can be distinguished. One approach is focused on the extraction of explicitly stated direct relationships between genes and other biomedical concepts. Early proposed systems for this task were based on the co-occurrence of terms in texts [49, 89]. Currently, the grammatical structure in a sentence is typically used for the task of relation mining and a wide variety of techniques has been developed. These techniques range from the detection of simple patterns such as "protein A - action X - protein B" [32, 33], to the complete parsing of whole sentences [36, 37]. The other approach is focused on the identification of indirect associations between concepts, such as genes. For instance, two genes can be found to have an association, because they are described in separate papers to be involved in the same biological process. To retrieve such indirect associations, the explicit, direct associations of the genes are compared. In this approach, syntactic structures are typically ignored, and only the statistics of occurrences and co-occurrences of words or terms in a text come into play.

Here we focus on the second approach. Several co-occurrence based methods have been developed for the analysis of DNA microarray data. GEISHA [90] took a cluster of genes from a DNA microarray data analysis. The system annotated this cluster with the most discriminant terms, and also retrieved relevant co-occurrences, sentences, and abstracts. The system was word-based but automatically identified common word combinations and treated them as single concepts. Shatkay et al. [50] used a kernel document to represent a gene, and used this document to retrieve a set of similar documents. A list of keywords was generated to summarize the recurring theme in the genes' sets of retrieved documents. Subsequently, genes were associated to each other by comparing the genes' sets of retrieved documents. Raychaudhuri et al. [91] analyzed a list of genes by identifying clusters

of genes that show "functional coherence" according to their literature-based neighbor divergence measure. We introduced the associative concept space (ACS) [92] as an aid to find associations between genes for microarray data analysis. The algorithm positioned concepts, in an iterative process, in a virtual space based on co-occurrence information. The idea behind the ACS is that concepts that are placed close to each other will be more likely to share an actual semantic relationship and the visualized ACS allowed browsing for associations between concepts, which is intuitively appealing. Several authors [58, 93–96] employed the vector space model, in which a gene is represented by means of a vector that characterizes a set of texts associated with the gene. The methods varied in the features, or dimensions, of the vector. Chaussabel and Sher [58] used a simple word-based approach to generate a list of co-occurring words for each gene. For the analysis of a list of genes, they attempted to bring to light interesting co-occurrence patterns by clustering both the genes and the co-occurring words. Glennison et al. [93] used concepts from a thesaurus as features, and identified terms in texts referring to thesaurus concepts. They used five thesauri to obtain different views on the associations of a gene and used clustering to find genes with similar profiles from a gene list. Others used factorization techniques to reduce the high dimensionality encountered when using words or concepts as features: Küffner et al. and Homayouni et al. used singular value decomposition [94, 95] and Chagoyen et al. employed non-negative matrix factorization [96]. The claim is that reduction of the dimensionality in this manner leads to a more robust data-analysis, which is less sensitive to sparse and noisy data [96].

From a user's perspective, the current approaches leave several requirements unfulfilled. For example, the ACS and Raychaudhuri methods suffer from a lack of transparency, i.e., a user will not easily understand how the programs come to their associations, which is important to know in an actual research setting. Transparency is also at stake when using factorization in vector space approaches, as it is not clear what the newly defined dimensions mean, or even whether the have a semantic interpretation at all. The methods described by Glennison and Chaussabel and Sher are transparent but use empirical methods for the weighting of concepts, which have problematic statistical properties (see Discussion section for more information). Also, it would be desirable for a user to have more control on which concepts or words are used to compute an association than is possible in the mentioned approaches.

Our aim in this paper is to create a text-mining system for the interpretation of gene lists derived from DNA microarray data that is transparent. Furthermore, in contrast to many earlier published text-mining systems, we will apply the system to actual research problems, in cooperation with molecular biologists. The approach we propose finds associations between genes by means of concept (co-) occurrence statistics and employs the vector space model, similar to Glennison et al. [93]. For each gene we generate a vector of weights, which we refer to as a concept profile. The features in the concept profile are thesaurus concepts that characterize a set of documents associated with the gene. A thesaurus concept is an entity with a definition and a set of terms that are used in texts, to refer to the concept. Every concept is also assigned a semantic type, such as "disease" or "gene". The set of concepts used in the concept profiles is filtered by semantic type using a user defined semantic filter. An important issue is the selection of the measure to weigh the association of a concept in a profile. The weight should distinguish between a concept that co-occurs through chance with the concept of interest and a concept with a semantically interesting association. With this in mind we adopted a test-based method

based on likelihood ratios [97], which has been successfully used for the identification of interesting collocations [98]. Compared to other test-based methods, the likelihood ratio does not require the data to have a normal distribution and is known to yield good results even on small samples. We developed a program called Anni to work with the concept profiles. With this program, genes associated with similar topics in literature are identified by hierarchical clustering of the corresponding gene concept profiles. Anni has a high degree of transparency. It provides for every identified cluster Anni a coherence measure, and also a p-value to illustrate how exceptional the cluster is, and a complete annotation of the underlying overlap of the concept profiles. Also, a link to the underlying texts is provided for all associations in the concept profiles.

We evaluated the method in two steps. Firstly, we present an evaluation based on a controlled test set and compare it to our earlier published ACS algorithm [92]. Secondly, we give a systematic analysis of the data from two DNA microarray experiments and evaluate the results together with domain experts.

3.2 Methods

Literature selection and indexing

We selected 2,585,901 abstracts with a Pubmed query for protein or gene mentioned together with mammals. MEDLINE titles, MeSH headings, and abstracts, if available, were indexed using Collexis software (`http://www.collexis.com` and [68]). In this context, indexing means the identification of references to thesaurus concepts in text and mapping these references to the concepts. Prior to indexing we removed stop words. All words are mapped to the uninflected form produced by the normalizer of the lexical variant generator [67]. The thesaurus we used for indexing was composed of two parts: MeSH and a human gene thesaurus derived from multiple databases [99]. For MeSH we used the UMLS semantic types [22] to select concepts that convey relevant biological information about genes. The filter was developed by molecular biologists and the selected semantic types are given in Additional file 4. This filtering facilitated the interpretation of the profiles and also slightly increased performance on our test set (data not shown). The gene thesaurus was expanded by rewrite rules to take into account common spelling variations [100]. For instance, numbers were replaced with roman numerals and vice versa, and hyphens before numbers at the end of gene symbols were inserted or removed (e.g. “WAF1” was rewritten as “WAF-1” and added as a synonym). Then, potentially highly ambiguous terms (less than five characters, none of them a digit) were removed in order to obtain a high precision on gene recognition. Gene symbols or full gene names that refer to more than one gene in the thesaurus were rejected as well.

ACS

The ACS algorithm has been described in detail before [61] and was developed to be applied for knowledge discovery. Briefly, it is a Hebbian-type of learning algorithm that in an iterative process positions the thesaurus concepts in a multidimensional Euclidean space. In this space the dimensions do not take a specific meaning, but just allow the positioning of the concepts relative to each other. The position of a concept follows from the mapping of co-occurrence relations (paths) between concepts to distances. A distance

between two concepts will not only reflect the co-occurrence of the two concepts, a one-step relation, but also indirect, multi-step relations between the two concepts. As the distance between concepts reflects the strength of both one- and multi-step co-occurrence paths between the concepts, it is possible that concepts are placed close to each other that do not have a direct co-occurrence. The idea behind the ACS is that we may postulate in such a case that there is an actual association between these concepts, which has not been reported in literature.

For the construction of the ACS we used a selection of literature. For the test set for each gene a maximum of 1000 randomly selected abstracts mentioning the gene are included. For the ACS we used a vector format to represent documents with term frequency * inverse document frequency weighting and standard algorithm settings [92].

Concept-profile generation

A concept profile of gene i is an M-dimensional vector $\mathbf{w_i} = (w_{i1}, w_{i2}, \ldots, w_{iM})$ where M is the number of concepts in the thesaurus. The weight w_{ij} for a concept j in this profile indicates the strength of its association to the concept i. The weights in a concept profile for concept i are derived from the set of documents in which concept i occurs. To obtain w_{ij} we employ the log likelihood ratio measure [98]. Two hypotheses are used: 1. The probability of occurrence of concept j is independent of the occurrence of concept i; 2. The probability of occurrence of concept j is dependent of the occurrence of concept i. For each hypothesis a likelihood is calculated based on the observed data using the binomial distribution. The ratio of these likelihoods tells us how much more likely one hypothesis is over the other, or, in other words, how sure we are that there is a dependency. A feature of the log likelihood ratio is that it behaves relatively well for sparse data [97], which is an advantage in our case.

The following equations give the likelihood ratio λ of concepts i and j:

$$\lambda\left(i,j\right) = \frac{L(n_{ij}, n_i, p)L(n_j - n_{ij}, N - n_i, p)}{L(n_{ij}, n_i, p_1)L(n_j - n_{ij}, N - n_i, p_2)}$$

with n_i and n_j the number of documents in which concepts i and j occur, n_{ij} the number of documents in which both concepts occur, N is the number of documents in the corpus, $p = \frac{n_j}{N}$, $p_1 = \frac{n_{ij}}{n_i}$, $p_2 = \frac{n_j - n_{ij}}{N - n_i}$, and $L(k,l,x) = x^k(1-x)^{l-k}$. A feature of likelihood ratios is that -2 times the log of the likelihood ratio is asymptotically χ^2 distributed [98], which can be used to test whether there is a statistically significant divergence from independence. The weight of concept j in the concept profile of concept i is given by:

$$w_{ij} = \frac{\log \lambda\left(i,j\right)}{L}$$

L is the theoretical maximum score of $\log \lambda$, which is obtained when a concept always and only occurs together with concept i. This factor normalizes for the effects of the occurrence rate of concept i, which is convenient when comparing weights between profiles.

For every concept co-occurring with concept i we calculated the log likelihood ratio, but in order for a concept to be included in the concept profile the null hypothesis (the occurrence of j is independent of the occurrence of i) has to be rejected at a significance level of 0.005. For efficiency reasons we included only the most significant concepts to a maximum of 200 concepts.

Associations between concepts are calculated based on concept profiles using cosine similarity scores [101].

The Anni system

In order to analyze a list of genes by means of their concept profiles we developed 'Anni'. The tool retrieves and displays the concept profile of a gene and can also characterize any combination of genes. The components of the Anni system are two databases and a web-based graphical user interface. The first database contains concept profiles for human genes. The second database contains the indexed literature underlying the concept profiles, which is used in the system to identify the documents supporting the associations in a concept profile. The interface provides the following functionality: 1. The user can specify a list of genes to analyze based Affymetrix, Entrez Gene or Swiss-Prot identifiers; 2. Groups of genes with similar profiles can be found using hierarchical clustering. As the input for the clustering algorithm, we use for each gene in the input list, the cosine scores between the concept profiles of this gene and the other genes. We used mean linkage hierarchical clustering with cosine as similarity metric; 3. An identified cluster of genes is given a coherence measure, the average of the cosine scores of all possible pairs within the cluster. To assess the significance of the average cosine score we give the probability that the same score or higher would be found in a randomly formed group of the same size. This probability was determined from the distribution of scores from a 10000-fold random sampling of groups of gene profiles; 4. A cluster of genes is characterized by showing the relative contribution of individual concepts as a percentage. In addition the weights of these concepts in the concept profiles are shown, which facilitates an easy assessment of the similarity of the profiles; 5. For every association in a concept profile a link to the underlying literature is provided.

For clarity, the only overlap between the Anni system and the ACS is the underlying database of indexed documents and the used thesaurus. Apart from this, the systems share no methodology.

To analyze gene lists in a standardized manner we used the following protocol. All clusters with a cosine coefficient greater than 0,15 and containing at least three genes were analyzed. The probability that the average cosine score was found by chance should be $< 0{,}005$. A cluster may be split into smaller, more consistent clusters, if there are smaller clusters with distinct common functions.

Evaluation

For comparison of the ACS and the concept profile method we used the test set and the evaluation procedure as described in [92]. The test set was made by pooling five groups of genes that share a biological relationship. Each group represented a different aspect of gene biology, being function, organelle, biological process, metabolic pathway, or association with a disease. Only human genes were taken into consideration. The selected groups are: spermatogenesis, 15 genes; lysosome; 10 genes; chaperone activity, 7 genes; breast cancer, 9 genes; glycolysis, 6 genes. For the evaluation, both the ACS and the concept profile method were employed to produce a ranking of the set of genes relative to one so-called seed gene. All genes in turn served as a seed, producing a ranking for each of the other 46 genes in our set. For the concept profile method, genes were rank-ordered

according to the cosine similarity scores [101] between the concept profile vector of the genes and the seed gene. Ties were ordered randomly. For the ACS, genes from the set were rank-ordered according to their Euclidean distances to the seed gene. For each gene a receiver operating characteristics (ROC) curve was then constructed [69]. The area under the curve (AUC) was used as a performance measure [70]. This value varies between 0 and 1. An AUC of 1 represents perfect ordering, i.e. all genes belonging to the group of the seed gene are at the top of the list followed by the other genes. The AUC has the useful property that a value of 0,5 represents random ordering [70]. This property provides us, in a way, with a built-in negative control.

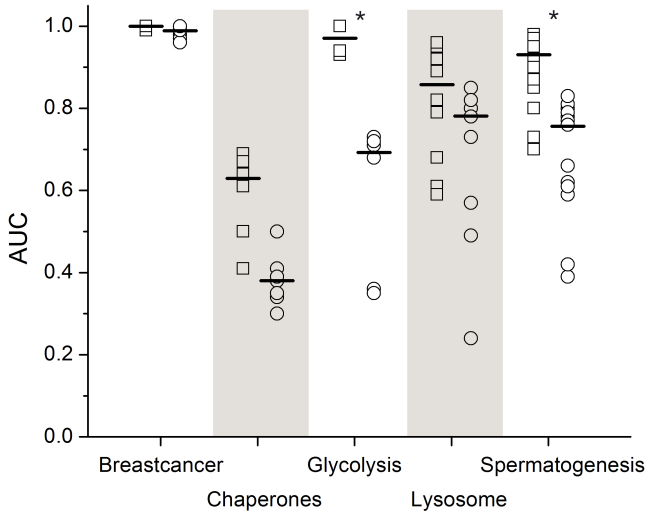
To determine whether the AUC scores differed significantly between the two methods, we used the non-parametric Wilcoxon signed ranks test. The test requires the AUC scores of the genes to be independent. Because this is not true in this case, we applied bootstrapping [71] to estimate the distribution of the Wilcoxon test statistic. We generated 100 new sets of genes by sampling genes from the original set with replacement. The sampling was stratified over the five gene groups to obtain groups of equal size as in the original set. AUCs were calculated for both methods, and the Wilcoxon signed ranks test was applied to measure the difference between the two methods per gene group. The results obtained for the 100 sets were used to determine if the two methods differ in performance at the 0,05 level.

Description DNA microarray data sets

The first set consisted of data from a recent study about prognostically useful gene-expression profiles in AML [10]. Gene expression in leukemic blast cells from 285 patients was measured. Clustering of the gene expression data resulted in 16 groups of patients with distinct profiles. For each cluster a profile of genes with the most distinguishing gene expression patterns was made with the significance analysis of microarray (SAM) method. For our analysis genes with a SAM score higher than 4 or lower than -4 were selected. Data acquisition and processing are described in detail in the original paper.

The second set consisted of differentially expressed genes following the agonistic stimulation of the androgen receptor in a prostate cancer cells. The androgen-dependent LNCaP prostate cancer cell line was maintained in RPMI media with 5% fetal calf serum and penicillin/streptomycin (Invitrogen, Merelbeke, Belgium). Before R1881 treatment, cells were androgen-deprived for 72 hours in a medium containing 5% dextran-filtered, charcoal-stripped fetal calf serum. After androgen deprivation, the medium was supplemented for 2, 4, 6 or 8 hours with 1 nM synthetic androgen R1881 or ethanol vehicle as the control. Three μg of total RNA was used for a T7 based linear mRNA amplification protocol [102]. Two micrograms of amplified RNA were used to produce Cy3- or Cy5-labeled cDNA. cDNAs from R1881-treated and control cells were compared directly by hybridization to the same microarray. This was done in duplicate with reversed Cy dye labeling. The cDNA microarrays were manufactured at the Central Microarray Facility of the Netherlands Cancer Institute (NKI, Amsterdam, The Netherlands) and contained over 18,000 features that have been selected from the Research Genetics Human Sequence Verified Library (Invitrogen). Normalization of spot intensities was performed using R-routines (Lowess method) using the NKI Microarray Normalization Tools (http://dexter.nki.nl). Genes were considered to be up or down-regulated by R1881 when both dye swaps gave a ratio larger than 1,62 (2log 0,7) for at least one time point. The data

Figure 3.1: Area under the curve scores for individual genes per group for the concept profile method (open boxes) and the ACS (open circles). An asterisk above a group indicates that the difference in performance of the two methods is statistically significant (at the 0.05 level).



have been deposited in NCBIs Gene Expression Omnibus (`www.ncbi.nlm.nih.gov/geo/`) and are accessible through GEO Series accession number GSE4027 and GSE1159.

3.3 Results

Performance evaluation on a controlled test set

The concept profile method and the ACS were compared based on a controlled test set, as described before [92]. The test set was made by pooling five groups of genes that share a biological relationship: chaperone activity (7 genes), glycolysis (6), breast cancer (9), spermatogenesis (15) and lysosome (10). A table with all 47 genes is given in Additional file 1. For each gene the methods were evaluated on their ability to distinguish between group members and non-group members. Receiver operating characteristics (ROC) curves were constructed for every gene and the area under the ROC curve (AUC) supplied the evaluation measure. As can be seen in Figure 2.1, the concept profile method has high AUC scores for 4 out of 5 gene groups. It significantly outperforms the ACS in 2 out of 5 groups and has higher median scores for the other groups as well. Overall, taking the genes from all groups together, the concept profile method significantly outperforms the ACS ($p<0{,}05$). As discussed in [92], the poor score for the chaperone group is caused by the scarce reference in the literature to this function. We examined with Anni the concept profiles of each gene group and looked for the ranking of the concept that characterizes the group's shared biological association. In their respective group annotation the concept "breast neoplasms" was ranked first, "lysosome" came second, "spermatogenesis" second, "molecular chaperones" first and "glycolysis" fifth. All groups, with the exception of the

chaperone group, had significant cohesion scores ($p<0,05$).

DNA microarray dataset 1: Gene expression profiles of acute myeloid leukemia patients

Based on gene-expression profiles of leukemic cells, 285 acute myeloid leukemia (AML) patients were separated into 16 groups [10]. Several of these groups coincided with known classes of AML patients. AML cases are classified by the occurrence of genomic aberrations in the leukemic cells. According to the report, group 5, one of the larger groups with 61 patients, does not associate with a known karyotypic abnormality and little is known about the background of the leukemic cells in this cluster [10]. The set of genes that characterize this patient group were analyzed with the literature-based clustering provided by Anni. We sought to find shared processes and other associations that could be indicative for the background of the leukemic cells.

A total of 42 gene clusters were found for the 992 genes in patient group 5 (the complete Anni analysis is included as Additional file 2). Based on this annotation we put forward the hypothesis of an association of patient group 5 to monocytes on the following grounds: Two clusters of genes were found to be involved in phagocytosis: a cluster of cathepsins and a cluster associated with respiratory burst. Of the cathepsins, *CTSS*, *CTSB* and *CTSL* are implicated in antigen presentation on the surface of cells from the monocytic lineage [103, 104]. Respiratory burst is a process characteristic for a sub-type of blood cells called phagocytes. From the group of phagocytes, we can exclude granulocytes as we identified a cluster associated with the major histocompatibility complex class 2 (MHC II). The presence of MHC II is a distinguishing factor between the myeloid cell types for it is absent in neutrophils, basophils and eosinophils [105]. This leaves us with monocytes.

Also within several other clusters genes were found to have an association with monocytes in their concept profile. Several of these genes indeed had a functional relationship with monocytes. A cluster of chemokines and chemokine receptors is associated with chemotaxis and macrophage inflammatory proteins. From this cluster *CCR1* and *CCR2* are involved in monocyte chemotaxis [106]. A cluster associated with antigens contained Cluster Differentiation genes, and *CD14* is a monocyte lineage specific marker. The immunologic receptor cluster contained a number of genes strongly associated with monocytes. One of these, *LILRB4* (*ILT3*) is a cell surface molecule selectively expressed by the myeloid antigen presenting cells of the monocytic lineage [107]. As we did not find clusters characteristic for other myeloid cell-types, such as erythrocyte precursors, we postulate that AML patient group 5 is associated with precursor cells from the monocytic lineage.

In the original paper by Valk et al. [10] morphological characteristics of the leukemic cells were presented by means of the widely used 8 subtypes of the French-American-British (FAB) classification system. Using this classification we could verify whether our postulate is in concordance with the cells' appearance. In the study, patient group 5 contained specimens with FAB M4 or M5 subtypes. Specimens with an M4 classification contain cells that show granulocytic or monocytic maturation, and those with M5 have cells classified as monoblastic or monocytic.

Finally, we verified the presence of the mentioned genes and clusters in the other patient groups (table 3.1). There is a considerable overlap with patient group 9, but not with other groups. According to the original paper, group 9 is indeed also composed of a mixture of the FAB classifications M4 and M5.

Table 3.1: **Occurrence of monocyte specific clusters in patient groups.**

	Patient groups										
Cluster descriptions	3	4	5	6	7	8	9	10	12	13	16
MHC 2	-	-	4↑	13↓	9↓	-	3 ↑	-	7↓	4↑	-
Cathepsins	-	11↓	9↑	-	3↓	-	-	4 ↓	3*	3↓	-
NADPH oxidase, respiratory burst	-	-	4↑	-	4↓	-	6↑	-	-	-	-
Gene names	3	4	5	6	7	8	9	10	12	13	16
CCR1	↓	↓	↑	-	-	-	↑	-	-	-	-
CCR2	↓	-	↑	-	-	-	↑	-	-	-	-
CD14	-	-	↑	-	-	-	↑	-	-	-	-
LILRB4	-	-	↑	-	-	-	-	-	-	-	↑

* The cathepsins of group 12 include 1 down-regulated and 2 up-regulated genes.

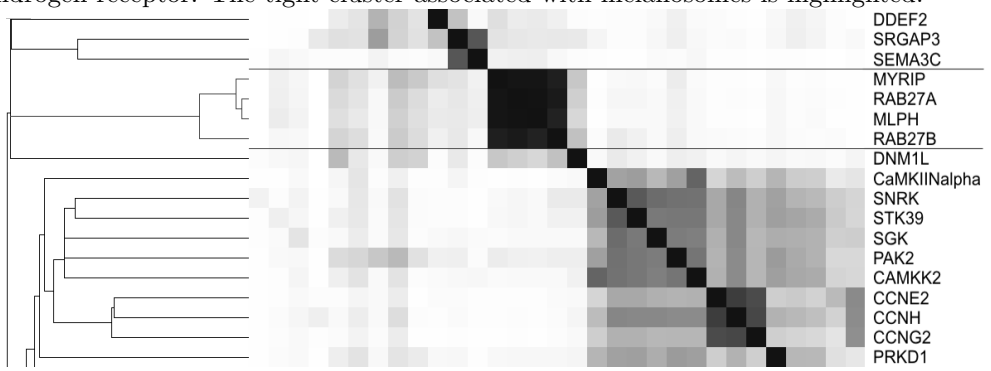
The upper half of the table shows for the patient groups the presence of the clusters of genes that were discussed for patient group 5. Several patient groups are not shown as the SAM analysis only yielded very few distinguishing genes. The size of the clusters is indicated and the arrows indicate if the genes are up- or down regulated. The lower half of the table shows the presence of the genes that were discussed in the text.

DNA microarray dataset 2: Agonistic stimulation of the androgen receptor

In the second evaluation experiment on microarray data, we used Anni for the analysis of the list of 221 differentially expressed genes as measured with a DNA microarray following the agonistic stimulation of the androgen receptor in a prostate cancer cell line. The androgen receptor is a transcription factor, activated by the androgens testosterone and dihydrotestosterone and is responsible for development and maintenance of the function of the normal prostate and for growth of early stage prostatic cancer [108]. The complete annotation of the mentioned gene list is given in Additional file 3.

The tightest cluster of genes consists of the genes *RAB27A, RAB27B, MYRIP* and *MLPH*, see Figure 3.2, and has an average cosine of 0.57, indicating a very strong within-cluster correlation. In table 3.2 we show which concepts contribute the most to this average cosine score. The four gene concepts themselves are in the top of this list, which implies that these genes are regularly co-published. Other notable concepts are several myosin related concepts, the concepts melanosomes and melanocytes, and the concepts exocytosis and secretory vesicles. According to the MeSH vocabulary definitions: Myosin Type V is involved in organelle transport and membrane targeting. Melanosomes are melanin containing vesicles found in melanocytes and they are involved in skin pigmentation. The concepts exocytosis and secretory vesicles are both associated with the cellular release of material with membrane-limited vesicles. With a manual check of the literature linked by Anni to the four genes, we verified that the genes are indeed involved in the same process and their biological activity is in concord with the calculated annotation: all genes are associated with in the transport of melanosomes to the cell surface by interaction with myosin type V [109–111]. Certainly, there is no pigmentation in the prostate, but what quickly becomes apparent from literature is that these genes more generally deploy their activity in secretory lysosomes, of which melanosomes are only one example [112]. Secretory lysosomes are modified lysosomes that can proceed to regulated secretion

Figure 3.2: Fragment of the hierarchical clustering tree and heatmap based on the concept profiles for the genes differentially expressed following the agonistic stimulation of the androgen receptor. The tight cluster associated with melanosomes is highlighted.



in response to external stimuli, with a special role for *RAB27A* [109, 112, 113]. Terms associated with lysosomal processing are also part of the annotation, but are not shown in Table 3.2 since their contribution was below 0,5%.

Secretory lysosomes may play their part in the major function of the prostate: the production and secretion of prostatic fluid. Several of the substances found in prostatic fluid point to a role for secretory lysosomes. Some of the secreted enzymes may be lysosomal; prostate acid phosphatase has for instance been localized in the lysosome [114]. Alternatively, *RAB27A* and associated proteins may be involved in the secretion of small vesicles called prostasomes. The latter hypothesis is supported by the identification of the *RAB27A* protein in prostasomes by proteome analysis [115]. It appears the potential roles of *RAB27A* and secretory lysosomes in the secretory processes of the prostate have currently not yet been investigated or reported. Semantic analysis of the literature associated with the genes differentially expressed in the microarray experiment, thus leads us to the novel hypothesis that secretory lysosomes are involved in the production of prostatic fluid and that their development and/or secretion are androgen-regulated processes.

3.4 Discussion

We evaluated our concept profiling method in two steps. Firstly, we applied it to a controlled test set and compared its performance to that of our previously published ACS method [61, 92]. The concept profiling method obtained high median scores for 4 of the 5 groups in the controlled test set, and performed significantly better than the ACS method for 2 groups, as well as overall. Secondly, we applied our method to actual research problems and annotated two DNA microarray datasets.

The first DNA microarray data set we analyzed, was the gene expression profile of the leukemic cells of a group of AML patients as identified in [10]. Little is known about the background of the leukemic cells in this cluster. With the Anni annotation and the underlying literature it was possible to identify several groups of genes and individual genes in the profile that indicate an association of the leukemic cells to cells of the monocytic lineage. This finding was in concordance with the morphological classification of the

Table 3.2: Concepts representative for the cluster RAB27B, MYRIP, MLPH, RAB27A as given by Anni.

		Weight in concept profile			
Concept Name	Contribution (%)	RAB27B	MYRIP	MLPH	RAB27A
RAB27A	52,17	0,61	0,74	0,73	1
MLPH	11,16	-	0,44	1	0,29
Myosin Type V	7,22	0,04	0,68	0,4	,0,22
Melanosomes	6,7	0,12	0,3	0,47	0,27
RAB27B	4,06	1	0,14	-	0,11
MYRIP	2,98	0,07	1	0,09	0,06
Melanocytes	2,73	0,13	0,14	0,28	0,17
Myosins	2,33	0,04	0,38	0,22	0,12
Myosin Heavy Chains	1,72	-	0,46	0,18	0,09
GTP Phosphohydrolases	1,31	0,17	0,23	0,04	0,08
Actins	1,17	0,05	0,32	0,12	0,06
Exocytosis	0,87	0,08	0,12	0,08	0,12
Secretory Vesicles	0,68	0,07	0,16	0,06	0,09
Carrier Proteins	0,59	-	0,11	0,17	0,09
Organelles	0,54	0,11	-	0,12	0,09
rab GTP-Binding Proteins	0,52	0,16	-	0,04	0,12

In the first column the concept names are shown, in the second the percentage contribution of this concept to the average cosine score (0,57) of this group. We limited the number of concepts to a contribution of 0,5 % to the average cosine score. The remaining columns show the weight of the concepts in the concept profiles of the genes whose names are shown in the column headings. These weights form the basis of the clustering of the 4 genes.

cells. The second data set consisted of a list of differentially expressed genes following the agonistic stimulation of the androgen receptor in a prostate cancer cell line. The Anni annotation revealed a cluster associated with, amongst others, melanosomes and secretory vesicles. Based on this finding and the underlying literature we formulated a hypothesis about the role of secretory lysosomes in prostate function. We conclude that Anni can be successfully used by molecular biologists studying DNA microarray datasets as a tool to automatically use the explicit and implicit information in literature.

The projected use of our method is the analysis of gene lists from high-throughput experiments. Our method is a useful addition to the current tool suite based on manual annotations or on automatic relation mining by analysis of the grammatical structure of sentences. Manual approaches, such as the GOA project, are limited in focus and tend to be incomplete due to the labor intensive annotation process. For example, in the case of the four melanosome-associated genes that we discussed, only *RAB27A* and *RAB27B* have, at the time of writing, a manual annotation by GOA. For these two genes the only curated annotation concerns their GTPase activity, even though there are numerous articles in Pubmed describing other features for which there are relevant Gene Ontology (GO) concepts, such as “melanosome”. The computerized extraction of relations suffers from the limitation that the systems need to be trained to retrieve specific relations and entities. Hence, if the extraction algorithm is not trained for a specific relation it is likely to miss it. For example, the company Ariadne Genomics has constructed a relation database based on extensive natural language parsing (see e.g. [116]). They focused on the recognition of proteins and small molecules and their relationships. For both entities, at the time of

writing, their database contains approximately 50,000 entries, but for biological processes there are only 263 entries which is a mere fraction of the more than 10,000 recognized in GO. The point is that the co-occurrence based method is simple and versatile. Associations can be retrieved between any two concepts once they can be recognized in text. Also the interpretation of associations differs from that of relationships. The association strengths in a concept profile for a concept A quantitatively reflect the statistical over-representation of concepts in texts in which concept A occurs. Hence, a concept profile of a particular concept can be seen as a view on the literature in which the concept is mentioned. This feature has value from an information retrieval point of view. The use of associations is also casting the net wide: not only are specific functional relationships retrieved, all significant associations between entities are retrieved, potentially even those not made explicit by the authors. This feature has been exploited for knowledge discovery purposes (see e.g. [41]).

Compared to other co-occurrence based approaches with similar objectives, our method may be considered an improvement on several points:

1. Anni was developed to be transparent, i.e. it is visible how the system comes to its associations. Transparency is a known problem with the ACS. The ACS was developed for knowledge discovery purposes and it uses an iterative algorithm to map concepts to a multi-dimensional space using concept co-occurrence data as input. In this space, the distance between concepts reflects the strength of one- and multi-step co-occurrence paths between the concepts. When applying the ACS, transparency was a problem for users of the system, as tracing distances between concepts back to the underlying literature was challenging. Compared to ACS, the Anni system is much more transparent: Anni provides a link to the underlying texts for every association between concepts. The system provides a coherence measure for a group of genes as well as the probability of a chance-occurrence of the group. Additionally, Anni illustrates the contribution of specific concepts to the coherence measure and shows the overlap between the concept profiles of the group members. It is, therefore, traceable why genes are clustered together. It is also trackable why certain concepts are associated with genes as the underlying articles can be accessed. In this aspect, Anni also contrasts favorably with, for instance, systems that use dimension reduction techniques [94–96]. Dimension reduction leaves the meaning of the dimensions unclear, and makes it difficult to verify, by consulting the underlying texts, whether the association between a gene and a dimension is true or relevant.

2. We used the controlled vocabulary Medical Subject Headings (MeSH) in addition to a gene thesaurus to identify concepts in texts. The use of thesauri allows the identification of multi-word concepts and the mapping of synonyms for the same concept, which reduces the noise caused by natural language variation. In addition, a thesaurus maps words or phrases to an abstract concept, thereby connecting it to all information available from other sources linked to this concept. For instance, a reference to a gene can be linked to its sequence or, as shown in this paper, semantic types can be used for filtering, and definitions of a concept can be used for interpretation. We used the semantic types associated with the biomedical concepts to focus the concept profiles on our area of interest. Several earlier approaches did not use a thesaurus for identifying biomedical concepts other than genes or proteins, e.g. [58]. The semantic filtering we used is more precise and adaptable than using different vocabularies as was done by [93].

3. The log-likelihood measure we use for the weighting of the associations between concepts is an important feature of our approach and has a sound statistical foundation.

Some of the empirical approaches described in literature have properties that can be considered problematic. For example, Glenisson *et al.* [93] took the normalized inverse document frequency as the weight for a concept in a document. To produce the weight of a concept in a concept profile based on a selected set of documents, they averaged the concept's weight over the set. However, this procedure favors more frequently occurring concepts. Suppose two concepts in a large set of documents occur with rates r_1 and r_2, with $r_1 < r_2$, and thus for their weights will hold $w_1 > w_2$ in individual documents. When averaging the weights in a given subset of documents in which, say, both concepts occur with the same rates r_1 and r_2, then the ratio of their original weights, $\frac{w_1}{w_2}$, will be reduced (by a factor $\frac{r_2}{r_1}$) in the resulting concept profile. This may result in the weight of the more common concept becoming higher than that of the rarer concept.

Our approach had several limitations. Firstly, the thesaurus had to be curated for unnecessarily ambiguous concepts. We chose to do this in order to achieve a better precision, but, especially for genes, this will have reduced our recall. Despite our curation efforts we encountered a small number of errors during our evaluation caused by polysemy, e.g. by gene symbols such as "protein s" as a synonym for the gene *PROS1*. More frequently we encountered errors in the thesaurus caused by errors in the underlying databases, such as "protein-tyrosine kinase" as a synonym for the gene *MUSK*. We expect our approach to further improve with a word-sense disambiguation module, as well as with progressive thesaurus curation. A second limitation in our study is the coverage of the thesaurus. New concepts arise constantly and may be very specifically used by a small group of specialists. Hence, to achieve optimal results for a thesaurus approach an up-to-date and domain-specific thesaurus is mandatory. A more flexible and dynamic approach to thesaurus construction is desirable. A third limitation is inherent in the use of co-occurrences to derive associations between concepts. Associations between concepts based on co-occurrences need not reflect actual biological relationships, even when their co-occurrence rate is far above the chance level.

3.5 Conclusion

Anni was applied to a controlled dataset and to two DNA microarray datasets. We conclude that our method can be used to efficiently analyze a DNA microarray dataset based on both explicit and implicit information in the literature and expect that our system can be useful for the interpretation of high-throughput experiments.

3.6 Acknowledgements

We would like to thank Natasja Dits, Peter Jan Roes and Roel Verhaak for technical assistance. We are greatly indebted to Renske Los and Gerard van Herpen for revising the style of the written English. This study was supported by the Erasmus MC Breedtestrategie and by the Dutch Cancer Society, grant number DDHK 2001-2455.

3.7 Additional Files

This article comes with 4 additional files: Additional file 1.rtf: The controlled test set; Additional file 2.rtf: Annotation of the first DNA microarray dataset; Additional file 3.rtf:

Annotation of the second DNA microarray dataset; Additional file 4.rtf: Semantic types used for filtering. The files can be found online at `http://www.biomedcentral.com/1471-2105/8/14/additional/`.

4

Weighting schemes for concept profiles

Literature-based concept profiles for gene annotation: the issue of weighting

R. Jelier, M. Schuemie, P. Roes, E.M. van Mulligen and J.A. Kors

Department of Medical Informatics
Erasmus MC, Rotterdam

Accepted for publication in the International Journal of Medical Informatics

Abstract

Text-mining has been used to link biomedical concepts, such as genes or biological processes, to each other for annotation purposes or the generation of new hypotheses. To relate two concepts to each other several authors have used the vector space model, as vectors can be compared efficiently and transparently. Using this model, a concept is characterized by a list of associated concepts, together with weights that indicate the strength of the association. The associated concepts in the vectors and their weights are derived from a set of documents linked to the concept of interest. An important issue with this approach is the determination of the weights of the associated concepts. Various schemes have been proposed to determine these weights, but no comparative studies of the different approaches are available. Here we compare several weighting approaches in a large scale classification experiment.

Three different techniques were evaluated: 1. weighting based on averaging, an empirical approach; 2. the log likelihood ratio, a test-based measure; 3. the uncertainty coefficient, an information-theory based measure. The weighting schemes were applied in a system that annotates genes with Gene Ontology codes. As the gold standard for our study we used the annotations provided by the Gene Ontology Annotation project. Classification performance was evaluated by means of the receiver operating characteristics (ROC) curve using the area under the curve (AUC) as the measure of performance.

All methods performed well with median AUC scores greater than 0.84, and scored considerably higher than a binary approach without any weighting. Especially for the more specific Gene Ontology codes excellent performance was observed. The differences between the methods were small when considering the whole experiment. However, the number of documents that were linked to a concept proved to be an important variable. When larger amounts of texts were available for the generation of the concepts' vectors, the performance of the methods diverged considerably, with the uncertainty coefficient then outperforming the two other methods.

4.1 Background

The number of scientific publications is increasing exponentially. In the fields of molecular biology and the biomedical sciences, scientists find themselves unable to read every publication of interest. Additionally, high-throughput experiments on genes and proteins, such as with DNA microarrays, have become common practice in these fields, causing a true information overload. The need for computational support to attempt to manage this information overload has become widely recognized and has spawned a lively area of research.

However, much of the knowledge on genes and proteins is locked in unstructured free text and cannot be used directly in computational systems. To save this several databases have become available that offer structured information on genes and proteins. These databases are either public, e.g. the databases offered by the Gene Ontology Annotation project [23] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) project [24], or commercial, e.g. as offered by GeneGO (www.genego.com) and Ingenuity (www.ingenuity.com). For a large part, these databases are filled with manually encoded information, generated by experts reading the scientific literature. Manual encoding is generally considered a reliable method for extracting information from the literature, but due to its labor-intensive nature, it is necessarily limited in scope and flexibility. Complementary to manual encoding, currently much research effort is spent in the field of text-mining: the development of computerized algorithms for extracting information from the scientific literature [26]. Automated methods have the advantage of speed and adaptability, though it is more difficult to achieve high precision and recall.

In text-mining two main approaches can be distinguished. One approach focuses on the extraction of precise relationships between genes and other biomedical concepts, using techniques varying from the detection of simple patterns such as "protein A - action X - protein B" [32, 33], to the complete parsing of whole sentences [36]. The second approach uses the occurrence and co-occurrence statistics of terms from a thesaurus or lexical features, such as words or bi-grams, in a set of documents.

Here we focus on the use of occurrence and co-occurrence information in text-mining. Despite its conceptual simplicity, the approach has proven quite effective in the field of information retrieval and information extraction in the biomedical domain. For example, several authors [49, 58, 91, 93] demonstrated the value of (co-)occurrence based systems for the analysis of DNA microarray data, and Stapley et al. [51] used weighted word counts to predict the sub-cellular location of proteins. The field of literature-based discovery, where the objective is to generate new hypotheses about relationships between concepts, makes ample use of occurrence and co-occurrence statistics (e.g. [40–42]). The approach has also been used to combine textual information with other types of information, typically to achieve specific tasks. For example, Xie et al. [117] use textual information together with sequence homology and information on protein domains to automatically assign Gene Ontology (GO) codes to proteins. Others combine gene expression data with text mining to identify disease genes [118].

In a number of text-mining approaches, concepts are represented by a set of texts related to the concept. Subsequently, concepts are related to each other by comparing the linked sets of texts. To make the comparison of two sets of texts, several authors [41, 51, 93, 119] have used the so-called vector space model to characterize a set of texts. Using this model, a concept is represented by a concept vector: a list of associated concepts, together

with weights that indicate the strength of the association. The associated concepts in the vectors and their weights are derived from the set of documents linked to the concept of interest. These concept-associated vectors, which we will call *concept profiles*, can be used to easily and transparently compare concepts based on underlying literature. Furthermore, patterns of similarity in a set of vectors can efficiently be found, for instance with clustering approaches. However, when using this approach, the determination of the weights in the concept profiles is an issue. Various weighting schemes have been proposed, with a wide range of motivations and statistical properties (see e.g. [58, 93, 120, 121]), but a comparative study of these weighting schemes is lacking. Here we compare three weighting schemes for generating concept profiles:

1. Weighting by averaging, an empirical approach. In this approach each document is characterized by a document vector, a (weighted) list of concepts found in the document. Glennison et al. [93] generated concept profiles by averaging document vectors.

2. The log likelihood ratio, a test-based measure. The log likelihood ratio has been used in statistical natural language processing for collocation discovery [98] and has recently been applied in text-mining [119].

3. The uncertainty coefficient, an information-theory based measure. The uncertainty coefficient is a normalized version of the mutual information measure, which is commonly used to measure stochastic dependence. An adapted mutual information measure was used by Wren [121] for his knowledge discovery system.

To compare the weighting schemes, they were applied in a system that annotates genes with GO codes, a task used before as a benchmark for text-mining systems (e.g. [122–124]). The Gene Ontology was designed to annotate gene products with their associated biological processes, cellular components and molecular functions in a species-independent manner [21]. As the gold standard for our study we used annotations of genes with GO codes as provided by the Gene Ontology Annotation project [23].

4.2 Methods

Corpus and Thesaurus

The corpus of literature for our experiments consisted of 3,072,396 MEDLINE abstracts, selected with the PubMed query "(protein OR gene) AND mammals". We used titles, MeSH headings, and abstracts. Stop words were removed and words were stemmed to their uninflected form by means of the normalizer of the lexical variant generator [67].

We used a thesaurus to identify concepts in texts. The use of a thesaurus allows the identification of multi-word terms and the mapping of synonyms to one concept. In addition, thesaurus concepts can be linked to useful information. For instance, we used the Unified Medical Language System (UMLS) [22] where every concept has been assigned a semantic type. These semantic types can be used to focus the set of identified concepts to an area of interest (see e.g. [41]).

The thesaurus was composed of two parts: the 2004AC version of the UMLS thesaurus [22] and a human gene thesaurus derived from multiple databases [99]. To exclude

irrelevant concepts, two molecular biologists created a list of UMLS semantic types (for the complete list see `www.biomedcentral.com/content/supplementary/1471-2105-8-14-s4.rtf`) relevant for biological information about genes. All concepts with other semantic types were removed from the thesaurus. Following Aronson [125], the UMLS thesaurus was also adapted for efficient natural language processing, avoiding overly ambiguous or duplicate terms, and terms that are very unlikely to be found in natural text. The gene thesaurus was expanded by rewrite rules to take into account common spelling variations [100]. For instance, numbers were replaced with roman numerals and vice versa, and hyphens before numbers at the end of gene symbols were inserted or removed (e.g. "WAF1" was rewritten as "WAF-1" and added as a synonym). Then, potentially highly ambiguous terms (less than five characters, none of them a digit) were removed to obtain a high precision on gene recognition. Gene symbols that refer to more than one gene in the thesaurus were rejected as well.

Concept Profile Methodology

A concept profile of concept i is an M-dimensional vector $\mathbf{w_i} = (w_{i1}, w_{i2}, \ldots, w_{iM})$ where M is the number of concepts in the thesaurus. The weight w_{ij} for a concept j in this profile indicates the strength of its association to the concept i. The weights in a concept profile for concept i are derived from the set of documents associated with concept i, D_i, which is a subset of the total set of documents D. The different weighting schemes that were employed to compute the weights w_{ij} are described below. Also the unit of measurement was varied for the different weighting schemes. Both the number of documents in which a concept occurs (document frequency), as well as the total number of occurrences of a concept in a set of documents (occurrence frequency) were considered. For efficiency reasons only the values of the 1000 highest ranking concepts in the concept profiles are used (their contribution always accounted for more than 99% of the vector length).

Binary Weighting

As a baseline we included a binary "weighting" scheme. This scheme simply gives concepts a weight of 1 if they are found in D_i, and 0 otherwise:

$$w_{ij} = \begin{cases} 1 & \text{if } n_{ij} > 0 \\ 0 & \text{if } n_{ij} = 0 \end{cases}$$

where n_{ij} is the number of documents in D_i in which concept j occurs.

Averaging Document Vectors

In the field of information retrieval a document is commonly represented as a vector $\mathbf{v_d} = (v_{d1}, v_{d2}, \ldots, v_{dM})$, with the weights v_{dj} as a measure of the relevance of concept j in document d. Typically v_{dj} is defined as the product of two factors: a local factor based on d only, e.g. f_{dj}, the number of times concept j occurs in d, and a global weight for concept j based on the whole corpus D. The latter is often defined as the inverse document frequency (IDF):

$$IDF_j = \log_2 \frac{N}{n_j}$$

with N the number of documents in D and $n_j = |\{d \in D | f_{dj} > 0\}|$, the document frequency of concept j in D. Following Glennison et al. [93], w_{ij} is obtained by simply averaging the concept weights of the document vectors of the set of documents D_i.

$$w_{ij} = \frac{1}{N_i} \sum_{d=1}^{N_i} v_{dj}$$

with N_i the number of documents in D_i. For document frequency as the unit of measurement we take IDF_j as v_{dj}, similar to [93].

$$v_{dj} = \begin{cases} IDF_j & \text{if } f_{dj} > 0 \\ 0 & \text{if } f_{dj} = 0 \end{cases}$$

For weighting based on occurrence frequency we chose the augmented normalized term frequency times the inverse document frequency [126]:

$$v_{dj} = \begin{cases} \left(0.5 + 0.5 \cdot \frac{f_{dj}}{f_d}\right) \cdot IDF_j & \text{if } f_{dj} > 0 \\ 0 & \text{if } f_{dj} = 0 \end{cases}$$

Here f_d is the maximum of f_{dj} over all concepts in d. This weighting employs both occurrence frequency and document frequency information, and the method is commonly used to represent documents as concept vectors for information retrieval purposes [127].

Log Likelihood Ratio

The log likelihood ratio measure [98] reflects whether the occurrence rate of concept j in D_i is significantly different from that in the complement of D_i. Two hypotheses are used: 1. The probability of occurrence of concept j is independent of D_i; 2. The probability of occurrence of concept j is dependent of D_i. For each hypothesis a likelihood is calculated based on the observed data, using the binomial distribution. The ratio of these likelihoods tells us how much more likely one hypothesis is over the other, or, in other words, how sure we are that there is a dependency. A feature of the log likelihood ratio is that it behaves relatively well for sparse data [97], which is an advantage in our case.

The following equations give the likelihood ratio λ of concepts i and j using document frequencies as the unit of measurement.

$$\lambda(i,j) = \frac{L(n_{ij}, N_i, p) L(n_j - n_{ij}, N - N_i, p)}{L(n_{ij}, N_i, p_1) L(n_j - n_{ij}, N - N_i, p_2)}$$

with $p = \frac{n_j}{N}$, the probability j occurs in a document irrespective of D_i, $p_1 = \frac{n_{ij}}{N_i}$, the probability j occurs in a document in D_i, $p_2 = \frac{n_j - n_{ij}}{N - N_i}$, the probability j occurs outside D_i, and $L(k,l,x) = x^k(1-x)^{l-k}$, the likelihood function according to the binomial distribution (the binomial coefficient can be ignored). A feature of likelihood ratios is that -2 times the log of the likelihood ratio is asymptotically χ^2 distributed [98], which can be used to test whether there is a statistically significant divergence from independence. The weight of concept j in the concept profile of concept i is given by:

$$w_{ij} = \frac{\log \lambda(i,j)}{T}$$

T is the theoretical maximum score of $\log \lambda$, which is obtained when a concept always and only occurs in D_i. This factor normalizes for the effects of the size of D_i, which is convenient when comparing weights.

For occurrence frequency as the unit of measurement, we replace all document frequencies by the corresponding concept occurrence frequencies, e.g. N is replaced by O, the number of concept occurrences in D: $O = \sum_{d=1}^{N} \sum_{j=1}^{M} f_{dj}$.

Uncertainty Coefficient

The uncertainty coefficient $U(X\,|Y\,)$ [128] for the stochastic variables X and Y is given by

$$U(X\,|Y\,) = \frac{H(X) - H(X\,|Y\,)}{H(X)}$$

where $H(X)$ is the entropy for X and $H(X\,|Y\,)$ is the entropy for X given Y. The numerator equals the mutual information measure $I(X;Y)$. The value varies between 0, when knowledge of Y does not provide information on X, and 1, when knowledge of Y completely predicts X. The measure gives the fraction of the entropy of X that is lost when Y is already known. For document frequency as the unit of measurement we have the following variables:

$$X_i(d) = \begin{cases} 1 & \text{if } d \in D_i \\ 0 & \text{if } d \notin D_i \end{cases}, \quad Y_j(d) = \begin{cases} 1 & \text{if } f_{dj} > 0 \\ 0 & \text{if } f_{dj} = 0 \end{cases}$$

X_i indicates if a document d is part of the subset of documents D_i, Y_j indicates if the concept j occurs in d. We have thus two asymmetric measures of uncertainty, $U(X_i\,|Y_j\,)$ and $U(Y_i\,|X_j\,)$, but as we consider both interesting we use the symmetric uncertainty coefficient $U(X_i, Y_j)$ [24] to derive the weight w_{ij}, which can be considered a weighted average of both:

$$w_{ij} = U(X_i, Y_j) = \frac{H(Y_j) + H(X_i) - H(X_i, Y_j)}{\frac{1}{2}\left(H(X_i) + H(Y_j)\right)}$$

The entropies are defined as follows:

$$H(X_i) = -\tfrac{N_i}{N} \ln \tfrac{N_i}{N} - \tfrac{N-N_i}{N} \ln \tfrac{N-N_i}{N},\ H(Y_j) = -\tfrac{n_j}{N} \ln \tfrac{n_j}{N} - \tfrac{N-n_j}{N} \ln \tfrac{N-n_j}{N}, \text{ and}$$
$$H(X_i, Y_j) = -\tfrac{n_{ij}}{N} \ln \tfrac{n_{ij}}{N} - \tfrac{n_j-n_{ij}}{N} \ln \tfrac{n_j-n_{ij}}{N} - \tfrac{N_i-n_{ij}}{N} \ln \tfrac{N_i-n_{ij}}{N} - \tfrac{N-n_j-N_i}{N} \ln \tfrac{N-n_j-N_i}{N}.$$

In the equations above document frequency was used as the unit of measurement. For occurrence frequency all document frequencies are replaced by the corresponding concept occurrence frequencies, identical as for the log likelihood ratio.

Assigning GO Codes to Genes

For our evaluation, we sought to reproduce the annotations from the Gene Ontology Annotation (GOA) project [23], using concept profiles for both GO codes and genes. The concept profile of a GO code is matched with the concept profiles of all genes, and a ranking of genes based on the matching score is generated. Using the GOA annotations

Table 4.1: First 20 concepts of the concept profiles for GO code 7601: visual perception. The concept profiles are based on 84 MEDLINE records. The unit of measurement is concept occurrence frequency. The AUC score for this GO code is 0.86 for averaging, 0.79 for the Log Likelihood Ratio, and 0.91 for the Uncertainty Coefficient.

Averaging		**Log Likelihood Ratio**		**Uncertainty Coefficient**	
Concept	Weight	Concept	Weight $\cdot 10^{-3}$	Concept	Weight $\cdot 10^{-3}$
Retina	2.81	Retina	5.17	Retinitis pigmentosa	3.05
Photoreceptors	1.96	Genes	3.74	Photoreceptors	2.89
Genes	1.88	Photoreceptors	3.33	SIX3	1.99
Eye	1.86	Retinitis pigmentosa	2.45	Rods	1.97
Sequence	1.72	Mutation	2.32	GUCA1A	1.96
Mutation	1.62	Rods	2.24	Cones	1.93
Encoding	1.56	Eye	2.22	MYO9A	1.83
Retinitis pigmentosa	1.53	Sequence	2.02	FSCN2	1.80
Complementary DNA	1.42	Cones	1.82	Vertebrate Photoreceptor	1.79
Exons	1.15	Encoding	1.46	GJA3	1.62
Retinaldehyde	1.11	Retinaldehyde	1.25	SIX6	1.53
Vertebrate Photoreceptor	1.09	Complementary DNA	1.24	RGS16	1.53
Retinal degeneration	1.09	Vertebrate Photoreceptor	1.17	TULP1	1.52
Acids	1.05	SIX3	1.05	GUCY2D	1.41
Rods	1.04	GUCA1A	1.04	VAX2	1.41
Introns	1.04	Retinal degeneration	0.98	Retina	1.40
Clone Cells	0.98	Exons	0.93	GJA8	1.35
Cones	0.95	MYO9A	0.91	Usher syndrome	1.33
Disease	0.91	FSCN2	0.91	Retinal degeneration	1.29
Chromosomes	0.87	Introns	0.91	EYA4	1.24

as the gold standard, a performance measure is then calculated based on the ranking. In order for a gene or a GO code to be included in our evaluation, it has to be used in at least one GOA annotation and at least one document has to be linked to the gene or GO code (see below).

From the GOA database (release 3-7-05), we retrieved for each GO code the set of human genes that had been annotated with that code, together with the PubMed IDs that referred to publications from which the annotations were derived. A gene/GO code combination may have an associated PubMed ID, but many combinations have none. Thus, each GO code was associated with a set of genes and a (smaller) set of documents. This document set was the basis for D_i (from which a concept profile for the GO code is generated). If a document is used for D_i, the gene from the associated gene/GO code combination is not used for testing the performance of our system for that GO code. We limited the number of documents used for D_i to ensure that no more than 2/3 of the genes for a GO code were excluded for performance evaluation. For a gene, D_i was taken as the set of documents in D in which a reference to the gene was found by our thesaurus-based indexation.

In the evaluation we took the true-path rule into account: if a gene is annotated with a GO code, then all the parent GO codes of this GO code are also valid annotations for that gene.

For each GO code, the genes were rank-ordered according to the cosine similarity scores [101] between the concept profile of the GO code and those of the genes. Ties

Table 4.2: Median AUC scores and interquartile ranges (between brackets) for the three weighting schemes and the binary method. Results are shown for concept document frequency and concept occurrence frequency as the unit of reference.

Weighting Schemes	Document Frequency	Occurrence Frequency
Binary	0.718 (0.585-0.871)	-
Averaging	0.883 (0.739-0.991)	0.895 (0.748-0.994)
Log Likelihood Ratio	0.841 (0.671-0.986)	0.862 (0.688-0.987)
Uncertainty Coefficient	0.875 (0.731-0.993)	0.889 (0.751-0.996)

were ordered randomly. Instead of setting a threshold to perform a single classification, we compared the ranking of the genes to the GOA gold standard. We divided the set of genes in positive cases, i.e. genes that were annotated with this GO code according to GOA, and negative cases. From the gene ordering a receiver operating characteristics (ROC) curve [69] was then constructed for each GO code. The area under the curve (AUC) [70] was used as a performance measure, in a similar way as described in [92]. The AUC varies between 0 and 1, a value of 1 representing perfect ordering, i.e., all genes annotated with the GO code according to the GOA are at the top of the list followed by the other genes. The AUC has the useful property that a value of 0.5 represents random ordering [70]. To determine whether the AUC scores differed significantly between the methods, we used the non-parametric Wilcoxon signed ranks test.

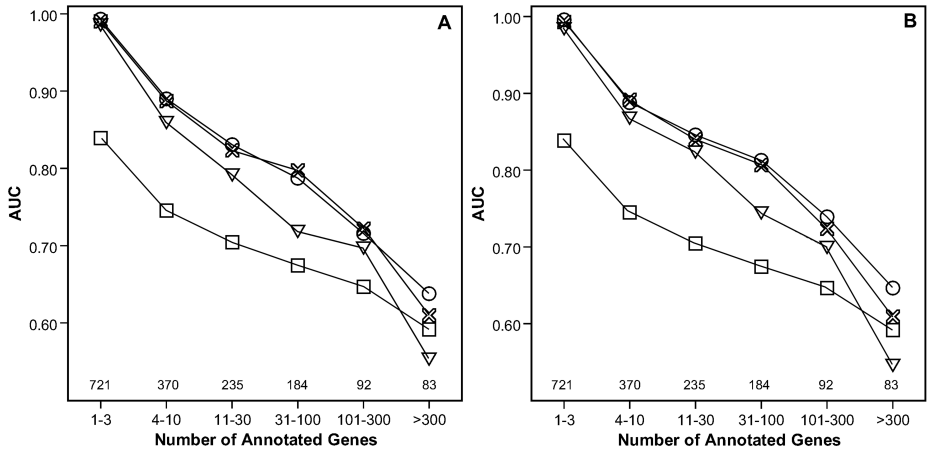
4.3 Results

The evaluation set consisted of 9283 genes and 1685 GO codes, including 777 biological processes, 236 cellular components, and 672 molecular functions. The number of genes annotated with a GO code showed a skewed distribution, where most of the GO codes had few gene annotations (median 5), and 399 codes (24%) only had 1 annotation. The amount of available MEDLINE records available for each concept profile also showed a skewed distribution. For genes the number of records ranged from 1 to 69482 (for tumor necrosis factor, TNF), and the number of genes with few records was overrepresented, with 2456 genes (26%) having 5 or less records. For GO codes the number of records ranged from 1 up to 589, and 1256 (75%) had 5 or less records.

Table 4.1 shows an example of concept profiles generated by the different schemes. The shown concept profiles are for the GO code "Visual Perception" for which 84 MEDLINE records were available, a comparatively large document set. For the averaging measure and the log likelihood ratio there is considerable overlap between the highest ranking concepts with many being general and frequently occurring, such as "Genes", "Mutation", "Eye", and "Complementary DNA". Contrastingly, in the concept profile for the uncertainty coefficient, many specific and less frequently occurring concepts such as gene names obtain higher weights. This pattern is observed for concept profiles for which more than a few records are available.

Table 4.2 shows the overall performance. All weighting schemes perform considerably and significantly better ($p<0.0001$) than the binary scheme. Between the weighting methods only small differences are apparent. For both occurrence and document frequency, the averaging measure and the uncertainty coefficient do not show significant differences

Figure 4.1: Median AUC scores for document frequency (panel A) and occurrence frequency (panel B) against the number of annotated genes per GO code. The averaging measure is depicted with ⊠, log likelihood ratio with ▽, uncertainty coefficient with ○, and binary weighting with □. The number of GO codes per category are indicated.



in performance, but both perform better ($p<0.0001$) than the log likelihood ratio. For all weighting schemes the use of occurrence frequency instead of document frequency causes a small but significant ($p<0.0001$) increase in performance.

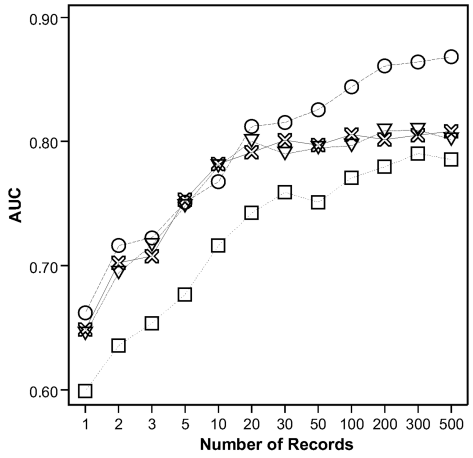
Figure 4.1 shows the median AUC scores against the number of annotated genes per GO code. There is a strong correlation between performance and the number of genes annotated by a specific GO code. As we use the true-path rule (see section 2.3), the number of annotated genes per GO code increases the closer a GO code is to the top of the GO hierarchy tree.

The number of MEDLINE records available for the generation of the concept profiles may affect the performance. To study the influence of only this factor, we took the 790 genes for which at least 500 records were available and varied the number of records for creating the gene concept profiles (see Figure 4.2). For increasing amounts of literature there is an increasing performance gap between the uncertainty coefficient and both averaging and the log likelihood ratio, whereas the difference in performance between the latter two and the binary method decreases.

4.4 Discussion

We compared three weighting schemes for generating concept profiles, using the annotation of genes with GO codes as an evaluation task. Our experiments illustrate the value of weighting by showing a considerable and statistically significant performance difference between our three weighting schemes and a simple binary scheme. The weighting schemes performed well for many GO codes, which underlines the utility of the concept profiling approach.

Figure 4.2: Median AUC scores against the number of MEDLINE records available for the gene concept profiles. For the unit of measurement only occurrence frequency is shown as the difference with document frequency was small (cf. Figure 4.1). A selection of genes was taken for which at least 500 records per gene were available and the amount of records used to build the concept profiles was varied. The averaging measure values are depicted with ⋈, log likelihood ratio with ▽, uncertainty coefficient with ○, and binary weighting with □.



Performance for a GO code was found to be negatively correlated with the number of genes annotated with that GO code (Figure 4.1). Since the number of annotations increases with the position of the code in the GO hierarchy, the number of gene annotations can be interpreted as a measure for how general a concept is. This suggests that our methodology performs better for specific GO codes than for general ones. One explanation for this phenomenon is that the relationships between genes and very general concepts, e.g. "Metabolism", are unlikely to be discussed in the MEDLINE records, because the described experiments tend to deal with more specific processes. Moreover, compared to specific concepts, the literature dealing with general concepts will contain more variation in the topics being discussed. The concept profile of a general concept is therefore likely to contain concepts related to a variety of topics, and may not match well with the "specific" concept profiles of genes.

Overall, the three weighting schemes show only small differences in performance (Table 4.2, Figure 4.1). The use of concept occurrence frequency in the weighting schemes instead of or together with the document frequency of concepts produces a statistically significant, but small increase in performance over using only document frequency. If we consider that most concepts occur infrequently (e.g. 52% of the identified concepts occurs 10 times or less in the corpus) and the small size of abstracts, a large dependency between concept frequency and document frequency may be expected. Hence a great deal of information would be conveyed by knowing whether a concept occurs in a document or not, and little additional information is conveyed by knowing its occurrence frequency in a document.

The overall results are dominated by the large number of genes and GO codes for which only a very small set of texts is available (e.g. 26% of genes have 5 or less records/documents; 75% of GO codes have 5 or less records). As illustrated in Figure

4.2, averaging and the log likelihood ratio on the one hand and the uncertainty coefficient on the other show an increasing performance gap when larger amounts of texts are used for the generation of the gene concept profiles. Averaging and log likelihood ratio yield similar concept profiles (Table 4.1), which are distinct from the concept profile of the uncertainty coefficient. The first two give general, more common concepts a higher weight, as observed generally in concept profiles for which a relatively high number of records was available. As these general concepts will appear in many concept profiles, this will make it harder to distinguish the concept profiles based on specific concepts. Since for specific GO codes specific concepts can be expected to be good distinguishing features, the high weights for general concepts may thus explain the effect shown in Figure 4.2.

The observed characteristics of the weighting schemes can be understood based on their definitions. With regard to the averaging method, suppose two concepts occur in a corpus with rates $r_1 = \frac{n_1}{N}$ and $r_2 = \frac{n_2}{N}$, with $r_1 < r_2$. Their IDF-based weights in individual documents (ignoring term frequency) will then be $v_{d1} = -\log_2 r_1$ and $v_{d2} = -\log_2 r_2$and $v_{d1} > v_{d2}$. Assuming that both concepts occur with the same rates r_1 and r_2 in D_i, the document set for concept i, then the ratio of their weights in the concept profile of i will become $\frac{w_{i,1}}{w_{i,2}} = \frac{r_1 v_{d1}}{r_2 v_{d2}}$. Therefore the ratio of their concept profile weights, $\frac{w_{i,1}}{w_{i,2}}$, will reduce by a factor $\frac{r_2}{r_1}$ as compared to their ratio of the document weights $\frac{v_{d1}}{v_{d2}}$. The weight in the concept profile of the rarer concept can therefore become smaller than that of the more common concept.

The log likelihood ratio method will give a higher weight to a concept in a concept profile, if the concept's occurrence rate in the document set is less likely to have occurred by chance. Common concepts can therefore also get a higher weight, because if we have more observations, we can be more certain that an observed dependency between the occurrence rate of a concept j and the document selection D_i did not occur by chance. This implies that a common concept will get a higher weight than a rarer concept when they are equally overrepresented in D_i (that is, for identical deviations from 1 of the ratio of the occurrence rate in D_i over the occurrence rate in its complement).

Different from the other two methods, the uncertainty coefficient tends to give rarer concepts a higher weight. The uncertainty coefficient is a measure of the strength of the dependence between two variables. The measure quantifies how much information is conveyed about the occurrence rate of the concept j by knowledge of the document selection D_i and vice versa. Given that D_i is generally relatively small, strong dependencies between commonly occurring concepts and D_i are rare.

Our study has several limitations. Firstly, we did not study all possible weighting schemes, though other most other described schemes are closely related to the tested schemes. Apart from averaging document vectors that are weighted with an IDF-based weighting scheme, other schemes that make use of IDF are conceivable. Srinivasan [41] did not use averaging, but an IDF-based normalizing approach. We tested this method, but only found small differences with the averaging approach (data not shown). Concerning the information theoretic measure, we decided not to use the mutual information measure as such, but a normalized variant. The unnormalized mutual information measure not only grows with the degree of dependence between two variables, but also according to the entropy of the variables. This would be an undesirable effect as in that case e.g. the size of the underlying document selections would have an influence on the matching score between concept profiles.

A second limitation is that we did not test the use of lexical features, such as words, n-

grams etc, instead of thesaurus concepts. Given that the characteristics of the weighting schemes, as discussed above, are not specifically associated with the use of thesaurus concepts, we expect that the behavior of the three weighting schemes will be similar when lexical features are used. A third limitation concerns the generalizability of our study. Our results are only based on concept profiles for GO codes and genes. Whether our findings also hold for other types of concepts, such as diseases, is still to be investigated.

4.5 Conclusions

The use of occurrence frequency over document frequency in the weighting schemes, results in only a small, though significant and consistent performance increase.

For small sets of literature a simple document vector averaging approach to generate concept profiles works fine. However, when more literature is available this weighting method tends to give very general concepts relatively high weights. In those cases the uncertainty coefficient is a more appropriate measure.

In our experiments better performance was observed for specific concepts than for general concepts. This can indicate concept profiles are less suitable to represent general concepts.

5

Literature-aided microarray data meta-analysis

Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease.

R. Jelier[1], P.A.C. 't Hoen[2], E. Sterrenburg[2], J.T. den Dunnen[2], G.J.B. van Ommen[2], J.A. Kors[1] and Barend Mons[1]

[1]Department of Medical Informatics, Erasmus MC, Rotterdam
[2]Department of Human Genetics, LUMC, Leiden

RJ and PH contributed equally.

Submitted for publication

Abstract

Comparative analysis of expression microarray studies can confirm findings from individual studies and identify interesting parallels between studies. However, such analyses are hampered by the large influences of design, technical and statistical factors on the found differentially expressed genes. Comparisons based on perturbed biological processes could be more robust as different genes may hint at the same process. We developed LASSO (literature-based association analysis) for this purpose. LASSO uses gene associations, automatically retrieved from Medline abstracts, to quantify the similarity between studies and to reveal overlapping biological processes. We tested our method and compared it to classical Gene Ontology group over-representation analysis through a comparative meta-analysis on 102 microarray studies published in the field of muscle development and disease.

The over-representation analysis did not perform well, due to limited sensitivity and the incompleteness of the manually curated gene annotations. LASSO retrieved many more biologically meaningful links between studies, even across species, microarray platforms and between studies that did not have any differentially expressed genes in common. Hierarchical clustering demonstrated limited influence of technical factors and correct grouping of muscular dystrophy, regeneration and myositis studies. As an example of new discoveries with our approach, cullin proteins, a class of ubiquitinylation proteins, were found associated with genes down-regulated during muscle regeneration. Interestingly, ubiquitinylation was previously reported to be activated during the inverse process: muscle atrophy.

Our approach facilitates finding common biological denominators in microarray studies, without raw data analysis or curated gene annotation databases.

5.1 Background

The comparative analysis of expression microarray studies can refine conclusions and interpretations from individual studies and can be used to identify previously uncharacterized parallels between studies [129, 130]. However, such analyses are hampered by the large influences of biological variation between specimens (see e.g. Eid-Dor et al.[131]), and technical differences between the studies [132–137] on the identified differentially expressed genes. The varying technical factors include: differences in experimental procedures for the collection of the biological material and for RNA amplification and labeling [136], differences in sampling times and the DNA microarray platform used (see Kuo et al. [137] for a recent platform comparison), and the applied statistical analysis [134].

To overcome this hurdle, it has been suggested that studies should be compared at the level of perturbed biological processes [138, 139]. This could be more robust as different genes may hint at the same process. To identify perturbed biological processes, methodologies have been developed in recent years by analysis of the correlated behaviour of groups of genes with a similar biological function [139–141]. A limitation when using these approaches is that to identify which genes share a biological function, we are currently largely dependent on the ontology-based annotation of genes in manually curated databases. Due to the labor intensive manual curation effort, these databases are necessarily highly focused and notoriously incomplete (see e.g. Khatri et al.[142]). The best known public databases are the Gene Ontology (GO) annotation project [23] for biological process, molecular function and cellular localization, and KEGG [24] for metabolic pathways.

In the present study we introduce an approach to compare expression profiling studies based on perturbed biological processes. Instead of using manually curated gene annotation databases we base our approach on gene associations automatically derived from literature. To identify gene associations, concept profiles are generated for all genes [119]. A concept profile is a weighted list of biological concepts that characterizes the set of documents associated to a gene. Subsequently concept profiles are compared to identify gene associations: pairs of genes strongly associated to the same biological concepts. Finally, DNA microarray datasets are compared based on the observed number of gene associations between the sets of differentially expressed genes. We call our approach Literature-based ASSOciation analysis (LASSO).

We evaluate our methodology on a compendium of 102 DNA microarray studies published in the field of muscle development and disease, and compare it to analyses based on gene overlap and the classical group overrepresentation analysis. The compendium contains a very diverse set of datasets: patient versus control studies for different myopathies; studies in animal disease models and studies in cultured muscle cells. The studies were performed on 22 different microarray chip types, and in three different organisms: human, mouse and rat. The considerable influence of the statistical analysis on the identified differentially expressed genes [134], indicates that, ideally, a standardized statistical analysis should precede any comparison between datasets. Unfortunately, raw data is required for such an analysis and they are often unavailable (see also Larsson et al.[130]). Therefore, we relied on the reported lists of differentially expressed genes, which should be useful for initial comparisons of microarray studies [143, 144] and, at least, were judged by the authors to be biologically relevant. In our evaluation, we first take a directed approach: we measured to which extent the approaches could reproduce a manual clustering of a

selection of datasets. Second, we perform an exploratory clustering of all datasets, and characterize and interpret the identified clusters.

5.2 Methods

Data acquisition

In our meta-analysis, we included DNA microarray studies on skeletal muscle development and/or disease. The compendium was limited to studies in human, mouse, and rat. Studies were included till December 2005. From each paper, lists of up- and downregulated genes were extracted from the tables reported in the paper or in the supplementary data. The compendium is not complete. For some of the studies, data could not be retrieved and requests for gene lists to the authors were unsuccessful. Since a full list of genes interrogated by each platform was essential for statistical analysis, studies on home-made arrays for which this information was not available had to be omitted as well. All probes on the array were mapped to Entrez Gene IDs. To be able to compare gene lists from the different organisms we mapped homologous genes to each other based on NCBI's HomoloGene database [66].

Comparing DNA microarray experiments based on gene identity

The similarity between two datasets based on gene identity was measured using the kappa statistic [145]. To use this measure we consider DNA microarray experiments to assign every gene on the microarray platform a tag: upregulated, downregulated or the remainder category. The kappa statistic is defined as

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times that the two experiments give the same tag to a gene and $P(E)$ is the proportion of times that we would expect the experiments to give the same tag by chance. When calculating kappa we only consider the genes that are present on both platforms. If the DNA microarray datasets show identical results ($P(A) = 1$) then $\kappa = 1$, and if there is only chance agreement ($P(A) = P(E)$) then $\kappa = 0$.

Recognizing references to concepts in texts

The corpus of literature for our experiments consisted of 3,160,002 MEDLINE abstracts, selected with the PubMed query "(protein OR gene) AND mammals". We used titles, MeSH headings, and abstracts. Stop words were removed and words were stemmed to their uninflected form by the LVG normalizer [67].

We used a thesaurus to identify concepts in texts. The thesaurus was composed of two parts: the 2006AC version of the UMLS thesaurus [22] and a gene thesaurus derived from multiple databases. The gene thesaurus was a combination of gene names from the rat genome database [146], mouse genome database [147], and a human gene thesaurus from several databases [99]. Homologous genes between the three species were mapped to each other using NCBI's HomoloGene database 66. In order to exclude irrelevant concepts, two molecular biologists created a list of UMLS semantic types (see

Appendix 1 for the complete list) relevant for biological information about genes. All concepts with other semantic types were removed from the thesaurus. Following Aronson [125], the UMLS thesaurus was also adapted for efficient natural language processing, avoiding overly ambiguous or duplicate terms, and terms that are very unlikely to be found in natural text. The gene thesaurus was expanded by rewrite rules to take into account common spelling variations 100. For instance, numbers were replaced with roman numerals and vice versa, and hyphens before numbers at the end of gene symbols were inserted or removed (e.g. "WAF1" was rewritten as "WAF-1" and added as a synonym).

Concept profile methodology

For every gene in our thesaurus that we identified in at least 5 documents, we characterized the documents in which the gene occurs with a concept profile. A concept profile of a concept i, for instance a gene, is an M-dimensional vector $\mathbf{w_i} = (w_{i1}, w_{i2}, \ldots, w_{iM})$ where M is the number of concepts in the thesaurus. The weight w_{ij} for a concept j in this profile indicates the strength of its association to the concept i. The weights in a concept profile for concept i are derived from the set of documents in which concept i occurs, D_i, which is a subset of the total set of documents D.

To obtain the weight w_{ij} we apply the symmetric uncertainty coefficient $U(X_i, Y_j)$ [128]:

$$w_{ij} = U(X_i, Y_j) = \frac{H(Y_j) + H(X_i) - H(X_i, Y_j)}{\frac{1}{2}\left(H(X_i) + H(Y_j)\right)}$$

Here the stochastic variable X_i defines whether a document is in D_i, and Y_j gives the occurrence frequency of concept j. The entropies H are defined as follows:

$$H(X_i) = -\frac{O_i}{O} \ln \frac{O_i}{O} - \frac{O-O_i}{O} \ln \frac{O-O_i}{O},$$

$$H(Y_j) = -\frac{o_j}{O} \ln \frac{o_j}{O} - \frac{O-o_j}{O} \ln \frac{O-o_j}{O},$$

$$H(X_i, Y_j) = -\frac{o_{ij}}{O} \ln \frac{o_{ij}}{O} - \frac{o_j-o_{ij}}{O} \ln \frac{o_j-o_{ij}}{O} - \frac{O_i-o_{ij}}{O} \ln \frac{O_i-o_{ij}}{O} - \frac{O-o_j-O_i}{O} \ln \frac{O-o_j-O_i}{O}$$ where

O and O_i represent the number of concept occurrences in D and D_i resp.; o_j and o_{ij} represent the number of occurrences of concept j in D and D_i resp. The uncertainty coefficient is a normalized variant of the mutual information measure. The symmetric coefficient is the weighted average of the two assymmetric uncertainty coefficients: 1. the proportion of information in Y explained by knowledge of X and 2. the proportion of information in X explained by knowledge of Y.

Literature-based comparison of gene lists

The similarity of the concept profiles of two genes was measured with the cosine similarity score [101]. If the similarity score exceeded a threshold, then the two genes were considered to have an association. For our experiments here, we calculated the similarity score for all pairs of genes for which a concept profile was available; the top 1% of pairs were taken as associations. Subsequently, the found associations are used to compare two gene lists. To do this, two gene lists were considered as separate sets of nodes, and the number of associations between the two were counted. We assessed how uncommon the observed

number of associations was, by generating an empirical distribution based on Monte Carlo simulations. For each simulation we performed the following two steps: 1. For each gene list we randomly selected a number of genes equal to the size of the gene list. These genes were selected from the genes present on the appropriate DNA microarray platform for which we had a concept profile available. 2. The number of connections between the two new gene lists were counted. Using the empirical distribution we subsequently estimated the chance of observing the given number of associations or more.

For each DNA microarray experiment we retrieved two gene lists, the upregulated and the downregulated genes. When comparing two experiments, p-values were computed for the two up and down lists; the final score was obtained by multiplying the p-values.

In order to interpret the LASSO score between two datasets we developed a computer program. The program shows for every gene in one set the associations that connect it to the other set. The biomedical concepts that underlie the gene associations could readily be retrieved and traced back to the literature through an incorporated version of Anni, a tool we published earlier [119]. To annotate a cluster of datasets we calculate the percentual contribution to the number of annotations for every gene, averaged over all dataset comparisons between cluster members. Subsequently we identified descriptive concepts for the cluster by retrieving concepts strongly associated to the top-ranking genes through the Anni annotation view. For table 5.1 we used as cutoffs 0.2% for selecting the genes and concepts in the annotation view were selected when their contribution was larger than 1. For brevity genes were excluded from this table. Of partially redundant concepts (e.g. "heterogeneous nuclear ribonucleoproteins" and "heterogeneous nuclear ribonucleoproteins activity") only the highest scoring concept was shown.

High-throughput analysis of over-represented GO-terms

To evaluate if a set of differentially expressed genes shows an over-representation of genes belonging to a certain biological process, molecular function or cellular localization, as annotated by the Gene Ontology (GO) consortium [21], a hypergeometric test is commonly used, see e.g. the web tools DAVID [148] and GOTM [149]. We used the HyperGTest from the GOstats 2.0.4 package (`http://www.bioconductor.org/`). Only GO terms from the branch "Biological Process" subset of GO-terms were evaluated, since this was most relevant to the biological problem. To perform the test, an annotation package was built per species with the AnnBuilder 1.12.0 package in R, for the concatenated list of Entrez Gene identifiers represented by the relevant platforms. We analyzed up and down regulated gene lists separately. Similar to how we calculated the similarity of gene lists based on gene identifiers, the kappa statistic was used to calculate the similarity of significantly overrepresented GO-terms ($p < 0.05$) in the up- and downregulated gene sets from two microarray datasets.

Reproduction of a manual clustering

We tested to which extent a manual grouping based on studied biological phenomena (cf. Appendix 5.7) was reflected by the pair-wise similarity dataset scores by performing a classification experiment: The association measures were used to produce a ranking of the set of studies relative to one so-called seed study. All studies in turn served as a seed, producing a ranking for each of the other studies in the groups. Studies from the same

group as the seed study were considered positive, studies from other groups negative cases. Based on the sorted list of positive and negative cases we constructed for each study a receiver operating characteristics (ROC) curve [69]. The area under the curve (AUC) was used as a performance measure [70]. An AUC of 1 represents perfect ordering, i.e. the studies from the same group as the selected study hold the top ranks, and an AUC of 0.5 is the expected score for a random ordering [70].

Clustering DNA microarray data experiments

The DNA microarray studies can be compared to each other through the LASSO and kappa measures. To identify patterns in these associations we clustered the studies through agglomerative hierarchical clustering and subsequently annotated the identified clusters. For this purpose the LASSO scores were $-log_{10}$ transformed.

5.3 Results

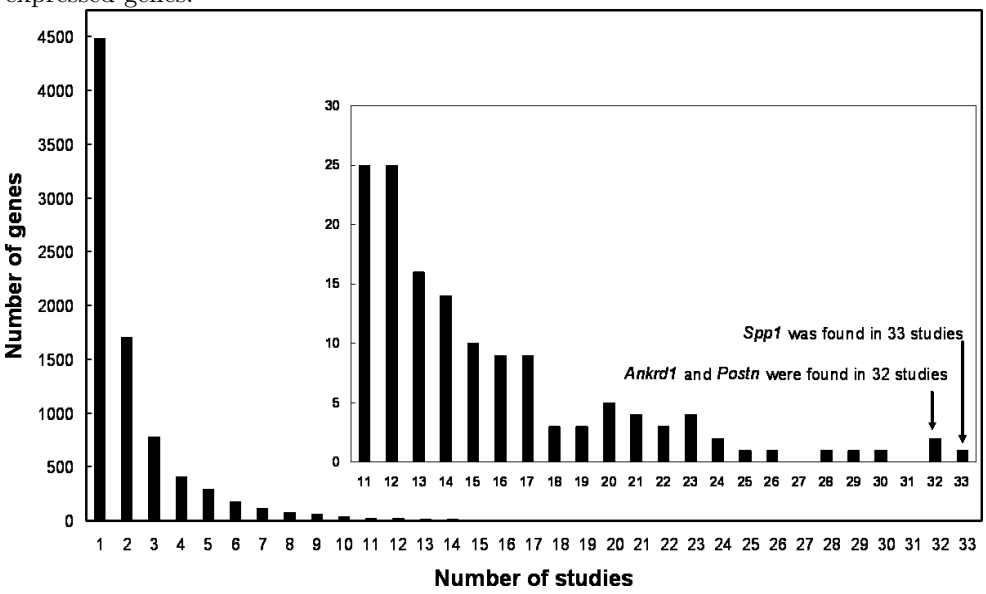
Study selection and data retrieval

The 102 microarray datasets in the compendium are represented and annotated in the appendix. They were extracted from 53 publications and 6 in-house studies. The datasets include studies on myoblast differentiation as an *in vitro* model for muscle development and regeneration, studies on gene expression differences between different types of skeletal muscles, skeletal muscle disease (including induced muscular atrophy), the effect of exercise and ageing and the treatment with drugs, growth factors or lipid infusion. The compendium was limited to studies in human (N=37), mouse (N=51), and rat (N=13), but included one study in monkey performed with a human DNA microarray platform. To allow for a direct comparison of datasets from different organisms, homologous genes were mapped to each other according to the NCBI's homologene database [66].

Frequency of differential expression per gene

After mapping of species-specific Entrez Gene IDs to Homologene, 8282 unique genes were identified as differentially expressed in at least one microarray study. Figure 5.1 displays the distribution of the number of microarray studies in which a gene was found differentially expressed. The majority of genes (4486) was found differentially expressed in only a single study. Indeed, the distribution implies the overlap between studies is limited, e.g. 84% of the genes occur in 3 or less studies, but they represent 54% of all gene occurrences. *Spp1* coding for osteopontin was the most frequently differentially expressed gene; it was found in 33 different studies described in 15 different papers coming from 8 different laboratories. *Spp1* was upregulated in animal models for muscular dystrophy and in human polymyositis and dermatomyositis, and can be regarded as an early marker for muscle inflammation [150, 151]. Conversely, it was found downregulated in presymptomatic *mdx* mice and during atrophy. *Ankrd1* and *Postn* were found in 32 different studies. Judged from the studies in which *Ankrd1* is differentially expressed, *Ankrd1* could be the most robust marker for muscular dystrophy with ongoing regeneration. *Postn* (periostin) was differentially expressed in many of the studies in which *Spp1* was found. A similar co-regulation was found in the heart and vasculature [152, 153], and both factors are involved

Figure 5.1: Distribution of the number of microarray studies in which a gene was found differentially expressed. A total of 102 studies was included with 8282 unique differentially expressed genes.



in tissue remodelling [151, 154, 155].

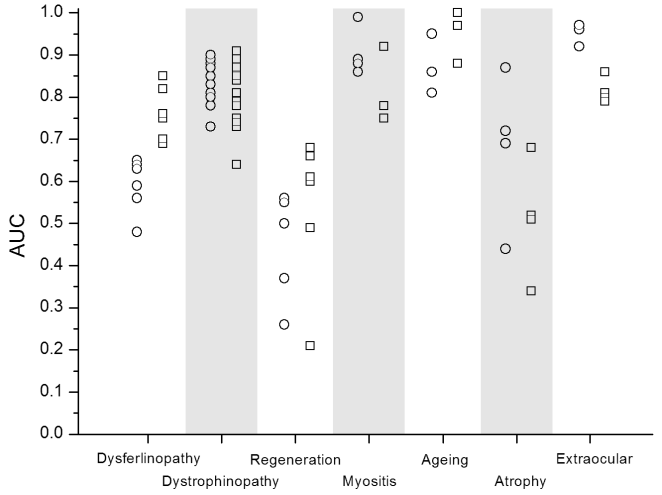
Pair-wise analysis of similarity between DNA microarray datasets

First, we compared the DNA microarray studies to each other based on gene identity. For every study, the genes interrogated by the DNA microarray platform were separated into three categories: upregulated, downregulated and not up- or downregulated. With the kappa statistic [145] we measured the chance-corrected level of agreement in the three categories between two studies. By performing a kappa statistic based test[156] we found that of the 5151 possible dataset pairs only 307 (6%) have an above chance level of agreement ($p<0.05$). This is in line with our conclusion of limited overlap in the previous section.

Second, our LASSO method found significant associations ($p<0.05$) between 2732 (53%) pairs of studies, which indicates that considerably more similarities between datasets are identified with our text-derived gene associations than based on gene list overlap.

Third, we compared datasets based on over-represented GO codes. We tested over-representation of biological processes in the up and down lists of our datasets separately with a cutoff of $p<0.05$. Subsequently, we compared the datasets through the kappa measure, and found 34% of scores to be very low (<0) and 918 (18%) of the dataset pairs had a significant overlap ($p<0.05$). The used GO over-representation test is overly permissive, as with the high number of tests, we should correct for multiple testing. But when we corrected for multiple testing (Benjamini and Hochberg's method [157], same chance level), we found no over-represented GO code for 43 of the 102 studies. For the dataset comparison as much as 85% of the kappa scores was low at 0 or lower and only

Figure 5.2: Performance for reproduction of the manual grouping by kappa (circles) and LASSO (squares).



212 dataset pairs (4%) had a significant association at the 0.05 level. The poor overlap between datasets is partly caused by the fact that it is less likely to identify an over-represented GO code if the gene list is small.

Reproduction of a manual clustering

To evaluate the performance of the methods, we manually grouped, before any development of new methodologies, the studies in the compendium based on similarities in the biological phenomena under study. We could identify 7 clusters for a subset of 50 studies: dystrophin-deficiency (human and mouse), dysferlin-deficiency (human and mouse), myositis, regeneration and differentiation, ageing, atrophy, and extraocular muscle (EOM)-specific expression profiles (cf. appendix 5.7). A classification experiment was performed to evaluate to which extent the kappa and LASSO association scores could reproduce 7 manually identified clusters, using the area under the ROC curve (AUC) statistic. Results are shown in figure 5.2. The performance varies from near perfect scores for the ageing group to near random classification performance for the "regeneration and differentiation" and atrophy groups. Indeed, the latter groups studied more diverse conditions. The LASSO method outperformed the kappa for the dysferlinopathy, regeneration and ageing subgroups. Conversely, the kappa performed better for the myositis and the extraocular subgroups. Both methods performed similarly for the dystrophinopathy group.

The dysferlinopathy group has much higher classification rates with LASSO than in the kappa analysis. The studies in this group were more heterogeneous than the other groups in several aspects: it contained human and mouse studies, different mouse strains and differently aged mice, and four different microarray platforms were used in the six studies contained in this group. The human study that compared limb-girdle muscular dystrophy (LGMD) type 2B patients to controls (dataset 16) was much better classified to the dysferlinopathy group with the LASSO based approach than with the

kappa based approach (AUC 0.73 vs 0.48). Datasets 16 (human) and 75a (dysferlin-deficient SJL mice versus controls) have no differentially expressed genes in common. Nevertheless, the LASSO score is the lowest possible score given the number of Monte Carlo simulations, which indicates there is a highly significant over-representation of gene associations. We identified "macrophage" as the most important shared concept between dysferlin-deficiency in humans and mice. Indeed, many of the identified associations are between genes known to be expressed in macrophages and macrophage infiltration is an important feature in both the LGMD patients pathology and the mouse model *[158].*

The slightly worse performance of the LASSO method in the classification of the myositis studies was due to strong associations with the group of dystrophinopathy studies. The two groups were found connected through concepts pertaining to inflammatory processes. This is reflective of the pronounced inflammatory component in both dystrophinopathies [150, 151, 159–161] and myositis patients. For the extraocular group the lower performance for LASSO is explained by the comparatively poor scores between datasets 4a and both 14a and b, while datasets 4b, 14a and 14b had high pairwise scores. Dataset 4a only contains 13 genes up-regulated genes, which limits the power for the LASSO analysis. The list shared only 2 genes with 14a and b, but still the kappa score was comparatively high due to the limited overlap overall. Classification for the GO-based over-representation analysis ($p<0.05$, no correction for multiple testing) showed poor results: performance was considerably worse than based on gene list overlap for 5 of the 7 groups, a similar score was obtained for the dysferlinopathy group and a slightly better score for the ageing group. These results and a view on the shared GO codes indicated the used test condition was too lenient and spurious GO codes were assigned. Based on these results and the poor results presented in the previous section, we do not discuss the clustering based on the method here, but include the classification results and the clustering as an appendix.

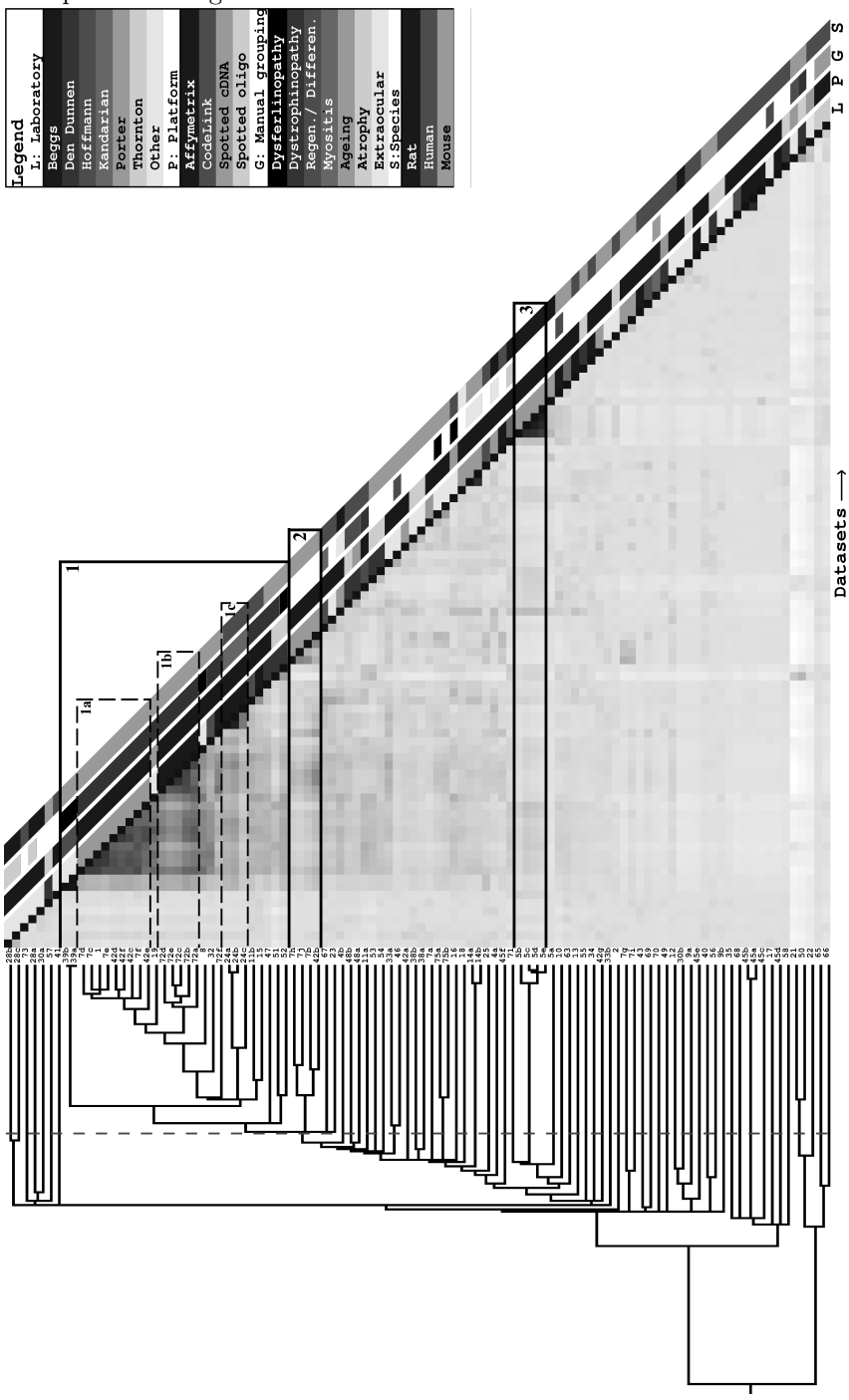
Classification of new studies

To demonstrate the utility of our approach for the interpretation of gene lists from new experiments, we compared to our manual grouping the gene lists from a recent paper on dy/dy mice [162]. These mice have a muscular dystrophy as a consequence of a genetic defect in alpha-2 laminin. Our LASSO-approach classifies this study with high confidence in the dystrophinopathy group (AUC=0.83). This is correct given the pathology of these mice and the two genetic deficiencies affecting the same macromolecular protein complex. The shared biological concepts between this dataset and a dataset from *mdx* mice (dataset 1; [151]), were the infiltration of macrophages and differential expression of collagens, metalloproteinases, cathepsins, and HLA-antigens.

Dataset clustering based on kappa statistic

To get an overall view on the identified connections between studies, a hierarchical clustering of the microarray studies was performed using the kappa value as a similarity score (figure 5.3). One big cluster (indicated as cluster 1) and several smaller clusters were identified. Cluster 1 contains comparisons of the gene expression profiles between dystrophic subjects and healthy controls (dystrophin-deficient *mdx* mice (datasets 1, 7c-f, 42c-f, 19, 32 47, 51, 72a-f), dysferlin-deficient SJL mice (datasets 8, 39ab), patients with Duchenne muscular dystrophy (DMD, datasets 11b, 15)), as well as studies in human myositis pa-

Figure 5.3: Kappa-based hierarchical clustering and heatmap. The dotted pink line indicates the used clustering cutoff and the identified clusters are indicated in addition to relevant subclusters. The dataset ids are shown between the tree and the heatmap. The colored bars provide background information on the datasets.



tients (datasets 24a-c, cluster 1c). Similar to our note on the LASSO classification in the previous section, muscular dystrophy and myositis expression profiles have considerable overlap. Some muscular dystrophy studies unexpectedly fall outside cluster 1 and have only limited overlap to the datasets in cluster 1 (dataset 16, 67, 75a). We believe this to be at least partly attributable to technical factors. The color bars on the side of figure 5.3 illustrate that studies tend to cluster on microarray platform or laboratory. For example, the similar studies in the *mdx* mouse by Porter et al. (datasets 1, 7, and 42) and by Haslett et al. (dataset 72) do not cluster in a way that makes sense biologically, that is by age and muscle type-dependent severity of the disease. Instead they cluster by laboratory (cluster 1a - Porter; cluster 1b - Haslett).

Apart from the large dystrophy/myositis cluster of studies, there is only very limited overlap between the gene lists from the studies, as expected based on the pair-wise analysis presented above. Cluster 2 contains two studies investigating the spared EOM muscle in dystrophin-deficient *mdx* mice and the expression profiles of the diaphragm and hind limb muscles of presymptomatic *mdx* mice. Cluster 3 contains 4 highly overlapping studies (datasets 5b-d) from the same paper on developmental changes in the EOM muscle. Again, clusters 2 and 3 contain only studies done by the same group on the same platform.

Dataset clustering based on LASSO

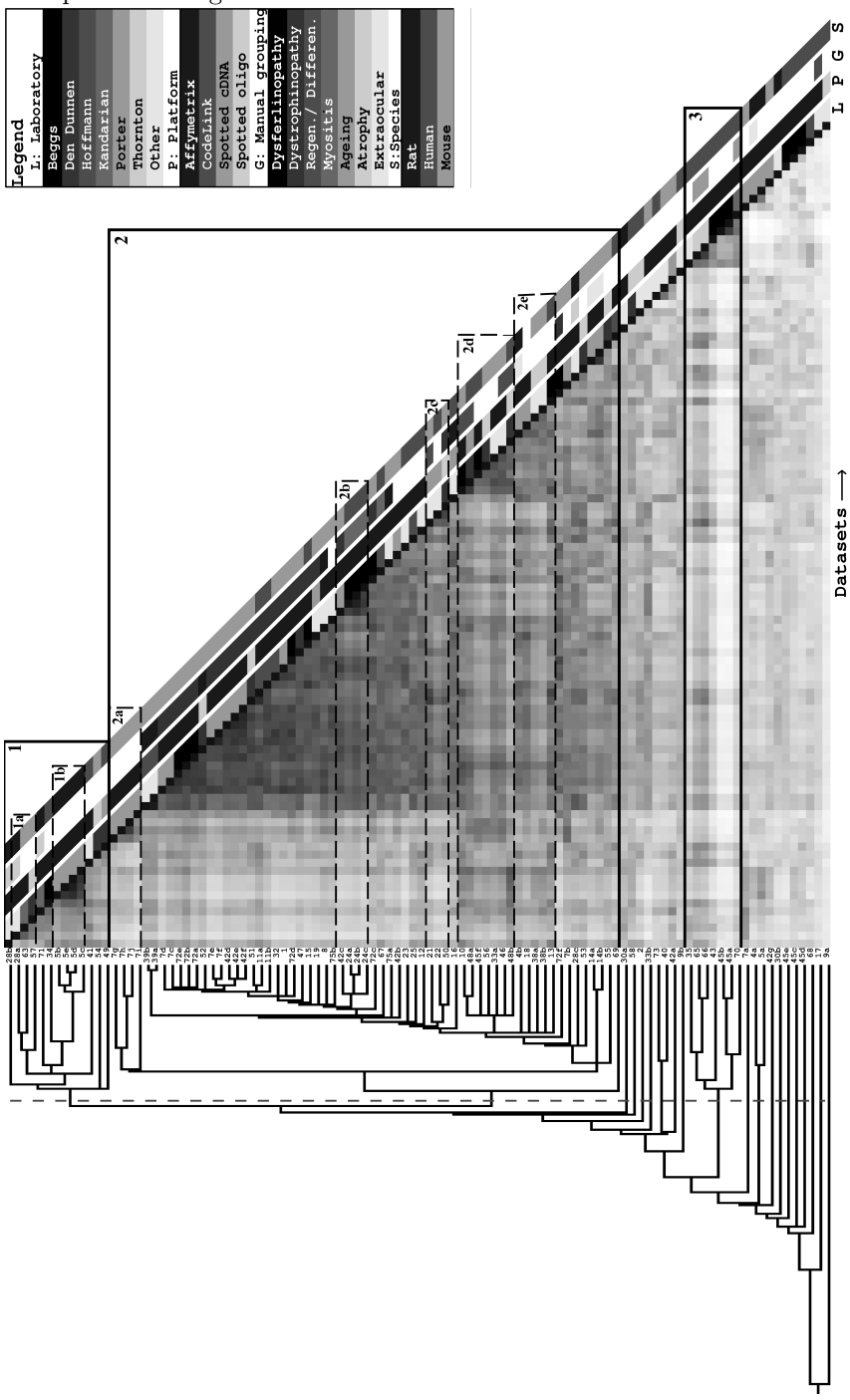
The LASSO-based hierarchical clustering revealed more clusters and significant associations than the kappa-based clustering (figure 5.4). The side bars show that the LASSO-based clustering is less governed by technical factors like microarray platform and laboratory, and is better able to connect studies investigating the same biological phenomenon in different species or biological systems (cell culture or tissue; see below).

Cluster 2 shows large overlap with cluster 1 in the kappa-based clustering, but contains many more studies. This cluster now contains all the studies on affected muscles in symptomatic *mdx* mice, all mouse models for LGMD, and all studies in human muscular dystrophy, including DMD, LGMD, and facioscapulohumeral dystrophy (FSHD), and myositis patients. The myositis patient profiles are closely associated with *mdx* mice of 23 days (dataset 42c) (subcluster 2b). We analysed the gene associations between dataset 24a (inclusion body myositis) and dataset 42c and found that the biological concept that contributes most to the associations is "chemokines". Indeed, at the analyzed age of 23 days the secretion of chemokines in the muscles of the *mdx* mice is maximal [151].

Table 5.1 shows the biological concepts underlying the gene associations in the different groups. For cluster 2, the most prominent biological terms for the upregulated genes were metalloproteinase activity (involved in extracellular matrix remodeling during fibrosis) and tr**itogen activated protein kinases, insulin, ERK1 acitivity, phosphorylatio**oponins. Metalloproteinases and troponins have been identified before to be important in muscular dystrophy (e.g. [151, 163, 164]) and in muscle regeneration. Studies on muscle regeneration are also included in cluster 2 (subcluster 2c). Remarkably, subcluster 2d contains all *in vitro* myoblast differentiation studies, both in primary human myoblast and in transformed mouse C2C12 myoblasts, whereas only a very small number of genes overlapped between the studies.

As apparent from table 5.1, the concept "cullin proteins" formed the most significant link between the downregulated genes from studies on muscle regeneration (datasets 21, 22, 50). Since cullin proteins are ubiquitin ligases, it seems that ubiquitinylation is shut

Figure 5.4: LASSO-based hierarchical clustering and heatmap. The dotted pink line indicates the used clustering cutoff and identified clusters are indicated in addition to relevat subclusters. The dataset ids are shown between the tree and the heatmap. The colored bars provide background information on the datasets.



down during regeneration. This is an interesting discovery since ubiquitinylation activity was previously shown to be activated in the inverse condition, muscular atrophy [165, 166].

Cluster 1 was not found by the kappa-based clustering. Analysis of the underlying concept associations revealed similarities between the molecular processes during induced atrophy in cultured myoblasts (dataset 63) and *in vivo* models for muscular atrophy, i.e. during hind limb suspension or in space-flown rats (datasets 28a, 28b and 71). Amongst others, there is an interesting set of non-overlapping members of the semaphorin family shared between atrophy studies. Semaphorins are presumably involved in cell-cell contacts in neuronal cells[167] (during axon regeneration) but also in fusing myoblasts [168]. For cluster 1 we observe an increase in metabolic activity (both glycolytic and fatty acid oxidation) and a downregulation of extracellular matrix proteins: These processes seem to be revelant to age-related changes in EOM muscle (datasets 5b-d; subcluster 1b), and diverse myopathies mitochondrial encephelo myopathy (dataset 41), nemaline myopathy (dataset 34), and oculopharyngeal muscular dystrophy (OPMD; dataset 57).

Cluster 3 contains all the ageing and sarcopenia studies. Interestingly, also a cell model for over-expression of the polyadenylation factor PABPN1 (also responsible for OPMD, dataset 35) is found in this cluster. In this case, the differences in RNA metabolism induced in the cell model aid the interpretation of the molecular phenotype observed in the ageing studies. Differences in RNA processing and splicing during ageing were also noted by the authors of the ageing studies [169, 170].

5.4 Discussion

The overlap between the gene lists of the microarray datasets in our compendium was limited, even though the studied phenomena were closely related. In addition, studies performed in the same laboratory or on the same microarray platform were more likely to demonstrate overlap than studies where more heterogeneous technologies and analysis approaches were used. The comparative analysis of the datasets through literature-derived gene associations resulted in the finding of many biologically relevant associations, and was more biology- than technology-driven. Both the analysis of the hierarchical clusterings and the reproduction of the manual clustering revealed that the LASSO method identified useful associations between datasets that were not retrieved by looking at gene overlap. Our method found these associations through correctly retrieved shared biological processes between the datasets.

Standard exploratory analysis based on an over-representation analysis of GO categories was not very powerful for our compendium, as shown by the lack of overlap between studies and the poor classification results. In general, the hypergeometric test will not often identify over-represented GO categories when short gene lists are analyzed. Yet our association-based method was still able to find useful associations between datasets, even in cases where not a single shared over-represented GO code was found. Also, the associations we use cover a much broader range than GO. Indeed, not all of the concepts in table 5.1 are covered by the GO thesaurus (e.g. leptin). In addition, even if an appropriate GO term exists, it may not have been assigned any genes yet (e.g., cullin deneddylation).

Our broad network of associations increases our sensitivity for identifying interesting associations between datasets. We chose to use associations derived from literature to optimize for serendipity, but the network of associations could be taken from any source, including GO. An important feature of this approach is that by modulating the asso-

Table 5.1: Characterizing concepts for clusters identified through LASSO analysis. Concepts are shown separately for the down and up regulated gene lists. The column "Characteristic" gives a description of the studied phenomena in the cluster.

Cluster	Subcluster	Characteristic	Biological Concepts (Up)	Biological Concepts (Down)
1	**overall**	**Atrophy**	**-**	**Cyclins**
1	1A	Atrophy - PABPN1 overexpression	Amino acyl tRNA synthetases, spermidine, polyamines, spermine, eukaryotic initiation factors	Platelet-derived growth factor, transforming growth factor-beta, insulin-like growth factor binding proteins
1	1B	EOM-specific	Adipocytes, acyl CoA dehydrogenase	Cyclins, keratin, cyclin-dependent kinases
2	**overall**	**Dystrophy/myositis**	**Troponin, matrix metalloproteases**	**Mitogen activated protein kinases, insulin, ERK1 acitivity, phosphorylation**
2	2A	Dystrophin defiency in EOM muscle	Troponin	-
2	2B	Myositis	Chemokine, chemokine receptor	-
2	2C	Regeneration	T-lymphocyte, phosphotransferases, phosphorylation, mitogen-activated protein kinases, integrins, cell cycle	Cullin proteins, mitogen-activated protein kinases, ligase
2	2D	Differentiation	Troponin, tropomyosin, nemaline myopathies, sarcomeres, myosin heavy chain, calsequestrin	Inhibitor of differentiation proteins, E2F transcription factors, proteoglycan, cell cycle proteins
2	2E	Ky-mutant / diverse	Leptin, desaturase, myosin heavy chains, neural cell adhesion molecules	Mitogen-activated protein kinases
3	**overall**	**Ageing**	**Heterogeneous nuclear ribonucleoproteins, protein sumoylation, small nuclear ribonucleoprotein**	**-**

ciations that are taken up in the network, the specificity and sensitivity of the found associations between datasets can be controlled.

Tomlins et al. [11] performed an over-representation analysis on gene lists representing the different stages of prostate cancer and used identified over-represented gene groups to compare the disease stages. The basis for their analysis was a database of 14000 groups of genes that share a relevant characteristic, or "molecular concept". They do not report low recall or lack of overlap of over-represented "molecular concepts". The likely explanation is that their meticulous sample preparation and highly standardized data generation and analysis avoided lack of overlap at the gene level and short gene lists. Clearly access to the raw data is commendable for exploratory studies, and standardized data generation is extremely useful given the high levels of variance observed with microarray experiments. It should be noted though that, besides the limited availability of raw data, the statistical models for the analysis of the raw data are hard to standardize. The choice for a statistical model depends on the design of the study, e.g. time course experiments or group comparisons, and technical factors, such as whether a one or two color microarray is used. Therefore perhaps the strongest point of our approach is that even with a wide range of study designs and statistical evaluations, across various platforms and species, a useful and insightful exploratory study was possible.

The issue of comparability between studies has been addressed for meta-analyses with a different objective than ours; the aggregation of information obtained from different DNA microarray studies [171–174]. It has been suggested that DNA microarray datasets could well be compared by the use of rankings of genes based on the level of significance of differential expression [174, 175]. Exploring the approach to compare datasets based on rankings would be an interesting extension of the current work. A rank-based approach could be adapted to incorporate our text-based associations between genes. Also in this case, information on the statistical ranks is, however, frequently unavailable.

A limitation of the current LASSO analysis is that it relies on simulations to derive a measure to compare datasets. Simulations are computationally intensive, and have a resolution proportional to and limited by the number of performed iterations. The results presented here can be considered a proof of the utility of our approach, and a logical next step is to derive a model-based approximation as an alternative to our simulation-based measure.

5.5 Conclusions

The compendium of studies showed limited overlap on gene ids, and a bias towards higher overlap between studies with technical similarities. The over-representation analysis based on GO categories was not very helpful in comparing studies, due to limited sensitivity and the incompleteness of the manually curated gene annotations. Compared to these approaches LASSO provided more biology- than technology-driven results and identified more biologically relevant associations between datasets. As the shared biological processes between studies could also be easily recognized, we believe LASSO is a powerful approach for the comparative meta-analysis of DNA microarray datasets.

5.6 Acknowledgments

This work was conducted within the Centre for Medical Systems Biology (CMSB), established by the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research (NGI/NWO). RJ was supported by the ErasmusMC Breedtestrategie.

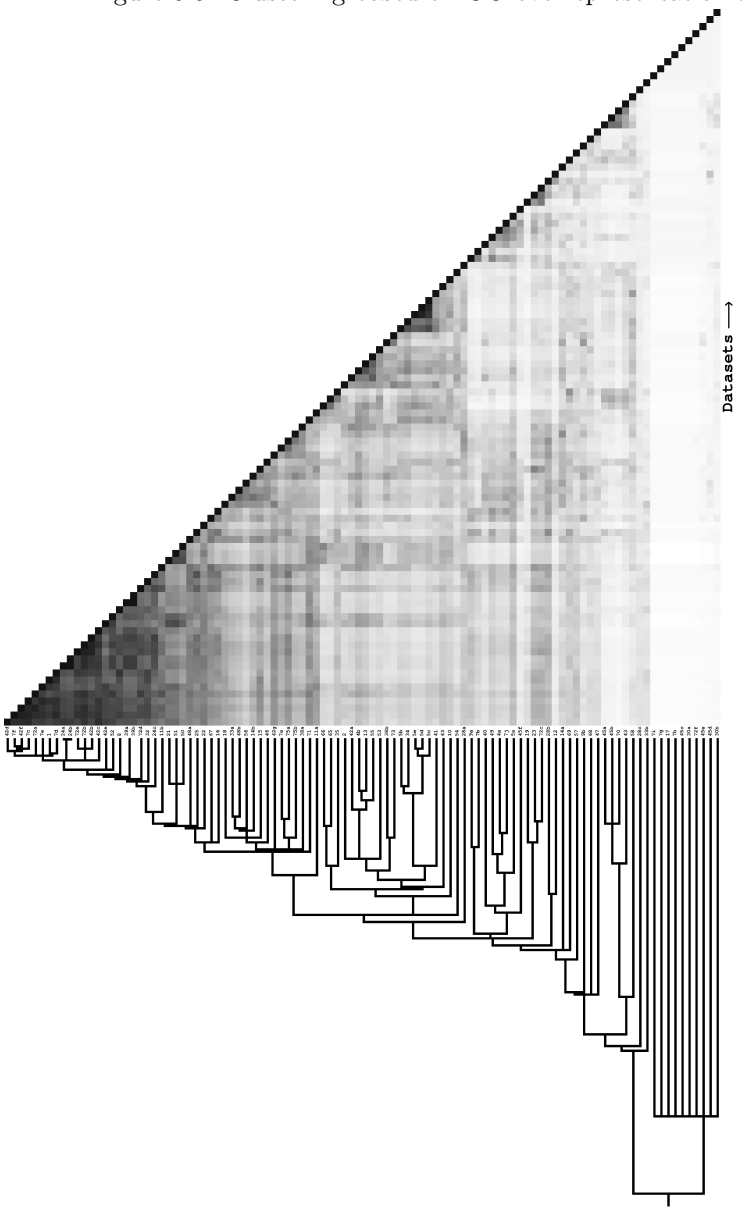
5.7 Appendix

Clustering and classification based on GO-overrepresentation analysis.

Table 5.2: Median classification scores (AUC) for the kappa and GO overrepresentation measure ($p<0.05$)

Group	kappa	GO
Dysferlinopathy	0.61	0.6
Dystrophinopathy	0.85	0.75
Regen. and diff.	0.53	0.5
Myositis	0.89	0.85
Ageing	0.86	0.96
Atrophy	0.71	0.44
Extraocular	0.97	0.51

Figure 5.5: Clustering based on GO overrepresentation analysis ($p<0.05$).



Description of datasets in the compendium.

The table gives for each dataset: 1. Pubmed ID (if available); 2. Year of publication; 3. Species (h =human, m=mouse, r=rat); 4. Platform specification: Affy = Affymetrix, Home = home spotted array, and LGTC = Leiden Genome Technology Center spotted arrays; 5. Studied cell type and specification: sm =skeletal muscle, mb=myoblast, mt=myotube, lc=lung carcinoma cells; 6. Studied condition; 7. Treatment; 8. Ex-

pert grouping (see section 5.2): 1=dystrophinopathy, 2=extraocular, 3=dysferlinopathy, 4=ageing, 5=myositis, 6=regeneration and differentiation, 7=atrophy.

ID	PubmedID	Y	S	Chip ID	T	Condition	Treatment	E
1	11823445	02	m	Affy U74A	sm: gastrocnemius, quadriceps	dystrophin	no	1
2	15983191	05	h	Home Sparks	sm: vastus lateralis	insulin-sensitive	high fat diet	
4a	12832294	03	r	Affy U34A	sm: extraocular	no	no	2
4b	12832294	03	r	Affy U34A	sm: extraocular	no	no	2
5a	15138310	04	r	Affy U34A	sm: extraocular	no	age 7d	
5b	15138310	04	r	Affy U34A	sm: extraocular	no	age 14d	
5c	15138310	04	r	Affy U34A	sm: extraocular	no	age 21d	
5d	15138310	04	r	Affy U34A	sm: extraocular	no	age 28d	
5e	15138310	04	r	Affy U34A	sm: extraocular	no	age 45d	
7a	12874102	03	m	Affy U74A	sm: hindlimb	dystrophin	no	
7b	12874102	03	m	Affy U74A	sm: hindlimb	dystrophin	no	
7c	12874102	03	m	Affy U74A	sm: hindlimb	dystrophin	no	1
7d	12874102	03	m	Affy U74A	sm: hindlimb	dystrophin	no	1
7e	12874102	03	m	Affy U74A	sm: hindlimb	dystrophin	no	1
7f	12874102	03	m	Affy U74A	sm: hindlimb	dystrophin	no	1
7g	12874102	03	m	Affy U74A	sm: extraocular	dystrophin	no	
7h	12874102	03	m	Affy U74A	sm: extraocular	dystrophin	no	
7i	12874102	03	m	Affy U74A	sm: extraocular	dystrophin	no	
7j	12874102	03	m	Affy U74A	sm: extraocular	dystrophin	no	
8	14506282	03	m	Affy U74A	sm: gastrocnemius, quadriceps	dysferlin	no	3
9a	15687108	05	h	Home Pennington	sm	no insulin sensi. change pre training	exercise	
9b	15687108	05	h	Home Pennington	sm	no insulin sensi. change post training	exercise	
10	14688207	04	m	Affy U74A+C	mb: C2C12	no	differentiation	6
11a	11121445	00	h	Affy HuGeneFL	sm	dystrophin	no	1
11b	11121445	00	h	Affy HuGeneFL	sm	dystrophin	no	1
12	14519683	03	h	Affy HuGeneFL	sm: biceps	FSHD	no	
13	12677001	03	h	Affy U95A+B	sm	nemaline myopathy	no	
14a	11572940	01	m	Affy U74A	sm: extraocular	no	no	2
14b	11572940	01	m	Affy U74A	sm: extraocular	no	no	2
15	12415109	02	h	Affy U95A	sm: quadriceps	dystrophin	no	1
16	12471055	02	h	Home Bologna	sm	dysferlin	no	3
17	14519196	03	h	Affy U95A	sm: vastus lateralis	no	exercise	6
18	15326121	04	h	Affy U133A	sm: orbital layer of ocular muscle	no	no	
19	12176734	02	m	Home INSERM DD	sm: hindlimb	dystrophin	no	1
21	12477723	03	m	Home NIH Caltech	sm: tibialis anterior	no	cardiotoxin	6
22	15126512	04	h	Home Viguerie	sm: vastus lateralis	no	epinephrine	
23	15598661	05	h	Affy U133A	sm: vastus lateralis	no	lipid infusion	
24a	12391344	02	h	Affy U95A	sm	inclusion body myositis	no	5
24b	12391344	02	h	Affy U95A	sm	polymyositis	no	5
24c	12391344	02	h	Affy U95A	sm	dermatomyositis	no	5
25	12937035	03	h	Affy U95A	sm: vastus lateralis	no	ec-&concentric contraction	
28a	14715702	04	r	Affy U34ABC	sm: gastrocnemius	no	space flight	7
28b	14715702	04	r	Affy U34ABC	sm: gastrocnemius	no	tail suspension	7
28c	14715702	04	r	Affy U34ABC	sm: gastrocnemius	no	denervation	7
30a	14618091	03	r	Affy U34A	sm: gastrocnemius	no	dexamethasone	
30b	14618091	03	r	Affy U34A	sm: gastrocnemius	no	dexamethasone + IGF	
32	12133862	02	m	Affy U74A	sm: gastrocnemius	dystrophin	no	1
33a	15475267	04	m	Affy U74ABC	mb: C2AS12	no	IGF	
33b	15475267	04	m	Affy U74ABC	mb: C2AS12	no	PDGF	
34	15056467	04	h	Affy U95AB	sm	nemaline myopathy non-typing	no	
35	15755682	05	h	Affy U133A	lc: A549tTA	Ala expansion in overexpr.Pabpn1	no	

ID	PubmedID	Y	S	Chip ID	T	Condition	Treatment	E
38a	15036332	04	m	Home Brown	sm: soleus (affected)	unknown - ky	no	
38b	15036332	04	m	Home Brown	sm: EDL (unaffected)	unknown - ky	no	
39a	15811552	05	m	Uniset mouse I	sm: quadriceps	dysferlin	no	3
39b	15811552	05	m	Uniset mouse I	sm: quadriceps	dysferlin	no	3
40	15272020	04	m	Affy U74A	sm: quadriceps	FOXO1	no	
41	15728662	05	h	Affy U133A	sm	mitochondrial encephalo-myopathy	no	
42a	14681298	04	m	Affy U74A	sm: diaphragm	dystrophin	no	
42b	14681298	04	m	Affy U74A	sm: diaphragm	dystrophin	no	
42c	14681298	04	m	Affy U74A	sm: diaphragm	dystrophin	no	1
42d	14681298	04	m	Affy U74A	sm: diaphragm	dystrophin	no	1
42e	14681298	04	m	Affy U74A	sm: diaphragm	dystrophin	no	1
42f	14681298	04	m	Affy U74A	sm: diaphragm	dystrophin	no	1
42g	14681298	04	m	Affy U74A	sm: diaphragm	no	no	
43	14625377	04	h	Affy U95A	sm: quadriceps, others	no	autopsy	
45a	15962335	05	h	Affy U133A	sm: deltoid	no	no	
45b	15962335	05	h	Affy U133A	sm: deltoid	no	no	
45c	15962335	05	h	Affy U133A	sm: deltoid	no	no	
45d	15962335	05	h	Affy U133A	sm: gastrocnemius	no	no	
45e	15962335	05	h	Affy U133A	sm: gastrocnemius	no	no	
45f	15962335	05	h	Affy U133A	sm: quadriceps	no	no	
46	12837262	03	m	Affy U74A	mb: C2C12	no	differentiation	6
47	12206806	02	m	Affy U74A	sm: hindlimb	dystrophin and dystrophin x utrophin	no	1
48a	15336692	04	h	LGTC muscle	mb: primary	no	differentiation	6
48b	15336692	04	h	LGTC Hu19K	mb: primary	no	differentiation	6
49	16679024	05	h	LGTC muscle	mb: primary	dystrophin	differentiation	
50	16011810	05	m	LGTC Mu7.5K	sm: quadriceps	dystrophin	no	1
51	16306063	05	m	LGTC Mu22K	sm: quadriceps	different severe muscular dystrophies	no	1
52	16306063	05	m	LGTC Mu22K	sm: quadriceps	dysferlin	no	3
53	-	05	m	LGTC Mu7.5K	sm: quadriceps	beta-sarcoglycan	no	
54	-	05	m	LGTC Mu7.5K	sm: quadriceps	gamma-sarcoglycan	no	
55	-	05	m	LGTC Mu22K	sm: psoas and quadriceps	calpain3	no	
56	-	05	m	LGTC Mu22K	sm: quadriceps	fxr	no	
57	-	05	m	LGTC Mu22K	mb: immortalized IM2	Ala expansion in overexpr. Pabpn1	no	
58	-	05	h	LGTC Hu19K	sm: quadriceps	Ala expansion in PABPN1 gene	no	
63	15608089	05	m	Affy MOE430A	mt: C2C12	no	starvation / atrophy	
65	12783983	03	h	Affy U133AB	sm: vastus lateralis	no	ageing	4
66	15036396	04	h	Affy U133AB	sm: vastus lateralis	no	ageing	4
67	11937576	02	h	Affy HuGeneFL	sm	juvenilie dermatomyositis	no	5
68	15079007	04	h	Affy U95A+U133A	sm: vastus lateralis	spastic paraplegia	no	
69	14755723	04	h	Affy U95A	sm	acute quadriplegic myopathy	no	
70	15687482	05	h	Affy U133A	sm: vastus lateralis	no	sarcopenia in older subjects	4
71	12844509	03	r	Affy U34A	sm: hindlimb	no	disuse	7
72a	16261416	05	m	Affy U74A	sm: diaphragm	dystrophin	no	1
72b	16261416	05	m	Affy U74A	sm: EDL	dystrophin	no	1
72c	16261416	05	m	Affy U74A	sm: gastrocnemius	dystrophin	no	1
72d	16261416	05	m	Affy U74A	sm: quadriceps	dystrophin	no	1
72e	16261416	05	m	Affy U74A	sm: soleus	dystrophin	no	1
72f	16261416	05	m	Affy U74A	sm: tibialis anterior	dystrophin	no	1
73	16263771	05	h	Affy U133AB	sm	dystrophin	oxandrolone	
75a	16288871	05	m	Affy U74ABC	sm: quadriceps, tibialis anterior	dysferlin	no	3
75b	16288871	05	m	Affy U74ABC	sm: quadriceps	no	no	

6
Anni 2.0

Anni 2.0: A multipurpose text-mining tool for the life sciences

R. Jelier[1], M.J. Schuemie[1], A. Veldhoven[1], G. Jenster[2], L.C.J. Dorssers[3] and J.A. Kors[1]

Departments of [1]Medical Informatics, [2]Urology and [3]Pathology
Erasmus MC, Rotterdam

Submitted for publication

Abstract

The large amount of biomedical literature and the high frequency at which new papers are published, seriously challenges the biomedical researcher to keep abreast of the current developments. Here we introduce a tool, Anni 2.0, designed to aid the researcher with a broad range of information needs. Anni provides an ontology-based interface to Medline and retrieves documents and associations for several classes of biomedical concepts, including genes, drugs and diseases, with proven text-mining technology. In this paper we illustrate Anni's usability by applying the tool to two user cases. First, a set of genes differentially expressed between localized and metastatic prostate cancer was interpreted. Based on our analysis we put forward a new hypothesis on the deregulated processes in metastatic prostate cancer. Second, a published literature-based knowledge discovery was reproduced: the application of thalidomide for the treatment of chronic hepatitis C. In a small number of steps we were able to reproduce this discovery and in the final query chronic hepatitis C was given the sixth rank. In addition, for two higher scoring diseases, publications have since been published that suggest potential therapeutic use of thalidomide. Anni is available online (`http://biosemantics.org/anni/`), and we believe she is a highly versatile and useful addition to the biomedical researcher's toolbox.

6.1 Background

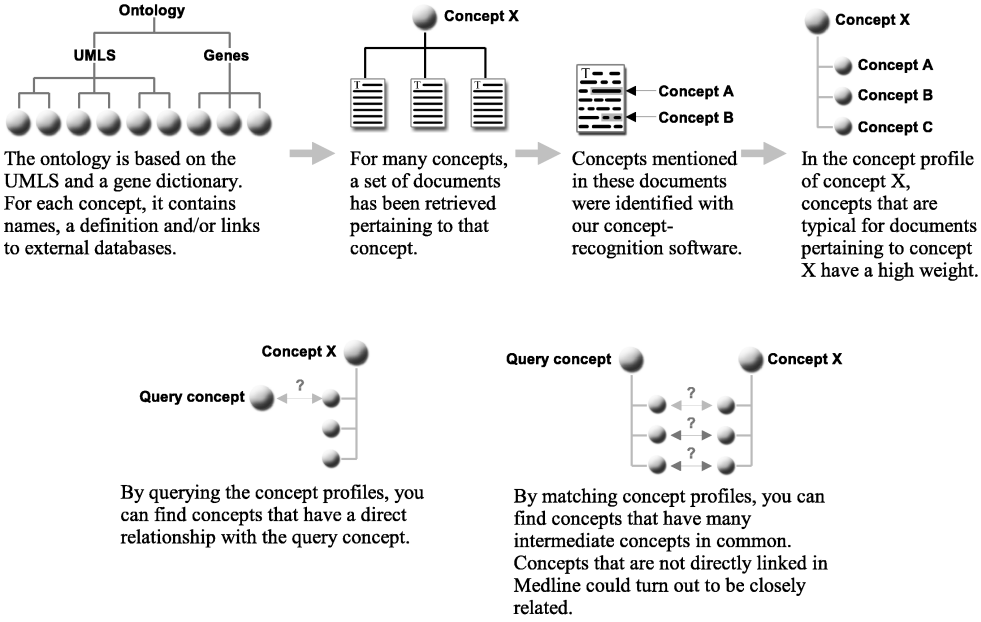
The amount of biomedical literature is vast and growing rapidly. It has become impossible for researchers to read all publications in their field of interest, which forces them to make a stringent selection of relevant articles to read. To keep abreast of the available knowledge, a wide range of initiatives has been deployed to mine the literature: from manual encoding of gene relations by the Gene Ontology Consortium [23], to automatic extraction of specific information such as transcript diversity [34], to the use of literature data for the prediction of disease genes [52, 118] (see [19, 27] for recent reviews). One of the emerging approaches is text-mining, which infers associations between biomedical entities by combining information from multiple papers. Text-mining approaches typically rely on occurrence and co-occurrence statistics of terms and have been successfully applied to a number of problems. The classic application is for literature-based knowledge discovery, which attempts to link disjunct sets of literature, in order to derive promising new hypotheses [40–44]. Swanson (see e.g. [45]) was a pioneer in this field and was able to publish several new hypotheses derived with the help of literature mining. His well known first example was the hypothesis that Raynaud's disease could be treated with fish oil [46], which was later corroborated experimentally [47]. Another field to which text-mining has been successfully applied is the analysis of DNA microarray data [58, 91, 92]. With these experiments, lists of hundreds of genes can be identified that are relevant to the studied phenomenon. The interpretation of such gene lists is challenging as for a single gene there can be hundreds or even thousands of articles pertaining to the gene's function. Text-mining can alleviate this complication by revealing the associations between the genes that are apparent from literature. This was the primary focus of the earlier version of Anni [119].

Here we present Anni 2.0, an ontology-based interface to the literature. An ontology contains concepts or defined entities, such as genes, diseases and drugs. Concepts come with a definition, a semantic type, a list of synonymous terms and can be linked to, for instance, online databases. References to concepts in texts can be identified with our indexing engine Peregrine [176]. The idea behind Anni is to relate or associate concepts to each other based on their associated sets of texts. Texts can be linked to a concept through automatic indexing, but also through manual efforts. The texts associated to a concept are characterized by a so-called concept profile [119] (see Figure 6.1 for an introduction into the technology behind Anni). A concept profile consists of a list of related concepts and each concept in the profile has a weight to signify its importance. Concept profiles have been succesfully used before to infer associations between genes [93, 119], Gene Ontology (GO) codes and genes [177], and to infer new associations between genes and the nucleolus [178] and between drugs and diseases [41].

Anni 2.0 is designed to be suitable for a broad range of applications in the biomedical domain. The tool provides concepts and concept profiles covering the full scope of the Unified Medical Language System[22], a biomedical ontology. In addition, the user is given extensive control to query for associations, to match concept profiles, and to explore the results. An important feature of Anni is transparency: all results can be traced back to the supporting documents.

Previously, we illustrated the utility of concept profiles to retrieve functional and relevant associations between various types of concepts [119, 177, 178]. Here, we evaluate our tool through two user cases. First we use Anni to analyse a DNA microarray dataset.

Figure 6.1: The technology behind Anni at a glance. Yellow balls indicate ontology concepts.



Second, we attempt to reproduce and expand a published literature-based knowledge discovery.

6.2 Implementation

Information sources

Anni is a Java client-server application and communicates with our server through Remote Method Invocation (RMI). She uses three information sources:

1. An ontology composed of the 2006AC version of the Unified Medical Language System (UMLS) ontology [22] and a gene thesaurus derived from multiple databases [99]. Following Aronson [125], the UMLS thesaurus was adapted for efficient natural language processing, avoiding overly ambiguous or duplicate terms, and terms that are very unlikely to be found in natural text. The gene thesaurus contains genes from three species: human, mouse and rat. Homologs from these three species were mapped through NCBI's Homologene database [66]. In addition, genes with identical nomenclature were mapped to each other.

2. A database with textual references to ontology concepts in Medline abstracts (from 1980 on) as identified by our concept-recognition software, Peregrine [176]. Apart from mapping synonomous terms to one concept as identified by the ontology, Peregrine attempts to disambiguate words or phrases that refer to multiple terms based

on contextual information. Abstracts were indexed together with the Medical Subject Headings (MESH) concepts. MESH is a controlled vocabulary and concepts are manually assigned to abstracts to facilitate docuement retrieval. The registry number field (RN field) contains information on chemicals to which the abstract refers and was also incorporated in the analysis. Indexation recall for genes was increased by taking common spelling variations into account [100].

3. A database with concept profiles based on the Medline indexation. The basis of a concept profile is a set of abstracts associated to a concept. For GO terms we used the papers associated to the term by the GO annotation consortium [23]. For genes the set of abstracts in which the gene occurs was taken, but from a subset of Medline containing documents on mammalian genes, selected by the Pubmed query "(gene OR protein) AND mammals". For the other concepts we relied on the complete Medline indexation. The weights in the concept profiles were derived by means of the symmetric uncertainty coefficient [128, 177]. For efficiency, we excluded from the concept profiles concepts with an association score lower than 10^{-8} and concepts that occurred only once in the Medline indexation.

Design paradigms

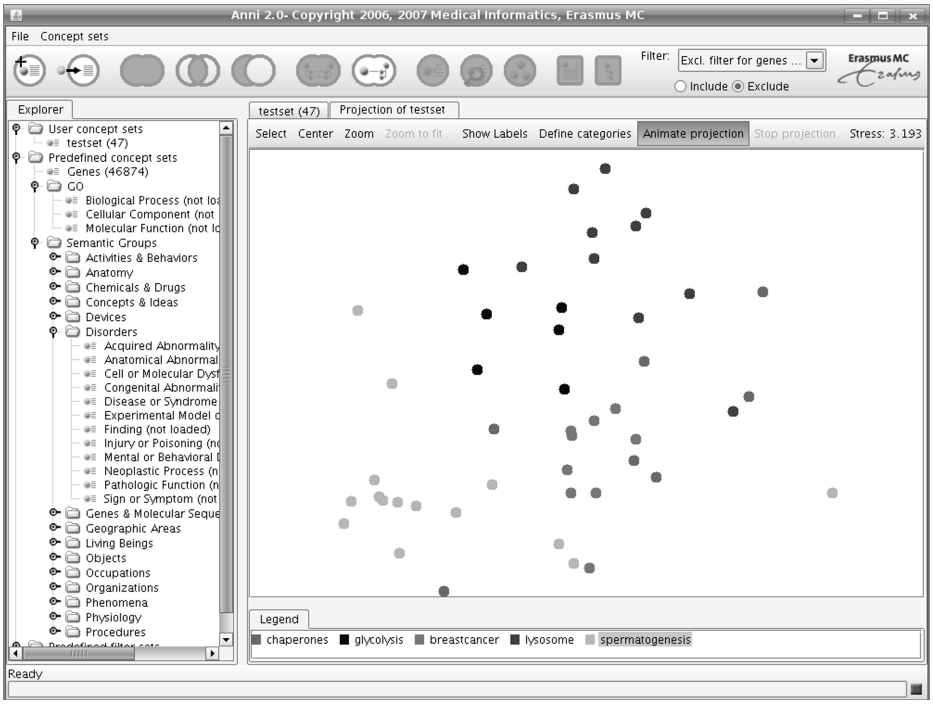
Anni is organized through concept sets, which are displayed in a tree view. Upon startup a range of predefined concept sets are loaded: the three branches of the Gene Ontology [21], the set of genes, and the semantic types as defined by the UMLS, e.g. "Disease or Syndrome" or ""Biologically Active Substance". Users can manipulate concept sets through basic set operations such as intersection, union and substraction, or they can create a new concept set and add concepts through an input panel. With the input panel the user can provide concept names or identifiers from several databases (a.o. Entrez Gene, Swissprot and Gene Ontology identifyers) through typing, pasting or loading a text file, and map them to concepts.

Wherever in the application concepts are shown, they can be selected and through a dropdown menu several options are available: show concept definition and semantic types; transfer concepts to a new concept set; show concept profile (if available).

In Anni, many concepts have a concept profile. Concept profiles can be both queried and matched. A query on concept profiles will retrieve concept association scores based on the concepts' co-occurrences, e.g. a query with the concept "prostate cancer" on the set of all genes will retrieve the genes mentioned together with this concept in abstracts, sorted by strength of association as measured by the uncertainty coefficient. Queries are performed with a query concept profile and query concepts can be individually weighted by the user. The table with the query results allows the user to sort on concept profiles that contained all the query concepts. In addition, the co-occurrence rate between concepts as observed in the Medline database can be shown in the query result table. The query result table can be explored through 2d hierarchical clustering and a heatmap.

Concept profiles can be matched to identify similarities between concept profiles, for instance to identify genes associated with similar biological processes. As a matching score we use a scaled inner product score between concept profiles. The user can use a filter to control which concepts are used for matching. Concept sets can be used as an inclusive filter, only the concepts in the concept set are used for matching, or as an exclusive filter, all concepts are used for matching except the concepts in the filter concept set.

Figure 6.2: Screenshot of Anni showing an MDS projection of a test set of 47 genes, organized in 5 groups through a shared commonality (see legend and [92, 119]). In the Explorer tab to the left, concept sets are organized in a tree. The toolbar on top provides concept set options and shows the current filter for matching concept profiles. The shown MDS view on a concept set can be used to get an overview of associations between the concepts, as used for instance in [178]. Groups of nodes can be selected and the similarities between their concept profiles analyzed in the annotation view. Nodes are colored based on user-defined features.



The associations between concept profiles within a concept set can be explored through hierarchical clustering or a Multi-Dimensional Scaling (MDS) projection (see Figure 6.2). Additionally, two concept sets can be matched, which will result in a matrix of association values. Similar to the query result table, the direct co-occurrence frequency can be shown. Concepts with a high association score but no Medline co-occurrences could indicate a new discovery: an association between concepts implicit in the literature but not yet explicitly described. The matrix can also be explored through 2d hierarchical clustering and a heatmap.

To provide transparency, Anni is equipped with an annotation view to evaluate the similarity within a group of concept profiles. The view provides a coherence measure, the average of the inner product scores of all possible pairs within the group. To aid the interpretation of the inner product scores also the probability is given that the same score or higher would be found in a randomly formed group of the same size. In addition, the percentual contributions of individual concepts to the coherence score are shown as well as the weights of these concepts in the individual concept profiles. Finally, every

Table 6.1: A selection of identified relevant clusters in the set of differentially expressed genes between metastatic and localized prostate cancer. The most descriptive concepts are shown as given by the Anni annotation view. The two left-most columns depict how many genes in the cluster were either up or down regulated in metastasized prostate cancer.

Up	Size	Descriptive concepts
A	24	kinetochores; mitosis; anaphase-promoting complex
B	5	nuclear proteins; tumor markers, biological
C	3	unfolded protein response
Down	**Size**	**Descriptive concepts**
A	7	complement system proteins
B	7	calponin; smooth muscle myosins
C	4	myosin phosphatase; smooth muscle (tissue)
D	7	extracellular matrix proteins
E	5	transcription factor; proto-oncogene proteins c-fos
F	4	cyclin-dependent kinases
G	3	melanosomes; membrane protein traffic; exocytosis
H	5	membrane transport proteins; symporter

association in a concept profile can be traced to the supporting documents.

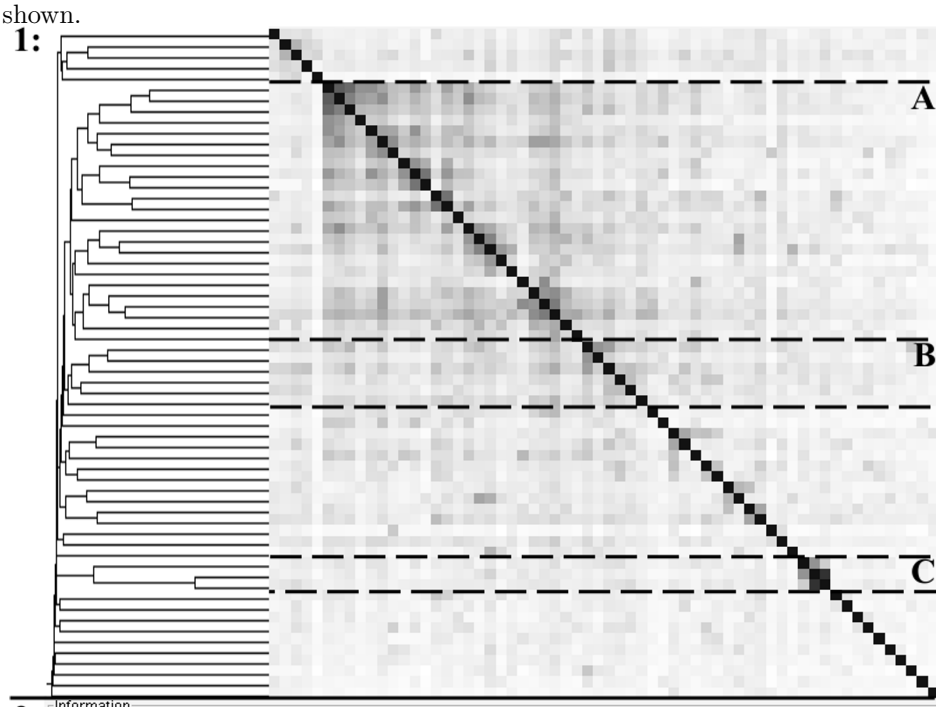
6.3 Results

User case 1: Analysis of a DNA microarray dataset

For this user case we applied Anni 2.0 to analyze a set of genes differentially expressed between localized and metastasized prostate cancer to unravel genes and pathways responsible for the progression of prostate cancer to metastatic disease. The dataset was generated based on three published studies [11, 179, 180]. Data from these studies was processed as in the original papers. For inclusion in our set, genes had to be in the top differentially expressed genes in at least 2 of the 3 studies. The set contained 69 genes expressed higher in metastasized cancer compared to local prostate cancer and 130 genes with lower expression (see appendix). As a first step we investigated if there were genes known to be associated to prostate cancer. We performed a query for the concept "malignant neoplasm of the prostate". 68 genes had a direct association through co-occurrence, which is a highly significant over-representation ($p = 2.04 \cdot 10^{-8}$) given the number of genes associated with this concept in the predefined concept set "Genes" .

To identify shared associated concepts between the genes in general, we clustered the up and down regulated genes separately using a broad semantic filter [119] (the filter is included as a predefined concept set). Figure 6.3 shows the clustering for genes up regulated in metastases and table 6.1 shows all identified clusters. First, we consider the analysis of genes down regulated in metastases. Two of the clusters are characterized by concepts apparently pertaining to the prostate stroma, such as "smooth muscle myosins" and "extracellular matrix proteins". This is expected as organ confined tumors contain stroma, whereas metastases, mainly from lymph nodes, are free of prostate stromal cells. Other gene clusters with lower expression in metastases pertain to the level of differentiation of

Figure 6.3: Panel 1 shows the clustering of genes up-regulated in prostate metastases. The clustering is based on the similarity of the concept profiles of the genes. Panel 2 depicts a fragment of the annotation for cluster A. The annotation view displays for a cluster a group cohesion score with a p-value, and a list of concepts with their percentual contribution to the score. In addition, the weights of the concepts in the concept profiles are shown.



2:

Information

Group size: 24

Average score: 0.0199

P(Average score = 0.0199): 1.0E-25 Estimate Parameters

Concept	Contribution (%)	BUB1B	CENPE	UBE2C	CENPF	AURKA	PTTG1
CDC2	6.1801	0.0002	0	5.814E-5	0.0003	0.0005	0.0007
CENPE	5.827	0.0077	0.0805	0	0.0076	5.886E-5	0
AURKA	4.2868	0.0011	0.0001	0	6.134E-5	0.1	0.0002
Kinetochores	4.2497	0.0142	0.0269	4.973E-5	0.0107	0.0005	0.0006
CDC25B	2.7352	0	0	0	0	0.0007	0
BUB1B	2.7143	0.0662	0.0091	0	0.003	0.0006	0.0006
Cyclin B	2.6059	0.0002	0.0002	0.0011	0.0001	0.0003	0.0025
KIF23	2.3484	0	0.0005	0	0	0	0
PTTG1	2.2478	0.0009	0	0.001	6.912E-5	7.044E-5	0.1027
Mitosis	2.1824	0.0021	0.0018	0.0003	0.0012	0.002	0.0012
Anaphase	2.0035	0.0023	0.0042	0.0009	0.0008	0.0006	0.0052
anaphase-promoting comple›	1.9781	0.0029	0.0007	0.0067	0.0003	0.0004	0.0054

the cancer cells and hence the grade of the cancer. Lower grade prostate tumors contain more differentiated epithelial cells that are involved in the secretion of prostatic fluid, which is reflected by clusters characterized by concepts such as: "membrane transport proteins" and "exocytosis" [119].

The clustering of the genes higher expressed in metastatic prostate cancer is dominated by the large cluster associated with kinetochores, anaphase-promoting complex and mitosis (see Figure 6.3). In this cluster, subclusters associated with "kinetochores", "mitotic checkpoint" and "anaphase promoting complex" indicate the cluster is not just a signature of proliferation, but shows associations to a specific phase in mitosis: the spindle checkpoint. Indeed, the concept "spindle checkpoint acgtivity" was the 13th concept (not counting genes) in the annotation for this cluster. The spindle checkpoint prevents a dividing cell to advance from metaphase into anaphase before all kinetochores are correctly attached to the mitotic spindles. A kinetochore is the protein structure assembled on the centromere that links the chromosome to the microtubules of the mitotic spindle. The anaphase promoting complex (APC) ubiquitin ligase plays an important role in controlling the progression to anaphase by triggering the appropriately timed, ubiquitin-dependent proteolysis of mitotic regulatory proteins. A perturbation involving the APC is apparent, as a query on "anaphase promoting complex" reveals that 11 of the up-regulated genes have a strong association ($> 10^{-5}$), a highly significant overrepresentation ($p < 5 \cdot 10^{-11}$). Using the links in the application to the underlying literature and the Entrez Gene database we can easily confirm the associations. For instance for the genes shown in Figure 6.3, panel 2: CENPE is a kinetochore protein and CENPF is essential for kinetochore attachment [181], BUB1B is a mitotic checkpoint protein interacting with the APC [182], PTTG1 and AURKA are substrates of the APC [183, 184] and UBE2C is one of the two ubiquitin-conjugating enzymes used by the APC [185, 186].

Deregulation of APC *in vitro* can result in defects in chromosome segregation, chromosomal instability, aneuploidy and increased sensitivity for tumorigenesis (see for a review [187]). Also, changed levels of APC regulators and substrates have been found to be correlated with cancer malignancy and, for some cancers, with tumor aggresiveness [188]. A causal relation between deregulation of APC and malignancy or tumor aggresiveness has been suggested to exist through a higher mutation rate. Nevertheless, causality is not established *in vivo*, and observed APC deregulation could also be a consequence of tumorigenesis and genomic instability. Interestingly, Lehman et al. [188] did not find an APC mitotic cluster in prostate cancer and attributed this observation to the low aggressiveness of prostate cancers. As they studied organ confined prostate cancer, this is in line with our observation here. Therefore it appears that also in prostate cancer, APC deregulation is correlated with tumor aggressiveness. Deregulation of the APC could have clinical consequences as some anti-neoplastic agents, such as nocodazole and taxol, work by activation of the spindle checkpoint [188]. Deregulation of the APC could therefore reduce the effectivity of these drugs. For instance, overexpression of UBE2C can cause the nocodazole induced mitotic blockade to be bypassed [189].

Concluding, with Anni we were able to functionally annotate a DNA microarray dataset. Genes published to be associated with prostate cancer were easily retrieved. We identified clusters with genes with lower expression levels in metastases likely associated with stroma and differentation features of cancer cells. Amongst the genes higher expressed in metastases we identified a cluster associated with the spindle checkpoint and the APC. This is a previously unknown feature of metastasized prostate cancer and may

be an indicator for the aggressiveness of the cancer.

User case 2: Literature-based knowledge discovery

Here, we illustrate Anni's knowledge discovery potential by reproducing a published literature-derived hypothesis. When looking for new therapeutic uses of the drug thalidomide, Weeber et al. [40] suggested, amongst others, that chronic hepatitis C could be treated with thalidomide. We selected this hypothesis as experimental evidence has recently emerged that appears to substantiate the claim [190, 191]. Weeber et al. took the following approach: First they retrieved from the MEDLINE database concepts of the UMLS semantic type "immunological factors" that occurred together in a sentence with thalidomide. Based on an expert's opinion they selected from the retrieved list the concept "interleukin-12" at position 7 in their list, to represent a biological process modulated by thalidomide, and queried for occurrence of this term in the same sentence with concepts of the semantic type "disease or syndrome". From the query results, diseases known to be associated with thalidomide were removed and after some additional manual curation, a shortlist was analyzed by an expert to identify diseases that could benefit from thalidomide treatment.

For reproducing this experiment we used the set of MEDLINE records published upto the time point given by Weeber et al. (07-2000), and generated concept profiles based on this set of records. In the following three simple steps we could reproduce Weeber et al.'s query:

1. Based on the predefined concept sets available in Anni we can readily select concepts belonging to a semantic type of choice. To reproduce Weeber et al.'s first filtering we select the predefined concept sets "Genes" and "Immunological factors", merge them and set the resulting set as an inclusive filter[1]. With this filter, the concept profile of thalidomide contained "interleukin-12", incidentally also at the 7th rank, which reproduces the first step of their approach.

2. As the next step, we queried the 8152 concepts of the predefined concept set "Disease or syndrome" for which a concept profile is available. Weeber et al. [40] describe the biological process they query for as follows: "Thalidomide has strong inhibitory effects on mononuclear cell production of IL-12 and a stimulatory effect on IL-10 production". Through these effects, thalidomide influences the balance of T-helper 1 versus T-helper 2 cells. Based on this description we generated the following query: "IL-12", "IL-10", "Th1 cells", "Th2 cells" and "peripheral mononuclear cells". All concepts in the query were given equal weight, and all concepts are required to occur in the disease concept profile.

3. As we are only interested in diseases not previously associated with thalidomide, all diseases mentioned with thalidomide in a MEDLINE record, up to 07-2000, were removed automatically from the resulted ranking (the query view can show MEDLINE co-occurrence rates). After this, some simple additional manual cleanup is required on the query result to create a shortlist for the expert[2].

The top 10 of our result is shown in Table 6.2. Chronic hepatitis C appears on the

[1] We include "genes" as genes in the UMLS thesaurus were removed in favour of our custom made gene

Table 6.2: Final ranking and scores for the query for "IL-12", "IL-10", "Th1 cells", "Th2 cells" and "peripheral mononuclear cells" on the concept set "Diseases or Syndromes".

Rank	Disease name	Score
1	Leishmaniasis	0.002417946
2	Schistosoma mansonii infection	5.68E-04
3	Extrinsic asthma	5.44E-04
4	Listeriosis	4.88E-04
5	HTLV-I Infections	3.44E-04
6	Hepatitis C, Chronic	3.43E-04
7	Tropical Spastic Paraparesis	3.17E-04
8	Epstein-Barr Virus Infections	2.73E-04
9	Hepatitis B, Chronic	2.38E-04
10	Filarial Elephantiases	2.38E-04

6th rank. Interestingly, of the higher scoring diseases we found that PubMed now contains preliminary studies into the use of thalidomide for the treatment of Leishmaniasis [192] and Listeriosis [193]. On closer inspection, an association between Leishmaniasis could actually have been found before 2000, because the parasite underlying the disease, Leishmania, had been mentioned in connection with thalidomide [194].

6.4 Discussion

Anni was applied to two very different user cases with good results: a new hypothesis on the progression of localized prostate cancer to metastatic disease and reproduction and extension of a previously published literature-based discovery. The tool has several innovative and useful features:

1. Anni uses a concept-based approach. In the application, definitions are available for the concepts, as well as links to external databases. In addition, when references to concepts are identified in texts, synonymous terms are mapped to the same concept. For this process, we pursued a high level of precision through a carefully curated ontology and by applying automatic homonym disambiguation. This is especially relevant for genes, as gene terminology is rich in synonymous and ambiguous terms [28, 195] and is also an important feature of information retrieval tools like iHop [196].

2. Anni can compare concepts based on similarities in the documents associated with these concepts. Therefore implicit relations between concepts can be found. In

thesaurus.

[2]The following curation was performed: 1. Diseases closely related to previously filtered diseases that had a known association to thalidomide were removed, e.g. "severe combined immunodeficiency" since thalidomide has been used to treat wasting in AIDS; 2. Unpractically broad disease concepts were removed, such as "parasitic infection"; 3. Closely related diseases were mapped to one, to reduce redundancy, e.g. "cutaneous leishmaniasis", "leishmaniasis" and "visceral leishmaniasis" were mapped to "leishmaniasis"; 4. Animal diseases were removed, e.g. "toxoplasmosis, animal".

addition the user has complete control over which concepts are taken into account during the comparison. Combined these features are very useful for knowledge discovery [41]. The approach also allows concepts to be included that are very hard to find in documents, such as GO codes, which are usually described with long, systematic terms.

3. Anni is a highly interactive application and offers a range of options to interactively explore the implicit and explicit associations between concepts. Query and match results can be viewed in a textual represenation or in a graphical form through hierachically clustered heatmap or MDS projection visualizations. In addition the tool provides a high level of transparency, which further improves the usability of the tool.

Anni is a multi-purpose text-mining tool and the modular set-up and broad range of biomedical concepts allow many more tasks than the ones presented. The broad applicability of Anni is a distinguishing factor when compared to previously published text-mining tools, which tend to focus on one application, such as knowledge discovery [44, 197] or the analysis of DNA microarray data [91, 93, 119]. Arrowsmith for example [44] can compare two document sets to each other at a time, which is well suited for knowledge discovery, but impractical when looking for associations between a group of genes.

The Anni system has some limitations. First of all, the system works with co-occurrence based associations. These associations not always reflect functional relations or facts. In addition, Anni relies on an ontology and automatic concept recognition in texts and both are not error free. For these reasons Anni was built to be transparent and all results can be traced back to the underlying documents. Another limitation is that only genes from mouse, rat and human are covered; support for other species is a development point.

In conclusion, Anni provides an innovative ontology-based interface to the literature, and builds on advanced and well evaluated text-mining technology. Anni is a highly versatile tool, applicable to a broad range of tasks. She is available online at www.biosemantics.org.

6.5 Acknowledgements

We gratefully acknowledge dr. Marc Weeber for help with the second user case 2. We thank our user group that patiently provided feedback that proved essential for the development of Anni 2.0. RJ was supported by an ErasmusMC Breedtestrategie grant. AV was supported by INFOBIOMED, 6th R&D Framework, EC (IST 2002 507585).

6.6 Appendix

Differentially expressed genes localized vs metastasized prostate cancer

Gene symbol	Entrez gene	Value	Gene symbol	Entrez gene	Value
ATF3	467	-1.92	WIF1	11197	-0.72
SRD5A1	6715	0.14	IQGAP2	10788	-0.90
EGR2	1959	-2.02	FLJ39822	151258	-1.12

Gene symbol	Entrez gene	Value	Gene symbol	Entrez gene	Value
DUSP5	1847	-1.14	MFAP4	4239	-1.34
MMP7	4316	-2.29	RAB27A	5873	-1.14
KRT17	3872	-2.03	SLC2A10	81031	-1.31
HSPC150	29089	1.75	PGM5	5239	-1.61
SLC4A4	8671	-0.97	EGR1	1958	-2.27
ARHE	390	-1.07	CDC2	983	1.19
NDUFV1	4723	0.47	MGC24665	116028	2.46
CYR61	3491	-1.77	PTTG1	9232	1.92
KIF11	3832	0.86	MGP	4256	-1.71
PCOLCE2	26577	-3.02		400047	-1.86
C10orf3	55165	1.22	ZIC2	7546	1.85
ZNF185	7739	-2.15	FRAS1	80144	0.44
NEXN	91624	-1.69	PDE5A	8654	-0.72
FBLN1	2192	-1.03	PTN	5764	-1.73
RAB27B	5874	-1.13	FOLH1	2346	-0.57
STC1	6781	0.64	SERPING1	710	-0.76
FABP7	2173	0.31	GTSE1	51512	0.66
BUB1B	701	1.07	MGST1	4257	-0.19
GREM1	26585	-0.95	FBXO32	114907	-0.15
SMC4L1	10051	1.33	UBE2M	9040	0.32
DNAJC3	5611	-0.66	HLA-DQB1	3119	-1.02
F3	2152	-3.31	CTGF	1490	-2.36
GAS1	2619	-0.75	FLJ23259	79782	0.75
B3GNT3	10331	0.58	ELOVL6	79071	0.10
BF	629	-1.38	KIAA0934	22982	-0.08
HMMR	3161	0.55	Pfs2	51659	1.16
CAV1	857	-0.95	KRT5	3852	-1.58
TMP21	10972	-0.71	CDCA5	113130	0.91
SHMT2	6472	0.99	APP	351	-1.02
RGS1	5996	-0.84	FLJ38507	389136	-1.22
CTSO	1519	-1.03	FHL1	2273	-2.14
CENPE	1062	0.83	ALDH1A2	8854	-0.93
AF1Q	10962	0.95	ATAD1	84896	-0.59
DLG7	9787	1.10	SULT1C1	6819	0.72
PDE8B	8622	-1.42	TGM4	7047	-2.08
GGH	8836	1.22	KIAA0056	23310	-1.33
PCP4	5121	-1.37	SPG20	23111	-0.68
MKI67	4288	1.01	FHL2	2274	-1.57
CKS2	1164	0.96	EZH2	2146	1.18
LMNB1	4001	1.80	PPP1R12B	4660	-0.31
EXOSC9	5393	0.83	C14orf24	283635	-0.34
UBE2C	11065	1.50	C1S	716	-0.97
TITF1	7080	1.32	PRC1	9055	0.96
COL12A1	1303	-0.49	ALDH1A3	220	-1.04
EYA4	2070	-0.91	MYH11	4629	-0.89
DOC1	11259	-0.99	MGC5178	79008	0.80
DKFZP564O0823	25849	-0.75	COLEC12	81035	-0.97
CSRP1	1465	-2.42	CNN1	1264	-2.78
UHRF1	29128	2.86	PLA2G2A	5320	-1.36
OGN	4969	-1.54	FLJ14681	84910	-1.00
CD9	928	-0.90	ACSL3	2181	-0.78
EVA1	10205	-1.28	EAF2	55840	-1.24
DIXDC1	85458	-0.80	MYLK	4638	-1.50
PDE4D	5144	-0.56	TNFSF10	8743	-1.42
CIT	11113	0.46	LIM	10611	-0.72
CDK6	1021	-0.53	SYNPO2	171024	-2.63
P2RY5	10161	-1.19	KCNMB1	3779	-1.34
C10orf48	283078	-1.29	DST	667	-0.33
RARRES1	5918	-0.68	RAB23	51715	-1.07
ZMPSTE24	10269	-0.71	PKIB	5570	0.10
	440423	0.82	TOX	9760	0.85
COL8A1	1295	-0.63	SOAT1	6646	-1.08
NEGR1	257194	-1.17	NAV1	89796	0.55
ST18	9705	0.00	CENPF	1063	0.93
PEG10	23089	1.39	CDK5R1	8851	-0.10
SMARCA1	6594	-0.81	ACTA2	59	-1.12
CDC25B	994	1.13	MRVI1	10335	-1.74
CYP1B1	1545	-1.42	TAGLN	6876	-3.22
TOP2A	7153	1.59	TRIM29	23650	-1.16
PROX1	5629	0.56	BICD1	636	-0.16
IGF1	3479	-1.55	TGFB1I4	8848	-0.35
ANTXR2	118429	-1.52	ABCC4	10257	-1.79
KIAA0830	23052	-1.33	MEIS2	4212	-0.84
TTK	7272	2.09	NUSAP1	51203	1.75
PLN	5350	-1.42	ENO2	2026	1.21
SSX3	10214	0.60	HOXB2	3212	1.72
LPL	4023	0.94	NPTX1	4884	0.84
DMXL1	1657	-0.88	ACTG2	72	-2.54
SSPN	8082	-1.09	ARL6IP6	151188	0.08
EDNRA	1909	-1.01	MAP4	4134	0.32
MELK	9833	2.43	SLC15A2	6565	-1.63
MT1K	4499	-2.00	FMOD	2331	-1.87
CCNB2	9133	0.89	ELOVL5	60481	-0.59
PTGS2	5743	-1.47	VIK	79027	-1.34
NEFH	4744	-1.94	DDIT4	54541	1.27

Gene symbol	Entrez gene	Value	Gene symbol	Entrez gene	Value
DIO2	1734	-0.91	NR4A1	3164	-1.09
KIAA1411	57579	-1.14	ASAH1	427	-0.52
SFRP1	6422	-1.10	FLJ11029	55771	1.41
CYBRD1	79901	-0.70		440945	1.90
STC2	8614	1.15	LTF	4057	-3.08
CPVL	54504	-1.44	TFRC	7037	0.37
PGM3	5238	-1.10	SQSTM1	8878	0.93
SLC22A3	6581	-1.05	RNASEH2A	10535	1.89
STK6	6790	1.29	TES	26136	-0.92
SEMA3C	10512	-1.32	H19	283120	1.07
KIF23	9493	0.85	ADAMTS1	9510	-1.08
RNASE4	6038	-0.96	SERPINA3	12	-1.75
CRIP2	1397	1.14	HELLS	3070	0.58

7 Discussion

This thesis presents the development and evaluation of co-occurrence based text-mining algorithms for several biomedical research problems. Two methods were presented and evaluated, the ACS and concept profiles. In this chapter I will discuss the main findings, the encountered problems, the limitations of the methods and indicate possible directions for further research.

7.1 Evaluation of the ACS

In chapter 2 we evaluated the ACS to aid in the interpretation of a list of differentially expressed genes. The first challenge was the lack of established evaluation procedures for this task, with several authors resorting to ad-hoc evaluations (see e.g. [58, 95]). To quantitatively evaluate the performance of the ACS and compare it to a simple gene co-occurence method, we made the assumption that the main task is the retrieval of functional associations between the genes. To evaluate performance on this task a test set was made of 5 groups of functionally related genes. Devising this test set proved a difficult task due to two factors: First, several high-confidence annotations provided by the Gene Ontology Annotation project [23] proved to be wrong and required us to go back to the literature and check if the remaining annotations were indeed correct. Second, preliminary studies with the ACS revealed relationships between members of different groups, and required us to perform a pilot study to assess the amount and impact of intergroup relationships. Eventually, the ACS was evaluated on the test set and achieved good scores for four of the five groups, even with small amounts of documents available per gene. The ACS performed significantly better than a simple gene co-occurence method. Our main conclusion was that the ACS would be a useful tool for the interpretation of gene lists. But the evaluations also revealed several characteristics of the methodology that limit its usability. One issue is that the ACS suffers from a lack of transparency and that users cannot easily verify why concepts are placed close together. Another issue arises with the assumption that the strongest associations between concepts, as derived from the high-dimensional dataset, can be mapped meaningfully to distances in a Euclidean ACS

space. Unfortunately, this cannot be guaranteed. When the associations between concepts are less clustered and more evenly distributed, the ACS will have more problems with accurately representing them. In the worst case this can result in a random positioning. Therefore an important further development for the ACS would be the design of a measure, potentially per concept, of how well the concept positioning reflects the associations in the high-dimensional space.

7.2 Concept profiles

In chapter 3 we introduce concept profiles to retrieve functional associations between genes. We first evaluated the method onthe controlled test set introduced in chapter 2. Concept profiles were found to perform better than the ACS. As a next step we wanted to take the evaluation of the methodology a step further, and apply it to actual DNA microarray data together with domain experts. To do this, the conceptual simplicity of concept profiles was exploited to design a transparent user interface.The tool, Anni, allowed researchers to quickly identify similarities between the concept profiles of a set of genes with hierarchical clustering and to evaluate the similarities underlying the found clusters. In addition, relevant associations from the concept profiles could be traced back to the literature. This transparency greatly facilitated the user cases, as it speeded up the annotation of clusters of genes and the verification of unexpected associations. The tool also facilitated the identification of errors in the machinery, for instance in the thesaurus. As case studies, two DNA microarray datasets were annotated with the tooland in both cases the analysis led to new insights. We concluded that through concept profiles a set of differentially expressed genes can be efficiently annotated based on both explicit and implicit information retrieved from the literature.

In chapter 4 we further explored the characteristics of concept profiles. The main focus was to evaluate different weighting schemes for the concepts' weights in the profiles. To compare the weighting schemes we chose to apply concept profiles to assign Gene Ontology codes to genes and constructed a test set based on the Gene Ontology Annotation project database [23]. The use of this test set had several advantages. First, the test set is very large, which allowed for more subtle effects to be retrieved. Second, the use of GO codes allowed to explore how suitable concept profiles are to retrieve functional associations between genes and other concepts. On the down side, as also illustrated in chapters 2 and 3, the GOA annotation database is incomplete and contains errors. In the evaluation, we compared two previously published weighting schemes, an averaging approach and the log-likelihood ratio measure that we proposed in chapter 3, and the uncertainty coefficient, an information theoretic measure. All methods performed well, and scored considerably higher than a binary approach without any weighting. When larger amounts of texts were available for the generation of the concept profiles, the performance of the methods diverged considerably, with the uncertainty coefficient then outperforming the two other weighting schemes. Overall, better performance was observed for specific concepts than for general concepts. The literature dealing with general concepts will contain more variation in the topics being discussed. It is possible that such a variety of topics cannot be usefully summarized in a concept profile. In that case, one road to improvement would be to explore options that leave more of the topic structure intact, such as k-nearest approaches or an approach which would allow multiple profiles per concept.

The positive experiences with concept profiles described in chapters 3 and 4, and else-

where [178], in combination with improvements in our thesauri and concept recognition software [100, 176] prompted us to bring concept profile based text-mining to a broader audience. In chapter 6 we present Anni 2.0, an online text-mining tool that provides an ontology-based view on the literature through concept profiles. Anni 2.0 is designed to be versatile and highly interactive. In Anni, concept profiles are available for concepts spanning the full breadth of the UMLS. The tool is organised around concept sets. Concept sets allow for a modular organization of the exploratory options and also provide the user with a lot of control in a straightforward manner, for instance to filter the concepts used in concept matching. Anni provides several options to explore and visualize concept associations, such as hierarchical clustering and a map projection through multi-dimensional scaling. The user cases in chapter 6, the reproduction of a literature-based discovery and the annotation of a set of differentially expressed genes, demonstrate the potential of the tool to aid biomedical researchers in obtaining an overview on the current literature and to generate new hypotheses.

One limitation of Anni is that it doesn't allow the user to change the set of available concepts, for instance if a concept of interest is lacking. To partly circumvent this issue, the user could be allowed to define a document set (e.g. with a pubmed query) that would then be characterized with a concept profile. This concept profile could then be used both as a representation of a missing concept, as well as a summary of the document set. Another logical extension of Anni would be to incorporate triplet relations (subject-relation-object) either retrieved from the same texts as the concept occurrence statistics, or from other information sources. Especially for knowledge discovery it would be very informative to explore the relations together with concept profile matching scores, for instance in a map view such as provided by cytoscape [198].

7.3 LASSO

Comparisons between DNA microarray studies are complicated by the large influence of biological variation and technical factors on the datasets. A potentially more robust approach is to compare studies based on the perturbed biological processes. In chapter 5 we introduced LASSO (literature-based association analysis) for this purpose. LASSO employs associations between genes, retrieved by concept profiles, to incorporate background knowledge for comparing DNA microarray datasets. The methodology was evaluated on a large compendium of 102 DNA microarray datasets taken from the field of muscle development and disease, and was compared to approaches based on gene overlap or GO code overrepresentation analysis. The LASSO based comparisons allowed for the identification of more parallels between the studies, and aids in the interpretation of these parallels.

The idea to compare DNA microarray datasets at the level of biological processes based on gene associations is novel and could be further extended in several directions. A logical first extension would be to further develop the methodology to work with normalized or raw data. In the compendium study, LASSO was most useful to retrieve associations between heterogenous experiments, across species and platforms. Given that concept profiles can be used to associate a broad range of concept types to each other, another extension would be to apply the methodology to compare datasets from different high-throughput techniques, e.g. transcriptomics and metabolomics experiments. Finally, the technology could be incorporated in a tool to allow the researcher to interactively explore similarities between datasets by modulating the associations being used. An Anni-like

tool could be used, e.g. to allow manipulation of the concepts used in the concept profile matching to retrieve gene associations. In addition, different other types of data could be used as a source of gene associations, such as GO annotations, metabolic networks, and protein-protein interactions.

7.4 Improving concept profiles

Concept profiles provide a simple framework to represent and compare concepts based on associated document sets. Several routes can be envisioned to improve the quality of concept profiles. More and more journals adopt an open access policy and make available full-text publications. More information can be retrieved from full-text publications than from abstracts, though at a lower density (see e.g. [199]). Therefore full-text scientific documents could be used to increase precision and recall for retrieving functional associations.

Another route to improve concept profiles would be to improve the named entity recognition. Three directions can be distinguished: 1. Improve the disambiguation of terms. Our machinery could for instance be extended to employ more syntactical information, acquired by part-of-speech tagging and shallow parsing, to increase precision. 2. Improve the recognition of spelling variations of terms. Proper handling of the flexibility of language can have a large impact on the recall [100, 200]. 3. Improve the thesaurus.

The quality of the thesaurus is an important factor for the performance of the named entity recognition system [31, 176], but it is difficult to acquire an adequate set of descriptors for the concepts. UMLS and gene databases, the sources currently used to obtain the descriptors, require a lot of manual curation to remove errors as well as strings of little value as descriptors (see also [125, 201]). Gene databases contain easily retrieved errors, such as "open reading frame" as a gene name, but also more subtle errors like a gene name which actually represents a class of enzymes, e.g. "aldehyde dehydrogenase". UMLS and the gene thesaurus also contain poor descriptors, such as unnecessarily ambiguous strings (e.g. "period" as a synonym for "menstruation", or "Other") or concepts that are practically indistinguishable from each other (UMLS distinguishes 7 concepts for "Cold")[1]. Another type of error is that thesauri tend to be incomplete. Apart from the huge number of concepts that can be distinguished in a knowledge domain (UMLS contains more than a million concepts), the meaning of words changes over time and new words are continuously being invented. Most language systems though, are generated by organizations with limited manpower. A solution for this latter problem may come from community-based knowledge repositories such as wikipedia (`www.wikipedia.org`): user communities that adapt and improve the ontology and the descriptor set and that can immediately profit from their efforts by text-mining.

The use of ontology concepts as features has advantages. The use of an ontology allows a certain standardization of language. Semantic smoothing can be achieved through the mapping of synonyms, the recognition of multi-word terms and resolution of the meaning of ambiguous words. Another advantage is that ontologies allow to link to other information about the concepts. In Anni, for instance, link-outs are provided to external databases, concept definitions are provided, and categorizations of the concepts are offered

[1] To clean up the ontology and the descriptor set we eventually used 37 general filtering patterns, 5 rewrite rules, and applied thousands of concept-specific filtering and mapping steps.

through semantic types. But also some disadvantages of thesaurus concepts have been noted, especially in the context of information retrieval. In information retrieval, the task of retrieving the appropriate documents for an information need, the standard approach relies on lexical features (words, bi-grams). It has been reported by us (see e.g. [202, 203]) and by others [204, 205] that performance for queries only based on concepts can go up in individual cases, but will go down on average or at least not improve. Small average performance gains have been reported when concepts are combined with words as features [204, 205]. Several causes can be distinguished. First, good named entity recognition requires a complicated machinery and errors will occur, for instance in the disambiguation of terms. Second, closely related concepts can be distinguished, but their relationship is not used by the system. This can result in lower recall, for instance when for a query on "acute myeloid leukemia", the relation between this concept and "pediatric acute myeloid leukemia" is not taken into account. Third, incompleteness of the ontologies and the set of descriptors reduce recall. Consequently, also in the context of concept profiles it should be explored how the use of simple lexical features can be combined with concept tagging to improve performance. Also, ontological relationships are currently not taken into account in the matching process. Use of these or other information sources to incorporate the relatedness of concepts could increase the robustness of concept matching.

7.5 Text-mining and the curse of dimensionality

From a machine-learning perspective, text mining involves a huge amount of features in comparison to the amount of data (the curse of dimensionality). That is, to learn, without generalizations, how these features should behave to signify, for instance, a functional association between genes, much more data would be required than is available. To circumvent this phenomenon, several authors proposed dimension reduction techniques to identify functional associations between genes [94–96]. The authors of these papers claim better performance, but evaluation of these methodologies has been scarce and requires further research. The main disadvantage of the proposed methods is that they make the grounds for the predictions less traceable, and given the ambiguities and errors involved in text-mining, a high level of interactivity with users will likely remain important. Another disadvantage is that the dimension reduction is derived from the data. This implies that the concepts mappings not necessarily make sense semantically. Another route to identify a manageable set of features is to use generalizations about concepts based on background information, such as, as mentioned earlier, ontological relations like parent-child relations or semantic categorizations like the UMLS semantic types. The semantic mapping could be optimized similar to conventional dimension reduction techniques to maintain a maximum of variation between features.

7.6 Outlook

The wealth of information stored in literature and the comparative ease at which text-mining based associations can be retrieved and used, make text-mining a prime candidate to be combined with other data sources to achieve specific goals. For example, Xie et al. [117] used text mining together with sequence homology and protein domain information to assign Gene Ontology (GO) codes to proteins. Others combine gene expression data

with text mining to identify disease genes [52, 118]. For these specific tasks, the use of statistical techiques to combine the information sources may be the most promising. Torvik et al. [206] for example used a logistic regression model to better prioritize the most "relevant" terms that connect two disjunct sets of documents (the B-terms in the closed knowledge discovery model). They employed 8 features of the B-terms, such as a measure to indicate whether the concept is general or specific, the date of the first occurence of the concept in Medline, and a measure that indicates how characteristic the term is for the two document sets.

Language is flexible, ambiguous, complex, and requires a lot of background knowledge to understand. Retrieving knowledge from texts with computational methods is therefore difficult. One wonders why scientists choose to publish their research in a format from which it is hard to recover, and which makes their findings difficult to integrate with what was previously known. We still publish very similarly to the way we did before computers became available. Is this an anachronism? A recent editorial [207] in Nature heralded the database revolution and indeed, more and more data is being shared in online databases, and peers, journals and funding agencies put pressure on researchers to participate. Apart from data, new initiatives are arising to unlock higher-level information from unstructured text. Gerstein et al. [208] proposed that authors should supply with their publications a standardized encoding of the results in triplet relationships. Another Nature editorial [209] asked attention for an initiative to start a wikipedia for professionals, an online encyclopedia for scientific topics made and maintained by scientists themselves. Several other community-based knowledge repositories have recently sprung into existence, for example to link biological network resources (`http://pgrc.ipk-gatersleben.de/BINCO-wiki`), or to share workflows and experimental protocols (`http://myexperiment.org`). Whatever the form of the system, the goal of the current developments is the online and real-time sharing of data and knowledge, and not only in a textual format. Still, it is hard to imagine scientific discussion without human language, which is so close to human thinking. Therefore, whatever the future form of scientific publishing may be, I think text-mining is here to stay.

8 Summary

High-throughput technologies, such as DNA microarrays, have become common practice in molecular biology. The technologies allow for the measurements of large numbers of mRNAs, proteins or metabolites in parallel. The interpretation of the datasets generated with high-throughput technologies can be problematic as hundreds of the measured entities can be found to be associated with the experimental variable. The analysis of these datasets can therefore confront researchers with unprecedented information needs. Unfortunately the required information is often published primarily in the unstructured free-text form of scientific publications, which cannot be used directly in computational systems. In addition, the large amount of biomedical literature and the high frequency at which new papers are published, seriously challenges the biomedical researcher to keep abreast of the current developments. Computerized algorithms can be used to help extract and use information from the scientific literature. One of the emerging approaches is text-mining, which infers associations between biomedical entities by combining information from multiple papers. Text-mining typically uses the occurrence and co-occurrence statistics of concepts or lexical features, such as words or bi-grams. Text-mining can be used to retrieve associations between concepts that are explicitly mentioned in documents, but also implicit associations can be found: associations never mentioned, but inferred from other associations. This thesis presents the development, evaluation and application of co-occurrence based text-mining algorithms for molecular biology and biomedicine, with a special focus on the analysis of data generated by high-throughput technology.

In chapter 2, the associative concept space (ACS) is evaluated for the retrieval of functional relationships between genes. The ACS is a Euclidean space in which thesaurus concepts are positioned and where the distances between concepts indicate their relatedness. The ACS uses co-occurrence of concepts in documents as a source of information. Next to direct co-occurrences, indirect relations are used to determine the relatedness of concepts. To assess the performance of the ACS we composed a test set of five groups of functionally related genes. With the ACS, good scores were obtained for four of the five groups. When compared to a simple gene co-occurrence method, the ACS is capable of revealing more functional biological relations and can achieve results with less literature available per gene. The performance of ACS proved to be affected though, by the addition

of large amounts of randomly selected literature. Hierarchical clustering was performed on the ACS output, as a potential aid to users, and was found to provide useful clusters. The results suggest that the algorithm can be of value for researchers studying large numbers of genes.

Chapter 3 presents concept profiles to retrieve functional associations between genes. A concept profile summarizes the context in which a gene is mentioned in literature and consists of a list of associated concepts together with weights that indicate the strength of the association. The weight of each concept in the profile is based on a likelihood ratio measure. Gene concept profiles are clustered to identify related genes. The experimental validation was performed in two steps. The method was first applied on the controlled test set previously used to test the ACS and was found to perform better than the ACS. Next, the datasets from two DNA microarray experiments were systematically annotated and the results were evaluated by domain experts. The first dataset was a gene-expression profile that characterizes the cancer cells of a subgroup of acute myeloid leukemia patients. For this group of patients the biological background of the cancer cells is largely unknown. Using our methodology we found an association of these cells to monocytes, which agreed with other experimental evidence. The second data set consisted of differentially expressed genes following androgen receptor stimulation in a prostate cancer cell line. Based on the analysis, we put forward a hypothesis about the biological processes induced in these studied cells: secretory lysosomes are involved in the production of prostatic fluid and their development and/or secretion are androgen-regulated processes.

In chapter 4 three schemes are evaluated for weighting the associations in a concept profile: 1. Weighting based on averaging, an empirical approach; 2. The log likelihood ratio, a test-based measure; 3. The uncertainty coefficient, an information-theory based measure. The weighting schemes were applied in a large-scale classification system that annotates genes with Gene Ontology codes. As the gold standard for our study, we used the annotations provided by the Gene Ontology Annotation project. All methods performed well, and scored considerably higher than a binary approach without any weighting. Especially for the more specific Gene Ontology codes excellent performance was observed. Overall, the differences between the methods were small, however the number of documents that were linked to a concept proved to be an important variable. When larger amounts of texts were available for the generation of the concept profiles, the performance of the methods diverged considerably, with the uncertainty coefficient then outperforming the two other weighting schemes.

Comparative analysis of expression microarray studies can confirm findings from individual studies and identify interesting parallels between studies. However, such analyses are hampered by the large influences of design, technical and statistical factors on the found differentially expressed genes. Comparisons based on perturbed biological processes could be more robust as different genes may hint at the same process. Chapter 5 describes and evaluates LASSO (literature-based association analysis) for this task. LASSO uses gene associations, derived from concept profiles, to quantify the similarity between studies and to reveal overlapping biological processes. The method was evaluated and compared to classical Gene Ontology group over-representation analysis through a comparative meta-analysis on 102 microarray studies published in the field of muscle development and disease. The over-representation analysis did not perform well, due to limited sensitivity and the incompleteness of the manually curated gene annotations. LASSO retrieved many more biologically meaningful links between studies, even across

species, microarray platforms and between studies that did not have any differentially expressed genes in common. Hierarchical clustering demonstrated limited influence of technical factors and correct grouping of muscular dystrophy, regeneration and myositis studies. LASSO facilitates finding common biological denominators in microarray studies, without raw data analysis or curated gene annotation databases.

Chapter 6 introduces Anni 2.0, a tool designed to aid the researcher with a broad range of information needs, based on concept profiles. Anni provides an ontology-based interface to Medline and retrieves documents and associations for several classes of biomedical concepts, including genes, drugs and diseases. Anni is evaluated through two user cases. First, a set of genes differentially expressed between localized and metastatic prostate cancer was interpreted. Based on our analysis we put forward a new hypothesis on the deregulated processes in metastatic prostate cancer. Second, a published literature-based knowledge discovery was reproduced: the application of thalidomide for the treatment of chronic hepatitis C. In a small number of steps we were able to reproduce this discovery and in the final query chronic hepatitis C was given the sixth rank. In addition, for two higher scoring diseases, publications have since been published that suggest potential therapeutic use of thalidomide. Anni is available online.

9 Samenvatting

Recent beschikbaar gekomen technologieën, zoals DNA microarrays, maken het mogelijk om in een enkel experiment de concentratie van vele mRNA moleculen, eiwitten of metabolieten te meten. In korte tijd zijn deze technieken deel gaan uit maken van het standaard repertoire in de moleculaire biologie. Echter, bij het gebruik van deze methoden worden grote hoeveelheden gegevens gegenereerd. Bij de analyse van de data blijken soms honderden van de gemeten moleculen geassocieerd met de bestudeerde experimentele variabele. Voor de interpretatie van het experiment is daarom doorgaans veel achtergrondinformatie nodig. Dit is in de praktijk problematisch, onder andere omdat veel van de benodigde informatie opgeslagen is in de vorm van ongestructureerde tekst in wetenschappelijke publicaties. Deze informatie is daarmee niet direct beschibaar voor de analyse. Een mogelijke oplossing is het gebruik van computationele technieken om automatisch informatie uit tekst te extraheren. Eén van de aanpakken is text mining, waarbij associaties tussen biomedische concepten worden afgeleid door het combineren van informatie uit meerdere wetenschappelijke publicaties. Text mining is over het algemeen gebaseerd op de analyse van de frequenties waarmee woorden of (referenties naar) concepten worden gebruikt in een set teksten. De methodologie kan worden gebruikt om associaties terug te vinden die expliciet worden genoemd in de bestudeerde teksten, maar kan ook worden gebruikt om impliciete associaties te vinden: associaties die niet worden genoemd, maar die worden afgeleid uit overige associaties. In dit proefschrift wordt verslag gedaan van de ontwikkeling, de evaluatie en de toepassing van text mining algoritmes voor de moleculaire biologie en de biomedische wetenschappen, met een speciale focus op de analyse van data gegenereerd met DNA microarrays.

In hoofdstuk 2 wordt de evaluatie beschreven van de "associative concept space" (ACS) voor het identificeren van functioneel gerelateerde genen op basis van een set wetenschappelijke publicaties. De ACS is een Euclidische ruimte waarin concepten worden gepositioneerd. Concepten zijn entiteiten die gedefinieerd zijn in een thesaurus of ontologie, in dit geval genen en andere biomedische concepten zoals medicijnen en ziekten. De afstand tussen twee concepten in de ACS reflecteert de sterkte van associatie. De afstand tussen twee concepten volgt niet alleen uit directe of expliciete relaties zoals die in een tekst worden beschreven, maar ook uit indirecte relaties, bijvoorbeeld als beide concepten sterk

geassocieerd zijn met een derde concept. Om te evalueren hoe goed afstanden tussen genen in de ACS kunnen worden gebruikt om functionele relaties te identificeren, wordt een test set gebruikt van vijf groepen functioneel gerelateerde genen. Voor vier van de vijf groepen kwamen de afstanden tussen de genen in de ACS goed overeen met de bekende functionele relaties tussen de genen. De ACS is ook vergeleken met een methode die alleen gebruik maakt van de directe relaties tussen twee genen in documenten en bleek meer functionele relaties terug te vinden met minder documenten beschikbaar per gen. De prestaties van de ACS worden echter wel negatief beïnvloed te worden door de toevoeging van grote hoeveelheden random geselecteerde documenten. Als een mogelijk hulpmiddel voor gebruikers is hierarchisch clusteren toegepast op de ACS. Met deze methode zijn bruikbare clusters van functioneel gerelateerde genen gevonden. Op basis van deze resultaten wordt geconcludeerd dat de ACS van waarde kan zijn bij de interpretatie van DNA microarray datasets.

In hoofdstuk 3 worden concept profielen geïntroduceerd en geëvalueerd voor het identificeren van functioneel gerelateerde genen. Een concept profiel karakteriseert de context waarin een gen wordt genoemd in de literatuur en bestaat uit een lijst van geassocieerde concepten samen met een gewicht dat de sterkte van de associatie aangeeft. De gewichten in een concept profiel worden bepaald met een op een waarschijnlijkheids ratio gebaseerde maat. Concept profielen kunnen makkelijk worden vergeleken en met behulp van clustering technieken kunnen vergelijkbare profielen worden gegroepeerd. Om te evalueren of zo ook functioneel gerelateerde genen kunnen worden gevonden is een experimentele validatie in twee stappen uitgevoerd. Eerst is de methode toegepast op de in hoofdstuk 2 geïntroduceerde test set. Bij de evaluatie werd met concept profielen een significant beter resultaat gehaald dan met de ACS. Vervolgens is de methode gebruikt voor de interpretatie van twee DNA microarray datasets, in samenwerking met domein experts. De eerste dataset betrof een gen expressie profiel dat de kankercellen van een subgroep van acute myeloïde leukemie patiënten karakteriseerd. Voor deze groep van patiënten is de biologische achtergrond van de kankercellen niet bekend. Met behulp van de analyse met concept profielen kon een associatie van deze cellen met monocyten worden gevonden, welke overeen bleek te komen met andere experimentele bevindingen. De tweede dataset betreft genen waarvan de expressie verandert als gevolg van de stimulatie van de androgeen receptor in een prostaatkanker cellijn. Op basis van de analyse is een nieuwe hypothese gevormd omtrent de geïnduceerde biologische processen in de bestudeerde cellen: secretoire lysosomen zijn betrokken bij de productie van prostaatvloeistof en hun ontwikkeling en/of secretie wordt gereguleerd door androgenen.

In hoofdstuk 4 worden drie methoden geëvalueerd om de sterkte van associatie te wegen in concept profielen: 1. Het middelen van document profielen; 2. De waarschijnlijkheids ratio; 3. De genormalizeerde wederzijdse informatie, een informatietheoretische maat. Als evaluatie is een grootschalig classificatie experiment gebruikt, waarbij genen zijn geannoteerd met concepten uit de Gene Ontology. Als gouden standaard voor deze evaluatie zijn de annotaties van het Gene Ontology Annotation project gebruikt. Met alle drie de methoden zijn goede resultaten gehaald, en allen scoorden aanzienlijk beter dan een aanpak zonder weging. Met name voor specifieke Gene Ontology concepten zijn uitstekende scores mogelijk. Over de hele test set genomen zijn de verschillen tussen de methoden klein. Echter de grootte van de set documenten die geassocieerd is met een concept en ten grondslag ligt aan het concept profiel, is een belangrijke variabele. Als meer documenten beschikbaar zijn voor het afleiden van de concept profielen verschillen

de prestaties voor de methoden aanzienlijk en presteert de genormalizeerde wederzijdse informatie maat het best.

Voor de analyze van DNA microarray experimenten kan het nuttig zijn de gegevens te vergelijken met ander experimenten. Zo kunnen bijvoorbeeld resultaten van individuele studies worden bevestigd, maar kunnen er ook interessante parallellen tussen studies worden gevonden. Zulke analyses worden echter belemmerd door de grote invloed van studie ontwerp, evenals technische en statistische factoren op de gevonden veranderingen in genexpressie. Het zou robuuster kunnen zijn studies te vergelijken op basis van de veranderde biologische processen, aangezien expressie veranderingen in meerdere genen kunnen wijzen op een verandering van hetzelfde biologisch proces. In hoofdstuk 5 wordt LASSO (literature-based association analysis) beschreven. Met LASSO worden associaties tussen genen, geïdentificeerd met concept profielen, gebruikt om studies te vergelijken en biologische processen die veranderd zijn in beide studies te herkennen. De methode is geëvalueerd door middel van een vergelijkende meta-analyse van 102 microarray studies met betrekking tot spierontwikkeling en spierziekten. Ter vergelijking is ook de klassieke methode van een op Gene Ontology gebaseerde groep overrepresentatie analyse op het compendium toegepast. De overrepresentatie analyse gaf echter geen goede resultaten, voornamelijk vanwege beperkte sensitiviteit en de onvolledigheid van de manueel gecureerde annotaties. Met LASSO werden meer biologisch relevante associaties tussen studies te gevonden, ook tussen verschillende organismen, microarray platformen, en tussen studies die geen enkel differentieel tot expressie gekomen gen deelden. Uit een groepering van de datasets op basis van een hierarchische clustering bleek een beperkte invloed van technische factoren en een correct clusteren van spierdystrofie, regeneratie en myositis studies. Concluderend, LASSO faciliteert het vinden van parallellen tussen microarray studies op basis van de veranderde biologische processen, zonder heranalyse van ruwe data en zonder gebruik van manueel gecureerde gen annotaties.

In hoofdstuk 6 is Anni 2.0 geïntroduceerd, een applicatie dat met behulp van concept profielen onderzoekers kan helpen met een breed scala aan informatiebehoeften. Anni geeft een concept gebaseerde blik op de wetenschappelijke literatuur en kan de gebruiker helpen documenten en associaties te vinden voor een breed scala van biomedische concepten, inclusief genen, medicijnen en ziekten. Anni is geëvalueerd aan de hand van twee toepassingen. Ten eerste is een set genen geanalyseerd die in expressie veranderen bij de overgang van gelocaliseerde prostaatkanker tot metastaserende prostaatkanker. Op basis van de analyse kon een nieuwe hypothese worden geformuleerd omtrent de verstoorde processen die bijdragen tot metastaserende prostaatkanker. Ten tweede is Anni gebruikt om een gepubliceerde literatuur gebaseerde ontdekking uit 2003 te reproduceren: de toepassing van thalidomide voor de behandeling van chronische hepatitis C. Met een klein aantal handelingen kon de ontdekking worden gereproduceerd. In de uiteindelijke ordening kreeg chronische hepatitis C de zesde rang. Bovendien zijn voor twee ziekten met een hogere score sindsdien publicaties verschenen waarin een mogelijke therapeutische toepassing van thalidomide wordt gesuggereerd. Anni is vrijelijk beschikbaar via het internet.

Bibliography

[1] Solé RV, Munteanu A, Rodriguez-Caso C, Macía J (2007) Synthetic protocell biology: from reproduction to computation. Philos Trans R Soc Lond B Biol Sci

[2] Heinemann M, Panke S (2006) Synthetic biology–putting engineering into biology. Bioinformatics 22:2790–2799

[3] Carninci P (2007) Constructing the landscape of the mammalian transcriptome. J Exp Biol 210:1497–1506

[4] Claverie JM (2001) Gene number. What if there are only 30,000 human genes? Science 291:1255–1257

[5] Chen M, Ying W, Song Y, Liu X, Yang B, Wu S, Jiang Y, Cai Y, He F, Qian X (2007) Analysis of human liver proteome using replicate shotgun strategy. Proteomics 7:2479–2488

[6] Beecher C (2003) Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis, chap. The Human Metabolome, pp. 311–319. Kluwer Academic Publishers (Boston)

[7] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487

[8] Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

[9] Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: understanding cancer using microarrays. Nat Genet 37 Suppl:S38–S45

[10] Valk PJM, Verhaak RGW, Beijen MA, Erpelinck CAJ, van Waalwijk van Doorn-Khosrovani SB, Boer JM, Beverloo HB, Moorhouse MJ, van der Spek PJ, Löwenberg B, Delwel R (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 350:1617–1628

[11] Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM (2007) Integrative molecular concept modeling of prostate cancer progression. Nat Genet 39:41–51

[12] Hendriksen PJM, Dits NFJ, Kokame K, Veldhoven A, van Weerden WM, Bangma CH, Trapman J, Jenster G (2006) Evolution of the androgen receptor pathway during progression of prostate cancer. Cancer Res 66:5012–5020

[13] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. Cell 102:109–126

[14] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313:1929–1935

[15] Galperin MY (2007) The Molecular Biology Database Collection: 2007 update. Nucleic Acids Res 35:D3–D4

[16] Yi Y, Mirosevich J, Shyr Y, Matusik R, George AL (2005) Coupled analysis of gene expression and chromosomal location. Genomics 85:401–412

[17] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucleic Acids Res 35:D760–D765

[18] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM (2001) The Stanford Microarray Database. Nucleic Acids Res 29:152–155

[19] Jensen LJ, Saric J, Bork P (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7:119–129

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

[21] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

[22] Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32:D267–D270

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

[24] Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27–30

[25] Los RK, Roukema J, van Ginneken AM, de Wilde M, van der Lei J (2005) Are structured data structured identically? Investigating the uniformity of pediatric patient data recorded using OpenSDE. Methods Inf Med 44:631–638

[26] Cohen AM, Hersh WR (2005) A survey of current work in biomedical text mining. Brief Bioinform 6:57–71

[27] Erhardt RAA, Schneider R, Blaschke C (2006) Status of text-mining techniques applied to biomedical text. Drug Discov Today 11:315–325

[28] Schijvenaars BJA, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA (2005) Thesaurus-based disambiguation of gene symbols. BMC Bioinformatics 6:149

[29] Schuemie MJ, Kors JA, Mons B (2005) Word sense disambiguation in the biomedical domain: an overview. J Comput Biol 12:554–565

[30] Koike A, Takagi T (2004) Gene/Protein/Family Name Recognition in Biomedical Literature. In: L Hirschman, J Pustejovsky (eds.) HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases, pp. 9–16. Association for Computational Linguistics, Boston, Massachusetts, USA

[31] Hirschman L, Yeh A, Blaschke C, Valencia A (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6 Suppl 1:S1

[32] Blaschke C, Andrade MA, Ouzounis C, Valencia A (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. In: Proc Int Conf Intell Syst Mol Biol, pp. 60–67

[33] Sekimizu T, Park H, Tsujii J (1998) Identifying the interaction between genes and gene products based on frequently seen verbs in Medline Abstracts. Genome Inform Ser Workshop Genome Inform 9:62–71

[34] Shah PK, Jensen LJ, Boue S, Bork P (2005) Extraction of transcript diversity from scientific literature. PLoS Comput Biol 1:e10

[35] Wattarujeekrit T, Shah PK, Collier N (2004) PASBio: predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics 5:155

[36] Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17 Suppl 1:S74–S82

[37] Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 20:604–611

[38] Appelt D, Israel D (1999) Introduction to information extraction technology. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence. Stockholm, Sweden

[39] Ananiadou S, Kell DB, ichi Tsujii J (2006) Text mining and its potential applications in systems biology. Trends Biotechnol 24:571–579

[40] Weeber M, Vos R, Klein H, Berg LTWDJVD, Aronson AR, Molema G (2003) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inform Assoc 10:252–259

[41] Srinivasan P (2004) Text mining: generating hypotheses from MEDLINE. JASIST 55:396–413

[42] Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR (2004) Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics 20:389–398

[43] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74:289–298

[44] Smalheiser NR, Torvik VI, Bischoff-Grethe A, Burhans LB, Gabriel M, Homayouni R, Kashef A, Martone ME, Perkins GA, Price DL, Talk AC, West R (2006) Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators. J Biomed Discov Collab 1:8

[45] Swanson DR (1990) Medical literature as a potential source of new knowledge. Bull Med Libr Assoc 78:29–37

[46] Swanson D (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 30:7–18

[47] DiGiacomo RA, Kremer JM, Shah DM (1989) Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. Am J Med 86:158–164

[48] Weeber M, Klein H, de Jong-van den Berg L, Vos R (2001) Using concepts in literature-based discovery, simulating Swanson's Raynaud-fish oil and migrain-magnesium discoveries. Journal of the American Society for Information Science 52:548–557

[49] Jenssen TK, Laegreid A, Komorowski J, Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28:21–28

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

[51] Stapley BJ, Kelley LA, Sternberg MJE (2002) Predicting the sub-cellular location of proteins from text using support vector machines. In: Pac Symp Biocomput, pp. 374–385

[52] Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. Nat Genet 31:316–319

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

[54] Zhang B, Schmoyer D, Kirov S, Snoddy J (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 5:16

[55] Masys DR, Welsh JB, Fink JL, Gribskov M, Klacansky I, Corbeil J (2001) Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics 17:319–326

[56] Shatkay H, Feldman R (2003) Mining the biomedical literature in the genomic era: an overview. J Comput Biol 10:821–855

[57] Pustejovsky J, Castaño J, Zhang J, Kotecki M, Cochran B (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput pp. 362–373

[58] Chaussabel D, Sher A (2002) Mining microarray expression data by literature profiling. Genome Biol 3:Research0055

[59] Becker KG, Hosack DA, Dennis G, Lempicki RA, Bright TJ, Cheadle C, Engel J (2003) PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics 4:61

[60] Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 27:1210–4, 1216–7

[61] Van der Eijk C, Van Mulligen EM, Kors JA, Mons B, Van den Berg J (2004) Constructing an associative concept space for literature-based discovery. JASIST 55:436–444

[62] Swanson D, Smalheiser N (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91:183–203

[63] Raychaudhuri S, Altman RB (2003) A literature-based method for assessing the functional coherence of a gene group. Bioinformatics 19:396–401

[64] Wren JD, Garner HR (2004) Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. Bioinformatics 20:191–198

[65] Keen JC, Davidson NE (2003) The biology of breast carcinoma. Cancer 97:825–833

[66] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, et al. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35:D5–12

[67] McCray AT, Srinivasan S, Browne AC (1994) Lexical methods for managing variation in biomedical terminologies. In: Proc Annu Symp Comput Appl Med Care, pp. 235–239

[68] Van Mulligen EM, Van der Eijk CC, Kors JA, Schijvenaars BJA, Mons B (2002) Research for research: tools for knowledge discovery and visualization. In: Proc AMIA Symp, pp. 835–839

[69] Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293

[70] Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36

[71] Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician 37:36–48

[72] Sammon J (1969) A nonlinear mapping for data structure analysis. IEEE trans Comput C-18:401–409

[73] Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863–14868

[74] Metz C (1978) Basic princliples of ROC analysis. Semin Nucl Med 8:283–298

[75] Gwynn B, Eicher EM, Peters LL (1996) Genetic localization of Cd63, a member of the transmembrane 4 superfamily, reveals two distinct loci in the mouse genome. Genomics 35:389–391

[76] Nykjaer A, Dragun D, Walther D, Vorum H, Jacobsen C, Herz J, Melsen F, Christensen EI, Willnow TE (1999) An endocytic pathway essential for renal uptake and activation of the steroid 25-(OH) vitamin D3. Cell 96:507–515

[77] Lisi S, Pinchera A, McCluskey RT, Willnow TE, Refetoff S, Marcocci C, Vitti P, Menconi F, Grasso L, Luchetti F, Collins AB, Marino M (2003) Preferential megalin-mediated transcytosis of low-hormonogenic thyroglobulin: a control mechanism for thyroid hormone release. Proc Natl Acad Sci U S A 100:14858–14863

[78] Kuwasako K, Shimekake Y, Masuda M, Nakahara K, Yoshida T, Kitaura M, Kitamura K, Eto T, Sakata T (2000) Visualization of the calcitonin receptor-like receptor and its receptor activity-modifying proteins during internalization and recycling. J Biol Chem 275:29602–29609

[79] Gagnon AW, Kallal L, Benovic JL (1998) Role of clathrin-mediated endocytosis in agonist-induced down-regulation of the beta2-adrenergic receptor. J Biol Chem 273:6976–6981

[80] Martínez A, Vos M, Guédez L, Kaur G, Chen Z, Garayoa M, Pío R, Moody T, Stetler-Stevenson WG, Kleinman HK, Cuttitta F (2002) The effects of adrenomedullin overexpression in breast tumor cells. J Natl Cancer Inst 94:1226–1237

[81] Oehler MK, Norbury C, Hague S, Rees MC, Bicknell R (2001) Adrenomedullin inhibits hypoxic cell death by upregulation of Bcl-2 in endometrial cancer cells: a possible promotion mechanism for tumour growth. Oncogene 20:2937–2945

[82] Zudaire E, Martínez A, Cuttitta F (2003) Adrenomedullin and cancer. Regul Pept 112:175–183

[83] Fernandez-Sauze S, Delfino C, Mabrouk K, Dussert C, Chinot O, Martin PM, Grisoli F, Ouafik L, Boudouresque F (2004) Effects of adrenomedullin on endothelial cells in the multistep process of angiogenesis: involvement of CRLR/RAMP2 and CRLR/RAMP3 receptors. Int J Cancer 108:797–804

[84] Weeber M, Schijvenaars BJ, Mulligen EMV, Mons B, Jelier R, Eijk CCVD, Kors JA (2003) Ambiguity of human gene symbols in LocusLink and MEDLINE: creating an inventory and a disambiguation test collection. AMIA Annu Symp Proc pp. 704–708

[85] Liu H, Lussier YA, Friedman C (2001) Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. J Biomed Inform 34:249–261

[86] Resnik P, Yarowsky D (2000) Distinguishing systems and distinguishing senses: new evaluation tools for word sense disambiguation. Natural Language Engineering 5:113–133

[87] Podowski RM, Cleary JG, Goncharoff NT, Amoutzias G, Hayes WS (2004) AZuRE, a scalable system for automated term disambiguation of gene and protein names. Proc IEEE Comput Syst Bioinform Conf pp. 415–424

[88] Hatzivassiloglou V, Duboué PA, Rzhetsky A (2001) Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics 17 Suppl 1:S97–106

[89] Stephens M, Palakal M, Mukhopadhyay S, Raje R, Mostafa J (2001) Detecting gene relations from Medline abstracts. In: Pac Symp Biocomput, pp. 483–495

[90] Blaschke C, Oliveros JC, Valencia A (2001) Mining functional information associated with expression arrays. Funct Integr Genomics 1:256–268

[91] Raychaudhuri S, Chang JT, Imam F, Altman RB (2003) The computational analysis of scientific literature to define and recognize gene expression clusters. Nucleic Acids Res 31:4553–4560

[92] Jelier R, Jenster G, Dorssers LCJ, van der Eijk CC, van Mulligen EM, Mons B, Kors JA (2005) Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics 21:2049–2058

[93] Glenisson P, Coessens B, Vooren SV, Mathys J, Moreau Y, Moor BD (2004) TXTGate: profiling gene groups with text-based information. Genome Biol 5:R43

[94] Homayouni R, Heinrich K, Wei L, Berry MW (2005) Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics 21:104–115

[95] Küffner R, Fundel K, 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

[96] Chagoyen M, Carmona-Saez P, Shatkay H, Carazo JM, Pascual-Montano A (2006) Discovering semantic features in the literature: a foundation for building functional associations. BMC Bioinformatics 7:41

[97] Manning C, Schütze H (1999) Foundation of statistical natural language processing. The MIT press, Cambridge MA

[98] Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. Computational Statistics 19:61–74

[99] Kors J, Schuemie M, Schijvenaars B, Weeber M, Mons B (2005) Combination of genetic databases for improving identification of genes and proteins in text. In: Biolink Conference

[100] Schuemie MJ, Mons B, Weeber M, Kors JA (2007) Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. J Biomed Inform 40:316–324

[101] Salton G (1989) Automatic text processing: The transformation, analysis, and retrieval of information by computer. Addison-Wesley, Reading, MA.

[102] Baugh LR, Hill AA, Brown EL, Hunter CP (2001) Quantitative analysis of mRNA amplification by in vitro transcription. Nucleic Acids Res 29:E29

[103] Hsing LC, Rudensky AY (2005) The lysosomal cysteine proteases in MHC class II antigen presentation. Immunol Rev 207:229–241

[104] Lennon-Duménil AM, Bakker AH, Maehr R, Fiebiger E, Overkleeft HS, Rosemblatt M, Ploegh HL, Lagaudrière-Gesbert C (2002) Analysis of protease activity in live antigen-presenting cells shows regulation of the phagosomal proteolytic contents during dendritic cell activation. J Exp Med 196:529–540

[105] Hoffbrand AV, Pettit JE (1993) Essential Haematology. Blackwell Science, Oxford.

[106] Mukaida N, Harada A, Matsushima K (1998) Interleukin-8 (IL-8) and monocyte chemotactic and activating factor (MCAF/MCP-1), chemokines essentially involved in inflammatory and immune reactions. Cytokine Growth Factor Rev 9:9–23

[107] Cella M, Döhring C, Samaridis J, Dessing M, Brockhaus M, Lanzavecchia A, Colonna M (1997) A novel inhibitory receptor (ILT3) expressed on monocytes, macrophages, and dendritic cells involved in antigen processing. J Exp Med 185:1743–1751

[108] Jenster G (1999) The role of the androgen receptor in the development and progression of prostate cancer. Semin Oncol 26:407–421

[109] Stinchcombe J, Bossi G, Griffiths GM (2004) Linking albinism and immunity: the secrets of secretory lysosomes. Science 305:55–59

[110] Chen Y, Samaraweera P, Sun TT, Kreibich G, Orlow SJ (2002) Rab27b association with melanosomes: dominant negative mutants disrupt melanosomal movement. J Invest Dermatol 118:933–940

[111] El-Amraoui A, Schonn JS, Küssel-Andermann P, Blanchard S, Desnos C, Henry JP, Wolfrum U, Darchen F, Petit C (2002) MyRIP, a novel Rab effector, enables myosin VIIa recruitment to retinal melanosomes. EMBO Rep 3:463–470

[112] Fukuda M (2005) Versatile role of Rab27 in membrane trafficking: focus on the Rab27 effector families. J Biochem (Tokyo) 137:9–16

[113] Tolmachova T, Anders R, Stinchcombe J, Bossi G, Griffiths GM, Huxley C, Seabra MC (2004) A general role for Rab27a in secretory cells. Mol Biol Cell 15:332–344

[114] Warhol MJ, Longtine JA (1985) The ultrastructural localization of prostatic specific antigen and prostatic acid phosphatase in hyperplastic and neoplastic human prostates. J Urol 134:607–613

[115] Utleg AG, Yi EC, Xie T, Shannon P, White JT, Goodlett DR, Hood L, Lin B (2003) Proteomic analysis of human prostasomes. Prostate 56:150–161

[116] Yuryev A, Mulyukov Z, Kotelnikova E, Maslov S, Egorov S, Nikitin A, Daraselia N, Mazo I (2006) Automatic pathway building in biological association networks. BMC Bioinformatics 7:171

[117] Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L (2002) Large-scale protein annotation through gene ontology. Genome Res 12:785–794

[118] Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33:1544–1552

[119] Jelier R, Jenster G, Dorssers LCJ, Wouters B, Hendriksen P, Mons B, Delwel R, Kors JA (2007) Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. BMC Bioinformatics 8:14

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

[121] Wren J (2004) Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics 5:145

[122] Blaschke C, Leon EA, Krallinger M, Valencia A (2005) Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics 6 Suppl 1:S16

[123] Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 6 Suppl 1:S17

[124] Raychaudhuri S, Chang JT, Sutphin PD, Altman RB (2002) Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. Genome Res 12:203–214

[125] Aronson AR (2006) Filtering the UMLS metathesaurus for MetaMap. Tech. rep., National Library of Medicine

[126] Croft W (1983) Experiments with representation in a document retrieval system. Information Technology: Research and Development 2:1–21

[127] Wilbur WJ, Yang Y (1996) An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Comput Biol Med 26:209–222

[128] Goodman L, Kruskal W (1979) Measures of association for cross classifications. Springer-Verlag, New York

[129] Rhodes DR, Chinnaiyan AM (2005) Integrative analysis of the cancer transcriptome. Nat Genet 37 Suppl:S31–S37

[130] Larsson O, Wennmalm K, Sandberg R (2006) Comparative microarray analysis. OMICS 10:381–397

[131] Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21:171–178

[132] Tan PK, Downey TJ, Spitznagel EL, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC (2003) Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 31:5676–5684

[133] Mah N, Thelin A, Lu T, Nikolaus S, Kühbacher T, Gurbuz Y, Eickhoff H, Klöppel G, Lehrach H, Mellgård B, Costello CM, Schreiber S (2004) A comparison of oligonucleotide and cDNA-based microarray systems. Physiol Genomics 16:361–370

[134] Shi L, Tong W, Fang H, Scherf U, Han J, Puri RK, Frueh FW, Goodsaid FM, Guo L, Su Z, Han T, Fuscoe JC, Xu ZA, Patterson TA, Hong H, Xie Q, Perkins RG, Chen JJ, Casciano DA (2005) Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. BMC Bioinformatics 6 Suppl 2:S12

[135] Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, Kohane IS (2003) Reproducibility of gene expression across generations of Affymetrix microarrays. BMC Bioinformatics 4:27

[136] Draghici S, Khatri P, Eklund AC, Szallasi Z (2006) Reliability and reproducibility issues in DNA microarray measurements. Trends Genet 22:101–109

[137] Kuo WP, Liu F, Trimarchi J, Punzo C, Lombardi M, Sarang J, Whipple ME, et al. (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. Nat Biotechnol 24:832–840

[138] Manoli T, Gretz N, Gröne HJ, Kenzelmann M, Eils R, Brors B (2006) Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics 22:2500–2506

[139] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550

[140] Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20:93–99

[141] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA (2003) Global functional profiling of gene expression. Genomics 81:98–104

[142] Khatri P, Done B, Rao A, Done A, Draghici S (2005) A semantic analysis of the annotations of the human genome. Bioinformatics 21:3416–3421

[143] Cahan P, Ahmad AM, Burke H, Fu S, Lai Y, Florea L, Dharker N, Kobrinski T, Kale P, McCaffrey TA (2005) List of lists-annotated (LOLA): a database for annotation and comparison of published microarray gene lists. Gene 360:78–82

[144] Finocchiaro G, Mancuso F, Muller H (2005) Mining published lists of cancer related microarray experiments: identification of a gene expression signature having a critical role in cell-cycle control. BMC Bioinformatics 6 Suppl 4:S14

[145] Fleiss J (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76:378–382

[146] (2006). Rat Genome Database Web Site, Medical College of Wisconsin, Milwaukee, Wisconsin. World Wide Web

[147] (2006). Mouse Genome Database (MGD), Mouse Genome Informatics Web Site, The Jackson Laboratory, Bar Harbor, Maine. World Wide Web

[148] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4:P3

[149] Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33:W741–W748

[150] Turk R, Sterrenburg E, van der Wees CGC, de Meijer EJ, de Menezes RX, Groh S, Campbell KP, Noguchi S, van Ommen GJB, den Dunnen JT, 't Hoen PAC (2006) Common pathological mechanisms in mouse models for muscular dystrophies. FASEB J 20:127–129

[151] Porter JD, Khanna S, Kaminski HJ, Rao JS, Merriam AP, Richmonds CR, Leahy P, Li J, Guo W, Andrade FH (2002) A chronic inflammatory response dominates the skeletal muscle molecular signature in dystrophin-deficient mdx mice. Hum Mol Genet 11:263–272

[152] Li P, Oparil S, Feng W, Chen YF (2004) Hypoxia-responsive growth factors upregulate periostin and osteopontin expression via distinct signaling pathways in rat pulmonary arterial smooth muscle cells. J Appl Physiol 97:1550–8; discussion 1549

[153] Wang D, Oparil S, Feng JA, Li P, Perry G, Chen LB, Dai M, John SWM, Chen YF (2003) Effects of pressure overload on extracellular matrix expression in the heart of the atrial natriuretic peptide-null mouse. Hypertension 42:88–95

[154] Kii I, Amizuka N, Minqi L, Kitajima S, Saga Y, Kudo A (2006) Periostin is an extracellular matrix protein required for eruption of incisors in mice. Biochem Biophys Res Commun 342:766–772

[155] Trueblood NA, Xie Z, Communal C, Sam F, Ngoy S, Liaw L, Jenkins AW, Wang J, Sawyer DB, Bing OH, Apstein CS, Colucci WS, Singh K (2001) Exaggerated left ventricular dilation and reduced collagen deposition after myocardial infarction in mice lacking osteopontin. Circ Res 88:1080–1087

[156] Siegel S, Castellan N (1988) Nonparametric statistics for the behavorial sciences. McGraw-Hill, New York

[157] Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. Stat Med 9:811–818

[158] Nemoto H, Konno S, Nakazora H, Miura H, Kurihara T (2007) Histological and immunohistological changes of the skeletal muscles in older SJL/J mice. Eur Neurol 57:19–25

[159] Chen YW, Nagaraju K, Bakay M, McIntyre O, Rawat R, Shi R, Hoffman EP (2005) Early onset of inflammation and later involvement of TGFbeta in Duchenne muscular dystrophy. Neurology 65:826–834

[160] Deconinck N, Dan B (2007) Pathophysiology of duchenne muscular dystrophy: current hypotheses. Pediatr Neurol 36:1–7

[161] Turk R, Sterrenburg E, de Meijer EJ, van Ommen GJB, den Dunnen JT, 't Hoen PAC (2005) Muscle regeneration in dystrophin-deficient mdx mice studied by gene expression profiling. BMC Genomics 6:98

[162] van Lunteren E, Moyer M, Leahy P (2006) Gene expression profiling of diaphragm muscle in alpha2-laminin (merosin)-deficient dy/dy dystrophic mice. Physiol Genomics 25:85–95

[163] Bakay M, Zhao P, Chen J, Hoffman EP (2002) A web-accessible complete transcriptome of normal human and DMD muscle. Neuromuscul Disord 12 Suppl 1:S125–S141

[164] Boer JM, de Meijer EJ, Mank EM, van Ommen GB, den Dunnen JT (2002) Expression profiling in stably regenerating skeletal muscle of dystrophin-deficient mdx mice. Neuromuscul Disord 12 Suppl 1:S118–S124

[165] Cao PR, Kim HJ, Lecker SH (2005) Ubiquitin-protein ligases in muscle wasting. Int J Biochem Cell Biol 37:2088–2097

[166] Glass DJ (2003) Molecular mechanisms modulating muscle mass. Trends Mol Med 9:344–350

[167] Pasterkamp RJ, Verhaagen J (2006) Semaphorins in axon regeneration: developmental guidance molecules gone wrong? Philos Trans R Soc Lond B Biol Sci 361:1499–1511

[168] Ko JA, Gondo T, Inagaki S, Inui M (2005) Requirement of the transmembrane semaphorin Sema4C for myogenic differentiation. FEBS Lett 579:2236–2242

[169] Welle S, Brooks AI, Delehanty JM, Needler N, Thornton CA (2003) Gene expression profile of aging in human muscle. Physiol Genomics 14:149–159

[170] Welle S, Brooks AI, Delehanty JM, Needler N, Bhatt K, Shah B, Thornton CA (2004) Skeletal muscle gene expression profiles in 20-29 year old and 65-71 year old women. Exp Gerontol 39:369–377

[171] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 101:9309–9314

[172] Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. Bioinformatics 20:3166–3178

[173] Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clin Cancer Res 10:2922–2927

[174] DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R (2006) Combining results of microarray experiments: a rank aggregation approach. Stat Appl Genet Mol Biol 5:Article15

[175] Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. Nucleic Acids Res 30:e48

[176] Schuemie M, Jelier R, Kors J (2007) Peregrine: Lightweight gene name normalization by dictionary lookup. In: Proceedings of Biocreative 2

[177] Jelier R, Schuemie M, Roes P, Van Mulligen E, Kors J (2007) Literature-based concept profiles for gene annotation: the issue of weighting. accepted for publication in IJMI

[178] Schuemie M, Chichester C, Lisacek F, Coute Y, Roes PJ, Sanchez JC, Kors J, Mons B (2007) Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of MEDLINE. Proteomics 7:921–931

[179] Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A 101:811–816

[180] Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, Shah RB, Chandran U, Monzon FA, Becich MJ, Wei JT, Pienta KJ, Ghosh D, Rubin MA, Chinnaiyan AM (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 8:393–406

[181] Feng J, Huang H, Yen TJ (2006) CENP-F is a novel microtubule-binding protein that is essential for kinetochore attachments and affects the duration of the mitotic checkpoint delay. Chromosoma 115:320–329

[182] Jeganathan KB, van Deursen JM (2006) Differential mitotic checkpoint protein requirements in somatic and germ cells. Biochem Soc Trans 34:583–586

[183] Zou H, McGarry TJ, Bernal T, Kirschner MW (1999) Identification of a vertebrate sister-chromatid separation inhibitor involved in transformation and tumorigenesis. Science 285:418–422

[184] Honda K, Mihara H, Kato Y, Yamaguchi A, Tanaka H, Yasuda H, Furukawa K, Urano T (2000) Degradation of human Aurora2 protein kinase by the anaphase-promoting complex-ubiquitin-proteasome pathway. Oncogene 19:2812–2819

[185] Yu H, King RW, Peters JM, Kirschner MW (1996) Identification of a novel ubiquitin-conjugating enzyme involved in mitotic cyclin degradation. Curr Biol 6:455–466

[186] Peters JM (2002) The anaphase-promoting complex: proteolysis in mitosis and beyond. Mol Cell 9:931–943

[187] Baker DJ, Dawlaty MM, Galardy P, van Deursen JM (2007) Mitotic regulation of the anaphase-promoting complex. Cell Mol Life Sci 64:589–600

[188] Lehman NL, Tibshirani R, Hsu JY, Natkunam Y, Harris BT, West RB, Masek MA, Montgomery K, van de Rijn M, Jackson PK (2007) Oncogenic regulators and substrates of the anaphase promoting complex/cyclosome are frequently overexpressed in malignant tumors. Am J Pathol 170:1793–1805

[189] Reddy SK, Rape M, Margansky WA, Kirschner MW (2007) Ubiquitination by the anaphase-promoting complex drives spindle checkpoint inactivation. Nature 446:921–925

[190] Caseiro MM (2006) Treatment of chronic hepatitis C in non-responsive patients with pegylated interferon associated with ribavirin and thalidomide: report of six cases of total remission. Rev Inst Med Trop Sao Paulo 48:109–112

[191] Milazzo L, Biasin M, Gatti N, Piacentini L, Niero F, Poma BZ, Galli M, Moroni M, Clerici M, Riva A (2006) Thalidomide in the treatment of chronic hepatitis C unresponsive to alfa-interferon and ribavirin. Am J Gastroenterol 101:399–402

[192] Solgi G, Kariminia A, Abdi K, Darabi M, Ghareghozloo B (2006) Effects of combined therapy with thalidomide and glucantime on leishmaniasis induced by Leishmania major in BALB/c mice. Korean J Parasitol 44:55–61

[193] Guo TL, Chi RP, Karrow NA, Zhang LX, Pruett SB, Germolec DR, White KL (2005) Thalidomide enhances both primary and secondary host resistances to Listeria monocytogenes infection by a neutrophil-related mechanism in female B6C3F1 mice. Toxicol Appl Pharmacol 209:244–254

[194] Wolday D, Akuffo H, Demissie A, Britton S (1999) Role of Leishmania donovani and its lipophosphoglycan in CD4+ T-cell activation-induced human immunodeficiency virus replication. Infect Immun 67:5258–5264

[195] Sehgal AK, Srinivasan P (2006) Retrieval with gene queries. BMC Bioinformatics 7:220

[196] Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36:664

[197] Koike A, Takagi T (2007) Knowledge discovery based on an implicit and explicit conceptual network. JASIST 58:51–65

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

[199] Schuemie MJ, Weeber M, Schijvenaars BJA, van Mulligen EM, van der Eijk CC, Jelier R, Mons B, Kors JA (2004) Distribution of information in biomedical abstracts and full-text publications. Bioinformatics 20:2597–2604

[200] Aronson AR (1996) The effect of textual variation on concept based information retrieval. Proc AMIA Annu Fall Symp pp. 373–377

[201] Aronson A, Shooshan S (2006) Ambiguity in the UMLS Metathesaurus. Tech. rep., National Library of Medicine

[202] Trieschnigg D, Kraaij W, Schuemie M (2006) Concept based document retrieval for genomics literature. In: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)

[203] Kraaij W, Weeber M, Raaijmakers S, Jelier R (2004) MeSH based feedback, concept recognition and stacked classification for curation tasks. In: NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)

[204] Vallet D, Fernandez M, Castells P (2005) An ontology-bases information retrieval model. In: 2nd European Semantic Web Conference, ESWC 2005. Heraklion, Greece

[205] Aronson AR, Rindflesch TC, Browne AC (1994) Exploiting a large thesaurus for information retrieval. In: Proceedings of RIAO, pp. 197–216

[206] Torvik VI, Smalheiser NR (2007) A quantitative model for linking two disparate sets of articles in MEDLINE. Bioinformatics 23:1658–1665

[207] (2007) The database revolution. Nature 445:229–230

[208] Gerstein M, Seringhaus M, Fields S (2007) Structured digital abstract makes text mining easy. Nature 447:142

[209] Giles J (2007) Key biology databases go wiki. Nature 445:691

Curriculum Vitae

Rob Jelier was born in Dirksland, the Netherlands, on May 27, 1978. In 1996 he moved to Wageningen to study Bioprocess Engineering. During his study he participated in several research projects, including a study to identify proteins interacting with the human meiotic protein SPO11, and a study of the transition to rest state of the algae Haematococcus sp. with DNA microarray technology. In 2002 he obtained his M.Sc. degree with the thesis "Metabolic Engineering in the Post Genomic Era" which included a research proposal for optimizing folate production in lactic acid bacteria and an experimental study on the phenomenon protein burden in recombinant bacteria. In December 2002 he started as a PhD student in the Biosemantics group at the Erasmus MC, Rotterdam, on a project to develop and evaluate text-mining tools for the analysis of DNA microarray data, of which the results are reported in this thesis. In July 2007 he joined the Human Genetics department at Leiden University Medical Center as a post doctoral researcher.

Publications

1. **Jelier R.**, 't Hoen P.A.C., Sterrenburg E., Den Dunnen J.T., Van Ommen G.B., Kors J.A., Mons B. Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. Submitted.

2. **Jelier R.**, Schuemie M.J., Veldhoven, A., Jenster G., Dorssers L.C., Kors J.A., Anni 2.0: A multipurpose text-mining tool for the life sciences. Submitted.

3. **Jelier R.**, Schuemie M.J., Roes P., Van Mulligen E.M., Kors J.A. 2007. Literature-based concept profiles for gene annotation: the issue of weighting. *In press.* IJMI.

4. **Jelier R.**, Jenster G., Dorssers L.C., Wouters B.J., Hendriksen P.J., Mons B., Delwel R., Kors J.A. 2007. Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. BMC Bioinformatics 8, 14. 5.

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

6. **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, 2049-2058.

7. 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, 2597-2604.