Universiteit
Leiden
The Netherlands

In silico discoveries for biomedical sciences

Haagen, H. van

Citation
Haagen, H. van. (2011, September 21). *In silico discoveries for biomedical sciences*. Retrieved from https://hdl.handle.net/1887/17847

Version:	Corrected Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/17847

Note: To cite this publication please use the final published version (if applicable).

In silico discoveries for biomedical sciences

Herman van Haagen

Promotiecommissie

Promotor	prof. dr. J.T. den Dunnen	LUMC
Co-promotors	dr. B. Mons	NBIC
	dr. P.A.C. 't Hoen	LUMC
	dr. M.J. Schuemie	EMC Rotterdam
Overige leden	prof. dr. ir. M.J.T. Reinders	TU Delft
	prof. dr. J.N. Kok	LIACS Leiden
	prof. dr. F. A. H. van Harmelen	VU Amsterdam

Herman van Haagen
In silico discoveries for biomedical sciences

ISBN/EAN: 978-90-817676-0-6

©H. van Haagen, 2011

Cover design: Freek van Haagen

In silico discoveries for biomedical sciences

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit van Leiden

op gezag van Rector Magnificus prof. mr. P. F. van der Heijden

volgens besluit van het College van Promoties

te verdedigen op woensdag 21 september 2011

klokke 11.15 uur

door

Herman van Haagen

geboren te Breda in 1979

Table of content

Chapter	Title	Page
1	Introduction	3
2	In silico knowledge and content tracking	19
3	Novel protein-protein interactions inferred from literature context	32
4	*In silico* discovery and experimental validation of new protein-protein interactions	65
5	Finding gene-disease relations using implicit information in the scientific literature	98
6	General discussion	115
	Summary	127
	Samenvatting	130
	Curriculum vitae	134

Chapter 1

Introduction

Introduction

When a researcher starts a new research project, he performs a literature study. Let's say he starts with a new project in muscle diseases and he needs to collect information about Duchenne Muscular Dystrophy (DMD). He collects all papers about this topic that are relevant for him. A starting point would simply be to go to the local public library and ask if they have some textbook about DMD. He will also go to Google and enter the topic name or some keywords in the search box; thousands of WebPages pop-up. He is hoping that the first pages contain weblinks that are most relevant for him. Another solution would be to go to more field-specific databases like PubMed (www.pubmed.org). PubMed is the collection of scientific literature for life sciences. This is the place to be for biologists and bioinformaticians.

It is ironic today that the primary problem encountered in literature research is not finding information, but finding too much information. For instance, typing the search query Duchenne muscular dystrophy in PubMed results in more than 6000 hits. Reading 6000 articles is not an option. This problem occurs with other search queries as well. When you need information, in the form of text, you get it but it is simply too much information for any human being to process.

Information overload

High throughput experimental techniques, such as microarrays or next generation sequencing, and bioinformatics tools (e.g. sequence alignment techniques) have increased the pace at which biologists produce new information. This promotes the growth of scientific literature, which contains information on those experimental results in the form of published articles. PubMed, contains more than 20 millions articles published over the last 30 years and the number of published articles is growing at such a rate that scientists are not able to keep up even with the most current knowledge [9] (i.e., new articles added to PubMed every day). This growth is shown in figure in Figure 1. Lastly, more text information can also be found in blogs, Wikipedia or any website specific to the field of biology. This information explosion creates the need for automated approaches to processing biomedically meaningful information from large collections of text.

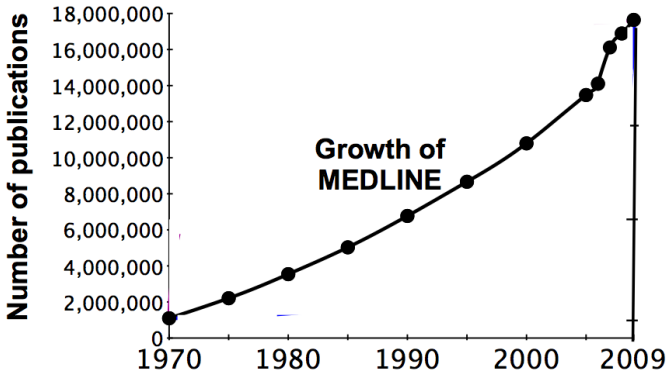


Figure 1. The growth of scientific literature over the last 40 years.

Text-mining

Text-mining is a specific sub-field of data-mining. It is the process of extracting meaningful information from human written text with a computer, for instance, the statement "Malaria is transmitted by mosquitoes". This introduction gives an overview of how text-mining had been developed the last two decades and how it has become an integral part of life science. First we described state of the art search engines and how they tackle the problem of finding relevant documents. Next we introduce the *concept* as a building block for extracting relevant relationships from text. We then describe how text-mining can be enriched using other non textual data sources. Finally we describe what is discussed in this thesis and coming chapters.

Searching for relevant documents

One of the first applications when handling textual data with a computer is the extraction of relevant documents from a large collection of documents. For instance search engines need to extract the relevant webpages from the internet. This process is commonly known as information retrieval (IR) [8]. Biologists can now do this via well known generic search engines like Google or Yahoo, and also by querying collections specific to biomedical sciences such as PubMed/MEDLINE. The success rate of retrieving relevant documents is dependent on the keywords in text and the search query. Keywords are words in text that are specific to the content and important points of the document and is the basis upon which the document should be found. To avoid the ambiguities of natural languages, keywords may be listed explicitly by the author or curator of the article using standard vocabularies. For instance in PubMed this is done using

Mesh Terms. Figure 2 shows schematically what happens in a very simple information retrieval system.

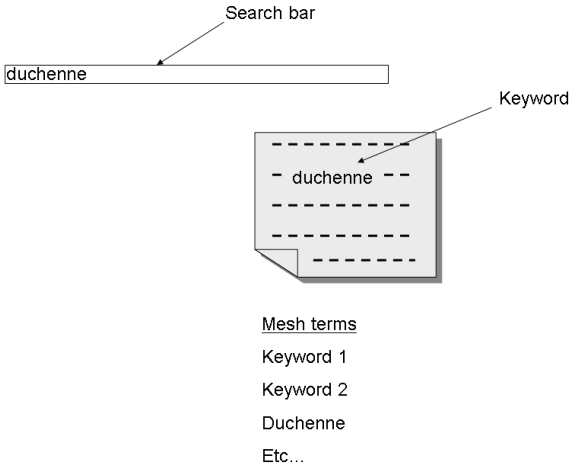


Figure 2. Schematic drawing of a simple text-mining system

In the search bar a query is entered, in this case “duchenne”. The word “duchenne” is scanned in all documents and the documents in which “duchenne” appears are returned. If the document does not contain the word “duchenne”, but the article is about this topic, then the Mesh Terms keyword “duchenne” might be added to the keyword list for this article. This allows the document to be retrieved using a keyword search alone.
The exact structure of the search query is very important for the results that are returned. State of the art machines help the scientist in defining this query. First they can handle typographical errors. When somebody types “duchene” then the system (*e.g.* think of Google) suggests: “Did you mean duchenne?” Second, the search engine can make suggestions on what search query is going to be entered. This is called an *auto complete function*. It works by finding searches done by other visitors that are similar to the search you are making. When “duch” is typed a pull down menu pops up with the words “duchenne”, “duchenne muscular” and many more. This example is shown in figure 3.

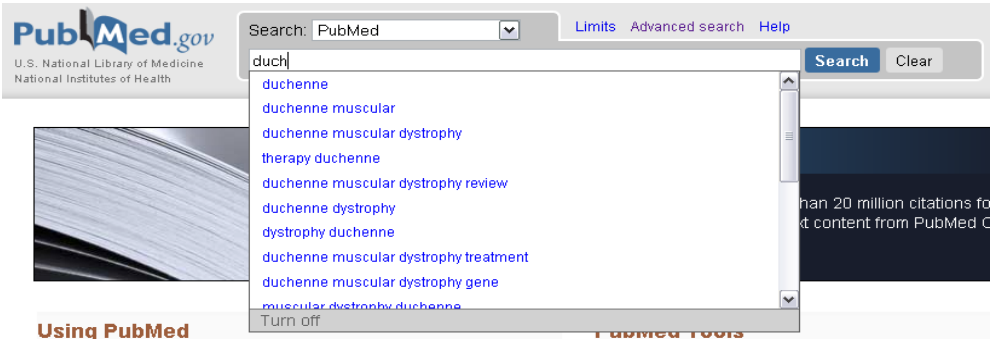


Figure 3. Example for a search query when only the first four letters in a search bar are entered.

Box1: Text-mining jargon

Indexing

Indexing is the process of scanning all documents for relevant keywords and storing the keywords per document in a database.

Concept

A concept is the smallest, unambiguous unit of thought. People reach consensus on the same meaning of the concept. In text-mining a concept is uniquely identifiable.

Thesaurus

A thesaurus is a list containing the concepts and all synonyms. In addition it contains accession numbers that are used in databases like Uniprot and Entrez Gene. Most often used thesauri for the biomedical field is UMLS (Unified Medical Language System)[1] and Biothesaurus[3]. Sometimes a combination of different thesauri is made to make the thesaurus more complete and that is covers more terms[4]. For instance UMLS contains less information for proteins. Therefore UMLS information can be complemented with protein information taken from Entrez Gene and Uniprot. An example for Duchenne Muscular Dystrophy is given in figure 4.

Concept recognition software (CRS)

The concepts in text are recognized with concept recognition software (CRS)[5, 6]. A CRS scans a document for words that are stored in the thesaurus. The software recognizes a word and normalizes it. For instance the word mosquitoes is a plural and it is normalized to mosquito.

Ambiguity

A term is called ambiguous if its meaning is not uniquely defined [7]. For instance the abbreviation PSA in PubMed has approximately 180 meanings. It could for instance mean Puromycin-sensitive aminopeptidase or prostate specific antigen.

Based on the context in which a term appears the CRS needs to disambiguate the term and map it to a concept.

Concept Unique identifier (CUI)

A concept that is uniquely identified in text is assigned a CUI. This is a number that uniquely represents the concept and is used to exchange the concept over different platforms and databases. A CUI normally is specific for the thesaurus that is used. For instance the CUI for Duchenne muscular dystrophy in UMLS is C0013264 (Figure 3)

A specific search query can still result in thousands of retrieved articles that have to be read manually. If a query results in a thousand hits, one might ask whether or not all these documents are equally relevant. Should all documents be read or only a selection? Can the relevance of the documents be prioritized? A first option is then to increase the specificity of the search by adding more search terms to the query.

Is there is a redundancy between articles, in other words, do they share the same information? Redundancy is normally the case, especially in the introduction of the article. A substantial amount of information is repetition of previous articles. Little new knowledge is added per new published article. This is called 'organized plagiarism' (quote by Jan Velterop[10]). Reading the same information, though rhetorically useful, is far too time consuming.

0|NDFRT;DXP;CSP;MTHICD9;COSTAR;MEDLINEPLUS;MSH|47| **Muscular Dystrophy**, Duchenne;duchenne muscular dystrophy;Duchenne muscular dystrophy;Muscular dystrophy, Duchenne;duchenne's muscular dystrophy;Duchenne Muscular Dystrophy;Dystrophy, Duchenne Muscular;Pseudohypertrophic Muscular Dystrophy, Childhood;muscular dystrophy, pseudohypertrophic, childhood;Childhood Muscular Dystrophy, Pseudohypertrophic;Childhood Pseudohypertrophic Muscular Dystrophy;Muscular Dystrophy, Childhood, Pseudohypertrophic;Muscular Dystrophy, Pseudohypertrophic, Childhood;Pseudohypertrophic Childhood Muscular Dystrophy;Progressive Muscular Dystrophy, Duchenne Type;Duchenne-Type Progressive Muscular Dystrophy;Duchenne Type Progressive Muscular Dystrophy;Muscular Dystrophy, Pseudohypertrophic;Dystrophies, Pseudohypertrophic Muscular;Muscular Dystrophies, Pseudohypertrophic;Pseudohypertrophic Muscular Dystrophies;Dystrophy, Pseudohypertrophic Muscular;Pseudohypertrophic Muscular Dystrophy?An X-linked recessive muscle disease caused by an inability to synthesize DYSTROPHIN, which is involved with maintaining the integrity of the sarcolemma. Muscle fibers undergo a process that features degeneration and regeneration. Clinical manifestations include proximal weakness in the first few years of life, pseudohypertrophy, cardiomyopathy (see MYOCARDIAL DISEASES), and an increased incidence of impaired mentation. Becker muscular dystrophy is a closely related condition featuring a later onset of disease (usually adolescence) and a slowly progressive course. (Adams et al., Principles of Neurology, 6th ed, p1415)**|13264**

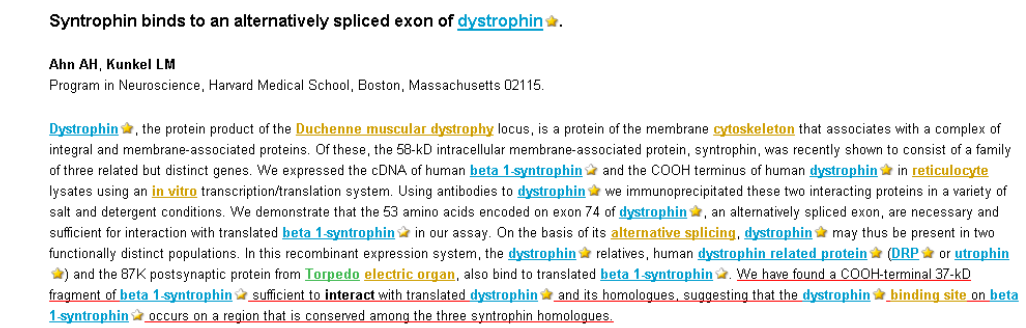
Figure 4. Example of an entry in UMLS for DMD. The field contains descriptions, synonyms and a unique identifier 13264 (last field).

Concepts and relationships

More sophisticated systems are those that are able to extract relevant sentences and phrases from text instead of simply counting words and retrieving whole documents. Automatic information extraction from text is more difficult than indexing keywords. Text is structured in such a way that makes it straightforward for humans to read, but very difficult for computers to interpret automatically. An example of a sentence that can be extracted from text is "malaria is transmitted by mosquitoes". A computer actually needs to understand the meaning of the sentence as we humans do. Processing a sentence like this involves two steps.

1. Recognizing single concepts in text.
2. Mining the relationship into a concept and assertion.

PubMed uses a so called WORD based approach for scanning the literature. Its counterpart is called the *concept based* approach (see Box 1 for definitions). Concepts in text are recognized using concept recognition software (CRS) and a thesaurus. One of the most important tasks of the CRS is to disambiguate a word (see Box1) and map it to its concept unique identifier (CUI). Once a document is tagged and all concepts are recognized, the CUI of the concepts are stored in a database. Figure 5 shows an example of a document in PubMed tagged by IHOP[11]. IHOP is a text-mining tool based on concepts and is an abbreviation for 'Information hyperlinked over proteins'. It tags documents especially for proteins and links them if they appear in the same document. IHOP can be found on http://www.ihop-net.org/.



Syntrophin binds to an alternatively spliced exon of dystrophin.

Ahn AH, Kunkel LM
Program in Neuroscience, Harvard Medical School, Boston, Massachusetts 02115.

Dystrophin, the protein product of the Duchenne muscular dystrophy locus, is a protein of the membrane cytoskeleton that associates with a complex of integral and membrane-associated proteins. Of these, the 58-kD intracellular membrane-associated protein, syntrophin, was recently shown to consist of a family of three related but distinct genes. We expressed the cDNA of human beta 1-syntrophin and the COOH terminus of human dystrophin in reticulocyte lysates using an in vitro transcription/translation system. Using antibodies to dystrophin we immunoprecipitated these two interacting proteins in a variety of salt and detergent conditions. We demonstrate that the 53 amino acids encoded on exon 74 of dystrophin, an alternatively spliced exon, are necessary and sufficient for interaction with translated beta 1-syntrophin in our assay. On the basis of its alternative splicing, dystrophin may thus be present in two functionally distinct populations. In this recombinant expression system, the dystrophin relatives, human dystrophin related protein (DRP or utrophin) and the 87K postsynaptic protein from Torpedo electric organ, also bind to translated beta 1-syntrophin. We have found a COOH-terminal 37-kD fragment of beta 1-syntrophin sufficient to **interact** with translated dystrophin and its homologues, suggesting that the dystrophin binding site on beta 1-syntrophin occurs on a region that is conserved among the three syntrophin homologues.

Figure 5. Screenshot of a PubMed abstract. The words highlighted in color are recognized as concepts by IHOP.

The second step is to extract the relationship from a sentence. The sentence that we use as an example is "malaria is transmitted by mosquitoes". Every complete thought, or relationship is described as a triplet. A triplet starts with a subject (malaria), then the type of relationship which is called the predicate (is transmitted by), and finally the object (mosquitoes). Figure 6 is a schematic of this

triplet. Another group of biomedically relevant triples are the protein-protein interactions (PPIs). For instance the protein Dystrophin (subject) physically interacts (predicate) with the protein Ankyrin-2 (object).

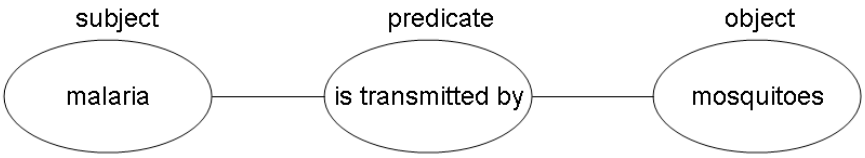


Figure 6. Schematic drawing of a relationship in triplet format.

Extracting relationships can be done in two ways namely:

1. Natural Language Processing (NLP)
2. Co-occurrences in some defined region of text.

NLP is the field within text-mining that studies how a computer analyses a sentence into its building blocks like nouns (*e.g.* the subject and the object) and verbs (*e.g.* the predicate). For instance PIE [12] (http://pie.snu.ac.kr/) is an online webtool based on NLP. It is designed to predict PPIs from PubMed abstracts. A similar approach was used in [13], where they used Bayesian networks for finding novel PPIs.

An alternative method for relationship extraction is that of co-occurrences. PubMed contains more than 20 million abstracts online. With this amount of data it is possible to use a statistical approach to extract relations. The co-occurrence approach is to identify concepts that co-occur within abstracts, sentences or full documents[14, 15], assuming that frequently co-occurring concepts have meaningful association. In PPI, the predicate becomes in all cases "is associated with". The level of association can be calculated using well known statistical tests such as chi square test or Fisher exact test.

The co-occurrence approach has the advantage over NLP that it is less computationally demanding. Only concepts need to be recognized in text without any complex processing. On the other hand NLP has the advantage of extracting the type of relationship (*i.e.* the predicate must be a verb). Co-occurrence based methods only can conclude that two concepts are 'associated'. Second, NLP is able to handle negations like "Protein A does ***not*** interact with protein B". Note that the possibility to handle negations is one of the most difficult to solve in text mining.

Extending text-mining with other data sources: data-mining

The quality of extracted relationships from text can be improved by adding other data sources such as genome sequences, microarray expression data, and annotation databases like the Gene Ontology. This is called data-mining. Data-mining in general contains two steps:

1. Extract information from each database, either non-textual or text.
2. Combine the information from these databases into one statistical measurement.

Relationships established by a computer may become more reliable when several data sources are combined, producing an evidence factor for the relationship. There are systems available as online web applications that work on data integration for the extraction of relationships[16, 17]. One of them is STRING[18] (figure 7), where there is evidence for a relationship between the proteins DMD and SNTB1.

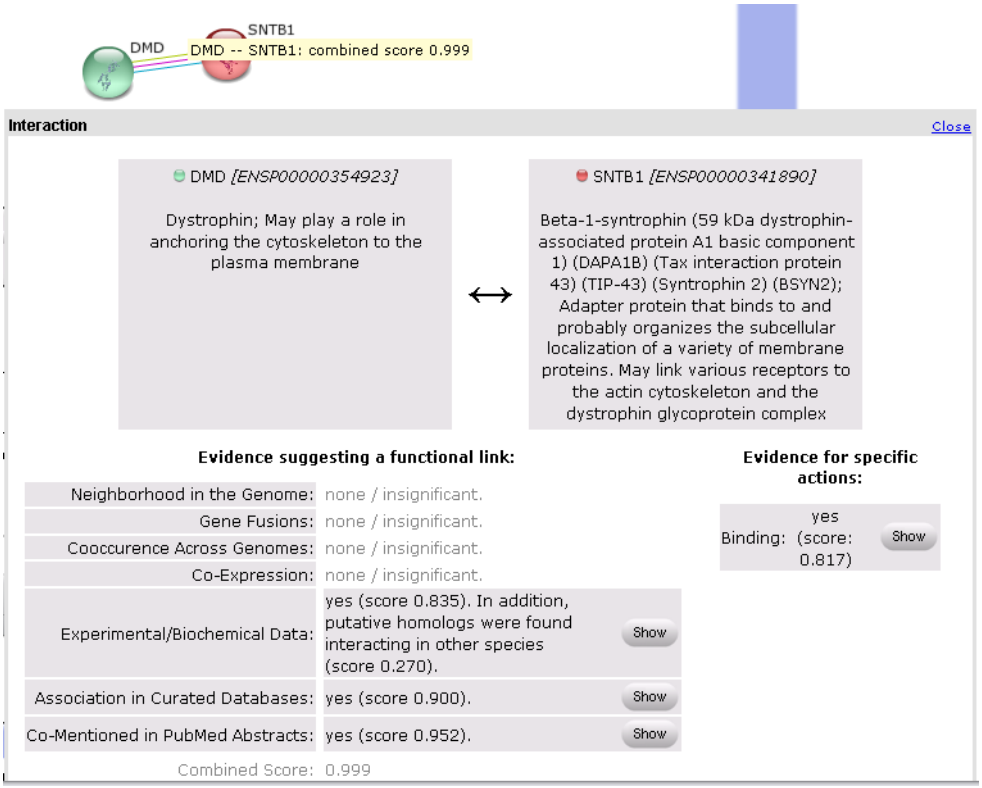


Figure 7. Screenshot of the STRING website. Here the evidence for DMD and STBN1 is mostly found in PubMed abstracts and curated databases.

Another application is when text-mining assists wetlab experiments in annotating the result. For instance in microarray experiments a set of differentially expressed genes is enriched with information from the literature to find gene functions. This is called gene set enrichment or functional enrichment[19-23]. We have used microarrays to complement text-mining for the prediction of PPIs. This is described in chapter 4.

A web of concepts

The next logical step in building triplets is to make a web of interrelated concepts. Currently the world wide web is evolving towards a concept web or semantic web[24-26] (also called web 3.0). Instead of retrieving documents or WebPages the concept web is a web of related concepts where the relationship is extracted from text and databases. The current web is a network of document whereas the concept web is a network of data (of linked data). One of the first applications in biology would be to generate a web of protein-protein interactions[27, 28]. This we can call the protein interaction space. Figure 8 shows an example of the interaction space surrounding the dystrophin protein generated by STRING.

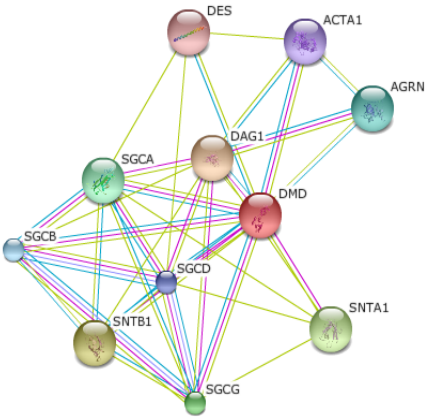


Figure 8. Example of a protein network surrounded around the dystrophin protein (DMD)

Beyond relationship extraction: Inferred relationships

There has been much progress in text-mining in the last two decades (reviewed in [2, 9, 29-32]). Nevertheless, text-mining can go beyond the relationship extraction and building networks. Google, PubMed and even advanced tools such as STRING are state of the art technologies for data analysis. However most of these technologies focus on information that is already known. A text-mining system is

able to find a relationship in less time or is able to find relationships that are overlooked. However, in theory and with great effort, a human being would have been able to extract the relationship by manual searching.
The goal is that a computer is able to find new relationships that no human being could ever find by hand. It might at first seem impossible for a computer to make discoveries on the basis of literature alone; after all, Information Extraction is only able to extract the facts that have already known (i.e., have been published). The principle of inferred relationship extraction is to use facts that have been extracted from several different publications and link them with each other (concept A affects concept B and B affects concept C). One of the first text-mining pioneers, Don Swanson, hypothesized that words in text can be linked with each other via intermediate words and the links would be something meaningful [33, 34]. Swanson found an example in the medical field where he inferred links between Reynaud's disease and fish oil based on the mutual association with concepts such as blood viscosity, platelet aggregation, and vascular reactivity. Later research confirmed that this disease can be treated with fish oil. Before the discovery the disease and the 'drug' had never been co-mentioned before in any article.

This finding was an inspiration and fundamental result for future work based on the same idea of the A-B-C triplet [35-38]. In most cases it concerns single examples of a biological discovery that was inferred using implicit links. Figure 8 shows a schematic drawing how inferred relationship extraction works. A large scale analysis that proves that this text-mining approach will work for many novel relationships has not been done yet. How much implicit information is there in text? and how effectively can it be inferred are burning questions.

The search space for all possible combinations of related concepts is typically huge. The human genome contains approximately 30,000 genes (and therefore more than 30,000 gene products, e.g. proteins)[39]. The search space for finding protein interaction pairs becomes >900 million possible combinations. Text-mining will not only be useful for knowledge discovery, but also assist a scientist in narrowing this search space to only the most informative protein pairs first.

Following the idea of Swanson, in this thesis we describe a text-mining technique called concept profiles[40]. The concept profile technique is based on the indirect links in text to link concepts with each other while they do not necessarily need to be co-mentioned together (Figure 9).

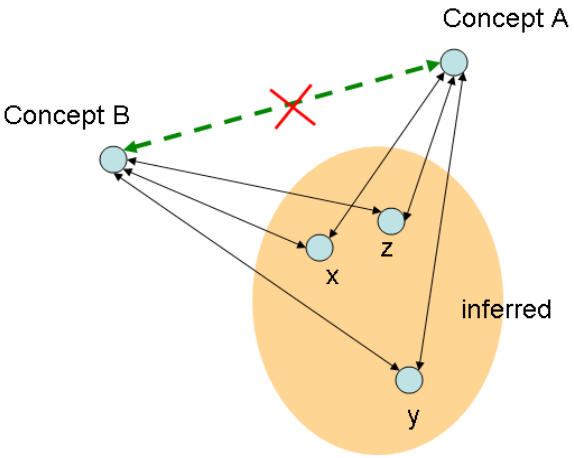


Figure 9. Principle of inferred relationship extraction. Concept A and B are never co-mentioned together in a document. Therefore they do not have a direct relationship. However, via the intermediate concepts X, Y, and Z an indirect relationship can be inferred.

We believe that the full discovery potential of text-mining tools will only be realized with the advent of data-mining approaches that integrate the literature with other large data sets such as genome sequences, microarray expression data, and annotation databases like the Gene Ontology.

However, these resources are generally not entirely independent from the published literature. For instance, the gene ontology (GO) consortium assigns functional annotations to genes that are usually based on evidence described in literature. Another example is microarray experiments where results are summarized in articles, as well as deposited in a database. Given the partial redundancy of literature and other data sources, the question arises as to what exactly is the added value is of other data sources other than text for the extraction of new relationships.

Lastly, we are interested in the predictive power of knowledge discovery algorithms for different kind of relationships. Is this different for protein-protein interactions than for gene-disease relationships?

Content of this thesis

This thesis is structured as follows. In chapter 2 we first describe the basic 'ingredients' that make up a text-mining system. The approach we use is concept based text-mining. Second we describe how to analyze text-mining systems using

ROC curves, retrospective studies and how to collect test data. Chapter 3 shows how implicit information extraction from PubMed abstracts works for protein-protein interactions in a large-scale dataset analysis. We compare the implicit information extraction method with the classical direct co-occurrence method. Also the WORD based method (used by Google and PubMed) is compared with the concept based method.
We extend the text-mining part in chapter 4 with other data sources, such as microarrays and Gene Ontology, and evaluate what is the added value of additional data sources. This chapter is therefore about data-mining. We also evaluate different methods to combine data sources and show the pros and cons of each one. We benchmark our system against the application STRING.
In chapter 5 we investigate another type of relationship namely gene mutation in relation to disease. Here, the implicit information extraction is described in detail and we show what the B part is in the A-B-C relationship.
Chapter 6 is the discussion where all findings are outlined and discussed in detail. We will discuss the power of text- and data-mining but also the limitations. We give future recommendations where data-mining, and in particular text-mining, can be improved.

1. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.
2. Hirschman, L., Park, J.C., Tsujii, J., Wong, L., and Wu, C.H., *Accomplishments and challenges in literature data mining for biology.* Bioinformatics, 2002. **18**(12): p. 1553-61.
3. Liu, H., Hu, Z.Z., Zhang, J., and Wu, C., *BioThesaurus: a web-based thesaurus of protein and gene names.* Bioinformatics, 2006. **22**(1): p. 103-5.
4. Kors, J.A., Schuemie, M.J., Schijvenaars, B.J.A., Weeber, M., and Mons, B., *Combination of genetic databases for improving identification of genes and proteins in text.* BioLINK, 2005
5. Schuemie, M.J., Jelier, R., and Kors, J.A. *Peregrine: Lightweight gene name normalization by dictionary lookup*. in *Biocrative 2 workshop*. 2007. Madrid.
6. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., et al., *Overview of BioCreative II gene normalization.* Genome Biol, 2008. **9 Suppl 2**: p. S3.

7. Mons, B., *Which gene did you mean?* BMC Bioinformatics, 2005. **6**: p. 142.
8. Manning, C., Raghavan, P., and Schütze, H., *Introduction to Information Retrieval*. 2008: Cambridge University Press.
9. Rebholz-Schuhmann, D., Kirsch, H., and Couto, F., *Facts from text--is text mining ready to deliver?* PLoS Biol, 2005. **3**(2): p. e65.
10. Velterop, J., *Open Access: Science Publishing as Science Publishing Should Be.* Serials Review, 2004. **30**(4): p. 308-309.
11. Hoffmann, R. and Valencia, A., *A Gene Network for Navigating the Literature.* Nature Genetics, 2004. **36**: p. 664.
12. Kim, S., Shin, S.Y., Lee, I.H., Kim, S.J., Sriram, R., et al., *PIE: an online prediction system for protein-protein interactions from text.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W411-5.
13. Chowdhary, R., Zhang, J., and Liu, J.S., *Bayesian inference of protein-protein interactions from biological literature.* Bioinformatics, 2009. **25**(12): p. 1536-42.
14. J. DING, D. BERLEANT, D. NETTLETON, and WURTELE, E. *MINING MEDLINE: ABSTRACTS, SENTENCES, OR PHRASES?* . in *Pacific Symposium on Biocomputing*. 2003.
15. Lin, J., *Is searching full text more effective than searching abstracts?* BMC Bioinformatics, 2009. **10**: p. 46.
16. Alexeyenko, A. and Sonnhammer, E.L., *Global networks of functional coupling in eukaryotes from comprehensive data integration.* Genome Res, 2009. **19**(6): p. 1107-16.
17. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
18. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., et al., *STRING 8--a global view on proteins and their functional interactions in 630 organisms.* Nucleic Acids Res, 2009. **37**(Database issue): p. D412-6.
19. Jelier, R., t Hoen, P.A., Sterrenburg, E., den Dunnen, J.T., van Ommen, G.J., et al., *Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease.* BMC Bioinformatics, 2008. **9**: p. 291.
20. Minguez, P., Al-Shahrour, F., Montaner, D., and Dopazo, J., *Functional profiling of microarray experiments using text-mining derived bioentities.* Bioinformatics, 2007. **23**(22): p. 3098-9.
21. Jelier, R., Jenster, G., Dorssers, L.C., Wouters, B.J., Hendriksen, P.J., et al., *Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation.* BMC Bioinformatics, 2007. **8**: p. 14.

22. Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., et al., *Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line.* BMC Bioinformatics, 2006. **7**: p. 373.
23. Kuffner, R., Fundel, K., and Zimmer, R., *Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts.* Bioinformatics, 2005. **21 Suppl 2**: p. ii259-67.
24. Robu, I., Robu, V., and Thirion, B., *An introduction to the Semantic Web for health sciences librarians.* J Med Libr Assoc, 2006. **94**(2): p. 198-205.
25. Burger, A., Romano, P., Paschke, A., and Splendiani, A., *Semantic Web Applications and Tools for Life Sciences, 2008--preface.* BMC Bioinformatics, 2009. **10 Suppl 10**: p. S1.
26. Berners-Lee, T., Hendler, J., and Lassila, O., *The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* 2001.
27. Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T.P., et al., *Integrated network analysis platform for protein-protein interactions.* Nat Methods, 2009. **6**(1): p. 75-7.
28. Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E., *A literature network of human genes for high-throughput analysis of gene expression.* Nat Genet, 2001. **28**(1): p. 21-8.
29. Krallinger, M. and Valencia, A., *Text-mining and information-retrieval services for molecular biology.* Genome Biol, 2005. **6**(7): p. 224.
30. Cohen, K.B. and Hunter, L., *Getting started in text mining.* PLoS Comput Biol, 2008. **4**(1): p. e20.
31. Rzhetsky, A., Seringhaus, M., and Gerstein, M.B., *Getting started in text mining: part two.* PLoS Comput Biol, 2009. **5**(7): p. e1000411.
32. Rodriguez-Esteban, R., *Biomedical text mining and its applications.* PLoS Comput Biol, 2009. **5**(12): p. e1000597.
33. Swanson, D.R., *Fish oil, Raynaud's syndrome, and undiscovered public knowledge.* Perspect Biol Med, 1986. **30**(1): p. 7-18.
34. Swanson, D.R., *Medical literature as a potential source of new knowledge.* Bull Med Libr Assoc, 1990. **78**(1): p. 29-37.
35. Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., and Garner, H.R., *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-98.
36. Srinivasan, P. and Libbus, B., *Mining MEDLINE for implicit links between dietary substances and diseases.* Bioinformatics, 2004. **20 Suppl 1**: p. i290-6.
37. Swanson, D.R., *Migraine and magnesium: eleven neglected connections.* Perspect Biol Med, 1988. **31**(4): p. 526-57.

38. Swanson, D.R., *Somatomedin C and arginine: implicit connections between mutually isolated literatures.* Perspect Biol Med, 1990. **33**(2): p. 157-86.
39. Claverie, J.M., *Gene number. What if there are only 30,000 human genes?* Science, 2001. **291**(5507): p. 1255-7.
40. Jelier, R., Schuemie, M.J., Roes, P.J., van Mulligen, E.M., and Kors, J.A., *Literature-based concept profiles for gene annotation: the issue of weighting.* Int J Med Inform, 2008. **77**(5): p. 354-62.

Chapter 2

In silico knowledge and content tracking

H.H.H.B.M. van Haagen, B. Mons

Chapter 9, Methods in Molecular Biology: In Silico Tools for Gene Discovery, Springer 2010

Abstract
We give a brief overview of a text-mining pipeline and the techniques allow explicit and implicit knowledge to be extracted from large text collections. First, a given ontology is used to tag terms in text as machine-readable concepts. Second, concepts are associated with each other using 2x2 contingency tables and test statistics. Third, from the contingency tables informative pair-wise links between concepts can be recovered. These links may be explicitly stated or implied through indirect associations. Fourth, validation techniques such as ROC curves and retrospective studies can be used to quantify the performance of the information extraction and knowledge discovery process. Lastly, we discuss methods combining text information with various non-textual data sources such as microarray expression data.
We conclude with a brief look at future directions for text-mining and knowledge discovery on the internet at large.

Keywords: text-mining, data-mining, information retrieval, disambiguation, retrospective analysis, ROC curve, prioritizer, ontology, semantic web

1 Introduction

The amount of biomedical literature is growing tremendously. It has become impossible for researchers to read all publications in their moving field of interest, which forces them to make a stringent selection of relevant articles to read. For the actual knowledge discovery process, which is in essence a systematic association process over an expanding number of interrelated concepts, life scientists increasingly rely on the computer. This stringent reduction of the percentage of relevant articles that can actually be 'read' has the disadvantage that relevant information from non-selected articles can be missed. The largest database of recorded biomedical literature is PubMed, which contains over 14 million articles published in the last 30 years (from 1980 till 2010). Besides the literature there are many other resources ranging from curated databases to online blogs, digital books recorded in libraries, and any text information that can be found via a search engine like Google.

The field that deals with automated information extraction from text is called text-mining. Text-mining on its own is a challenging field of research that intensively has been further developed over the last years. Computer systems have been developed based on natural language processing; a method of processing any sentence into its building blocks such as the subject, verbs, and nouns. Other methods are based on word tagging. PubMed for instance uses the words in a

search query and matches it with words found in abstracts with no additional information how the words are related with other words in text. In this chapter we describe the concept based method for automated information extraction from text.

2 Concept based text-mining

For concept based text-mining three 'ingredients 'are needed: (1) text data (2) a word tagger, and (3) a terminology system, mostly controlled vocabulary, or ontologies.
For biomedical text data, normally the abstracts recorded in PubMed are chosen. Reasons for this are that this is the greatest source of recorded literature, the abstracts are publically available and free to download, and the information density of abstracts is higher than that of full text documents (*1*).
Words in text are recognized by a so called word tagger and mapped to a concept identifier (*2*). In order to do so we first need to understand what a concept is. A concept is a unit of thought meaning that people agree that they share information about one and the same thing. A concept has terms and other 'tokens' that 'refer' to it. It can have synonyms, abbreviations, but also for instance Uniform Resource Identifiers (URI's) or accession numbers.
For instance, there exists a protein called dystrophin. When the gene encoding for this protein is mutated it can cause diseases such a Duchenne muscular dystrophy or Becker muscular dystrophy. Dystrophin normally is abbreviated to DMD. DMD (either in italic) also refers to the gene or the disease. Dystrophin is stored in databases like Entrez Gene (http://www.ncbi.nlm.nih.gov/gene) with the accession number 1756 and Uniprot Knowledge Database (http://www.uniprot.org/) with accession number P11532. The words dystrophin, DMD, 1756, and P11532 all refer to one and the same concept (we treat a gene and a protein as the same concept). The tagger maps the words to the concept identifier for dystrophin.
Lastly the synonyms, abbreviations, accession numbers, and concept identifiers are stored in an Ontology. The most common vocabulary for the biomedical field is the unified medical language system (UMLS)(*3*). An Ontology may be field specific. If only drug information from text needs to be extracted a drug-vocabulary is used instead of the whole vocabulary with all medical concepts.

3 Classical direct relationship detection

Once a text-mining system has been developed and concepts in text are recognized and stored in a database the question becomes what to do with this tagged text data? The main question is which two concepts are significantly related. The relationship between two concepts can be of any kind. In biology these are the most common ones we chose as examples: (1) two proteins that have a molecular

interaction, (2) a mutated gene that causes a disease, (3) a protein that has a particular function and (4) a drug that treats a disease or has a (adverse) side effect. Any relationship between two concepts can be seen as a triplet with a subject, predicate and object. An example of a triplet is protein dystrophin (subject) interacts with (predicate) protein ankyrin 2 (object).
The statistical way to define the strength of relationship between two concepts is by making a 2x2 contingency table (or frequency table). The table below gives an example for concepts X and Y.

	X	Not X
Y	A	B
Not Y	C	D

A are the number of documents where both concept X and Y are co-mentioned. *B* are the number of documents where concept Y occurs but not concept X. *C* is the reverse version of *B* and *D* are the number of documents where X and Y are not mentioned. Any statistical test can be applied to this table such as the likelihood ratio test, chi-squared test or the uncertainty coefficient. If X and Y are frequently co-mentioned together (*e.g. A* is a relatively large number) and the concepts are not exceptionally generic so that they occur frequently in text (*e.g B* and *C* are small) then the two concepts may be significantly related. There are many text-mining systems available based on direct relationship detection such as IHOP**(*4*)**, PubGene**(*5*)**, and systems where text-mining is an integral part such as STRING**(*6*)**, FunCoup**(*7*)**, and Endeavour**(*8*)**.

4 Implicit information extraction via concept profiling

The classical direct relationship detection method has the disadvantage that concepts that are not co-mentioned together are missed, while they still might be related to each other. This could be due to the reason that related concepts are stored in full text (frequently not freely available for mining) and not in the abstract or that concepts are related but no-one made the link yet. Via indirect links between terms in text, terms can still be related to each other even when they have never been co-mentioned - **(*9*)**. This we call implicit information extraction. Swanson et.al. **(*10*)** were the first to demonstrate that this approach works by linking the treatment of Raynaud's disease with fish oil. Van Haagen et. al. **(*11*)** demonstrated this idea further by predicting protein-protein interactions. They predicted the physical interaction between calpain 3, which causes a form of muscular dystrophy, and parvalbumin B, which is found mainly in skeletal muscle. Those two proteins were strongly linked via the intermediate concept dysferlin, which is a protein.

Concept profiling contains the following steps (see Fig. 1). First for a concept X (*e.g* a gene, a chemical or drug) the documents are selected wherein X appears. Next all other concepts that are co-mentioned with X are processed using the direct relationship detection method described previously (Fig. 1b). The 2x2 table information for each concept pair is stored in a profile. This concept profile for X is basically a vector of N dimensions. N are the number of concepts that are co-mentioned with X. Each entry in the vector is a number associating concept X with another concept (taken from a 2x2 table, Fig. 1a). Computation of the 'conceptual association' between two concepts can now be performed by matching their respective concept profiles by vector matching (Fig. 1c). Any distance measure can be used for this matching**(*9*)** such as the inner product, cosine, angle, Euclidean distance or Pearson's correlation. If two concept profiles have many concepts in their individual concept profiles in common, *e.g.* many implicit links, then the two concepts may be related to each other. A webtool is available, dubbed 'Anni', for implicit information extraction by concept profiling**(*12*)**. In the next section we will describe how to validate text-mining approaches and the amount of relatedness.

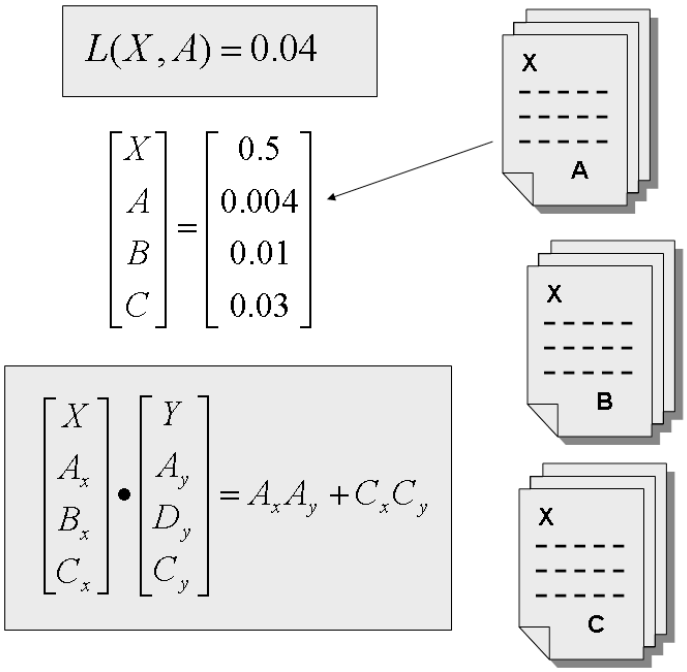


Figure 1. Basic scheme for concept based profiling. (a) Example of a likelihood function calculated between concept X and A. Information is taken from a 2x2 contingency table. The score reflects the strength of association between X and A. (b) Documents selected where concept X appears and is co-mentioned with other concepts. For a concept the documents are selected and transformed into a test statistic using a 2x2 contingency table. (c) The inner product score between two

concept profiles. The score is only calculated over the concepts the two profiles have in common.

5 Cross validation within text-mining and other performance measures

5.1 Defining a positive and a negative set.

In the previous sections we described how to extract relationships (content) between concepts from text either with direct relationship detection or concept profiling. Once a system is designed it needs to be tested to evaluate its performance in extracting or predicting relationships. To enable this step we need data to train the system and after training testing it. For instance data on protein function can be collected from the Gene Ontology **(*13*)** and data on gene-disease relationships from OMIM (http://www.ncbi.nlm.nih.gov/omim). Here we describe an example of the relationship type protein-protein interactions (PPIs). PPIs can be collected from online databases such as UniProt**(*14*)**, DIP**(*15*)**, BioGrid**(*16*)** and Reactome**(*17*)**. These samples of curated protein-protein interactions are labeled as positives instances. These positives instances are compared with negative instances to see if the text-mining system can discriminate between the two groups. In biology research no databases exists that stores samples of negatives instances, *e.g.* two proteins that have been confirmed not to interact. Normally generating negative instances is done by selecting random pairs from a group of proteins**(*18*)**.

5.2 Receiver operating characteristics curves

Receiver operating characteristics (ROC) curves are often used to evaluate the performance of a prediction algorithm **(*19*)**. A ROC curve is a graphical plot of the true positive rate (sensitivity) on the y-axis versus the false positive rate (1 − specificity) on the x-axis (see Fig. 2b). The ROC curve is defined for a binary classifier system (the positive and negative set described in section 5.1) as its discrimination threshold is varied. This measure is often used in information retrieval and it can be explained as a system design that collects as much information as possible (in terms of true positives) while at the same time reducing the noise (the false positives). A ROC curve is constructed as follows; in Fig. 2a the distributions of positive and negative instances are given and in Fig. 2b its corresponding ROC curve. The threshold that discriminates between the two groups is varied from the highest match score (x-axis Fig. 2a) value to the lowest. Each threshold corresponds to a true positive and false positive rate in ROC space. In Fig. 2a all the way up to the right on the x-axis is the threshold (around 7) where no true or negative instances pass this threshold. Therefore the true positive and false positive rates are both zero, resulting in the point (0,0) in ROC space (Fig. 2b bottom left corner). Then the threshold as a slider is moved to the right. At each

point a number of positive and negative instances will pass the threshold resulting in a point in ROC space anywhere between 0 and 1 on both axes. Finally the threshold reaches the extreme left point on the x-axis (around -2, Fig. 2a). Here all positive and negative instances pass this threshold. This corresponds with the point (1,1) in ROC space (top right corner Fig. 2b).
To translate the ROC space to a single measurement for performance we calculate the Area under the ROC curve (AuC). The AuC value normally varies between 0.5 and 1.
If a system shows a random behavior (*e.g.* two completely overlapping distributions, no discrimination between positive and negative set) the ROC space results in a straight line from the point (0,0) to (1,1). This corresponds with an AuC of 0.5. If a system behaves like a perfect classifier the ROC curve starts at point (0,0) and moves up to point (0,1) (*e.g.* first all positive instances are predicted) then it moves from point (0,1) to point (1,1) (*e.g.* all negative instances are predicted). This corresponds with an AuC of 1. The AuC for the example in Fig. 2 is 0.92.

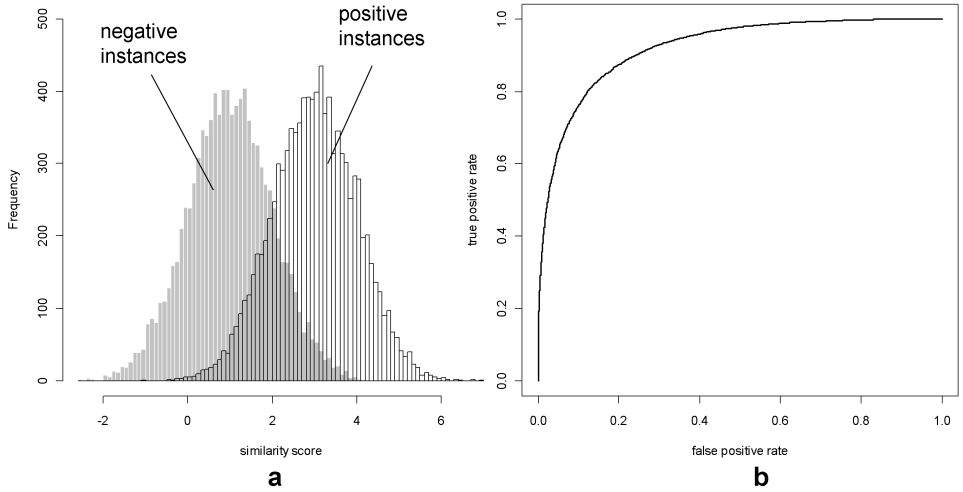


Figure 2. Histogram and its corresponding ROC plot. (a) the distribution of the positive and negative set. (b) a ROC curve with an AuC of 0.92.

5.3 Cross validation and bias

The performance of an associative *in silico* discovery system is tested using cross-validation(***20***). A system is first trained using training data. Then it is tested using test data. There is no explicit data for testing only, nor is their data used only for training. There is just data. Therefore a part of the data is selected for training and the remaining part for testing. The way to select the training and test data is arbitrary. Here we describe the most common approach of cross validation the 10-

fold CV. The first step (1) is to randomly shuffle the samples in your dataset (both positive and negative instances). (2) Then the dataset is divided into 10 equally sized subsets. Each piece contains samples of the positive and negative set. (3) In one iteration, 9 of the 10 subsets is used for training and the remaining subset is used for testing. (4) Step three is repeated until each subset is used once for testing.
An extremely important step during cross validation is to make sure that none of the test data is used during training. Else this would introduce a bias and gives an overestimation of the true performance. Within the field of text-mining and biology this seems virtually impossible. Most of the data stored in curated databases, such as protein-protein interactions or gene-disease relationships recorded in OMIM are based on published articles. This means that positives instances in the test set are based on articles that are also used to train a text-mining system. Other data sources also have this problem. For instance the Gene Ontology contains functional descriptions for a protein that are normally also based on literature evidence. In order to evaluate prediction performance it is therefore more appropriate to make use of a retrospective analysis

5.4 Retrospective validation

Before we explain the basics of retrospective validation, we need to distinguish between two types of prediction. The first one is prediction of current knowledge stored in databases. This knowledge is already known and the system recovers what is stored in these databases. For this, the cross-validation approach described above is useful.
The second one is the prediction of new and as yet unforeseen knowledge. This means 'implicit' knowledge that is not recorded in any database that cannot be explicitly found in text. To simulate the prediction of these 'hidden associations' a retrospective validation is done. First a time interval is defined when data is stored in a database. For PubMed this could for instance be all the abstracts of articles published between 1980 and January 2010. The second step would be to select test data published after a certain date, for instance all protein-protein interactions recorded in databases from January 2007 until January 2010. The third step is to train the text-mining system before that date using all data before January 2007. The last step is to evaluate what test samples were predicted before January 2007 that became only explicit (also in the literature) knowledge after January 2007. In other words, protein-protein interactions that could be found by simple co-occurrence before the 'closure date', but were not added to the databases yet, should not be counted as true predictions. In this evaluation there is no procedure to repeat these steps multiple times like with cross-validation. This means that no standard error on the performance can be calculated.

5.5 Prioritizers

Another way to view a ROC curve is as a prioritized list. The ROC curve is constructed by varying the threshold. The samples (*e.g.* protein pairs either a PPI or random) are ranked from the highest match score to the lowest. Going down in this ranked list from the top prediction to the lowest is done by walking over the ROC curve from point (0,0) to point (1,1). Experimental biologists are mainly interested in what is predicted in the top, *e.g.* the most likely predictions. Prioritizers are useful to evaluate where your test samples rank in the top. A ROC curve can also be plotted on the absolute scales of true positives and false positives by translating a prioritized list in a graphical way. Figure 3 shows an example of 20 ranked samples and its corresponding ROC curve. This curve is also called a ROC10 curve. It reflects the amount of true positive predictions (baits) at a fixed number of false positives (the costs), in this case 10. You can vary this threshold and define for instance a ROC50 or ROC100 curve.

sample	match score	label
s1	4.960111897	0
s2	4.662243252	0
s3	4.661449007	1
s4	4.589346313	0
s5	4.581602664	0
s6	4.438759168	0
s7	4.379399852	0
s8	4.377182023	0
s9	4.320084484	1
s10	4.303145807	0
s11	4.299505092	1
s12	4.259483679	0
s13	4.249962717	1
s14	4.230262428	1
s15	4.188873972	1
s16	4.179967181	1
s17	4.177931731	1
s18	4.176843766	0
s19	4.163993686	1
s20	4.118021526	1

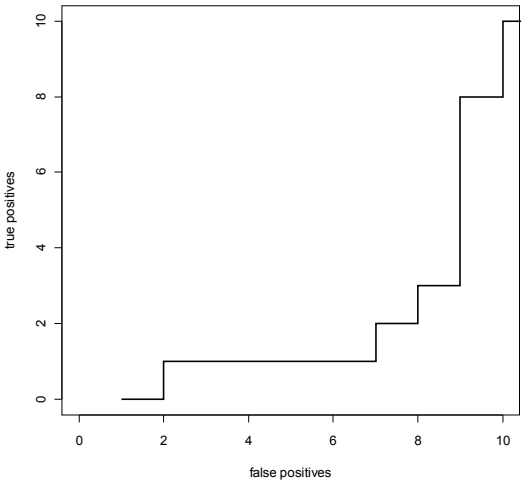


Figure 3. Prioritized list of 20 samples and its corresponding ROC10 curve.

6 Extending text-mining systems with other databases: data-mining

Text-mining actually is a subdivision of the broader field of data-mining. Data-mining is the field of research to extract any kind of information from a variety of resources. For instance, there are many data sources available for proteins. Besides the literature, there exists information in curated databases, microarray expression data(***21, 22***), domain interaction databases(***23***), functional annotations from the Gene Ontology, phylogenetic trees and sequence data. There are many tools and techniques available for data-mining on databases but they all share a common idea. To combine all information from several distinct data sources into one should reveal more information than can be recovered by the mining of each data source alone. Data-mining basically is a two step approach. The first step is to define a match or evidence score for every data source that is included in the system. For instances a microarray dataset may be transformed into a data matrix by calculating Pearson correlations between any two expression profiles for proteins or genes. The second step is to combine each evidence score for a data source into a single score. This can be done, for instance, using a Bayesian classifier. For protein-protein interactions there are several resources available based on data-mining techniques such as STRING(***6***), FunCoup(***7***), IntNetDB(***24***), and Prioritizer(***25***).

7 Beyond data-mining and scalable technology for the internet: the semantic web

Data-mining and text-mining are fields of technology that are used for the future web 3.0 technology: the semantic web (SW). The first trend in web technology (or web 1.0) included the static webpages that made the first version of the internet. No information exchange was possible, just readable plain text pages. The second trend (web 2.0) made it possible for users to interact with the internet. Think of uploading movies to YouTube, or writing your blog online and online shopping with a credit card. Web 2.0 is really the most unstructured and scattered form of information. Therefore, the new trend became web 3.0. It will structure the internet into a network of concepts and relationships between these concepts. Other terms for web 3.0 are the concept web or the semantic web. One of the goals of the web is to present information in a computer readable compact format instead of the current webpages that are retrieved after a search query. The predictions that are made using concept profiles or other technologies will be part of this SW.

The best known data model for the SW is RDF (resource description framework). RDF is used to translate any kind of date into a triple format. The ontologies used in webtechnology are mainly built using OWL (Web Ontology Language). The semantic web project is extremely large and it is very difficult to keep it scalable. There is now an ongoing project called the Large Knowledge Collider (LarKC). It builds the semantic web with all the current state of the art technology that is out

there (machine learning, information theory, pattern recognition, first order logic). All information on LarKC can be found on http://www.larkc.eu.

8 References

1. 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.
2. Schuemie, M. J., Jelier, R., and Kors, J. A. (2007) Peregrine: Lightweight gene name normalization by dictionary lookup, in *Biocrative 2 workshop*, pp 131-140, Madrid.
3. Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res 32*, D267-270.
4. Hoffmann, R., and Valencia, A. (2004) A Gene Network for Navigating the Literature, *Nature Genetics 36*, 664.
5. Jenssen, T. K., Laegreid, A., Komorowski, J., and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression, *Nat Genet 28*, 21-28.
6. Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms, *Nucleic Acids Res 37*, D412-416.
7. Alexeyenko, A., and Sonnhammer, E. L. (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration, *Genome Res 19*, 1107-1116.
8. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006) Gene prioritization through genomic data fusion, *Nat Biotechnol 24*, 537-544.
9. Jelier, R., Schuemie, M. J., Roes, P. J., van Mulligen, E. M., and Kors, J. A. (2008) Literature-based concept profiles for gene annotation: the issue of weighting, *Int J Med Inform 77*, 354-362.
10. Swanson, D. R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspect Biol Med 30*, 7-18.
11. van Haagen, H. H. H. B. M., t Hoen, P. A. C., Botelho Bovo, A., de MorrÃ©e, A., van Mulligen, E. M., Chichester, C., Kors, J. A., den Dunnen, J. T., van Ommen, G.-J. B., van der Maarel, S. r. M., Kern, V. c. M., Mons, B., and Schuemie, M. J. (2009) Novel Protein-Protein Interactions Inferred from Literature Context, *PLoS ONE 4*, e7894.

12. Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G., and Kors, J. A. (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences, *Genome Biol 9*, R96.
13. Gene Ontology, C. (2000) Gene ontology: tool for the unification of biology, pp 25 - 29.
14. (2009) The Universal Protein Resource (UniProt) 2009, *Nucleic Acids Res 37*, D169-174.
15. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update, *Nucleic Acids Res 32*, D449-451.
16. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006) BioGRID: a general repository for interaction datasets, *Nucleic Acids Res 34*, D535-539.
17. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009) Reactome knowledgebase of human biological pathways and processes, *Nucleic Acids Res 37*, D619-622.
18. Ben-Hur, A., and Noble, W. (2006) Choosing negative examples for the prediction of protein-protein interactions, p S2, BMC Bioinformatics.
19. Fawcett, T. (2003) ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, *Hewlett-Packard Company*.
20. Wessels, L. F., Reinders, M. J., Hart, A. A., Veenman, C. J., Dai, H., He, Y. D., and van't Veer, L. J. (2005) A protocol for building and evaluating predictors of disease state based on microarray data, *Bioinformatics 21*, 3755-3762.
21. Obayashi, T., Hayashi, S., Shibaoka, M., Saeki, M., Ohta, H., and Kinoshita, K. (2008) COXPRESdb: a database of coexpressed gene networks in mammals, *Nucleic Acids Res 36*, D77-82.
22. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc Natl Acad Sci U S A 101*, 6062-6067.
23. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Griffith-Jones, S., Haft, D., Hermjakob, H., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Orchard, S., Pagni, M., Peyruc, D., Ponting, C. P., Servant, F., and Sigrist, C. J. (2002) InterPro: an

integrated documentation resource for protein families, domains and functional sites, *Brief Bioinform 3*, 225-235.

24. Xia, K., Dong, D., and Han, J. D. (2006) IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model, *BMC Bioinformatics 7*, 508.
25. Lage, K., Karlberg, E. O., Storling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tumer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders, *Nat Biotechnol 25*, 309-316.

Chapter 3

Novel protein-protein interactions inferred from literature context

H.H.H.B.M. van Haagen[1], P.A.C. 't Hoen[1], A. Botelho Bovo[2], A. de Morree[1], E.M. van Mulligen[1], C. Chichester[1], J.A. Kors[1], J.T. den Dunnen[1], G.J.B. van Ommen[1], S.M. van der Maarel[1], V. Medina Kern[2], B. Mons[1], M.J. Schuemie[1]

1 Biosemantics Association, Department of Human Genetics, Leiden University Medical Center, Leiden, and Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
2 Post-Graduate Program in Knowledge Engineering and Management (EGC), Federal University of Santa Catarina (UFSC), Florianópolis, Brazil

PLoS One. 2009 Nov 18;4(11):e7894.

Abstract
We have developed a method that predicts Protein-Protein Interactions (PPIs) based on the similarity of the context in which proteins appear in literature. This method outperforms previously developed PPI prediction algorithms that rely on the conjunction of two protein names in MEDLINE abstracts. We show significant increases in coverage (76% versus 32%) and sensitivity (66% versus 41% at a specificity of 95%) for the prediction of PPIs currently archived in 6 PPI databases. A retrospective analysis shows that PPIs can efficiently be predicted before they enter PPI databases and before their interaction is explicitly described in the literature. The practical value of the method for discovery of novel PPIs is illustrated by the experimental confirmation of the inferred physical interaction between CAPN3 and PARVB, which was based on frequent co-occurrence of both proteins with concepts like Z-disc, dysferlin, and alpha-actinin. The relationships between proteins predicted by our method are broader than PPIs, and include proteins in the same complex or pathway. Dependent on the type of relationships deemed useful, the precision of our method can be as high as 90%. The full set of predicted interactions is available in a downloadable matrix and through the webtool Nermal, which lists the most likely interaction partners for a given protein. Our framework can be used for prioritizing potential interaction partners, hitherto undiscovered, for follow-up studies and to aid the generation of accurate protein interaction maps.

Introduction
Protein-protein interactions (PPIs), which we define as proteins that physically interact, are crucial in most complex biological processes. Experimental high-throughput methods such as yeast two-hybrid screens have been used to make large inventories of PPIs and to create protein interaction maps[1-6]. However, it is well known that these methods merely show physical interaction under experimental condition and not necessarily indicate a common involvement in a biological process. Computational methods for the prediction of PPIs could theoretically aid the discovery of candidate biological interaction partners. There are many different sources of information that can be used in PPI prediction[7], including protein structures, phylogenetic distribution, interactions between homologous proteins in other organisms, genomic neighborhood, and gene fusions. In this article, we will focus on one source of information, which is arguably the most comprehensive, but also the least structured: biomedical literature itself. Until now text mining techniques are mainly used to rediscover PPIs explicitly described in literature. Often, the now 18 million freely available abstract records of MEDLINE are used for this purpose. PPIs extracted this way have been shown to improve the accuracy of predicted biological networks[8, 9]. Structured information on explicit PPIs

extracted from MEDLINE and other sources is freely available in the STRING database[10], or can be found by querying the iHOP website[11].

However, text mining can go one step further; by combining known associations, previously unknown PPIs can be inferred. Because most text mining research, including this study, limits itself to MEDLINE abstracts, these 'previously unknown' interactions also include interactions that are effectively known, but not explicit in MEDLINE as they are only mentioned in a full text article. Swanson[12, 13] *et al.* were the first to demonstrate that text mining can lead to the discovery of new knowledge (e.g. the treatment of Raynaud's disease by fish oil). Other studies in the biomedical domain verified the importance of implicit information for knowledge discovery[14-16]. Whereas Swanson used a word-based approach, linking entities by intermediate words that appeared frequently in the contexts of both entities, in our work we use a concept-based approach: different terms denoting the same concept (*i.e.* synonyms) are mapped to a single concept identifier, and ambiguous terms, e.g., identical terms used to indicate different concepts (*i.e.* homonyms) are resolved by a disambiguation algorithm. Such an approach is essential given the wide diversity and many ambiguities in gene and protein nomenclature[17, 18].
In order to predict PPIs, we summarize the typical context in which each protein appears into *concept profiles[15, 16, 19]*. We hypothesize that a high similarity between the concept profiles of two proteins is indicative for an actual biological interaction. For example, if two proteins are consistently mentioned together with a particular disease, the probability that these proteins interact is higher than the a priori probability of two randomly selected proteins[20, 21]. This probability should increase further when they are also frequently co-mentioned with a particular pathway, a sub-cellular localization, or other proteins.
In this article, we first demonstrate the added value of a concept-based approach over a traditional term-based approach in detecting explicitly described relations. We proceed to show the added value of the concept profile-based approach over classical direct relation extraction, including the text-mining techniques used in the STRING database. Subsequently, we show the predictive power of our method by doing a retrospective study; we demonstrate that we can employ the literature available in 2005 to predict 52% of the PPIs newly described in Swiss-Prot in 2007 at a specificity level of 95%. We show that in addition, some of the PPIs that we predicted but are not yet recorded in any database represent indirect protein interactions and have biological relevance. Finally, we confirm one of the many predicted PPIs in three wet lab experiments, supporting our claim that the concept profiling method is capable of previously unknown PPI prediction from current literature.

These predictions will be useful for (i) the ranking of potential PPIs for more specific experimental analysis, and (ii) complementing other types of data such as co-expression and yeast two-hybrid data when using an integrative systems biology approach.

Results

Improved PPI detection using concept profiles

We compared the performance of different PPI prediction approaches in detecting known human PPIs in MEDLINE. The online human-curated databases Biogrid, DIP, HPRD, MINT, Reactome, and UniProt/Swiss-Prot were used to establish a set of 61,807 known human PPIs. A set of probable Non-Interacting Protein Pairs (NIPPs) was generated from all pairs of proteins that do not occur in the above databases nor in the IntAct[22] database, which includes, in addition to all PPIs recorded in UniProt/Swiss-Prot, many non-curated PPIs from high-throughput experiments. We compare four approaches:

- *Word-based direct relation*. This approach uses direct PubMed queries (words) to detect if proteins co-occur in the same abstract. This is the simplest approach and represents how biologists might use PubMed to search for information.
- *Concept-based direct relation*. This approach uses concept-recognition software to find PPIs, taking synonyms into account, and resolving homonyms. Here two concepts (in our case two proteins) are detected if they co-occur in the same abstract.
- *STRING[10]*. The STRING database contains a text mining score which is based on direct co-occurrences in literature.
- *Concept profile-based relation*. This approach uses the similarity in literature context. Here two proteins (concepts) can also be indirectly related via the concepts in their profiles. More detail on concept profiles and their construction can be found in the Methods section.

The word-based and concept-based direct relation methods could find at least one abstract containing both proteins for respectively 33% and 32% of the pairs in the PPI set. A text mining score from STRING could be obtained for 30% of the PPIs, in line with the co-occurrence based approach used to create STRING. Thus, a majority of the known PPIs cannot be found explicitly in MEDLINE. For the concept profile-based approach, we could create concept profiles and calculate a similarity score for 76% of the PPI set.

Similar to STRING, the other three approaches can also be used to calculate a continuous score that indicates the strength of the relation between two proteins. Figure S1 displays the distribution of the similarity scores of the concept profile-based method for the PPI and NIPP sets. This figure shows that the scores for the PPI set are higher although there is also overlap between the two distributions. The

continuous scores can be used to rank protein pairs. After ranking the pairs in the PPI and in the NIPP set, we calculated the sensitivity at a specificity of 99% and 95%, and the Area under the Curve (AuC), which is often used in the evaluation of classifiers, and expresses the area under the Receiver Operator Characteristics (ROC) curve (see supplement S5). An AuC of 0.5 indicates a random classifier; an AuC of 1 indicates a perfect classifier. For this analysis, we limited ourselves to those pairs in the PPI and NIPP set for which all methods could make a prediction. We analyzed 44,920 pairs in the PPI set, and 58,388,409 pairs in the NIPP set.
The results show that, using concept profiles, we can detect 43% of the known PPIs, with a specificity of 99%, and 66% of all known PPIs with a specificity of only 95%. In contrast, the direct relations methods and STRING show much lower scores (Table S1).

Table 1. Performance of different PPI prediction approaches on detecting known PPIs in MEDLINE. CDR stands for Concept-based Direct Relation method.

	Word-based	CDR	Concept profiles	STRING
Sensitivity at spec = 99%	28%	37%	43%	39%
Sensitivity at spec = 95%	33%	41%	66%	41%
Area under Curve	0.62	0.69	0.90	0.69

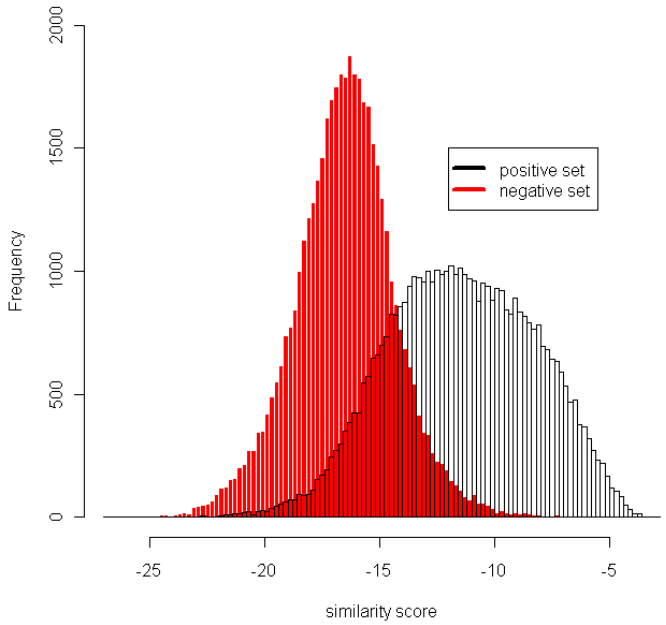


Figure 1. Histogram of the distributions of similarity scores of the concept profile-based method for the PPI and NIPP sets. A log transformation is applied to the similarity scores for better visualization.

Proteins connected via one intermediate protein

The results reported in the previous section indicate that not all proteins with high similarity scores are known to interact according to the combined protein databases. One possible explanation for this is that the proteins are related in another way, *e.g.* they could be involved in the same pathway or be part of the same protein complex, but do not physically interact. To determine whether this occurs, we also tested both concept-based approaches on the detection of known connections via one intermediate protein. For instance, if the protein pairs A-B and B-C are recorded as PPIs in databases, we form the additional protein pair A-C. In total we were able to create 1,028,265 of such pairs to serve as an independent test set. When the pairs are filtered on coverage by all methods the remaining set contains 790,245 pairs. At a specificity level of 99% and 95% the sensitivity level of the different methods was determined for those pairs. The results are given in Table S2 and indicate that the concept profile-based approach is indeed superior in predicting relationships between proteins potentially present in the same complex or pathway.

Table 2. Performance on predicting proteins that are connected via an intermediate protein.

	Concept-based	CDR	STRING
Sensitivity at spec = 99%	8%	9%	8%
Sensitivity at spec = 95%	13%	29%	12%
Area under Curve	0.54	0.78	0.53

Average prediction performance per protein

Most researchers will not be interested in all PPIs, but only in those interactions involving a (set of) protein(s) of interest. Therefore, for each protein we created a top 10, top 100, and top 1,000 best matching proteins according to the concept-based direct relation, the concept profile method, and STRING. In these lists, we calculated the number of PPIs that are either (i) part of the PPI set, or (ii) described in the IntAct database, or else (iii) part of the pairs that are connected through intermediate proteins as described in the previous section. We limited our analyses to the 10,812 proteins that were detected in at least five MEDLINE abstracts (covered by the concept profiles method). The averages of these performance measures in terms of precision and recall are shown in Table S3. For comparison, the average total number of pairs per protein in each set is provided in the third column. For instance, on average each protein is involved in 8.73 interactions according to the PPI set, of which on average 6.34 are found in the top 1,000 of the concept profile method (precision and recall of 0.006 and 0.73 respectively) , and only 3.93 and 3.83 in the top 1,000 of the concept-based direct relation method and STRING respectively. The latter two methods show a slightly better performance for the top 10. Thus, it appears that co-occurrence-based methods can detect a smaller number of PPIs with a somewhat higher accuracy, but the concept profile method, by including indirect evidence, can predict more PPIs and is therefore likely to be more valuable for actual knowledge discovery.

Table 3. Analysis of the top 10, 100, and 1,000 returned by the Concept Profile (CP) method, the Concept-based Direct Relation (CDR) method, and by STRING. The analysis shows the precision and recall of protein pairs that are in the PPI set, of additional pairs

			Top 10		Top 100		Top 1,000	
	Method	Total	Precision	Recall	Precision	Recall	Precision	Recall
PPI	CP	8.73	0.096	0.110	0.033	0.37	0.006	0.73
	CDR	8.73	0.108	0.124	0.026	0.30	0.004	0.45
	STRING	8.73	0.112	0.128	0.026	0.30	0.004	0.44
IntAct	CP	1.61	0.009	0.056	0.002	0.12	0.000	0.29
	CDR	1.61	0.009	0.056	0.002	0.11	0.000	0.24
	STRING	1.61	0.008	0.050	0.002	0.11	0.000	0.24
Indirectly connected	CP	190.21	0.105	0.006	0.080	0.042	0.048	0.25
	CDR	190.21	0.137	0.007	0.068	0.036	0.027	0.14
	STRING	190.21	0.100	0.005	0.062	0.033	0.026	0.14

Retrospective prediction of currently known PPIs

Protein annotation databases are struggling to stay up-to-date with the literature, and there is often a substantial time lag between the first publication of a finding, and the time the PPI is entered in a database. It could therefore be postulated that many of the unknown PPIs predicted today are in fact correct, but may not be entered in a database for several years. We have performed a retrospective study to answer the question: how many of the PPIs that would have been predicted by the different methods in 2005 were confirmed in 2007?

Both direct relation and concept profile method-based PPI prediction scores were created using a MEDLINE corpus with publication dates up to February 2005. We ranked the PPIs according to the scores, and set a cut-off value at the 95% and 99% specificity levels based on PPIs present in Swiss-Prot 2005 (this is the only database for which historic versions are available). We subsequently evaluated how many of the 3,295 PPIs that were added to Swiss-Prot between 2005 and 2007 were above these cut-off values in 2005. These are the sensitivity values reported in Table S4. We also calculated the AuC based on Swiss-Prot 2007 alone.

The prediction performance is much better for concept profiles (52% versus 38% for a specificity level of 95%). This indicates that the majority of currently known PPIs were not yet explicitly described in MEDLINE at our testing point, but would have been predicted at a specificity rate of 95%. We postulate that this finding is indicative for the assumption that based on the full current literature a meaningful percentage of the 'unknowns' that pass the prediction threshold will be actual pairs worth studying in more detail.

Table 4. Results of the retrospective prediction of PPIs added to Swiss-Prot between 2005 and 2007. PPIs are ranked based on MEDLINE up to 2005, and specificity levels are based on Swiss-Prot 2005.The sensitivity is determined on Swiss-Prot 2007

	Concept-based	Concept profiles
Sensitivity at spec = 99%	27%	33%
Sensitivity at spec = 95%	38%	52%
Area under Curve	0.70	0.84

Case Studies

The next logical step was therefore to investigate whether this method can only predict PPIs that are 'known' but not explicit in the literature corpus used, or whether it would also be able to effectively predict unknown, but real PPIs. We investigated this in two case studies. We generated predicted interactions for proteins with two proteins that are intensively investigated in our group: (i) Dystrophin (DMD), a structural protein causing Duchenne muscular dystrophy

when defective, and (ii) Calpain 3 (CAPN3), a protease when mutated causing Limb-girdle muscular dystrophy (LGMD).

DMD

We presented the list of predicted interacting proteins with DMD ordered by descending association scores, to two experts for evaluation. At a specificity of 99%, there are 196 proteins predicted to interact with DMD. This list was too long to manually evaluate and we therefore restricted the human curation analysis to the 99.8% specificity level (top 42 proteins, Table S5). The full list is presented as Table 7 in the supplementary file. The 42 proteins include 7 of the 19 dystrophin-interacting proteins that are known from curated databases (sensitivity of 37% at this very high specificity level). The remaining established interaction partners generally rank high in the list of literature-predicted targets (13/19 in the top 196, p-value from Kolmogorov-Smirnov test for comparison with overall ranking: $3.4 \cdot 10^{-10}$). There are three proteins in the predicted set with at least indirect evidence in the literature for a physical interaction with DMD (CAV3, SPTB, ACTN2). One protein (SLMAP) may well interact given its distribution and localization but this needs experimental testing. Ten proteins in the list are found in the same protein complex as DMD but do not interact directly as far as known. Four proteins in the list were found wrongly associated with DMD due to homonym problems during literature indexing.

The remaining 17 proteins in the list are associated with DMD for other reasons (e.g. also involved in muscular dystrophy, or structural or functional homology) but are not likely to physically interact. If we only allow direct physical interaction pairs as true positives (11 proteins) the estimated precision is 26%. If predictions of protein pairs in a complex also are counted as true positives (21 proteins in total), the estimated precision would be 50%. Since also conceptually-related proteins that do not physically interact may be of interest to the biologist, the overall precision of our prediction method may be as high as 90%.

Table 5. Top 42 ranked proteins with DMD. In total 10,812 proteins were matched against DMD. 7 proteins as known to interact with DMD. Only 4 proteins are real false positives due to homonyms problem resulting in a precision over 0.9.

Rank	Protein symbol	Swiss-Prot id	Log similarity score	Direct relations	In PPI set	False positives (homonym)
1	**UTRN**	P46939	-5.14	214	x	
2	SGCA	Q16586	-6.13	119		
3	**DAG1**	Q14118	-6.22	139	x	
4	SGCB	Q16585	-6.60	54		
5	SGCD	Q53XA5	-6.95	46		
6	FCMD	O75072	-7.05	29		
7	DYSF	O75923	-7.19	43		
8	**DTNA**	Q9BS59	-7.31	17	x	
9	DRP2	Q13474	-7.34	9		
10	SSPN	Q0JV68	-7.45	17		
11	LAMA2	P24043	-7.46	25		
12	GK1	P32189	-7.56	33		x
13	CAPN3	P20807	-7.93	28		
14	CAV3	P56539	-7.95	24		
15	**SNTA1**	Q13424	-7.97	8	x	
16	EIF3S12	Q9UBQ5	-8.05	91		x
17	BEST1	O76090	-8.13	26		x
18	SPTB	P11277	-8.15	15		
19	FKRP	Q9H9S5	-8.16	4		
20	MEB	6988	-8.17	7		
21	SLMAP	Q14BN4	-8.20	4		
22	**SNTB1**	Q13884	-8.20	6	x	
23	NEB	P20929	-8.33	16		
24	SGCE	O43556	-8.35	10		
25	SGCG	Q13326	-8.46	305		
26	ACTN2	P35609	-8.49	11		
27	POMT1	Q5JT03	-8.50	3		
28	LOC130074	Q6NZ40	-8.50	16		x
29	CMD1K	14541	-8.50	27		
30	FER1L3	Q9NZM1	-8.51	3		
31	NOS1	P29475	-8.53	42		
32	IKBKAP	O95163	-8.63	10		
33	MACF1	Q5T3B3	-8.66	9		
34	AQP4	P55087	-8.67	13		
35	CKM	P06732	-8.70	11		
36	FSHMD1A	3966	-8.74	8		
37	TCAP	O15273	-8.75	7		

38	**DTNB**	O60941	-8.76	9	x	
39	LOC619409	619409	-8.82	5		
40	VCL	P18206	-8.87	36		
41	LGMD1A	6574	-8.88	3		
42	**SNTG1**	Q9NSN8	-8.90	5	x	

CAPN3

For CAPN3, an evaluation of the precision is more difficult since there is, compared to an intensively studied protein such as DMD, not enough established knowledge about its regulatory partners and substrates. Table S6 summarizes the currently known interaction partners for CAPN3: 13 interactions have been described in the literature (not necessarily in the abstracts that were used for our predictions, see column 'direct relation') and of those, six interactions have been entered in PPI databases. These known interaction partners generally rank high in the list of literature-predicted targets (Table S6, p-value from Kolmogorov-Smirnov test: $5.7 \cdot 10^{-5}$). Interestingly, the concept profiling method correctly predicted the interaction between myosin light chain 1 (MYL1) and CAPN3 on the basis of conceptual overlap in MEDLINE abstracts (specificity > 99%), although this interaction was only described in a full text paper[23] and not in any MEDLINE abstract used to generate the concept profiles.

Apart from literature based rediscovery of known interactions, we also set out to actually find new interaction partners for CAPN3. We selected predicted interaction partners that have not been entered in PPI databases so far and that do not have a direct co-occurrence in MEDLINE. The top ranked conceptual match is with Sarcoglycan-epsilon (SGCE), which is the smooth muscle counterpart of SGCA. Like for CAPN3, mutations in SGCA cause LGMD, but as far as we know, the protein is not expressed in skeletal muscle.

The second highest ranking protein was deemed to be an interesting candidate by the experts: Parvalbumin B (PARVB). The concept profiling method yielded a high association score because both proteins are described to have a physical interaction with dysferlin (DYSF)[24, 25], and with α-actinin (ACTN2)[26, 27], and they are both located at the Z-disc[28, 29]. For this predicted protein pair, we experimentally demonstrated a physical interaction, using three different set-ups.

First, it was shown that immobilized GST-fused PARVB could pull down recombinant T7-CAPN3 from bacterial lysates. Second, immobilized GST-PARVB could pull down endogenous CAPN3 from IM2 mouse myoblasts, and vice versa (Figure S2).

CAPN3 is hypothesized to act as a cytoskeleton remodeler and has been shown to interact with other focal adhesion proteins like Talin and Vinexin[30] (see Table S6). Ectopic CAPN3 over-expression results in cell rounding and cleavage and loss of co-expressed Talin and Vinexin[30]. This suggests that CAPN3 is a modulator of focal adhesions. Like CAPN3, PARVB is predominantly expressed in skeletal

muscle, where it plays a role in cell spreading and localizes to focal adhesions[26] (for a review, see [31]). The predicted interaction is coherent with this hypothesis, and substantiates the evidence for a role for CAPN3 outside the sarcomere.
This showcase is just one example of a correct and meaningful PPI prediction using concept profiles. This exemplary case study can not be seen proof that many of the other high ranking predictions will also be true physical and biologically relevant interactions. However none of the other consulted applications (STRING, iHOP) predicted this pair of interacting proteins. As the predictions using concept profiling are based on conceptual relatedness rather than an explicit co-occurrence in MEDLINE, this case study is indicative of the power of concept profiles to discover new, implicitly related pairs of interacting proteins. The statistics presented in this paper support the conclusion that predicted PPIs using our method, especially the subset that remains after expert analysis of the top ranking list are likely to be very significantly enriched for proteins that are worthwhile studying in wet lab experiments.

Table 6. List of proteins known to interact with Calpain-3. In total 10,812 proteins known to have a concept profile are matched against Calpain-3.

Name	Symbol	In PPI set	In literature (full text)	Direct relation (abstract)	Rank in literature-based prediction	Significant at specificity of 95 %
Dysferlin	DYSF	x	x	x	2	x
Titin	TTN	x	x	x	4	x
Filamin C	FLNC	x	x	x	27	x
Alpha-actinin	ACTN2		x	x	43	x
Calpastatin	CAST		x	x	55	x
IkappaBalpha	NFKBIA	x	x	x	126	x
Myosin light chain 1	MYL1		x		398	x
Alpha-spectrin	SPTAN1	x	x		426	x
Filamin A	FLNA		x		853	
Ezrin	VIL2		x		2739	
Vinexin	SORBS3		x		3301	
Talin	TLN1		x		4725	
AHNAK	AHNAK		x	No (*)	7371	
YWHAQ	YWHAQ	x			7617	

(*) paper describing this interaction in the abstract appeared in June 2008 and was not in the literature corpus used for the prediction

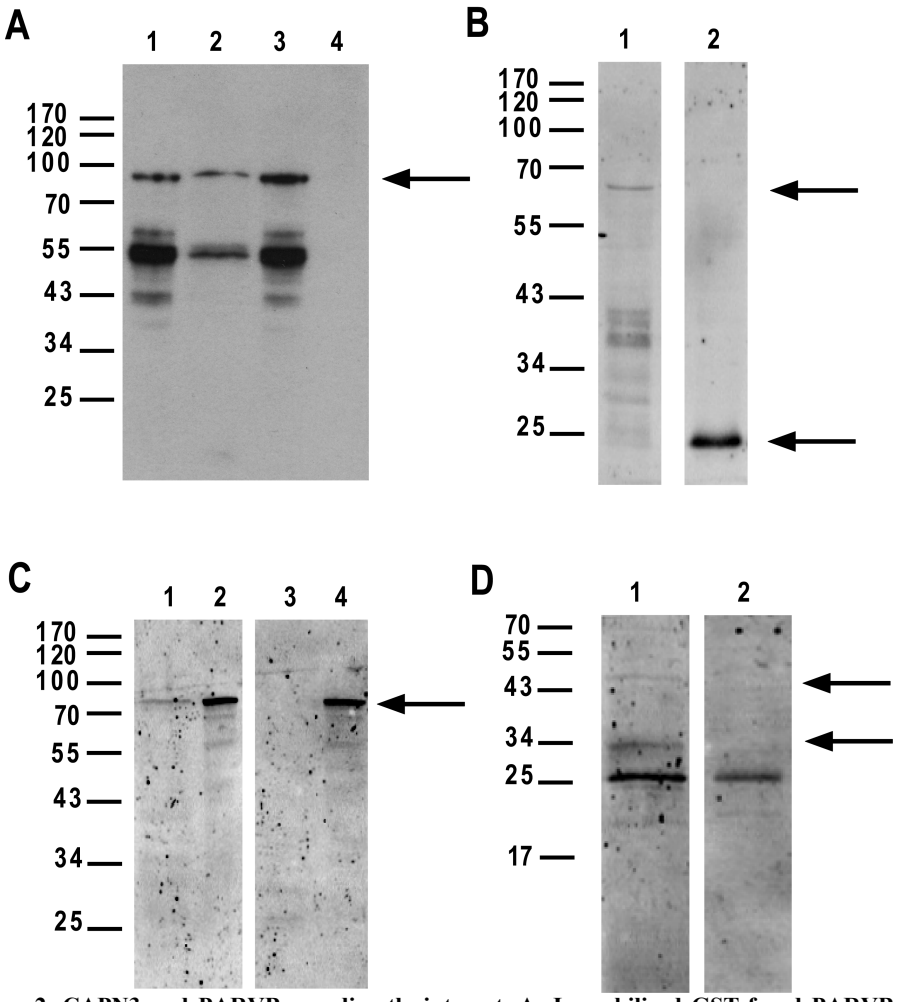


Figure 2. CAPN3 and PARVB can directly interact. A: Immobilized GST-fused PARVB can pull down recombinant CAPN3 from a bacterial T7-tagged CAPN3 lysate (Lane 2 vs 1), where unfused GST cannot (Lane 4 vs 3). As CAPN3 is an unstable protein that outside skeletal muscle rapidly autolyses we used the active site mutant C129S48. All fractions were resolved on SDS-PAGE gel and analyzed by immunoblotting with anti-CAPN3. The lanes represent: GST-PARVB non-bound fraction (1), GST-PARVB bound fraction (2), GST non-bound fraction (3), GST bound fraction (4). B: Equal loading was confirmed with anti-GST (Lane 1 GST-PARVB, Lane 2 GST). C: GST-fused PARVB can pull down endogenous full-length CAPN3 from an IM2 lysate (Lane 1 vs 2), contrary to unfused GST (Lane 3 vs 4). Lane 1 GST-PARVB bound fraction, Lane 2 non-bound fraction, Lane 3 GST bound fraction, Lane 4 non bound fraction. D: Likewise, GST-CAPN3 can pull down endogenous PARVB (Lane 1), contrary to GST (Lane 2). Both PARVB translation products bind. Here we used the Δ6 variant of Capn3 that does not autolyse yet retains function30, 49, and is expressed in the proliferating IM2 myoblasts. The arrows indicate the detected proteins and in all panels a molecular marker is depicted on the left.

Discussion

Scientists in general and scientific annotators in particular derive their knowledge on PPIs not directly discovered by their own experiments from the literature. However, as we show here, only 32% of the known PPIs covered by curated PPI databases can be found in MEDLINE abstracts (Table S1), the resource that is most commonly used for concept searches in the biomedical domain. This is despite the use of a sophisticated synonym expansion and homonym disambiguation systems . It is likely that many of these interactions are only mentioned in the full text of articles, or that the interactions have never been explicitly described in literature but were directly submitted to a database. In either case, the applicability of the most commonly used approach for PPI detection - the direct relation method in publicly available literature - appears to be severely limited.

The specificity and sensitivity levels achieved by our novel prediction method appear to be very promising. However, when we predict interaction partners for a specific protein, the estimated precision levels (*i.e.* how many of the predicted proteins are true interaction partners) are still seemingly quite moderate. A first consideration is that we are intrinsically unable to determine an accurate 'true false positive rate' for the predicted PPIs, due to the fact that many PPIs have simply not been discovered and described yet. This unavoidable complication most certainly will lead to an underestimation of precision levels. The case study of CAPN3 and PARVB signifies this point; initially this pair would have been classified as a 'false positive'.

For a realistic estimation of the precision of our prediction method, effectively each predicted protein pair should be validated in a wet lab experiment, which is out of the realistic scope of this study. For this reason we developed Nermal. (http://biosemantics.org/nermal). In Nermal, researchers can enter the UniProt identifier of a protein of interest, and the tool will return a ranked list of proteins that are most likely to interact with the query protein, in combination with information on whether the PPI has already been described explicitly in MEDLINE and/or in one of the protein databases.

A second complicating factor is the size of the 'negative' set (>50 million) compared to the 'positive' set (44,920) . This aspect is illustrated by the average prediction performance for each protein in Table S3 and by the case study with DMD in Table S5, where the top 42 proteins yielded a precision of only 26%, whilst the specificity was 99.8%. We are currently working on a further improvement of the precision by including data sources other than the literature in the PPI prediction algorithms. A final consideration is that our predictions are yielding more conceptual connections than physically interacting proteins only. Conceptual overlap obviously can indicate a variety of other types of relations

between proteins. For instance, we demonstrate that many proteins with high concept profile similarity do not interact directly, but are connected through intermediary proteins and are potentially part of the same complex or pathway. Therefore, the precision is to a certain extent dependent on the definition of a useful prediction. When other relationships than direct physical interactions are also deemed of interest, the precision of our method can become as high as 90%. The practical use of concept profiles will be in knowledge discovery in general, which is much broader than discovery of PPIs alone. In fact the hypothetical connection between any given pair of concepts can be calculated using our method.

To allow researchers to incorporate conceptual overlap data into their own analyses, we have made the concept profile similarity scores publicly available in two forms; first, a table containing similarity scores between all human proteins can be downloaded from our website; second, the previous mentioned web tool dubbed Nermal.

We conclude that concept profile similarity is a significantly better literature based predictor of PPIs than co-occurrence based methods. These improved predictions can be used to increase the biological interpretation and accuracy of interaction maps generated by high-throughput experiments, or can be used to prioritize proteins for further testing. In further studies, we will evaluate whether the use of concept profiles can also be applied in the prediction of other types of relations, for instance between drugs and diseases, and between genes and diseases.

Methods

Direct relation detection

Direct relations are typically extracted from literature based on co-occurrence[32]; if two proteins are mentioned in the same sentence or document more often than can be expected by chance, they are presumably related. We evaluated two alternatives for the detection of protein occurrences: a word-based approach and a concept-based approach. The word-based approach consists of combining the names of two proteins in an 'AND' query in the PubMed search engine. For the concept-based approach we have used the concept-recognition software Peregrine[33, 34], which includes synonyms and spelling variations[35] of concepts and uses simple heuristics to resolve homonyms. For this, Peregrine uses a protein ontology that was constructed by combining several gene and protein databases[36]. Even though a previous study has shown that Peregrine achieves state-of-the-art performance (75% precision and 76% recall on the BioCreactive II gene normalization testset[33, 34]), the concept recognition process is still error prone.

We used the likelihood ratio[19] to indicate the strength of the relation between two proteins. This ratio increases with the likelihood of there being a dependency between the occurrence of two proteins. Two hypotheses are used: (i) the occurrence of one protein is statistically dependent on the occurrence of the other protein; (ii) the occurrences are statistically independent. 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. The following equations give the likelihood ratio λ of concepts *i* and *j*.

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

where N is the total number of documents in the corpus, n_i , n_j, and n_{ij} are the number of documents containing *i, j,* and both *i* and *j*, respectively. $p=\frac{n_j}{N}$, the probability *j* occurs in an abstract irrespective of *i,* $p_1=\frac{n_{ij}}{n_i}$, the probability *j* occurs in an abstract containing *i*, $p_2=\frac{n_j-n_{ij}}{N-n_i}$, the probability *j* occurs in a document not containing *i*, and $L(k,l,x)=x^k(1-x)^{l-k}$, the likelihood function according to the binomial distribution.

Concept profile-based relation detection

To calculate the similarity of the contexts in which proteins appear in literature, we summarize the context of each protein in a concept profile. This profile contains all concepts that have a direct relation with a protein as found using the direct relation method described above. We evaluated two possible ways of applying this method: (i) using co-occurrences within a sentence, and (ii) using co-occurrences within an abstract. As shown in supplement S6, co-occurrence within an abstract yields a slightly higher AuC on predicting PPIs. We therefore used the abstract-based method in our study. The concepts in a profile include, in addition to proteins, all other concepts described in the Unified Medical Language System (UMLS) [37], such as diseases, symptoms, tissues, biological processes and many other types of concepts. We used the uncertainty coefficient[19] to calculate the weights of the concepts in the profiles. The uncertainty coefficient for the stochastic variables X and Y is given by

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

with *H(X)* is the entropy for *X* and *H(X|Y)* is the entropy for *X* given *Y*. X and Y can be any concept known in the ontology, e.g. drugs, proteins, diseases, disorders, chemicals, etc. The uncertainty coefficient is an information-theoretical measure that takes the a priori probability of direct relations into account. It gives extra weight to those concepts that are very specific for the set of documents belonging to the protein for which the concept profile is constructed. For a detailed description of concept profiles we refer to Jelier *et al*.[19].
The similarity score between two concept profiles A and B is taken as the inner product of the concept profile vectors, following Jelier *et al*.[38].

$$ip = \sum_{k=1}^{N} A_{uc(k)} B_{uc(k)}$$

with *uc(k)* the k^{th} uncertainty coefficient in the profile and *N* the total number of concepts the two profiles have in common. The inner product increases with increasing overlap in concept profiles. If two proteins co-occur, the inner product of their concept profiles is in general high. This is shown in supplement S4.

MEDLINE corpus
We extracted the title and abstract of subsections of MEDLINE. The corpus used in our main study has a time span from 1980 up to July 2007 and contains 12,098,042 citations. The corpus used for the retrospective study has a time span from 1980 up to February 2005 and contains 10,363,027 citations. This is an increase in time of 9.8% whereas the increase in published articles over the last two years is 17%.

Generation of the PPI and NIPP sets
There are many protein databases that describe PPIs. Not all of these use protein identifiers that could be linked to our protein ontology and the databases also show a high degree of overlap (see supplement S2). In our analysis we use BioGRID[39], DIP[40], HPRD[41], IntAct[42], MINT[43], Reactome[44], and Swiss-Prot[45] and only consider human proteins. Except for IntAct, all these databases are curated, meaning that they only contain PPIs that were judged to be correct according to strict criteria. IntAct, on the other hand, also contains unchecked results from high-throughput experiments which could contain many false positives. For a comparison of the prediction performance of our method on the individual databases we refer to supplement S3. The release dates and dates of download can be found in supplement S1.
For the construction of our set of known PPIs, we only rely on the curated databases; if a PPI was mentioned in one of these databases, we assumed it to be a true PPI. The resulting positive set contains 61,807 PPIs. After removing pairs that are not covered by all four prediction methods, 44,920 PPIs remain. Unfortunately, there is no database of proteins that are known not to interact. We can therefore only create a set of proteins which are less likely to interact. For our NIPP set we

took all pairs of human proteins that are not in the PPI set, and are not in the high-throughput part of the IntAct database. For computational reasons the calculation of the specificity and AuC was done on a random sample of 44,920 pairs of this set, setting both the positive and negative set size equal. Two randomly selected proteins form a pair and are checked if (i) they are not in the positive PPI set, (ii) not the same protein, e.g. proteins that interact with themselves are not taken into account, (iii) the protein pair is not already in the NIPP set, e.g. protein pairs can only occur once in a set. The random sample is actually quite small compared to the total NIPP set, however the ROC curve analysis is set size independent if the sample size is sufficiently large.
One last remark is that the positive set is incomplete. Therefore the creation of the NIPP set will introduce false negatives (PPIs that should have been in the positive set and recorded in a curated database). However the bias introduced by false negatives is negligible since the ratio of expected PPIs in human compared to the total set of formable protein pairs (~60 million) is very small[22].

STRING database
A copy of the STRING database, version 7.1, was downloaded from the STRING website. STRING is a pre-calculated database in PostgreSQL format. Only the text mining score table was used in our analysis.

Sensitivity, Specificity, Precision
In information retrieval terms like the sensitivity, specificity and precision are frequently used. The definitions are:

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

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

$$precision = \frac{TP}{TP + FP}$$

where TP are the number of true positives, FN number of false negatives, FP number of false positives, and TN number of true negatives. A perfect predictor has a specificity and sensitivity of 1.
When both set sizes are equal (#NIPP=#PPI) the precision equals the sensitivity. The specificity is sometimes confused with the precision. The distinction is critical when the classes are different sizes. A test with very high specificity can have a very low precision if there are far more true negatives than true positives, and vice versa.

Online web tool Nermal

Nermal is a web tool that prioritizes proteins that are most likely to be related with the protein you study. Given a query protein, the similarity scores are calculated between this protein and all other proteins in the ontology. The proteins are ranked on the similarity scores and presented in a table. Each row shows the similarity score between the two proteins, the databases in which the protein pair is known, and the sensitivity and (1-specificity) for that similarity score. These two rates should be interpreted as follows: given a similarity score between two proteins, (1-specificity) is the probability that a protein pair passing that score is a false positive. The sensitivity is the probability that you will miss a true PPI at that same score. Nermal can be found on http://biosemantics.org/nermal/. The full set of all protein pair match scores for human proteins can be downloaded at this link as well as the PPI and NIPP set used in the study.

DNA cloning

PARVB was amplified from proliferating IM2 myoblast cDNA with the following UTR primers: fw cgcactcgcttatgtcctc, rv ctccacatccttgtacttggtg. The ORF was amplified with a nested PCR introducing restriction sites for cloning into pET28aGST (modified pET28a vector with GST tag instead of T7 [46]). Primers were: fw aatatggatcctcctccgcgccaccacggt, rv atattctcgagctccacatccttgtacttgg. CAPN3 was similarly amplified with primers fw atgccaactgttattagtc, and rv ctaggcatacatggtaagc, and cloned into pET28aGST using fw tattacggatccatgccaactgttattagtc, and rv gtaatactcgagctaggcatacatggtaagc. The exon 6 deletion that does not autolyse was used for this experiment.
CAPN3c129s in pET28c was described previously[47]. All DNA constructs were verified by direct sequencing (LGTC, Leiden, The Netherlands), and subsequently transformed into BL21 (DE3)-RIL *E. coli* cells (Stratagene) for protein production.

Protein production and preparation of lysates

BL21 cells transformed with pET28aGST, pET28aGST-PARVB, pET28aGST-CAPN3 or pET28cCAPN3c129s were grown to log phase and stimulated with 1mM IPTG (Fermentas), and left to grow for 3 h at 37 °C. Next cells were spun down at 3,000 g and 4°C for 15 min. Pellets were dissolved in lysis buffer A (50 mM Tris-HCl pH 7.4, 1mM EDTA, 1.5 mg/ml lysozyme, 0.15 M NaCl, 1% Triton, Benozonase, 2x protease inhibitor cocktail tablet (Roche Molecular Biochemicals, Basel, Switzerland)), and sonicated on ice. Lysate was cleared by centrifugation at 13,000 g, and 4 °C for 30 min.
IM2 cells were grown at 33°C and 10% CO2 in DMEM 60196 (GIBCO-BRL, Grand-Island, NY) supplemented with 20% FCS, INFγ, glucose, pen/strep, glutamine and chick embryo extract. 15 cm plates (2x) were grown 75% confluent, washed 1x with PBS (37 °C) and lysed on ice with 1 ml lysis buffer B (50 mM

Tris-HCl pH 7.5, 150 mM NaCl, 0.2% Triton X-100, 2x protease inhibitor cocktail tablet). Lysate was spun down at 13,000 g and 4 °C for 30 min.

Pull-down

GST sepharose beads (4B, Amersham, Uppsala, Sweden) were washed with PBS (2x) and pre-equilibrated with lysis buffer (2x), and added to the cleared GST fusion lysates. Lysates were incubated at 4 °C and tumbling for 2 h. Next the lysates were spun down at 500 g, 4 °C for 5 min, and washed 3x with lysis buffer A. Separately, IM2 lysates were treated with washed and pre-equilibrated GST sepharose beads (buffer B). An aliquot of the GST fusion proteins was loaded on SDS-PAGE gel and Coomassie stained to confirm equal loading.
IM2 lysate, or T7-CAPN3c129s lysate, was added to the bait, and incubated O/N at 4 °C and tumbling. GST sepharose beads were spun down and the sup was stored as non-bound fraction. The beads were washed 5x with ice cold lysisbuffer (A or B, 3x short, 2x five minutes tumbling). All remaining sup was removed with an insulin syringe and proteins were eluted with 2x Laemmli sample buffer and boiled 5 min. An aliquot of the non-bound fraction was similarly prepared.

Western blot

Samples were loaded onto SDS-PAGE gels, separated and blotted to PVDF membrane. Blots were blocked in 4% skimmed milk PBS (Marvel) and incubated with primary antibody O/N at 4°C. Next morning blots were washed with 0.05% Tween in PBS, and incubated with secondary antibody for 1 h. Blots were washed again and scanned with an Odyssey scanner (Licor) or incubated with ECL plus (Amersham) and exposed to a Kodak XAR film. The following antibodies were used for Western detection: GaGST (1;10,000 Stratagene) MaCAPN3 (1;100, 12A2 Novocasta, Newcastle, UK), GaPARVB (1;200 Santa Cruz), GaMouseIRDye680 (1;5,000 Westburg, Leusden, NL), DaGIRDye800 (1;5,000 Westburg), RaMouseHRP (1;2,000 Dako Cytomation, Glostrup, Denmark), DaGoatHRP (1;10,000 Promega).

References

1. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
2. Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., et al., *Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.* Proc Natl Acad Sci U S A, 2000. **97**(3): p. 1143-7.

3. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.
4. Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.
5. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.
6. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.
7. Harrington, E.D., Jensen, L.J., and Bork, P., *Predicting biological networks from genomic data.* FEBS Lett, 2008. **582**(8): p. 1251-8.
8. Li, S., Wu, L., and Zhang, Z., *Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach.* Bioinformatics, 2006. **22**(17): p. 2143-50.
9. Kuffner, R., Fundel, K., and Zimmer, R., *Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts.* Bioinformatics, 2005. **21 Suppl 2**: p. ii259-67.
10. von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., et al., *STRING 7--recent developments in the integration and prediction of protein interactions.* Nucleic Acids Res, 2007. **35**(Database issue): p. D358-62.
11. Hoffmann, R. and Valencia, A., *A Gene Network for Navigating the Literature.* Nature Genetics, 2004. **36**: p. 664.
12. Swanson, D.R., *Fish oil, Raynaud's syndrome, and undiscovered public knowledge.* Perspect Biol Med, 1986. **30**(1): p. 7-18.
13. Swanson, D.R., *Medical literature as a potential source of new knowledge.* Bull Med Libr Assoc, 1990. **78**(1): p. 29-37.
14. Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., and Garner, H.R., *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-98.
15. Schuemie, M.J., Chichester, C., Lisacek, F., Coute, Y., Roes, P.J., et al., *Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of MEDLINE.* Proteomics, 2007. **7**(6): p. 921-31.
16. Jelier, R., Jenster, G., Dorssers, L.C., Wouters, B.J., Hendriksen, P.J., et al., *Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation.* BMC Bioinformatics, 2007. **8**: p. 14.

17. Tuason, O., Chen, L., Liu, H., Blake, J.A., and Friedman, C., *Biological nomenclatures: a source of lexical knowledge and ambiguity.* Pac Symp Biocomput, 2004: p. 238-49.
18. Chen, L., Liu, H., and Friedman, C., *Gene name ambiguity of eukaryotic nomenclatures.* Bioinformatics, 2005. **21**(2): p. 248-56.
19. Jelier, R., Schuemie, M.J., Roes, P.J., van Mulligen, E.M., and Kors, J.A., *Literature-based concept profiles for gene annotation: the issue of weighting.* Int J Med Inform, 2008. **77**(5): p. 354-62.
20. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., and Leunissen, J.A., *A text-mining analysis of the human phenome.* Eur J Hum Genet, 2006. **14**(5): p. 535-42.
21. Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.
22. Ben-Hur, A. and Noble, W., *Choosing negative examples for the prediction of protein-protein interactions*. 2006. p. S2.
23. Cohen, N., Kudryashova, E., Kramerova, I., Anderson, L.V., Beckmann, J.S., et al., *Identification of putative in vivo substrates of calpain 3 by comparative proteomics of overexpressing transgenic and nontransgenic mice.* Proteomics, 2006. **6**(22): p. 6075-84.
24. Matsuda, C., Kameyama, K., Tagawa, K., Ogawa, M., Suzuki, A., et al., *Dysferlin interacts with affixin (beta-parvin) at the sarcolemma.* J Neuropathol Exp Neurol, 2005. **64**(4): p. 334-40.
25. Huang, M., Zhu, X., Hao, Y., Payan, D.G., Qu, K., et al., *Discovering patterns to extract protein-protein interactions from full texts.* Bioinformatics, 2004. **20**(18): p. 3604-12.
26. Yamaji, S., Suzuki, A., Kanamori, H., Mishima, W., Yoshimi, R., et al., *Affixin interacts with alpha-actinin and mediates integrin signaling for reorganization of F-actin induced by initial cell-substrate interaction.* J Cell Biol, 2004. **165**(4): p. 539-51.
27. Ojima, K., Ono, Y., Doi, N., Yoshioka, K., Kawabata, Y., et al., *Myogenic stage, sarcomere length, and protease activity modulate localization of muscle-specific calpain.* J Biol Chem, 2007. **282**(19): p. 14493-504.
28. Sorimachi, H., Kinbara, K., Kimura, S., Takahashi, M., Ishiura, S., et al., *Muscle-specific calpain, p94, responsible for limb girdle muscular dystrophy type 2A, associates with connectin through IS2, a p94-specific sequence.* J Biol Chem, 1995. **270**(52): p. 31158-62.
29. Bendig, G., Grimmler, M., Huttner, I.G., Wessels, G., Dahme, T., et al., *Integrin-linked kinase, a novel component of the cardiac mechanical stretch sensor, controls contractility in the zebrafish heart.* Genes Dev, 2006. **20**(17): p. 2361-72.

30. Taveau, M., Bourg, N., Sillon, G., Roudaut, C., Bartoli, M., et al., *Calpain 3 is activated through autolysis within the active site and lyses sarcomeric and sarcolemmal components.* Mol Cell Biol, 2003. **23**(24): p. 9127-35.
31. Sepulveda, J.L. and Wu, C., *The parvins.* Cell Mol Life Sci, 2006. **63**(1): p. 25-35.
32. Cohen, A.M. and Hersh, W.R., *A survey of current work in biomedical text mining.* Brief Bioinform, 2005. **6**(1): p. 57-71.
33. Schuemie, M.J., Jelier, R., and Kors, J.A. *Peregrine: Lightweight gene name normalization by dictionary lookup*. in *Biocrative 2 workshop*. 2007. Madrid.
34. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., et al., *Overview of BioCreative II gene normalization.* Genome Biol, 2008. **9 Suppl 2**: p. S3.
35. Schuemie, M.J., Mons, B., Weeber, M., and Kors, J.A., *Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification.* J Biomed Inform, 2007. **40**(3): p. 316-24.
36. Kors, J.A., Schuemie, M.J., Schijvenaars, B.J.A., Weeber, M., and Mons, B., *Combination of genetic databases for improving identification of genes and proteins in text.* BioLINK, 2005
37. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.
38. Jelier, R., Schuemie, M.J., Veldhoven, A., Dorssers, L.C., Jenster, G., et al., *Anni 2.0: a multipurpose text-mining tool for the life sciences.* Genome Biol, 2008. **9**(6): p. R96.
39. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Research, 2006. **34**(Database): p. 535-539.
40. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
41. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.
42. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.
43. Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.

44. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.
45. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A., *UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase.* Methods Mol Biol, 2007. **406**: p. 89-112.
46. Huang, Y., Laval, S.H., van Remoortere, A., Baudier, J., Benaud, C., et al., *AHNAK, a novel component of the dysferlin protein complex, redistributes to the cytoplasm with dysferlin during skeletal muscle regeneration.* Faseb J, 2007. **21**(3): p. 732-42.
47. Huang, Y., de Morree, A., van Remoortere, A., Bushby, K., Frants, R.R., et al., *Calpain 3 is a modulator of the dysferlin protein complex in skeletal muscle.* Hum Mol Genet, 2008. **17**(12): p. 1855-66.

Supplementary information belonging to the article "Novel protein-protein interactions inferred from literature context"

S1 Downloaded protein database and release dates

In total seven protein databases are used in the study. The UniProt database consists of Swiss-Prot and TrEMBL.

Protein database	Date of download
Biogrid	September 28, 2007
DIP	September 20, 2007
HPRD	August 22, 2007*
IntAct	January 26, 2008
MINT	September 24, 2007*
Reactome	September 20, 2007
UniProt	February 14, 2008*

* For these databases it is possible to retrieve the original release dates. HPRD was released at January 9, 2007, MINT at June 28, 2007. Swiss-Prot and TrEMBL are combined in the database UniProt and have different release versions. UniProt release 12.0 contains Swiss-Prot release 54.0 and TrEMBL release 37.0. Both are dated from July 24, 2007.

S2 PPI overlap between the seven databases

Many of the PPIs appear in several databases. The following table shows the distribution and overlap over the seven protein databases.

	Biogrid	DIP	HPRD	IntAct	MINT	Reactome	Swiss-Prot
Biogrid	**16240**	205	15476	3006	2637	909	827
DIP		**365**	278	84	118	66	53
HPRD			**34957**	8031	7046	1401	1839
IntAct				**17456**	5754	595	3839
MINT					**10772**	375	650
Reactome						**29672**	290
Swiss-Prot							**3841**

S3 Performance on individual databases

The positive set is a combination of six protein databases. The databases vary in size and also the level of curation of each PPI. The following table gives the Area under the ROC (AuC) curve for each database individually. The last row is the AuC for the complete positive set.

Database	Concept profiles	Log likelihood	String
Biogrid	0.95	0.82	0.82
Dip	0.99	0.96	0.94
Hprd	0.93	0.79	0.78
Intact	0.71	0.57	0.56
Mint	0.87	0.72	0.70
Reactome	0.90	0.60	0.60
Swiss-Prot	0.84	0.71	0.71
Positive set	0.90	0.69	0.69

S4 Relationship between direct relation detection and concept profiles

The coverage in S3 shows that some PPIs have both overlap in concept profiles and a direct relation, while others have only concept profile overlap. The similarity score for proteins that share a direct relation is generally high. This is illustrated in figure 1.

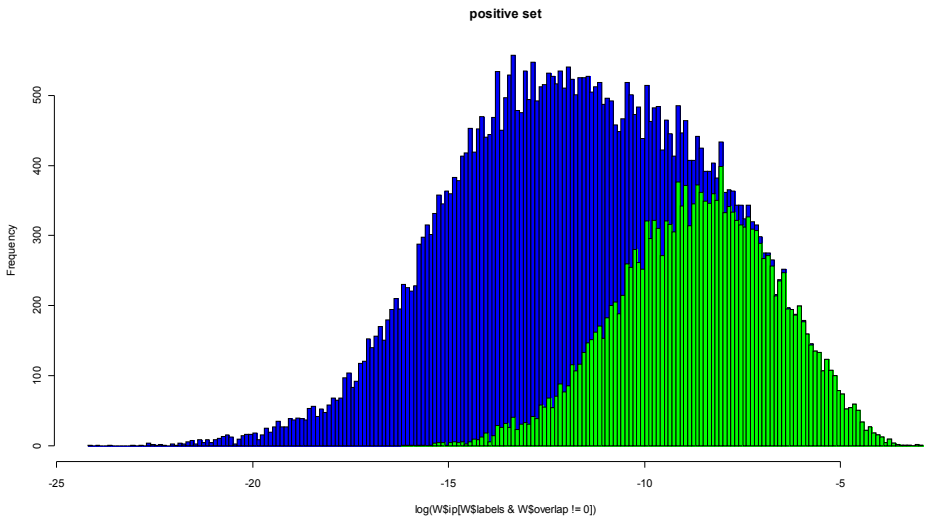
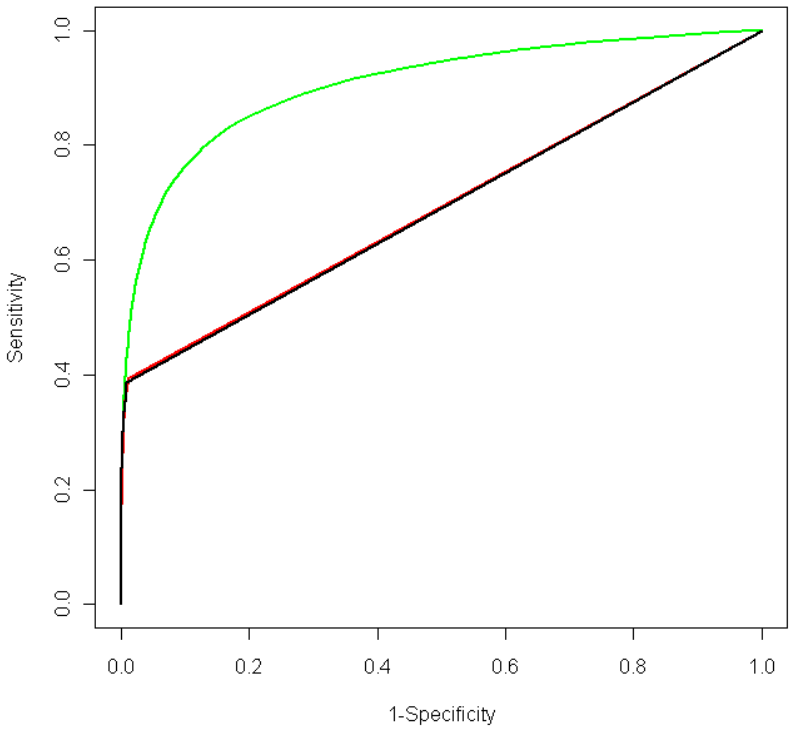


Figure 1. Histogram of the distribution of the similarity scores of: (blue) PPIs with concept profile overlap and no direct relation, and (green) PPIs with both a concept profile overlap and a direct relation.

S5 ROC curve analysis

The next figure shows the ROC curves for the concept profile similarity score (green), and the likelihood ratio of the direct relation method (red). For the direct relation method we discern two special cases: (i) each protein individual occurs in Medline but they are never mentioned together, and (ii) one of the proteins does not occur in MedLine at all. In the first case the likelihood score is –infinity, in the

second case the likelihood score is 0. These cases are quite frequent resulting in many duplicate values, and no natural ordering of the PPIs. We assume a perfect random ordering, resulting in the straight line at the end of the ROC curve in the figure (red for concept based method and black for the String database).



S6 Relation detection at the abstract and sentence level

For the construction of concept profiles, we investigated two options: assume two concepts are related when they co-occur (i) in the same sentence, and (ii) in the same abstract. For each option we evaluated the performance on the prediction of PPIs.

	Abstract level	Sentence level
*AuC**	0.93	0.91

The difference in results are neglectable. There is a very small decrease in performance using sentence based detection of relations.

* this analysis was done using a MedLine corpus up to April 2007 and using an older ontology.

S7 Ranked list of proteins predicted to interact with dystrophin (DMD)
The following table shows the proteins which similarity score with DMD have a specificity higher than 99%.

Rank	Protein symbol	Swiss-Prot id	Log similarity score	Direct relations	FP rate	TP rate	Biogrid	Dip	Hprd	Intact	Mint	Reactome	Swiss-Prot
1	UTRN	P46939	-5.14	214	0.003	0.856	0	0	1	0	0	0	0
2	SGCA	Q16586	-6.13	119	0.013	4.047	0	0	0	0	0	0	0
3	DAG1	Q14118	-6.22	139	0.013	4.047	0	0	1	0	0	0	0
4	SGCB	Q16585	-6.6	54	0.022	5.853	0	0	0	0	0	0	0
5	SGCD	Q53XA5	-6.95	46	0.032	8.168	0	0	0	0	0	0	0
6	FCMD	O75072	-7.05	29	0.034	8.62	0	0	0	0	0	0	0
7	DYSF	O75923	-7.19	43	0.039	9.65	0	0	0	0	0	0	0
8	DTNA	Q9BS59	-7.31	17	0.048	10.576	0	0	1	0	1	0	0
9	DRP2	Q13474	-7.34	9	0.049	10.625	0	0	0	0	0	0	0
10	SSPN	Q0JV68	-7.45	17	0.055	11.543	0	0	0	0	0	0	0
11	LAMA2	P24043	-7.46	25	0.055	11.543	0	0	0	0	0	0	0
12	GK1	P32189	-7.56	33	0.059	12.306	0	0	0	0	0	0	0
13	CAPN3	P20807	-7.93	28	0.08	15.06	0	0	0	0	0	0	0
14	CAV3	P56539	-7.95	24	0.08	15.06	0	0	0	0	0	0	0
15	SNTA1	Q13424	-7.97	8	0.081	15.274	0	0	1	0	0	0	0
16	EIF3S12	Q9UBQ5	-8.05	91	0.091	16.02	0	0	0	0	0	0	0
17	BEST1	O76090	-8.13	26	0.096	16.703	0	0	0	0	0	0	0
18	SPTB	P11277	-8.15	15	0.097	16.896	0	0	0	0	0	0	0
19	FKRP	Q9H9S5	-8.16	4	0.098	17.046	0	0	0	0	0	0	0
20	MEB	6988	-8.17	7	0.099	17.106	0	0	0	0	0	0	0
21	SLMAP	Q14BN4	-8.2	4	0.102	17.288	0	0	0	0	0	0	0
22	SNTB1	Q13884	-8.2	6	0.102	17.288	0	0	1	1	0	0	1
23	NEB	P20929	-8.33	16	0.117	18.497	0	0	0	0	0	0	0
24	SGCE	O43556	-8.35	10	0.117	18.497	0	0	0	0	0	0	0
25	SGCG	Q13326	-8.46	305	0.132	19.584	0	0	0	0	0	0	0
26	ACTN2	P35609	-8.49	11	0.137	19.754	0	0	0	0	0	0	0
27	POMT1	Q5JT03	-8.5	3	0.137	19.754	0	0	0	0	0	0	0
28	LOC130074	Q6NZ40	-8.5	16	0.138	19.925	0	0	0	0	0	0	0
29	CMD1K	14541	-8.5	27	0.138	19.925	0	0	0	0	0	0	0
30	FER1L3	Q9NZM1	-8.51	3	0.138	19.925	0	0	0	0	0	0	0
31	NOS1	P29475	-8.53	42	0.139	20.11	0	0	0	0	0	0	0
32	IKBKAP	O95163	-8.63	10	0.152	21.011	0	0	0	0	0	0	0
33	MACF1	Q5T3B3	-8.66	9	0.162	21.337	0	0	0	0	0	0	0
34	AQP4	P55087	-8.67	13	0.162	21.337	0	0	0	0	0	0	0
35	CKM	P06732	-8.7	11	0.167	21.668	0	0	0	0	0	0	0
36	FSHMD1A	3966	-8.74	8	0.172	21.859	0	0	0	0	0	0	0
37	TCAP	O15273	-8.75	7	0.173	22.153	0	0	0	0	0	0	0
38	DTNB	O60941	-8.76	9	0.173	22.153	0	0	1	0	1	0	0

39	LOC619409	619409	-8.82	5	0.181	22.675	0	0	0	0	0	0	0
40	VCL	P18206	-8.87	36	0.189	23.173	0	0	0	0	0	0	0
41	LGMD1A	6574	-8.88	3	0.192	23.273	0	0	0	0	0	0	0
42	SNTG1	Q9NSN8	-8.9	5	0.194	23.459	0	0	1	0	1	0	0
43	EMD	P50402	-8.94	12	0.201	23.864	0	0	0	0	0	0	0
44	GNE	Q6QNY6	-9	7	0.205	24.407	0	0	0	0	0	0	0
45	MYOZ2	Q9NPC6	-9.03	7	0.209	24.632	0	0	0	0	0	0	0
46	PGM5	Q15124	-9.04	3	0.212	24.733	0	0	1	0	0	0	0
47	CASQ1	P31415	-9.05	5	0.213	24.892	0	0	0	0	0	0	0
48	NR0B1	P51843	-9.06	18	0.218	25.047	0	0	0	0	0	0	0
49	SYNC1	Q9H7C4	-9.08	4	0.219	25.066	0	0	0	0	0	0	0
50	TTN	Q8WZ42	-9.08	7	0.22	25.157	0	0	0	0	0	0	0
51	DENR	O43583	-9.12	3	0.228	25.497	0	0	0	0	0	0	0
52	POMGNT1	Q8WZA1	-9.15	7	0.233	25.802	0	0	0	0	0	0	0
53	RAPSN	Q13702	-9.19	8	0.239	26.192	0	0	0	0	0	0	0
54	MYOT	Q9UBF9	-9.27	5	0.253	27.025	0	0	0	0	0	0	0
55	GDF8	O14793	-9.28	5	0.254	27.08	0	0	0	0	0	0	0
56	AIED	351	-9.3	2	0.256	27.193	0	0	0	0	0	0	0
57	TRIM32	Q13049	-9.31	3	0.256	27.193	0	0	0	0	0	0	0
58	MYH7	P13533	-9.36	18	0.265	27.894	0	0	0	0	0	0	0
59	LAMB1	P07942	-9.36	6	0.266	27.898	0	0	0	0	0	0	0
60	RP23	10277	-9.41	6	0.274	28.242	0	0	0	0	0	0	0
61	SNTG2	Q05AH5	-9.42	2	0.275	28.462	0	0	1	0	0	0	0
62	ACTN3	Q08043	-9.46	5	0.284	28.783	0	0	0	0	0	0	0
63	LMNA	P02545	-9.46	17	0.285	28.814	0	0	0	0	0	0	0
64	SPTBN4	Q9H254	-9.51	1	0.289	29.342	0	0	0	0	0	0	0
65	OTC	P00480	-9.55	8	0.298	29.59	0	0	0	0	0	0	0
66	DTNBP1	Q96EV8	-9.56	5	0.299	29.616	0	0	0	0	0	0	0
67	SNTB2	Q13425	-9.56	2	0.302	29.732	0	0	1	1	0	0	1
68	LGMD1B	6575	-9.57	0	0.304	29.838	0	0	0	0	0	0	0
69	SYNPO2	Q9UMS6	-9.57	3	0.307	29.862	0	0	0	0	0	0	0
70	RPGR	Q4VX65	-9.59	5	0.314	29.997	0	0	0	0	0	0	0
71	SPTBN1	Q01082	-9.59	7	0.314	29.997	0	0	0	0	0	0	0
72	GYPC	P04921	-9.6	3	0.318	30.105	0	0	0	0	0	0	0
73	TAZ	Q16635	-9.63	8	0.329	30.387	0	0	0	0	0	0	0
74	SNORD95	32757	-9.63	3	0.329	30.387	0	0	0	0	0	0	0
75	DMN	O15061	-9.64	3	0.33	30.483	0	0	0	0	0	0	0
76	SEPN1	Q9NZV5	-9.74	2	0.364	31.413	0	0	0	0	0	0	0
77	GATM	P50440	-9.76	2	0.37	31.678	0	0	0	0	0	0	0
78	MTM1	Q13496	-9.78	5	0.372	31.823	0	0	0	0	0	0	0

79	PLEC1	Q15149	-9.82	1	0.384	32.306	0	0	0	0	0	0	0
80	NRG4	Q0P6N4	-9.82	1	0.387	32.363	0	0	0	0	0	0	0
81	AAVS1	22	-9.83	4	0.389	32.414	0	0	0	0	0	0	0
82	MYOD1	O75321	-9.84	9	0.389	32.414	0	0	0	0	0	0	0
83	FLNC	Q14315	-9.87	3	0.398	32.802	0	0	0	0	0	0	0
84	VAULTRC3	12656	-9.88	1	0.4	32.846	0	0	0	0	0	0	0
85	CFC1	Q9GZR3	-9.89	16	0.401	32.965	0	0	0	0	0	0	0
86	IL1RAPL1	Q7Z2K4	-9.9	4	0.403	33.116	0	0	0	0	0	0	0
87	DYNLT3	P51808	-9.91	3	0.406	33.239	0	0	0	0	0	0	0
88	DTL	Q9NZJ0	-9.93	2	0.411	33.417	0	0	0	0	0	0	0
89	DMPK	Q09013	-9.93	5	0.411	33.417	0	0	0	0	0	0	0
90	MYOG	P15173	-9.94	8	0.414	33.444	0	0	0	0	0	0	0
91	DGKZ	Q13574	-9.95	2	0.417	33.614	0	0	1	0	0	0	0
92	SRRM2	O60382	-9.96	2	0.418	33.686	0	0	0	0	0	0	0
93	SMN1	Q16637	-10.04	3	0.441	34.576	0	0	0	0	0	0	0
94	MYL2	P10916	-10.05	2	0.445	34.75	0	0	0	0	0	0	0
95	MYLPF	Q6IB41	-10.09	2	0.457	35.062	0	0	0	0	0	0	0
96	PVALB	P02144	-10.1	22	0.464	35.21	0	0	0	0	0	0	0
97	COL6A1	P12109	-10.14	2	0.473	35.566	0	0	0	0	0	0	0
98	MYH7	P12883	-10.14	2	0.474	35.583	0	0	0	0	0	0	0
99	CAPN8	1485	-10.14	1	0.476	35.607	0	0	0	0	0	0	0
100	MEAX	6987	-10.15	2	0.477	35.638	0	0	0	0	0	0	0
101	POMT2	Q59GJ5	-10.15	0	0.479	35.702	0	0	0	0	0	0	0
102	AGRN	O00468	-10.18	3	0.483	35.963	0	0	0	0	0	0	0
103	DNPEP	Q9HAC6	-10.18	2	0.484	35.967	0	0	0	0	0	0	0
104	XIC	12809	-10.19	0	0.491	36.045	0	0	0	0	0	0	0
105	PDLIM3	Q53GG5	-10.2	2	0.499	36.198	0	0	0	0	0	0	0
106	COL6A2	P12110	-10.21	1	0.5	36.268	0	0	0	0	0	0	0
107	GAA	P10253	-10.21	7	0.501	36.3	0	0	0	0	0	0	0
108	LAMA1	P25391	-10.26	0	0.52	36.811	0	0	0	0	0	0	0
109	MYF6	P23409	-10.27	2	0.524	36.845	0	0	0	0	0	0	0
110	CHRNG	P07510	-10.29	1	0.531	37.012	0	0	0	0	0	0	0
111	SPTA1	O60686	-10.3	2	0.535	37.144	0	0	0	0	0	0	0
112	CSRP3	P50461	-10.3	3	0.542	37.216	0	0	0	0	0	0	0
113	EPB41	P11171	-10.31	4	0.548	37.322	0	0	0	0	0	0	0
114	PBDX	P55808	-10.32	1	0.548	37.322	0	0	0	0	0	0	0
115	LAMB2	P55268	-10.32	1	0.549	37.398	0	0	0	0	0	0	0
116	WDM	50988	-10.34	1	0.561	37.627	0	0	0	0	0	0	0
117	HHG	4902	-10.35	1	0.563	37.68	0	0	0	0	0	0	0
118	RPS4Y1	P22090	-10.35	2	0.563	37.68	0	0	0	0	0	0	0

119	ITGA7	Q13683	-10.35	1	0.563	37.68	0	0	0	0	0	0	0
120	TNNT2	P45379	-10.37	4	0.574	37.877	0	0	0	0	0	0	0
121	CMD1B	2102	-10.37	2	0.574	37.877	0	0	0	0	0	0	0
122	FOSL2	P15408	-10.38	1	0.578	38.028	0	0	0	0	0	0	0
123	SFRS2	Q01130	-10.39	3	0.582	38.131	0	0	0	0	0	0	0
124	MIB2	Q0JSM5	-10.4	1	0.59	38.257	0	0	0	0	0	0	0
125	MSRB2	Q9Y3D2	-10.4	1	0.59	38.257	0	0	0	0	0	0	0
126	DNM1L	O00429	-10.4	1	0.59	38.257	0	0	0	0	0	0	0
127	XKR1	P51811	-10.41	2	0.593	38.307	0	0	0	0	0	0	0
128	XIST	12810	-10.42	0	0.605	38.416	0	0	0	0	0	0	0
129	TPM1	O15513	-10.42	5	0.606	38.439	0	0	0	0	0	0	0
130	COL6A3	P12111	-10.42	1	0.61	38.522	0	0	0	0	0	0	0
131	SUCLG1	P53597	-10.42	1	0.613	38.547	0	0	0	0	0	0	0
132	NRG3	P56975	-10.43	1	0.614	38.587	0	0	0	0	0	0	0
133	PPP1R10	Q96QC0	-10.43	1	0.614	38.587	0	0	0	0	0	0	0
134	RNPS1	Q15287	-10.44	2	0.623	38.738	0	0	0	0	0	0	0
135	MYEF2	Q9P2K5	-10.46	2	0.632	38.88	0	0	0	0	0	0	0
136	GAMT	Q14353	-10.48	1	0.646	39.13	0	0	0	0	0	0	0
137	TNNC1	P63316	-10.49	3	0.65	39.232	0	0	0	0	0	0	0
138	RP2	O75695	-10.51	3	0.661	39.446	0	0	0	0	0	0	0
139	MYL6	P60660	-10.51	1	0.663	39.469	0	0	0	0	0	0	0
140	CTSH	P09668	-10.51	2	0.664	39.48	0	0	0	0	0	0	0
141	CXADR	P78310	-10.53	4	0.681	39.609	0	0	0	0	0	0	0
142	ELOVL4	Q9GZR5	-10.53	1	0.682	39.635	0	0	0	0	0	0	0
143	MYF5	P13349	-10.57	3	0.703	40.025	0	0	0	0	0	0	0
144	FBXO32	Q969P5	-10.59	1	0.716	40.275	0	0	0	0	0	0	0
145	PRX	Q9BXM0	-10.6	2	0.716	40.275	0	0	0	0	0	0	0
146	BLOC1S1	P78537	-10.6	1	0.718	40.311	0	0	0	0	0	0	0
147	MUSK	O15146	-10.6	1	0.718	40.311	0	0	0	0	0	0	0
148	SMN2	Q16637	-10.62	1	0.731	40.491	0	0	0	0	0	0	0
149	IS2	282552	-10.62	0	0.734	40.516	0	0	0	0	0	0	0
150	DM1	2923	-10.62	2	0.735	40.521	0	0	0	0	0	0	0
151	DYNLL1	P63167	-10.63	1	0.737	40.567	0	0	0	0	0	0	0
152	PDAP1	Q13442	-10.63	1	0.737	40.567	0	0	0	0	0	0	0
153	INVS	Q5JS85	-10.65	3	0.748	40.76	0	0	0	0	0	0	0
154	PABPN1	Q86U42	-10.66	1	0.76	40.877	0	0	0	0	0	0	0
155	NOS1AP	O75052	-10.67	1	0.762	40.915	0	0	0	0	0	0	0
156	KCNJ10	P78508	-10.67	3	0.765	40.985	0	0	0	0	0	0	0
157	TCTA	P57738	-10.68	0	0.767	41.034	0	0	0	0	0	0	0
158	ACTA1	P68133	-10.68	2	0.769	41.074	0	0	1	0	0	0	0

159	CACNA1I	Q9P0X4	-10.68	1	0.776	41.155	0	0	0	0	0	0	0
160	MST4	Q8NC04	-10.71	1	0.786	41.424	0	0	0	0	0	0	0
161	KFSD	6313	-10.72	2	0.786	41.424	0	0	0	0	0	0	0
162	IGFBP5	P24593	-10.72	1	0.789	41.494	0	0	0	0	0	0	0
163	DST	O94833	-10.75	7	0.812	41.721	0	0	0	0	0	0	0
164	FRG1	Q14331	-10.76	0	0.82	41.913	0	0	0	0	0	0	0
165	CD5L	O43866	-10.77	0	0.823	41.983	0	0	0	0	0	0	0
166	ITPR1	Q14643	-10.79	1	0.83	42.123	0	0	0	0	0	0	0
167	PARVB	Q9HBI1	-10.8	0	0.838	42.202	0	0	0	0	0	0	0
168	RIMS1	Q5SZK2	-10.8	1	0.838	42.202	0	0	0	0	0	0	0
169	GAS2	O43903	-10.8	3	0.845	42.293	0	0	0	0	0	0	0
170	WAS	P42768	-10.8	238	0.846	42.308	0	0	0	0	0	0	0
171	CDH15	P55291	-10.81	2	0.849	42.363	0	0	0	0	0	0	0
172	ACTC1	P68032	-10.82	1	0.85	42.429	0	0	1	0	0	0	0
173	MLS	7145	-10.82	1	0.85	42.429	0	0	0	0	0	0	0
174	CACNA1S	Q13698	-10.82	3	0.851	42.507	0	0	0	0	0	0	0
175	ERF	P50548	-10.83	1	0.856	42.558	0	0	0	0	0	0	0
176	SFRS1	Q07955	-10.84	1	0.865	42.672	0	0	0	0	0	0	0
177	DCTN3	O75935	-10.87	1	0.894	42.976	0	0	0	0	0	0	0
178	DDX3Y	O15523	-10.87	1	0.894	42.976	0	0	0	0	0	0	0
179	SFRS5	Q13243	-10.87	1	0.896	43.018	0	0	0	0	0	0	0
180	ALG3	Q92685	-10.87	109	0.896	43.018	0	0	0	0	0	0	0
181	RYR1	O75591	-10.88	1	0.908	43.141	0	0	0	0	0	0	0
182	GAS2L1	Q99501	-10.9	1	0.919	43.308	0	0	0	0	0	0	0
183	COL4A5	P29400	-10.9	0	0.919	43.308	0	0	0	0	0	0	0
184	PTBP2	O95652	-10.91	0	0.926	43.393	0	0	0	0	0	0	0
185	MYH6	P13533	-10.91	4	0.928	43.444	0	0	0	0	0	0	0
186	IGFBP4	P22692	-10.92	3	0.933	43.582	0	0	0	0	0	0	0
187	SYNE1	Q5JV23	-10.93	1	0.934	43.62	0	0	0	0	0	0	0
188	ZNF91	Q05481	-10.93	1	0.939	43.637	0	0	0	0	0	0	0
189	SP1	P08047	-10.93	7	0.94	43.669	0	0	0	0	0	0	0
190	PTPN22	Q5TBC0	-10.93	1	0.941	43.671	0	0	0	0	0	0	0
191	LOC619511	619511	-10.94	1	0.943	43.701	0	0	0	0	0	0	0
192	EIF4EBP1	Q13541	-10.95	3	0.953	43.838	0	0	0	0	0	0	0
193	MYOZ1	Q9NP98	-10.97	0	0.977	44.029	0	0	0	0	0	0	0
194	BSN	Q2NLD3	-10.98	0	0.988	44.167	0	0	0	0	0	0	0
195	FBXO11	Q86XK2	-10.99	1	0.999	44.248	0	0	0	0	0	0	0
196	ZBTB20	Q9HC78	-10.99	1	0.999	44.248	0	0	0	0	0	0	0

Chapter 4

In silico discovery and experimental validation of new protein-protein interactions

Herman H.H.B.M. van Haagen, Peter A.C. 't Hoen, Antoine de Morrée, Willeke M.C. van Roon-Mom, Dorien J. M. Peters, Marco Roos, Barend Mons, Gert-Jan van Ommen, Martijn J. Schuemie

Manuscript accepted to Proteomics at November 25, 2010

Abstract
We introduce a framework for predicting novel protein-protein interactions (PPIs), based on Fisher's method for combining probabilities of predictions that are based on different data sources, such as the biomedical literature, protein domain and mRNA expression information. Our method compares favorably to our previous method based on text-mining alone and other methods such as STRING. We evaluated our algorithms through the prediction of experimentally found protein interactions underlying Muscular Dystrophy, Huntington's Disease, and Polycystic Kidney Disease, which had not yet been recorded in protein-protein interaction databases. We found a 1.74 fold increase in mean average prediction precision for dysferlin and a 3.09 fold for huntingtin when compared to STRING. The top 10 of predicted interaction partners of huntingtin were analysed in depth. Five were identified previously, and the other five were new potential interaction partners. The full matrix of human protein pairs and their prediction scores is available for download. Our framework can be extended to predict other types of relationships such as proteins in a complex, pathway or related disease mechanisms.

Introduction
The biomedical literature and domain-specific databases contain a wealth of background information, which should aid biomedical researchers in the design and interpretation of their experiments. Many databases compile information from several resources for use as reference and lookup. Databases such as KEGG, STRING, and IntNetDB are examples that are useful for studying protein-protein relations. These resources represent existing knowledge well. However, of particular interest is the potential to reveal genuinely novel relations by data mining algorithms [1]. In previous work we showed the ability to predict protein-protein interactions using information contained in literature alone [2]. With the so called concept profile technology, we found novel protein interaction pairs that could not have been found by a simple MEDLINE query. This was illustrated by the prediction of the physical interaction between calpain 3, which causes a form of muscular dystrophy, and parvalbumin B, which is found mainly in skeletal muscle. However, this method does not exploit the full potential of information available for data mining. Combining data sets beyond literature may increase coverage and the reliability of our predictions.
Combining data from different sources for extracting relevant knowledge is a general objective in bioinformatics. Here, we distinguish data concatenation and evidence score combination. Data concatenation merely summarizes the results of queries to a number of individual databases (*e.g.* www.genecards.org [3]). The summaries are provided to an investigator for interpretation. When investigating PPIs, a summary may contain information on the presence of certain PPIs in a curated PPI database, Gene Ontology terms that are shared, and co-expression of

genes in certain tissues or cellular compartments. No additional algorithms are provided to predict and highlight putatively novel relations.
Evidence score combination provides a score to order the information from a combination of data sets. For each set, the score reflects the contribution of the set to the overall result of a query across a number of data sets. The individual scores are combined using one of several combination techniques. Evidence score combination can be used to predict new relationships between biological concepts, including protein-protein interactions (PPIs).
Several web tools are available that provide some form of data integration and evidence score combination for the extraction of PPIs [4-6]. (i) STRING[6], which is maintained by EMBL, contains functional associations for over 600 species. STRING uses information on genomic content, high throughput experiments, co-expression, and co-mentioning in PubMed abstracts and recorded in public curated databases like KEGG or Reactome. STRING uses a combination technique based on the product of p-values to provide a confidence score for predicted PPIs. (ii) FunCoup[5] provides a predicted protein-protein network for eight eukaryotes. It uses information on PPIs, mRNA expression, sub-cellular co-localization, phylogenetic profiles, miRNA-mRNA targets, transcription factor regulation, protein expression, and protein domain interactions. The network is optimized using a Bayesian approach. (iii) IntNetDB v1.0.[4] is restricted to a few species and mainly focuses on human data. IntNetDB uses physical interactions, phenotype similarity, genetic interactions, shared GO annotation, domain-domain interactions, co-expression, and gene context in PubMed articles. IntNetDB uses a Naive Bayes classifier as combination technique. As stated previously, these web tools perform well on reproducing existing PPIs. STRING for instance aggregates known interactions from several databases and predictions made by several predicting methods. Their evidence score reflects how well supported an interaction or association is by these sources. Our aim is to develop a true interaction predictor, and a score that reflects the likelihood that the prediction is true. In contrast to STRING, our method will predict known as well as unknown interactions that can have equally high scores. The correspondence with known protein-protein interactions validates our approach.
The remainder of this article is structured as follows. We give a brief introduction of our framework which is based on Fisher's method for combining p-values based on different data sources. Next our framework was validated by evaluating three show cases. The first case is on dysferlin (*DYSF* encoded protein), its deficiency causing progressive Limb Girdle Muscular Dystrophy type 2B. We aimed for the discovery of dysferlin interaction partners by immunoprecipitation experiments and show how well we could predict these new interactions. The second case relates to the huntingtin protein which is associated with Huntington's disease. We took PPIs from the article by Kaltenbach et. al. [7], which had not been stored (yet) in PPI

databases and which had not been described in MEDLINE abstracts. They serve as a good test set where we simulate that our framework is able to predict proteins from these lists. The last show case is on Polycystic Kidney Disease caused by the mutated *PKD1* gene. This case illustrated how to solve homonym problems encountered in text by including additional expression data. We end with summarizing the results for each show case and conclude that we significantly improve the discovery of novel PPIs over previous methods.

Materials and methods

Performance measurement

For measuring performance we used receiver operating characteristics (ROC) curves and the area under the curve (AuC). Second, we used the mean average precision (MAP). Both are measurements often used in information retrieval. In the case studies the test sets used for dysferlin and huntingtin are labeled as positive instances. The rest of the proteins in our ontology are labeled as negative instances. The AuC values have a range between 0.5 and 1. 0. A value of 0.5 means that the systems is no different than a random ordering of the samples, *i.e.* the positive instances are equally distributed over the ordered list (ordered by match score) of all proteins. An AuC of 1 mean the system is a perfect predictor, *i.e.* al positive instances first rank at the top followed by all negative instances.

The mean average precision is a measurement more sensitive to samples size of both the positive and negative set. The MAP is calculated by averaging all precisions where each precision is calculated at the occurrence of a positive instance in an ordered list (ordered by match score).

Match scores for each individual database.

Concept profiles

To calculate the similarity of the contexts in which proteins appear in literature, we summarize the context of each protein in a concept profile. This profile for a protein contains all concepts that are co-mentioned with the protein as found in MEDLINE abstracts. To find concepts in text we have used the concept-recognition software Peregrine [8], which includes synonyms and spelling variations of concepts and uses simple heuristics to resolve homonyms. For this, Peregrine uses a protein ontology that was constructed by combining several gene and protein databases. Proteins from different species are fused together and we do not distinguish between a gene and a protein.

Each concept in the profile is assigned a weight. The weight reflects the strength of the association between the concept and the protein. The concepts that appear in both protein profiles are used to calculate a match score. The match score is the

inner product calculated over the weights for the shared concepts. For a detailed description of concept profiles and weight calculation we refer to [9].

Gene Ontology

Match scores defined for the Gene Ontology were investigated by Mistry *et al.* [10]. They compared the term overlap with other well known similarity measures adapted from the work of Resnik[11], Lin[12], and Jiang[13]. We did a ROC curve analyses on all four similarity measures and obtained the highest AuC value for the method by Resnik. The score we use is inferred from Resnik. Resnik originally defines the score to find the similarity between two GO terms, whereas we want to find the similarity between two proteins. First the information content for a GO term t_i is defined

$$IC(t_i) = -\log(p(t_i))$$

where $p(t_i)$ is the probability of a gene being annotated to that term. $p(t_i)$ can be calculated as follows

$$p(t_i) = \frac{\#genes_annot(t_i)}{\#genes_annot(rootnode)}$$

In words, the number of genes annotated to GO term t_i divided by all the genes under consideration. All the genes are annotated in the root node of the GO graph. The information content of the root node therefore is 0 as would be expected. Resnik's similarity measure is then calculated by taking the IC of the lowest common ancestor (LCA) shared between two proteins.

$$sim_{Resnik}(p1, p2) = IC(LCA)$$

With *p1* and *p2* the two proteins that form a pair (either random or a PPI).

Microarray data

Microarray co-expression values are pre-calculated for COXPRESdb [14] and can be used directly after download. For Gene Atlas [15] the human GNF1H chip is used. First, the log was taken from the MAS5.0 normalized expression values for each tissue (78 in total), and probes with the same EntrezGene IDs were averaged. Subsequently, a Pearson correlation was calculated for the gene expression values for all pairs of genes.

Tissue specificity

TiGER[16] contains expressed sequence tags that are defined for 30 tissues. For TiGER we evaluated a number of vector similarity measures namely, Pearson's correlation coefficient, inner product, cosine, euclidean distance, and the Tanimoto coefficient. The latter one showed the best prediction performance. The Tanimoto coefficient between two vectors A and B is defined as follows:

$$T(A,B)=\frac{A\cdot B}{\|A\|^2+\|B\|^2-A\cdot B}$$

Tanimoto coefficient values > 0.85 are generally considered similar to each other.

Domain-domain interactions

We used InterPro[17] to annotate each protein in our ontology with domains. Subsequently we used DOMINE[18] to determine which domains (one of protein A and the other of protein B) interact. The final score is simply the number of interacting domains.

Probable non interacting protein pairs

A null hypothesis was generated by choosing random protein pairs[19]. This null hypothesis is used to calculate a single sided p-value for Fisher's Method. The only constraint that we applied is that the protein pair should not be in a curated database nor in the high-throughput database IntAct [20]. The curated databases used are listed in the supplementary files. The complete random protein pair set consisted of over 500 millions proteins pairs (all possible combinations of two proteins). For computational reasons our analysis was limited to a random subset of 100,000.

Combined match score: Fisher's method

Fisher's method combines one sided p-values from different databases into one test statistics which follows a χ^2 distribution with $2*L$ degrees of freedom using the formula

$$\chi^2=-2\sum_{i=1}^{L}\log(p_i)$$

When the p-values tend to be small, the test statistic χ^2 will be large. The p-values are obtained from the random protein pairs distribution described earlier.

In the first version of Fisher, missing values for this combiner are also completely ignored. This is done by setting the p-value to 1. The log becomes 0 and the missing value does not contribute to the score. The degrees of freedom are fixed and are the same for each sample (a protein pair). The second version takes into account the degrees of freedom (dof). The dof is only taken for databases that have a match score. The last two variations are where the individual database scores are weighted. They are weighted with the AuC and MAP values and then the previous formula is applied. Fisher's method can be sensitive to databases if p-values become 0. Then the combined score is dominated by one database only. This could result in false positives. We added a small offset to each p-value of 10^{-4} to filter for this side effect when p-values are too small (or 0).

STRING

We benchmarked our system against the STRING database. We downloaded STRING version 8.1 that was last updated on October 18, 2009. A current version of STRING can be found online: http://string.embl.de. The databases used by STRING are:

- Neighborhood in the genome (nscore)
- Gene fusion (fscore)
- Co-occurrence across genomes (homology; pscore and hscore)
- (Co-expression (ascore)
- Experimental/biochemical data (escore),
- Association in curated databases (dscore)
- Co-mentioned in PubMed abstracts (tscore; text-mining based on direct co-occurrences)

String uses a combiner based on the product of probabilities using the following formula

$$S = 1 - \prod_{i}^{N} (1 - S_{i)}$$

With *Si* the probability score for database *i*, *S* the combined score, and *N* the total number of databases to be combined.

Dataset

The raw scores for each database, and the combined Fisher Method score, are merged together in a tab delimited text file which can be downloaded from our website http://www.biosemantics.org/ppi-prediction

Results and Discussion

Our previous approach used only text-mining for the prediction of PPIs [2]. We postulated that a combination of information indicative of protein-protein associations, such as co-expression and functional and structural similarities, increases the overall probability of a genuine PPI. Therefore, we included information from these five additional databases:

- Gene Ontology: manual functional annotation
- COXPRESdb[14]: mRNA co-expression over a wide range of conditions
- Gene Atlas[15]: mRNA co-expression in 78 tissues
- Tiger[16]: expressed sequence tags (EST) counts in 30 tissues
- InterPro/DOMINE[17, 18]: domain annotation and domain-domain interactions

Our hypothesis was that these data sources are valuable for the prediction or protein interactions since two interacting proteins should be expressed in the same tissue and cell, are likely to be co-regulated at the transcriptional level, and interact via a specific combination of protein domains. The selected databases are publically available and have suitable data formats for processing (xml, tab delimited files, Entrez Gene or Uniprot accession numbers, etc). For each database a score was defined that reflects a degree of association between two proteins. Individual scores were then combined to obtain a final score for a protein pair.
For combining the score we used a method developed by Fisher (see Materials and Method for detailed explanation of this method). This method is based on combining p-values taken from different predictions. Briefly, the match score for every database is converted into a single sided p-value. Then, the p-values are log transformed and summed resulting in a Fisher score with 2*N degrees of freedom (N the number of p-values to be summated). We made two variations of this Fisher method. In the first one, the degrees of freedom are fixed (missing values taken into account). In the second one, each p-value is weighted with AuC or MAP values, giving more weight to the data sources that are most important for the prediction. The AuC stands for Area under the ROC curve and MAP for Mean Average Precision. Both measures are well known in the field of information retrieval and data-mining. An AuC of 0.5 reflects a prediction with random behavior (like flipping a coin). An AuC of 1 correlates to a perfect prediction. MAP values range from 0 to 1 (perfect prediction). All performance measures (AuC and MAP values) are given in the supplementary files.
We choose Fisher's method after evaluating three other methods for combining databases (see supplementary files for the other methods and the evaluation). Fisher's method showed the best overall results both in AuC and MAP.
In the analysis we will benchmark our system against STRING. STRING is a web tool that has been intensively optimized and updated since 2000. It enables downloading of previous releases. STRING has the same approach for predicting PPIs, *e.g.* it defines evidence scores for several databases and combines them into a single score.

Example 1: Predicting proteins interacting with Dysferlin (*DYSF*, MIM: 603009)
Dysferlin is a 230 kDa C2-domain containing transmembrane protein. Dysferlin is highly expressed in skeletal muscle, but is also found in other tissues such as kidney, heart and monocytes. Mutations in dysferlin cause progressive muscular dystrophies like Limb Girdle Muscular Dystrophy type 2B (MIM: 253601), Miyoshi Myopathy (MIM: 254130) and Distal Anterior Compartment Myopathy (MIM: 606768), collectively referred to as dysferlinopathies [21]. From cellular studies it is known that dysferlin participates in membrane repair. Cultured

myotubes show a calcium-dependent accumulation of dysferlin at sites of membrane damage upon laser-inflicated membrane wounding [22]. In absence of calcium or dysferlin the muscle fiber cannot repair the damage, and undergoes necrosis [22].
We performed a high-throughput screen for proteins interacting with dysferlin and evaluated whether our PPI prediction algorithm could predict dysferlin's experimentally identified interaction partners. To date, nine physical interaction partners were described in literature, and all are believed to aid dysferlin in its membrane repair function. However, it is not completely understood how dysferlin functions, and possibly it does more than membrane repair alone.
We have developed a specific, robust and reproducible immunoprecipitation (IP) method to isolate dysferlin protein complexes from biological sources ([23], de Morrée et al in preparation). We *in vitro* differentiated mouse myoblasts to spontaneously contracting myotubes, and immunoprecipitated dysferlin protein complexes. Mass spectrometry analysis yielded a list of 352 putative interaction partners (manuscript in preparation), including the previously described *ANXA2, AHNAK, CAPN3, TRIM72* encoded proteins, underlining the validity of the method. The proteins already known to interact with dysferlin (recorded in a database) were omitted from this IP list. We created a prioritized list of 25,036 proteins with our Fisher combiner, by matching dysferlin against all other proteins known in our ontology, and compared the IP dataset with this list. Figure 1a shows that text-mining yields a high AuC of 0.78, indicating that implicit information contained in the literature is able to correctly predict interaction partners for dysferlin. As shown in figure 1a the other nine databases yield AuC's between 0.6 and 0.7, and as a result the Fisher combiner AuC does not differ much from text-mining alone. Thus, most predictive value is contained in text and to a lesser extent in gene expression and Gene Ontology. STRING gives an AuC of 0.63, close to random behavior, confirming that our system performs better than STRING. The MAP reflects how many IP partners are present in lists of predicted proteins, a useful measure for those interested in validation of candidates. In figure 1b the MAP are plotted for the IP partners. The MAP achieved by the AuC weighted Fisher combiner was 1.74 fold better than STRING's. Again, literature had the highest predictive value, and the addition of other databases to the prediction led to only small improvement in precision. Finally, we evaluated how many dysferlin interaction partners from the IP list were found in the top 50 of predicted interaction partners. As shown in table 4, the Fisher combiner yields 9 hits, whereas STRING finds only 6. The top 50 of predicted proteins are given in Supplementary table 6.

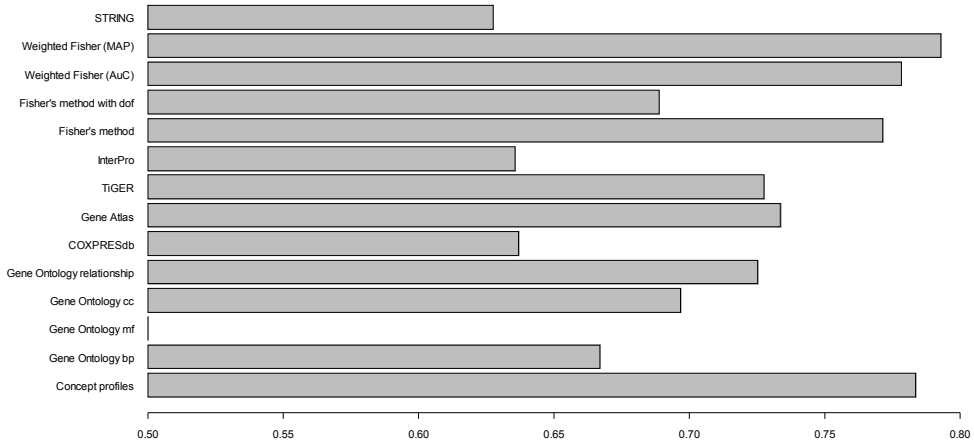


Figure 1a. AuC values (ranging from 0.5 till 1) for the individual databases, the Fisher methods, and STRING, for the dysferlin case study.

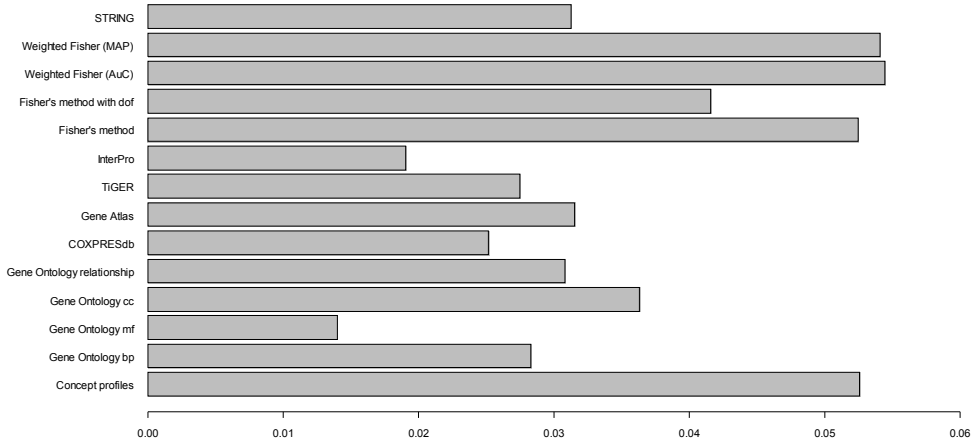


Figure 1b. MAP values for the individual date sources, the Fisher methods, and STRING, for the dysferlin case study.

Example 2: Predicting proteins interacting with Huntingtin (*HTT*, MIM: 613004)

Huntington's disease (HD, MIM: 143100) is a progressive autosomal dominant neurodegenerative disorder that is caused by a CAG repeat expansion in the *HTT*

gene, which results in an expansion of polyglutamines at the N-terminal end of the huntingtin protein, and the accumulation of cytoplasmic and nuclear aggregates in neurons. The polyglutamine expansion in the protein plays a central role in the disease and the size of this expansion has a direct link to the aggregation-proneness as well as the severity of pathology and clinical features [24]. When the mutation for HD was found, huntingtin was a protein of unknown function but extensive research over the past decade has revealed numerous functions for huntingtin and many cellular processes are affected in HD, such as transcriptional de-regulation, mitochondrial dysfunction, and vesicle transport dysfunction [25]. Although the precise underlying disease mechanism of HD is still unknown there is evidence to support a role for aberrant protein-protein interactions in HD pathogenesis [26].

A recent study by Kaltenbach *et al*. [7] identified a comprehensive set of huntingtin-interacting proteins. (1) With yeast two-hybrid screening (Y2H) 104 interacting proteins were identified and (2) affinity pull down followed by mass spectrometry identified 130 proteins. Subsequently, Kaltenbach *et al*. tested if the interacting proteins they had identified could influence mutant huntingtin toxicity. (3) An arbitrary sample of 60, out of the 234, proteins were tested in either over-expressing or partial loss of function *Drosophila* strains expressing the first 336 amino acids of the huntingtin protein containing an expanded 128 glutamines.

For the current study, the already known interacting proteins were omitted from these three datasets to serve as a test panel to examine if our framework can predict proteins from these lists, leaving 92 proteins from the Y2H experiment, 120 from the pull down experiments, and 42 from the Drosophila huntingtin-induced neurodegeneration. With our Fisher method (figure 2b), we obtained a MAP of 0.025 for the Drosophila interaction partners. This is a 3.09 fold increase compared to STRING. The Y2H, and IP experiments showed a 1.48, and 2.56 fold increase over the STRING method respectively. The top 50 of predicted proteins out of 25,036 proteins, are shown in table 3.

From the top 50 proteins identified by our system, 3 proteins namely syntaxin 1A (*STX1A* encoded protein), catenin beta 1 (*CTNNB1* encoded protein) and adaptor-related protein complex 2 (*AP2A1* encoded protein) were in the group of 60 that we-re tested in the *Drosophila* model (compared to 0 by STRING, table 1), and all three were confirmed to modify phenotype, validating that these PPIs are functional.

The interaction between huntingtin and syntaxin 1A has been proposed previously (PMID: 16162412) but the direct interaction between catenin beta 1 and huntingtin was a novel prediction in the Kaltenbach paper that was also high in our list (rank 16) of potential interacting proteins. This protein shows no co-occurrences in MEDLINE abstracts with the huntingtin protein (also not in STRING), but it has been reported in some papers that beta catenin overexpression protects cells from poly(Q) toxicity (PMID: 12097329). AP2A1 is part of the adaptor protein 2 (AP-2)

complex found in clathrin coated vesicles . Although AP2A1 has never been associated with HD previously, the AP-2 complex is involved in the clathrin mediated endocytosis of GABA(A) receptors (PMID: 17690529) and GABA(A) receptors are present on the class of striatal GABAergic neurons that are affected in Huntington's disease[27].

There are 5 proteins out of the top 10 predicted interacting partners for huntingtin that are new potential huntingtin-interacting proteins:

(1) **Platelet-activating factor acetylhydrolase 1b**, regulatory subunit 1 (*PAFAH1B1* encoded protein) inactivates Platelet-Activating Factor (PAF) by removing the acetyl group at the sn-2 position. It is required for induction of nuclear movement and control of microtubule organization[28]. *PAFAH1B1* is also known as *LIS1* [29]and deletions in *LIS1* cause Lissencephaly, a disorder of neuronal migration[30]. A possible link to HD might lie in the fact that PAF induces Clathrin-Mediated Endocytosis [31], which is a common pathway used by G protein-linked receptors to transduce extracellular signals. Both huntingtin interacting protein 1 (*HIP1* encoded protein) and huntingtin interacting protein 1 related (*HIP1R* encoded protein) have been implicated in this process (see below).

(2) **Adenomatous polyposis coli protein** (Protein APC or FPC, ranks position 7 in table 3) is a tumor suppressor protein that acts as an antagonist of the Wnt signaling pathway and has a role in regulating microtubules and actin in polarized epithelia [32]. The *APC* gene is highly expressed in the embryonic and postnatal developing brain. In addition, APC is present in astrocytes, although its role in astrocytes is, as yet, unknown [33].

(3) **Metabotropic glutamate receptor 3** (*GRM3* encoded protein) is an interesting protein because it has been implicated in Huntington's Disease (contributes 22.21% to the concept profile score, PMID: 9600992) while there was no evidence found in STRING (http://string.embl.de/). There is convincing evidence showing that glutamate-mediated excitotoxicity plays a role in HD pathology [34, 35] but there have been no reports to our knowledge directly implicating mGluR3 in HD.

(4) **Vesicle-associated membrane protein-associated protein B** (*VAPB* encoded protein) is a protein that plays an important role in protein folding [36]. To function efficiently, the endoplasmic reticulum relies on a system that detects a buildup of unfolded or misfolded proteins. The cell's response to prevent or correct this buildup is called the unfolded protein response. VAPB is implicated in the autosomal dominant adult-onset form of Amyotrophic Lateral Sclerosis 8 (*ALS8* encoded protein) and in this disease cytosolic aggregates were present in all cell types examined, including mouse and human nonneuronal cells[37]. Protein aggregates can

impair the ability of cells to function normally and huntingtin aggregates are a hallmark of HD[38].

(5) **The GABA(A) receptor-associated protein** (*GABARAP* encoded protein) protein clusters neurotransmitter receptors by mediating interaction with the cytoskeleton[39]. Although there were no co-occurrences for GABARAP and STRING did not find any functional links between GABARAP and huntingtin, it is highly likely that this protein is involved in HD, since GABA(A) receptors are present on the class of striatal GABAergic neurons that are affected in Huntington's disease[27].

Of the top 10 predicted interacting partners for huntingtin, there are 5 proteins that have been identified previously:

(1) **Syntaxin1A** (*STX1A* encoded protein) was identified by Kaltenbach *et al.* and when tested in an HD fruitfly model, STX1A influenced the phenotype [7]. Previous studies have shown that huntingtin enhances calcium influx by blocking STX1A inhibition of N-type calcium channels[40, 41].

(2) **Solute carrier family 1** (glial high affinity glutamate transporter) member 2 (*SLC1A2* encoded protein) was also identified by Kaltenbach *et al.* SLC1A2 is also called glutamate transporter 1 (GLT1). It is a membrane-bound protein that is the principal transporter clearing the excitatory neurotransmitter glutamate from the extracellular space at synapses in the central nervous system and was found to be increased in HD[42, 43].

(3) **Microtubule-associated protein tau** (*MAPT* encoded protein) promotes microtubule assembly and stability, might be involved in the establishment and maintenance of neuronal polarity. Tau is involved in several neurodegenerative disorders such as Alzheimer's disease (AD) and although AD and HD are both protein aggregation disorders, Tau has never been documented to interact with mutant huntingtin. However, it was recently suggested that the level of tau phosphorylation could limit the severity and/or progression of HD[44]. The tau protein in most cases could not be detected by our text-mining algorithm or by STRING resulting in no co-occurring hits with huntingtin. However this problem is solved by intermediate concepts that relate huntingtin with tau (Neurodegenerative Disorders, and Nerve Degeneration).

(4) **Dopamine receptor D2** (*DRD2* encoded protein) is a G-protein-coupled receptor that inhibits adenylyl cyclase activity. In HD there is a major loss of DRD2 binding in the caudate nucleus, putamen and globus pallidus externus[45].

(5) **Huntingtin interacting protein 1 related** (*HIP1R* encoded protein) has a role in clathrin-mediated endocytosis (CME)[46]. It binds to huntingtin interacting protein 1 (*HIP1* encoded protein) and links actin to clathrin[47].

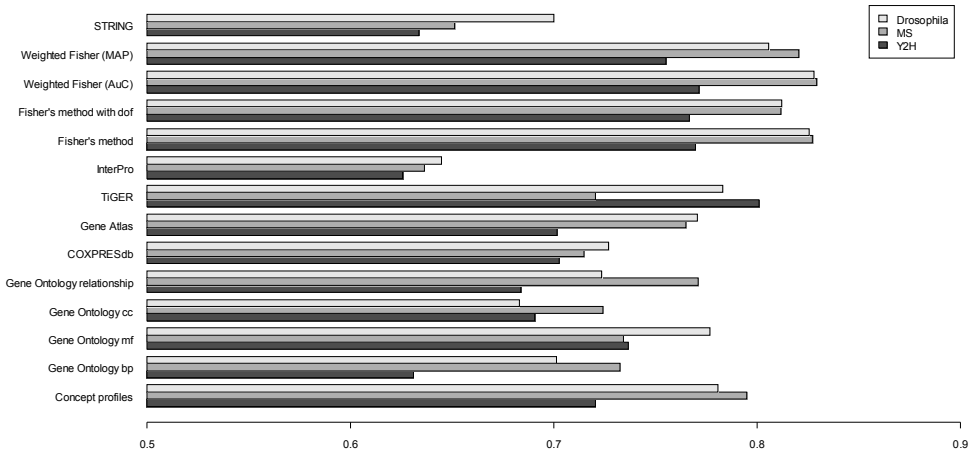


Figure 2a. AuC results for the huntingtin case study.

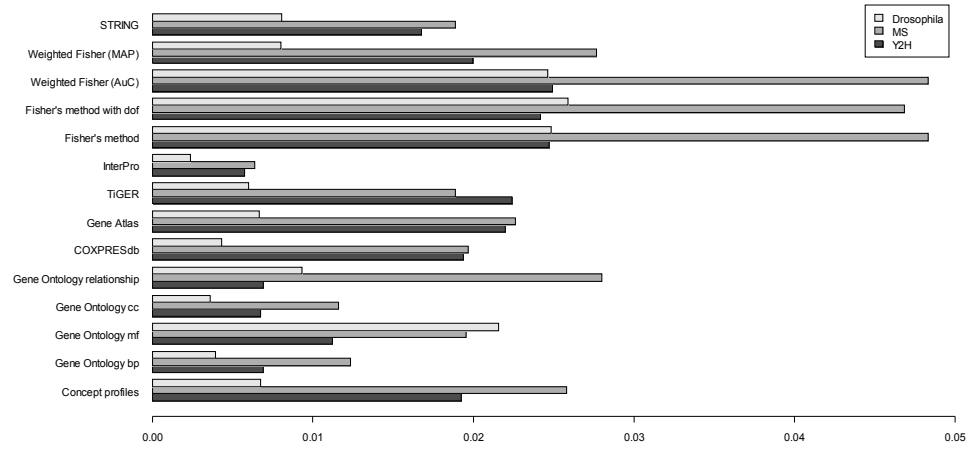


Figure 2b. MAP results for the huntingtin case study.

Showcase 3: polycystic kidney disease 1 (*PKD1*, MIM: 601313). Filtering by feature selection and solving homonyms

In specific cases, certain databases may add noise instead of valuable information. We evaluated a ranked list for the *PKD1* gene that causes polycystic kidney disease 1. The extracellular part of *PKD1* encoded protein contains many domains important for physical interactions with other proteins. The protein domain

information therefore dominated the prediction of PKD's interactions partners. This effect was undesired and therefore the InterPro/DOMINE score was left out.
The prediction of interaction partners for PKD1 by literature analysis alone was also not ideal. Although the literature remains the biggest information source, it is also the information source which requires the most preprocessing. Text-mining on its own is a challenging field of research with involves many steps such as extracting public articles, defining an ontology containing concepts and their synonyms, and disambiguating words in text using concept recognition software. Disambiguation is the process of mapping a word in text to a unique concept and labels it with a unique identifier. A term is considered to be ambiguous if it has multiple meanings. We investigated this homonym problem for PKD1. The first homonym problem is that the name 'polycystic kidney disease 1' itself can refer to the gene or the disease. When only concept profiles were used the top of most associated proteins with PKD1 showed six proteins that ranked high due to homonyms. Two proteins, protein kinase D1 (*PRKD1* encoded protein) and ectonucleotide pyrophosphatase/ phosphodiesterase 1 (*ENPP1* encoded protein), were caused by direct homonym problems. In literature, PRKD1 is also referred to as PKD1. ENPP1 has a synonym PC1 that is also used as a synonym for PKD1. The other four proteins had synonym problems in the overlapping concepts of their concept profiles. These can be seen as indirect homonym problems. In literature protein kinase D2(*PRKD2* encoded protein) is referred to as polycystic kidney disease 2 (*PKD2* encoded protein) which has a close relationship with PKD1. Protein kinase D3 (*PRKD3* encoded protein) is referred to as protein kinase C and has many relationships with PRKD1 in literature. The same holds for protein kinase C substrate 80K-H (*PRKCSH* encoded protein) which is referred to as protein kinase C substrate. Phosphoglycolate phosphatase (*PGP* encoded protein) is referred to as PRKD1 which on itself causes homonym problems. When the concept profiles are used in combination with expression data these homonyms can be suppressed.
For PKD1 we generated a ranked list while omitting the InterPro domain information from the prediction. We calculated the match score based on Fisher's method and checked if mentioned homonyms were suppressed since these proteins are not likely to be co-expressed with PKD1. The last column in table 2 shows that mentioned proteins with homonym problems indeed had much lower rankings than in the prediction based on literature only. Further manual curation by an expert showed that Fisher's method gives better associations with PKD1 in the top predictions when concept profiles are used in combination with microarray expression data and eliminating the InterPro domain information. In practice an expert should be able to choose which databases are being combined for the best prediction.

Table 1. Ranks of the homonyms associated with PKD1. The first rank is based on concept profiles, showing that the homonyms rank high. Fisher's methods suppressed these homonyms and the rank becomes lower

Gene symbol	Gene name	Rank Concept profiles	Rank Fisher's method
PRKD1	Serine/threonine-protein kinase D1	2	46
PRKD2	Serine/threonine-protein kinase D2	4	164
PRKD3	Serine/threonine-protein kinase D3	7	328
ENPP1	Ectonucleotide pyrophosphatase/phosphodiesterase family member 1	15	258
PRKCSH	Glucosidase 2 subunit beta	30	283
PGP	Phosphoglycolate phosphatase	36	1983

Concluding remarks

In this study we have shown that combining information from the biomedical literature and from different databases using Fisher's method significantly improves the prediction of novel protein interactions compared to previously applied methods. We evaluated three case studies on dysferlin, huntingtin, and polycystin-1 and predicted proteins previously not recorded in any protein interaction database. For huntingtin, besides the literature, other databases like Gene Atlas and The Gene Ontology contributed to the matchscore. An evaluation of the top 10 predicted huntingtin interacting proteins showed 5 proteins known to be associated with huntingtin. The other 5 were novel ones that have been curated and are potential interaction partners with huntingtin. From these top 10 proteins 5 could not be detected with a MEDLINE query, indicating that implicit knowledge extraction is possible.

For dysferlin we showed that the literature remains the biggest information source and that the other databases to a lesser extent contribute to the match score. Although for dysferlin the contribution of other databases to the literature alone seems low, the aid of other databases has been shown to be useful in solving homonym problems. This was shown in the PKD1 study. PKD1 showed 6 proteins that were caused by homonyms and these were suppressed when the concept profiles were combined with other databases. Thus the combination of literature and non-textual information makes our algorithm more robust.

Fisher's Method is a simple and robust method to combine several databases. In addition its interpretation is very intuitive. For every database you first define a p-value for a sample that needs to be evaluated. Fisher's methods then tells if the

combination of individual p-values (taken from different databases) for that sample is significant.
We made a list available of Fisher match scores between every two proteins in our ontology. The list can be downloaded from www.biosemantics.org/ppi-prediction.

Acknowledgements
This project was supported by the Biorange project SP 3.5.1 of the Netherlands Bioinformatics Center and the Center for Medical Systems Biology, both financed by the Netherlands Genome Initiative, and by the Dutch Prinses Beatrix Fonds.

Table 2. Prediction in the top 50

Hungtintin	Fisher fixed dof	Fisher variable dof	Weighted Fisher AuC	Weighted Fisher MAP	STRING
Y2H	2	3	2	0	0
IP	3	3	3	1	0
Drosophila	3	3	3	2	0
Dysferlin					
IP	9	9	9	6	6

Table 3. Top 50 for huntingtin predicted interacting partners

rank	name	Y2H	MS	Drosophila	PPI	cooccurrences
1	HTT	x	x			1131
2	STX1A		x			2
3	SLC1A2		x			2
4	PAFAH1B1					0
5	GABARAP					0
6	MAPT					0
7	FPC					0
8	DRD2					9
9	GRM3					0
10	VAPB					0
11	HIP1R				x	4
12	HIP1	x			x	23
13	KIF5B					0
14	MAPRE1					0
15	GSK3B					2
16	CTNNB1	x		x		0
17	ATN1					19
18	STX6					0
19	CLASP1	x				0
20	BID					0
21	TMED10					0
22	KIF1B					0
23	CDK5					5
24	NTRK2					0
25	HIPK2					0
26	MAP1S					0
27	AP2A1			x		0
28	CLASP2					0
29	RAE1					0
30	BBS4					0
31	GIPC1					0
32	PACSIN1	x		x	x	3
33	AKT1				x	2
34	KLC1					0
35	SYT1		x			0
36	NRCAM					0
37	ATXN1					4
38	BCL2L11					2
39	RAB3A					1
40	CDK5R1					1
41	ULK1					0
42	HIF1A					0
43	DIAPH1					0
44	SNCA					18
45	SOD1					5
46	YKT6					0
47	BDNF					38
48	AP2A2	x		x	x	4
49	TPPP					0
50	DYNC1I1					0

References

1. Swanson, D.R., *Fish oil, Raynaud's syndrome, and undiscovered public knowledge.* Perspect Biol Med, 1986. **30**(1): p. 7-18.
2. van Haagen, H.H.H.B.M., t Hoen, P.A.C., Botelho Bovo, A., de MorrÃ©e, A., van Mulligen, E.M., et al., *Novel Protein-Protein Interactions Inferred from Literature Context.* PLoS ONE, 2009. **4**(11): p. e7894.

3. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D., *GeneCards: integrating information about genes, proteins and diseases.* Trends Genet, 1997. **13**(4): p. 163.
4. Xia, K., Dong, D., and Han, J.D., *IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model.* BMC Bioinformatics, 2006. **7**: p. 508.
5. Alexeyenko, A. and Sonnhammer, E.L., *Global networks of functional coupling in eukaryotes from comprehensive data integration.* Genome Res, 2009. **19**(6): p. 1107-16.
6. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., et al., *STRING 8--a global view on proteins and their functional interactions in 630 organisms.* Nucleic Acids Res, 2009. **37**(Database issue): p. D412-6.
7. Kaltenbach, L.S., Romero, E., Becklin, R.R., Chettier, R., Bell, R., et al., *Huntingtin interacting proteins are genetic modifiers of neurodegeneration.* PLoS Genet, 2007. **3**(5): p. e82.
8. Schuemie, M.J., Jelier, R., and Kors, J.A. *Peregrine: Lightweight gene name normalization by dictionary lookup*. in *Biocrative 2 workshop*. 2007. Madrid.
9. Jelier, R., Schuemie, M.J., Roes, P.J., van Mulligen, E.M., and Kors, J.A., *Literature-based concept profiles for gene annotation: the issue of weighting.* Int J Med Inform, 2008. **77**(5): p. 354-62.
10. Mistry, M. and Pavlidis, P., *Gene Ontology term overlap as a measure of gene functional similarity.* BMC Bioinformatics, 2008. **9**: p. 327.
11. Philip, R., *Using information content to evaluate semantic similarity in a taxonomy*, in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*. 1995, Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada.
12. Dekang, L., *An Information-Theoretic Definition of Similarity*, in *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998, Morgan Kaufmann Publishers Inc.
13. Jiang, J.J. and Conrath, D.W. *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. in *International Conference Research on Computational Linguistics (ROCLING X)*. 1997.
14. Obayashi, T., Hayashi, S., Shibaoka, M., Saeki, M., Ohta, H., et al., *COXPRESdb: a database of coexpressed gene networks in mammals.* Nucleic Acids Res, 2008. **36**(Database issue): p. D77-82.
15. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.

16. Liu, X., Yu, X., Zack, D.J., Zhu, H., and Qian, J., *TiGER: a database for tissue-specific gene expression and regulation.* BMC Bioinformatics, 2008. **9**: p. 271.
17. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., et al., *InterPro: an integrated documentation resource for protein families, domains and functional sites.* Brief Bioinform, 2002. **3**(3): p. 225-35.
18. Raghavachari, B., Tasneem, A., Przytycka, T.M., and Jothi, R., *DOMINE: a database of protein domain interactions.* Nucleic Acids Res, 2008. **36**(Database issue): p. D656-61.
19. Ben-Hur, A. and Noble, W.S., *Choosing negative examples for the prediction of protein-protein interactions.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S2.
20. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., et al., *The IntAct molecular interaction database in 2010.* Nucleic Acids Res, 2009.
21. Laval, S.H. and Bushby, K.M., *Limb-girdle muscular dystrophies--from genetics to molecular pathology.* Neuropathol Appl Neurobiol, 2004. **30**(2): p. 91-105.
22. Bansal, D., Miyake, K., Vogel, S.S., Groh, S., Chen, C.C., et al., *Defective membrane repair in dysferlin-deficient muscular dystrophy.* Nature, 2003. **423**(6936): p. 168-72.
23. Huang, Y., Verheesen, P., Roussis, A., Frankhuizen, W., Ginjaar, I., et al., *Protein studies in dysferlinopathy patients using llama-derived antibody fragments selected by phage display.* Eur J Hum Genet, 2005. **13**(6): p. 721-30.
24. Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., et al., *The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease.* Nat Genet, 1993. **4**(4): p. 398-403.
25. Landles, C. and Bates, G.P., *Huntingtin and the molecular pathogenesis of Huntington's disease. Fourth in molecular medicine review series.* EMBO Rep, 2004. **5**(10): p. 958-63.
26. Harjes, P. and Wanker, E.E., *The hunt for huntingtin function: interaction partners tell many different stories.* Trends Biochem Sci, 2003. **28**(8): p. 425-33.
27. Waldvogel, H.J., Kubota, Y., Fritschy, J., Mohler, H., and Faull, R.L., *Regional and cellular localisation of GABA(A) receptor subunits in the human basal ganglia: An autoradiographic and immunohistochemical study.* J Comp Neurol, 1999. **415**(3): p. 313-340.
28. Arai, H., Koizumi, H., Aoki, J., and Inoue, K., *Platelet-activating factor acetylhydrolase (PAF-AH).* J Biochem, 2002. **131**(5): p. 635-640.

29. Hattori, M., Adachi, H., Tsujimoto, M., Arai, H., and Inoue, K., *Miller-Dieker lissencephaly gene encodes a subunit of brain platelet-activating factor acetylhydrolase [corrected].* Nature, 1994. **370**(6486): p. 216-218.
30. Reiner, O., Bar-Am, I., Sapir, T., Shmueli, O., Carrozzo, R., et al., *LIS2, gene and pseudogene, homologous to LIS1 (lissencephaly 1), located on the short and long arms of chromosome 2.* Genomics, 1995. **30**(2): p. 251-256.
31. McLaughlin, N.J.D., Banerjee, A., Kelher, M.R., Gamboni-Robertson, F., Hamiel, C., et al., *Platelet-Activating Factor-Induced Clathrin-Mediated Endocytosis Requires beta-Arrestin-1 Recruitment and Activation of the p38 MAPK Signalosome at the Plasma Membrane for Actin Bundle Formation.* The Journal of Immunology, 2006. **176**(11): p. 7039-7050.
32. Caldwell, C.M. and Kaplan, K.B., *The role of APC in mitosis and in chromosome instability.* Adv.Exp.Med Biol, 2009. **656**: p. 51-64.
33. Senda, T., Shimomura, A., and Iizuka-Kogo, A., *Adenomatous polyposis coli (Apc) tumor suppressor gene as a multifunctional gene.* Anat.Sci Int., 2005. **80**(3): p. 121-131.
34. Grunewald, T. and Beal, M.F., *Bioenergetics in Huntington's disease.* Ann.N.Y.Acad.Sci, 1999. **893**: p. 203-213.
35. Cicchetti, F., Prensa, L., Wu, Y., and Parent, A., *Chemical anatomy of striatal interneurons in normal individuals and in patients with Huntington's disease.* Brain Research Reviews, 2000. **34**(1-2): p. 80-101.
36. Kanekura, K., Suzuki, H., Aiso, S., and Matsuoka, M., *ER stress and unfolded protein response in amyotrophic lateral sclerosis.* Mol Neurobiol., 2009. **39**(2): p. 81-89.
37. Teuling, E., Ahmed, S., Haasdijk, E., Demmers, J., Steinmetz, M.O., et al., *Motor Neuron Disease-Associated Mutant Vesicle-Associated Membrane Protein-Associated Protein (VAP) B Recruits Wild-Type VAPs into Endoplasmic Reticulum-Derived Tubular Aggregates.* Journal of Neuroscience, 2007. **27**(36): p. 9801-9815.
38. Stefani, M. and Dobson, C.M., *Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution.* J Mol Med, 2003.
39. Wang, H., Bedford, F.K., Brandon, N.J., Moss, S.J., and Olsen, R.W., *GABAA-receptor-associated protein links GABAA receptors and the cytoskeleton.* Nature, 1999. **397**(6714): p. 69-72.
40. Romero, E., Cha, G.H., Verstreken, P., Ly, C.V., Hughes, R.E., et al., *Suppression of Neurodegeneration and Increased Neurotransmission Caused by Expanded Full-Length Huntingtin Accumulating in the Cytoplasm.* Neuron, 2008. **57**(1): p. 27-40.

41. Swayne, L.A., Chen, L., Hameed, S., Barr, W., Charlesworth, E., et al., *Crosstalk between huntingtin and syntaxin 1A regulates N-type calcium channels.* Molecular and Cellular Neuroscience, 2005. **30**(3): p. 339-351.
42. Miller, B.R., Dorner, J.L., Shou, M., Sari, Y., Barton, S.J., et al., *Up-regulation of GLT1 expression increases glutamate uptake and attenuates the Huntington's disease phenotype in the R6/2 mouse.* Neuroscience, 2008. **153**(1): p. 329-337.
43. Arzberger, T., Krampfl, K., Leimgruber, S., and Weindl, A., *Changes of NMDA receptor subunit (NR1, NR2B) and glutamate transporter (GLT1) mRNA expression in Huntington's disease--an in situ hybridization study.* J Neuropathol.Exp.Neurol, 1997. **56**(4): p. 440-454.
44. Caparros-Lefebvre, D., Kerdraon, O., Devos, D., Dhaenens, C.M., Blum, D., et al., *Association of corticobasal degeneration and Huntington's disease: Can Tau aggregates protect Huntingtin toxicity?* Movement Disorders, 2009. **24**(7): p. 1089-1090.
45. Glass, M., Dragunow, M., and Faull, R.L., *The pattern of neurodegeneration in Huntington's disease: a comparative study of cannabinoid, dopamine, adenosine and GABA(A) receptor alterations in the human basal ganglia in Huntington's disease.* Neuroscience, 2000. **97**(3): p. 505-519.
46. Gottfried, I., Ehrlich, M., and Ashery, U., *The Sla2p/HIP1/HIP1R family: similar structure, similar function in endocytosis?* Biochem Soc Trans. **38**(Pt 1): p. 187-191.
47. Engqvist-Goldstein, A.E., Kessels, M.M., Chopra, V.S., Hayden, M.R., and Drubin, D.G., *An actin-binding protein of the Sla2/Huntingtin interacting protein 1 family is a novel component of clathrin-coated pits and vesicles.* J Cell Biol, 1999. **147**(7): p. 1503-1518.

Supplementary information

S1 Databases for information extraction

Table 1 shows the databases that are used in our analysis and the date of download.

Table 1. Databases that are combined and their date of download.

Database	Date of download
Concept profiles	May 2009
Gene ontology	July 23, 2009
Gene Atlas	April, 2004
Coxpresdb	April 17, 2008
TiGER	February 19, 2009
InterPro	July 22, 2009
DOMINE	April 16, 2007
STRING version 8.1	October 18, 2009

S2 Curated Protein-Protein interaction databases

For training, testing and optimizing our system we constructed a set of known human protein-protein interactions (PPIs) taken from public, curated databases. We called this set of known PPIs the positive set. The databases used were Biogrid[1], DIP[2], HPRD[3], Mint[4], Reactome[5], and Uniprot/Swiss-Prot[6]. Table 2 shows the date of download for these databases. If a PPI was mentioned in one of these databases, we assumed it to be a true PPI. There is a level of redundancy between these databases meaning that some protein-protein interaction pairs occur in multiple databases, which is a good indication that it is a true PPI. These protein pairs count only once. We restricted our analysis to human proteins only. The resulting positive set contains 83,930 PPIs.

A negative set was constructed as described in the materials and method section of the paper. The negative set is the same as the null distribution used for the Fisher Method and has a size of 100.000 samples.

Table 2. Protein databases used for the positive set and their date of download.

Protein database	Data of download
BioGrid	July 1, 2009
DIP	October 15, 2008
HPRD	July 6, 2009
IntAct	July 11, 2009
MINT	July 23, 2009
Reactome	June 11, 2009
UniProt	June 17, 2009

S3 Cross validation

All performance measures (AuC and MAP) were calculated in a 5-fold cross validation loop. First the data consisting of positive and negative instances (*e.g* PPIs and random protein pairs) were splitted in five equally sized parts. Then at each iteration four parts were used for straining classifiers and combination methods and the remaining fifth part was used for testing. This was repeated until each part was used once for testing.

S4: Coverage and prediction accuracy of individual databases

In the analyses that follow we first defined a positive set that consists of protein-protein interactions recorded in six curated databases (see supplement S2), and a negative set of probable non-interacting random protein pairs. We evaluated how well each database covers samples from the positive and negative set. Table 3 shows the coverage for each individual database, the combination of databases and STRING. A protein pair is covered if at least one on the individual databases has a match score for that protein pair. Our combined databases cover almost the complete positive set. The coverage is similar to STRING.

To evaluate prediction performance for PPIs, we used the AuC and MAP criteria. A third measure is used to reflect the predictions made in the top of a ranked list. It counts the number of predicted true positives when the number of predicted false positives is fixed to 50. We refer to this measure as ROC50.

For each database the AuC and MAP were calculated and the results are given in table 4. The AuC and MAP were calculated in a 5-fold cross validation loop (see S3). Figure 1 shows the ROC50. Here it is shown that the concept profiles (cp) has the highest number of true positives (over 6,000). The STRING curated database score (dscore) also gives a performance of over 6,000 predicted true positives. This result is expected since this score is based on curated protein databases, some of which were also used to create our evaluation set. The Gene Ontology gives an overall best performance with a high AuC and high coverage.

Table 3. Coverage for each database and the databases combined. As a benchmark the coverage of STRING is given.

Database	Positive set (%)	Negative set (%)
Concept profiles	84	24
Gene Ontology biological process	94	34
Gene Ontology molecular function	95	39
Gene Ontology cellular component	95	43
Gene Ontology relationships	99	52
COXPRESdb	95	51
Gene Atlas	69	12
TiGER	72	25
InterPro/DOMINE	95	49
Combined system	99.97	67
All STRING databases	99.13	62

Table 4. AuC and MAP for the individual databases for a 5-fold cross validation. The standard errors are not shown because they were negligible small. The Gene Ontology is separated into the three main categories and the relationships.

Database	AuC	MAP
Concept profiles	0.88	0.90
Gene Ontology biological process	0.91	0.90
Gene Ontology molecular function	0.88	0.85
Gene Ontology cellular component	0.89	0.88
Gene Ontology relationships	0.91	0.88
COXPRESdb	0.82	0.79
Gene atlas	0.80	0.81
TiGER	0.78	0.75
InterPro/DOMINE	0.80	0.77
STRING DATABASES		
Neighborhood in the genome	0.69	0.59
Gene fusion	0.69	0.58
Cooccurrences across genomes	0.69	0.59
Coexpression	0.69	0.59
Experimental/biochemical data	0.81	0.82
Association in curated databases	0.82	0.82
Co-mentioned in PubMed abstracts	0.83	0.84

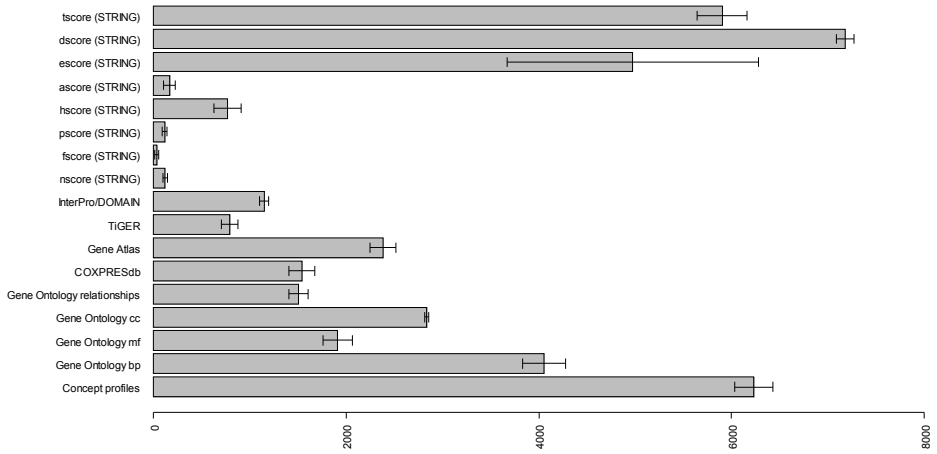


Figure 1. Number of true positives that are retrieved at 50 false positives for each individual database. The concept profiles (cp) retrieves the highest amount of true positives, reflecting that the literature is still the most important source of information. The assocation in curated database score (dscore) for STRING shows the best result as expected. The errorbars are the standard deviation around the mean calculated over 5-fold cross validation.

S5 Different combining techniques

Before we came to our final approach based on Fisher's Method we evaluated four other combining methods described below.

(1) Combining rules by Kuncheva

The first combiner is the one defined by Kuncheva [7]. In total there are five combining rules namely the product, sum, maximum, minimum, and majority vote. The combiners defined by Kuncheva are applied to the output of each classifier trained on a single database; hence this step requires training data. In our case we used a simple logistic regression classifier [8]. Each raw match score defined for a database is converted to a probability value between 0 and 1. The concept profiles score was first log transformed to produce more normal distributed classes. Since we evaluate a two class problem (the class of protein-protein interactions (PPI) and the class of non interacting protein pairs (NIPP)) the probability of the second class can be calculated once the probability of the first one is known. If *p1* is the probability for a sample in class one then *p2=1-p1* is the probability for that

sample in class two. On the output of each classifier the combining rule was applied. The product rule for a sample *x* is defined as follows

$$\mu_j = \prod_{i=1}^{L} p_{i,j}(x)$$

With μ_j the combined probability for class *j* and $p_{i,j}(x)$ the probability of x belonging to class j according to database i, and *L* the total number of databases to combine (our case *L=6*). After the rule is applied, the combined probabilities can be normalized to add up to 1. In the same way the sum, maximum and minimum rule can be defined. Missing values are completely ignored. If one database has a missing value the rule is applied to the remaining databases. If a sample has missing values for all the databases the probabilities are set to 0 and 1 for class one (PPIs) and two respectively.
The advantage is that these combiners do not require training data. The disadvantage is that the classifiers trained on each database in the first step make assumptions about your data, for instance that the classes follow a normal distribution. This could result in false predictions if the assumptions are not true.

(2) Rank combiners
Calculating a rank combiner is similar to the Kuncheva combiners. The same rules such as, product and sum, can be applied to ranks. For instance the rank product is a non-parametric statistic that is often used for gene expression profiling [9]. Here the formula for the rank product is given.

$$RP(x) = \left(\prod_{i=1}^{L} r_{x,i} \right)^{1/L}$$

With $r_{x,i}$ the rank obtained for database *i* for a sample *x*. In the same way the other combiners based on ranks can be derived. *L* are the number of databases with no missing value for that sample. The rank for samples where all values are missing is set to positive infinite. The advantage of this combiner is that it also does not require any training data. Furthermore, it does not put any constraints on the data. The disadvantage is that it is highly sensitive to the presence of poorly performing databases.

(3) Maximum AuC linear classifier (MALC)
Marrocco et. al. [10] describes a method where a linear classifier is trained such that the resulting trained classifier maximizes the AuC. Mainstream classifiers are designed to minimize the classification error, *e.g.* taken into account the false negatives, whereas the MALC is designed to minimize the false positives, *e.g.* maximizing the AuC. We implemented a different version of their algorithm which is both fast and robust. The features (match scores defined for databases) are combined in an iterative manner and at each iteration step two features are

combined and result in a new feature. Step one is to normalize the date between 0 and 1. The inner product score between concept profiles were first log transformed before normalization. Step two is to calculate a Pearson correlation matrix (all pair wise correlations between any two features). The two features with the lowest correlation are combined first. Step three is to apply the linear classifier to the features *h* and *k* which is given as

$$x_{lc} = \alpha x_h + (1-\alpha) x_k$$

where x_{lc} is the weighted sum of the two features, and alpha the weight parameter that needs to be optimized. Step four is to vary the alpha level between 0 and 1 in steps of 0.01 (or any other step size) and calculate the AuC for each alpha. Then choose the alpha level that corresponds with the highest AuC value. Step five is to replace the two features with xlc features and repeat steps two till five until all features are combined to a single feature.

4) Fisher's method

The Fisher method was described in the article. The advantage of this method is that it is robust, simple, and no information is needed on the positive set (PPIs) since it only uses the null distribution (negative set of probable non interacting protein pairs).

S6 Choosing the best combining method

The four different combining methods (and each with a number of variations) are compared which each other using the AuC and MAP as performance criteria in a 5-fold cross validation loop. The results for all combiners are given in table 3. Fisher's method and the MALC show the best results in both MAP and AuC. To further evaluate the accuracy of these combiners we looked at the ROC50 results. That is the number of true positives predicted when the number of false positives was fixed to 50. The results are given in figure 2. Here the Fisher method shows slightly better result than the MALC. Figure 3 shows the ROC curve for the Fisher method with fixed degrees of freedom. The histogram for the positive and negative set is given in figure 4.

Table 5. AuC and MAP for the different combining techniques. The standard errors are not shown.

Kuncheva combining rule	**AuC**	**MAP**
Product	0.94	0.92
Sum	0.91	0.86
Maximum	0.84	0.73
Minimum	0.90	0.86
Majority vote	0.90	0.83
Rank combiners		
Mean	0.94	0.83
Max	0.83	0.46
Min	0.93	0.93
Product	0.95	0.90
Fisher's method		
Fixed number of dof (=9)	0.97	0.97
Fisher with variable dof	0.97	0.96
Weighted Fisher with AuC	0.97	0.97
Weighted Fisher with MAP	0.97	0.97
Fisher +4 features from String	0.97	0.97
Maximize AuC		
MALC	0.97	0.96

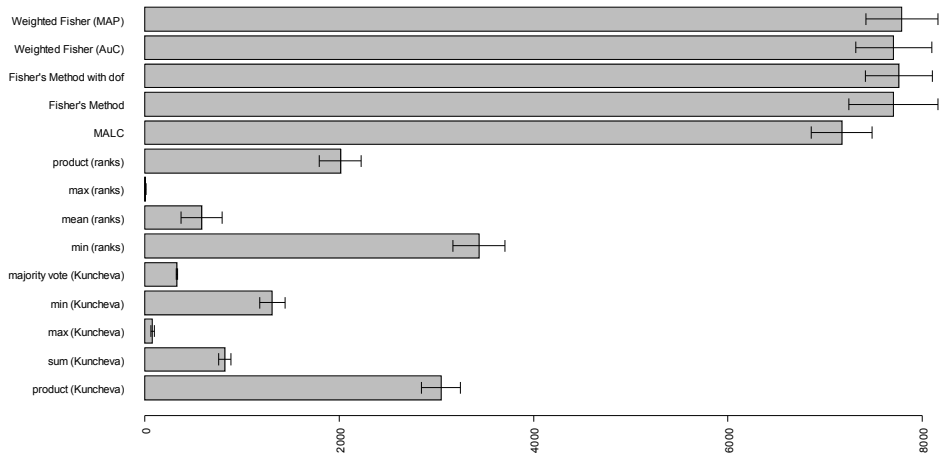


Figure 2. Number of true positives at 50 false positives for the different combination techniques

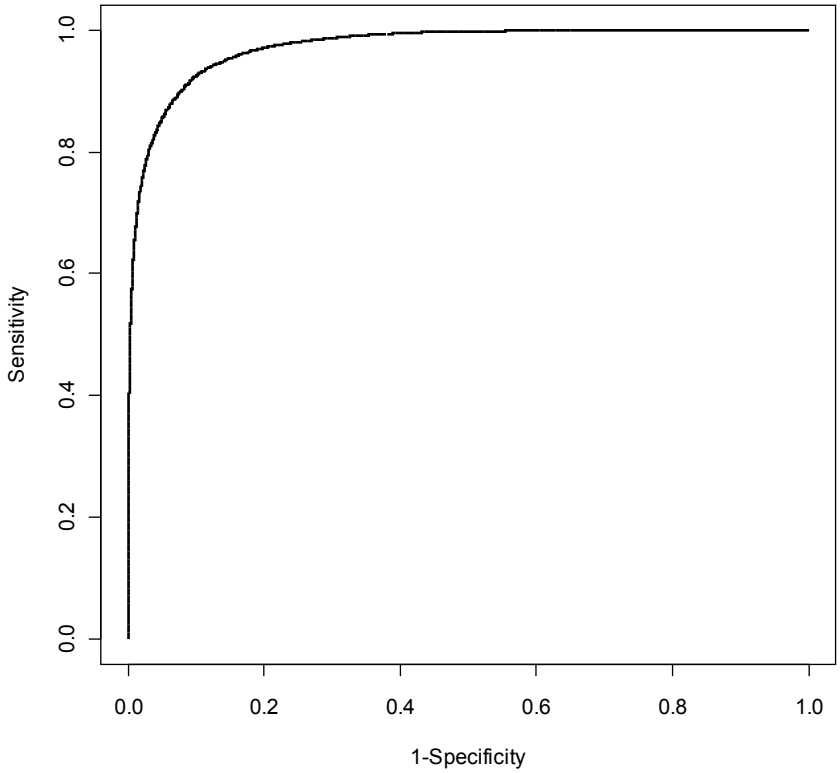


Figure 3. ROC plot of the Fisher method combiner.

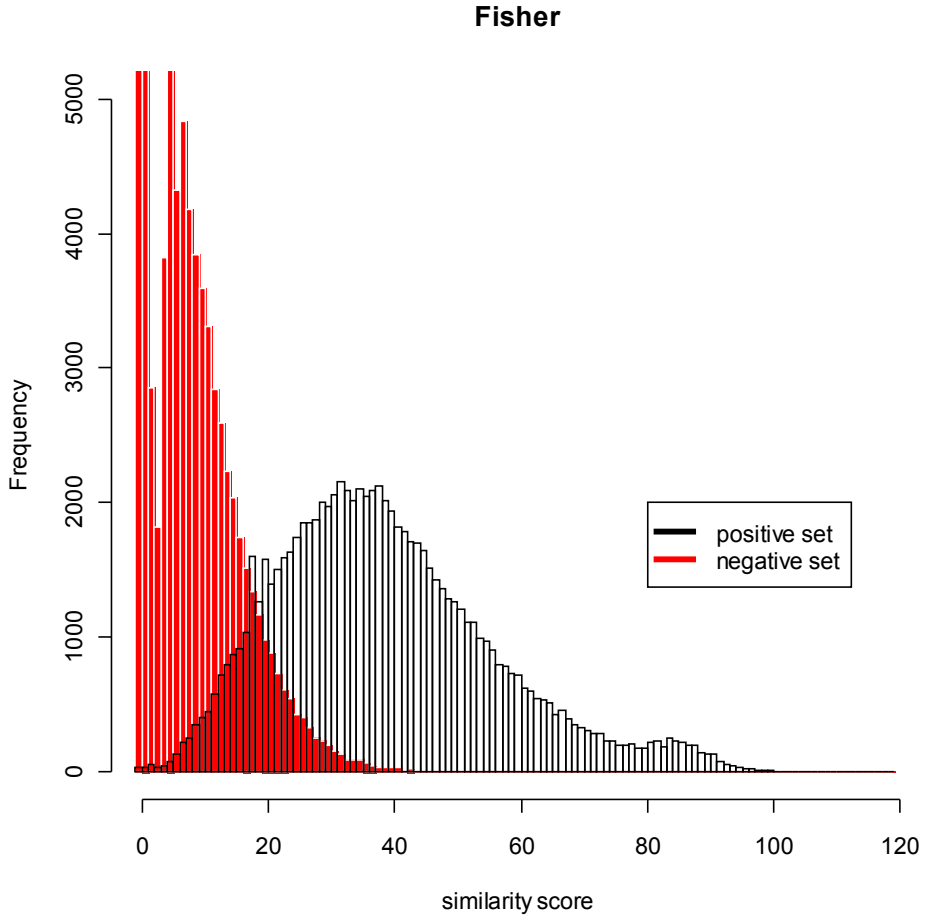


Figure 4. Histogram plot of the positive PPI set and the negative random set for the Fisher combiner.

Table 6. Top 50 of most associated proteins with Dysferlin.

rank	name	Co-occurrences	PPI
1	DYSF	246	
2	MYOF	13	
3	TGFB1	0	
4	RYR2	0	
5	TTN	2	
6	MYH7	1	
7	KCNQ1	0	
8	MYL3	0	
9	TNNC2	0	

10	SNTA1	2	
11	TCAP	13	
12	ADRBK1	0	
13	SGCA	18	
14	TPM1	0	
15	TNNC1	0	
16	MYH4	0	
17	IL1B	0	
18	MYOT	10	
19	C5AR1	0	
20	OTOF	6	
21	SSPN	1	
22	FKBP1B	0	
23	MYH2	0	
24	Cf5	0	
25	ACTN2	1	
26	UTRN	3	
27	TNNT1	0	
28	TNNT3	0	
29	CSF3R	0	
30	CACNA1H	0	
31	HMOX1	0	
32	MYBPC3	0	
33	DES	0	
34	GAA	0	
35	TPP1	0	
36	SNTB1	0	
37	KCNE1	0	
38	ACTA1	0	
39	HCK	0	
40	CAV3	35	X
41	CAPN3	44	X
42	FPR1	0	
43	RYR1	1	
44	KCNMA1	0	
45	MYLK2	0	
46	MYH6	0	
47	TNNI3	0	
48	NCF2	0	
49	NOS3	0	
50	CAV1	0	

1. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.

2. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
3. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
4. Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
5. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., et al., *Reactome knowledgebase of human biological pathways and processes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D619-22.
6. *The Universal Protein Resource (UniProt) 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D169-74.
7. Kuncheva, L.I., *Combining pattern classifiers*. 1 ed. 2004: Wiley-Interscience. 376.
8. Mitchell, T.M., *Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*, in *Machine Learning*. 2005. p. 1-17.
9. Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P., *Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.* FEBS Lett, 2004. **573**(1-3): p. 83-92.
10. Marrocco, C., Duin, R.P.W., and Tortorella, F., *Maximizing the area under the ROC curve by pairwise feature combination.* Pattern Recognition, 2008. **41**(6): p. 1961-1974.

Chapter 5

Finding gene-disease relations using implicit information in the scientific literature

Herman van Haagen[1], Emmelien Aten[1], Peter-Bram 't Hoen[1], Marco Roos[1,2], Tobias Messemaker[?], Erik A. Schultes[1], Barend Mons[1], Gert-Jan van Ommen[1], and Martijn Schuemie[3]

1. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

2. Institute for Informatics, University of Amsterdam, Amsterdam, The Netherlands

3. Biosemantics Group, Erasmus Medical Center, Rotterdam, TheNetherlands

Manuscript in preparation

Abstract
Despite large and ever-growing bioinformatic data sets, there is often no information that explicitly links genes to a disease in literature. Bioinformatic approaches have attempted to circumvent this problem by searching for genes similar to those already known to be associated with a disease [1-4]. However, this approach is frequently not useful because previous associated genes with a disease are not available. Here, we use concept profiles [5, 6], a vector-based description of terms, to discover implied relationships between genes and diseases for which no explicit link (co-occurrence) has been stated in either text or any other database. In a retrospective text mining analysis of scientific literature concept profiles were able to prioritize disease genes on average within the top 13 out of 200 genes located in a specified linkage interval at least one year before the publication of the landmark paper explicitly establishing the gene-disease relationship. Examination of the highly-ranked concepts shared between the gene and the disease in concept profiles was used by biomedical experts to evaluate the plausibility of the inferred relationships and rationalize potential biological mechanisms. By exploiting the implicit information in the literature, concept profiles performed two-fold better in prioritizing genes of polygenic diseases than the Endeavour gene prioritizer [2] using 26 data-mining resources. These results demonstrate the enormous untapped potential of implied information in scientific literature for biomedical discovery, and the application of concept profile technology in extracting new knowledge.

Introduction

Although linkage analysis, association studies, and next generation sequencing technology have produced voluminous amounts of genetic data that are essential for the characterization of disease mechanisms, isolating genes that cause or impact the etiology of a particular disease remains a time consuming and largely serendipitous task. Often, many interrelated factors must be considered. For example, individual genes may cause multiple diseases, distinct diseases may be caused by multiple genes, and different diseases will often have phenotypic overlap. To cope with these inherent complexities and with the size of large and rapidly growing datasets, bioinformatic tools have been developed combining text-mining and data-mining capabilities to automatically search for correlations among, and then prioritize, putative gene-disease pairs[7-14]. For example, the Endeavour web tool combines biomedical ontologies, text, and data from 26 distinct sources to prioritize genes for specific diseases [2]. Many of the prioritizers that integrate multiple data sources are based on so called ‘seed’ genes, which are genes having a known relation to a disease that help to find the next causative gene that results in the same phenotype. For instance, a novel gene for breast cancer may be found by using information about BRCA1 and BRCA2, genes already known to

cause the disease. However, for the majority of diseases recorded in OMIM, a causative gene is not yet known. In these cases, prioritizers based on seed genes do not work. Furthermore, for all those diseases or syndromes where the first gene has yet to be discovered, a prioritizer will be limited to text information only, yet before a landmark paper is published describing the disease causing gene, the disease and the gene tend to have few or no co-occurrence in the same abstract or article. Therefore text-mining systems based on direct co-occurrences will fail to predict the majority gene-disease relationships.

However, woven within the narrative of scientific literature there are a vast network of relations among terms that are to some degree left implicit by the authors. Implicit relations may arise as a consequence of new findings or as part of the scientific rational, and may or may not be intentional. Implicit information may be directly related to the immediate narrative or may have ancillary relations. Here, we used a text-mining method based on concept profiles to prioritize candidate genes by considering this large amount of implicit associative information in text. A concept profile for a given concept contains all other concepts that have a co-occurrence weighted by the Uncertainty Coefficient [5]. Concept profiles must be constructed uniquely for a given ontology and corpus [15, 16], but once they have been constructed, the similarity between any two concept profiles can be computed by taking the inner product of their corresponding weights, the so-called match score[17]. The statistical significance of the match score between the profiles of two concepts (i.e., gene and disease) can be evaluated by comparing the log transform of the match scores to that of a null distribution constructed from randomly chosen concept pairs (Fig 1). Hence, it is possible using concept profiles to establish a statistically significant association between concepts based on highly ranked concepts in their profiles, even when they do not have a co-occurrence (*i.e.* usually an explicit stated relationship) in the literature. Discovery of novel and informative associations between genes and diseases is thus not dependent on linkage analyses or seed genes.

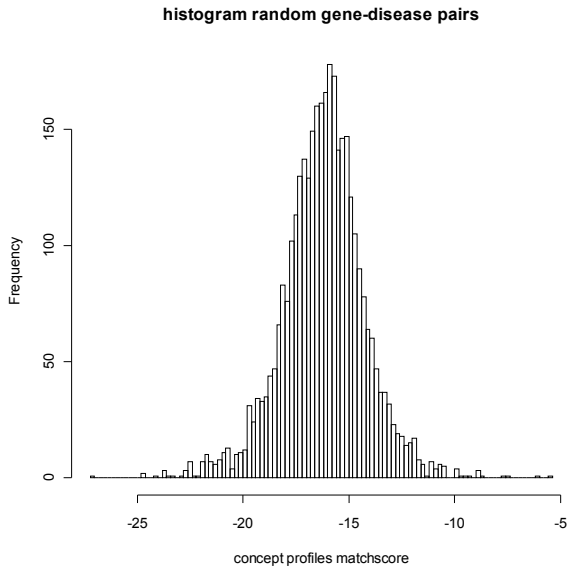


Figure 1. Distribution of concept profiles match scores calculated for randomly chosen gene-disease combinations. The MEDLINE abstract text corpus is from 1980 until May 2009.

Results and Discussion

We evaluated the effectiveness of concept profiles using 18 previously described gene-disease relationships taken from the Human Reference Protein Data base (HRPD)[18] (Table 1). Concept profiles for the genes and the diseases were constructed from all MEDLINE abstracts up to one year before the landmark paper explicitly describing the link between the gene and disease was published. This roll-back analysis used two time-delimited corpora: From 1980 to February 2005 (for landmark publications dating from February 2006 to December 2006) and from 1980 to August 2006 (landmark publications appearing after august 2007). For each of these gene-disease pairs, no co-occurrence was found between the gene and the disease before the landmark paper was published, both the gene and the disease appear in a minimum of five abstracts and the disease is currently considered to be monogenic. For each test gene an artificial linkage interval was arbitrarily set containing 200 genes (100 genes upstream and downstream of the test gene) following the approach by Aerts et. al. [7].

Table 1. Gene-disease pairs using concept profiles.

Gene	Disease	Landmark Publication Date	PMID	Rank	p-value
MFN2	Hereditary motor and sensory neuropathy VI	February 2006	16437557	2	0.0018
RECQL4	Baller-Gerold syndrome	February 2006	15964893	1	0.0046
KRT85	Ectodermal dysplasia, pure hair-nail type	March 2006	16525032	13	0.0093
ACVR1	Fibrodysplasia ossificans progressiva	May 2006	16642017	2	0.0052
TGFBI	Corneal dystrophy, epithelial basement membrane	June 2006	16652336	1	0.00045
IL10RB	Hepatitis B virus, susceptibility to	June 2006	16757563	16	0.062
IFN-AR2	Hepatitis B virus, susceptibility to	June 2006	16757563	3	0.0065
PLA2G6	Infantile neuroaxonal dystrophy 1	July 2006	16783378	11	0.043
TREX1	Aicardi-Goutieres syndrome 1	August 2006	16845398	105	0.93
CHRNG	Escobar syndrome	August 2007	16826520	66	0.77
DOK7	Myasthenia, limb-girdle, familial		16917026	NaN	NaN
SCN9A	Paroxysmal extreme pain disorder	September 2006	17145499	15	0.084
MYH11	FAA4	March 2006	16444274	4	0.013
TFAP2A	branchio-oculo-facial syndrome	May 2008	18423521	4	0.068
PIK3CA	Seborrheic keratosis	August 2007	17673550	9	0.076
VLDLR	Dysequilibrium syndrome	February 2008	18043714	27	0.18
BUB1B	PCS	February 2006	16411201	12	0.29
TRPV4	Brachyolmia	August 2008	18587396	44	0.76
			Average rank 20		Average p

The concept profile for each gene in the linkage area was matched with the disease profile resulting in a ranked list of the test gene among the 200 genes in the artificial linkage interval (Table 1). On average the test genes ranked within the top 20, and in two cases (epithelial basement membrane corneal dystrophy (EBMD)) and Baller-Gerold syndrome), the test genes ranked number one. However, the TGFBI gene is often co-mentioned with generic disease types, like hereditary corneal dystrophy and corneal dystrophy (column 3 in Table 2) suggesting that its high rank is not necessarily an indicator of a specific relation to EBMD. For the 'Myasthenia, limb-girdle, familial' there was not enough information for the test gene to build a concept profile. When prioritizing gene-disease pairs in practice, it is essential that the significance of the putative gene-disease relation be subject to evaluation. Hence one-sided p-values were calculated from the concept profile

match scores, using the null distribution (Fig 1 & Table 1). This p-value is indicative of the quality of the prediction, and therefore of the reliability of the ranking of the list: the cases with very high ranks could have been predicted based on the low p-value of the gene (Fig 2). If we had used a cut-off p-value of 0.02 to reject the prioritizer output the results for Aicardi-Goutieres syndrome 1, Brachyolmia and Myasthenia, limb-girdle, familial would have been rejected. Escobar syndrome (test gene ranks 66) with a p-value of 0.017 for the highest ranked gene would have remained as a reliable output. With three outliers rejected the average rank of the remaining 15 samples would becomes 12.5. Clearly concept profiles are highly effective in identifying gene-disease pairs deliberately using only the implicit information in MEDLINE prior to the landmark paper.

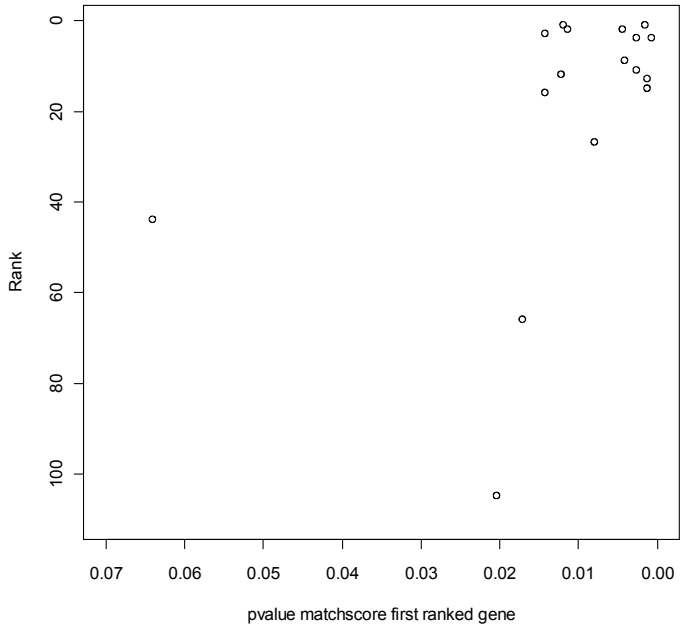


Figure 2. p-value of the highest rank gene versus the rank of the test gene.

In addition to prioritizing genes, concept profiles provide important biological insight revealing how the gene might be associated with a disease. However, by inspecting the highly ranked concepts in the two concept profiles that linking the gene and disease a biomedical expert would likely (example of gene *PIK3CA*

Table 1) choose for instance gene with rank nine over of the first eight for further investigation. To explore the utility of the information in concept profiles in rationalizing predicted gene-disease pairs we chose the three gene-disease pairs having the highest ranking: Hereditary motor and sensory neuropathy VI (HMSN VI), Baller-Gerold syndrome (B-G syndrome), and EBMD. Biomedical researchers with expertise in these genes and diseases evaluated the shared concepts in concept profiles for their biological significance. Table 2 shows the top five of overlapping concepts between the gene and disease concept profiles. For B-G syndrome the dominating concept is Rothmund-Thomson (R-T) syndrome (contributes more than 95% to the overall score). Two documents were found that support the association between B-G syndrome and R-T syndrome, PMID: 11045594 and 9934984. The first one gives information for the clinical phenotypic overlap between the two syndromes. The gene RECQL4 has been co-mentioned before with R-T syndrome as a gene that when mutated causes this syndrome (PMID: 12379465, 12601557, 12673665, 12734318, 12838562, 12915449, 12952869, 15221963, 15317757, and 15558713). Because of the phenotypic overlap between the two syndromes, RECQL4 would be the most likely candidate to investigate first. Indeed, the landmark paper (PMID: 15964893) reports precisely this reasoning: "Clinical overlap between BGS, Rothmund-Thomson syndrome (RTS), and RAPADILINO syndrome is noticeable. Because patients with RAPADILINO syndrome and a subset of patients with RTS have RECQL4 mutations, we reassessed two previously reported BGS families and found causal mutations in RECQL4 in both."

Table 2. Indirect concepts linking the gene with the disease.

	Baller-Gerold syndrome		**Hereditary motor and sensory neuropathy VI**		**Corneal dystrophy, epithelial basement membrane**	
Top	Overlapping concepts	Contribution to score	Overlapping concepts	Contribution to score	Overlapping concepts	Contribution to score
1	rothmund-thomson syndrome	95.79	opa1	40.37	hereditary corneal dystrophy	42.28
2	Poikiloderma	2.47	optic atrophy, autosomal dominant	35.32	Corneal dystrophy	41.43
3	online mendelian inheritance in man	0.45	OPA1	23.17	lattice corneal dystrophy	12.08
4	Growth deficiency	0.32	Axonal neuropathy	0.61	Dystrophy	2.73
5	Clinical variability	0.24	recessive inheritance	0.3	Corneal erosion	0.58

In the case of HMSN VI, this disease is caused by mutations in the MFN2 gene. The overlapping concepts in the top three are all a form of optic atrophy 1. The first concept opa1 is the gene in zebrafish, the second concept is a disease and the third concept the human gene. Together they contribute more than 98% to the overall score. The landmark paper (PMID: 16437557) clearly shows that this concept is a strong indirect link, stating: "It is intriguing that MFN2 shows functional overlap with optic atrophy 1 (OPA1), the protein underlying the most common form of autosomal dominant optic atrophy, and mitochondrial encoded oxidative phosphorylation components as seen in Leber's hereditary optic atrophy." The MFN2 gene ranked second place (Table 1). This means one false positive before the test gene is found. The gene that ranks first place is KIF1B where the top five concepts between it and the disease are hereditary motor and sensory neuropathies (65.61%), HMSN II (15.76%), hereditary liability to pressure palsies (7.8%), Axonal neuropathy (4.61%), and HMSN I (2.96%). In consulting the supporting documents for KIF1B it was found that mutations Charcot-Marie-tooth disease type 2A1 (CMT 2A1 or HMSN2A1). Intriguingly, HMSN VI is also known as Charcot-Marie-tooth disease type 6 (CMT6). Thus, it appears that KIF1B is not a false positive but a gene that causes a related disease.

Detailed expert analyses of the concept profiles in linking genes to Seckel syndrome are provided in the Supplementary Information.

Table 3. Endeavor Gene Prioritizer predictions for monogenic and polygeneic diseases. Averages are only calculated over the ranks that are both covered by Endeavour and concept profiles.

Disease (monogenic)	**Endeavour**	**Concept profiles**
arrthythmia	4	20
congenital heart disease	(3)	NaN
cardiomyopathy 1	2	2
parkinsons disease	(50)	NaN
charcot-marie-tooth disease	14	1
amyotrophic lateral sclerosis	27	16
klippel-trenaunay disease	(3)	NaN
cardiomyopathy 2	1	10
distal hereditary motor neuropathy	15	51
Cornelia de Lange syndrome	(9)	NaN
average ranking	11	17
Disease (polygenic)	**Endeavour**	**Concept profiles**

Rheumatoid_arthritis	11	24
Parkinson_disease	23	30
Atherosclerosis1	54	5
Atherosclerosis2	29	21
Crohn disease	71	11
Alzheimer disease	54	3
Average ranking	40	16

To gauge the performance of concept profiles against methods based on co-occurrence, we replicated a recent study [2] using the gene prioritizer Endeavour where gene-disease predictions were made for ten monogenic and six polygenic diseases (Table 3). We generated concept profiles for the diseases in these test sets and for the test genes in their corresponding linkage interval. We used the same roll back analyses as Endeavour, taking only literature information up to one year before the landmark paper was published. For the monogenic diseases there were three genes where there was not enough information to calculate a match score using concept profiles. Of the 7 remaining gene-disease pairs for monogenic conditions, the average performance of the two methods was comparable. However, in the case of polygenic diseases having inherently complex interrelations among numerous genes and other concepts, concept profiles outperformed Endeavour's ranking on average by two-fold. By drawing on the deep network of conceptual relations that inform the study of polygenic diseases but usually remain un-stated in the literature, concept profiles are uniquely suited for knowledge discovery in complex multifactorial systems.

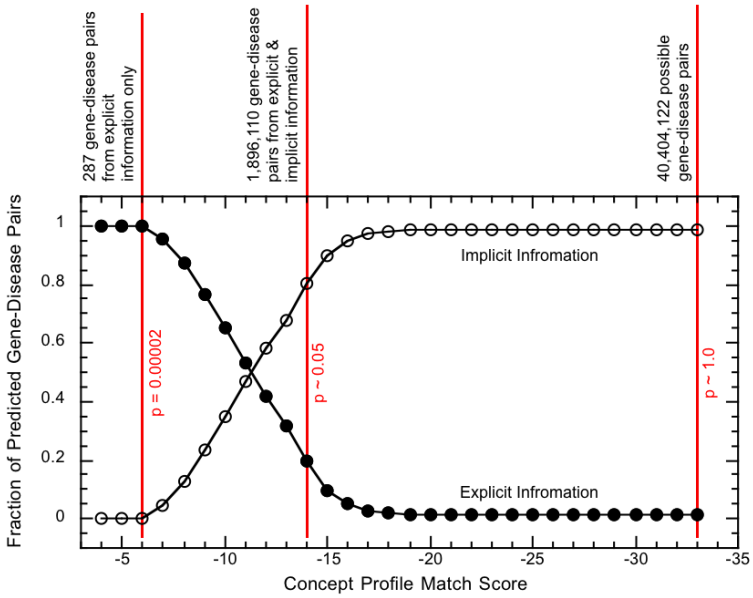


Figure 3. Estimation of implicit and explicit information. Co-occurrence methods can prioritize only 287 of the possible 40 million gene-disease pairs, while concept profiles can prioritize 5% at p= 0.05. Note the vast majority of textual information is implicit.

These results indicate the importance of implicit information in discovering new knowledge. Concept profiles can be used to estimate the relative proportions of implicit and explicit information. For example, given the number of genes and diseases meeting our minimal information criteria used herein, there are 40,404,412 possible gene-disease combinations. The match score and corresponding p-value for each these gene-disease pair can be calculated. For each p-value, the cumulative number of implicit and explicit gene-disease pairs (and then normalized to a percentage) can be computed (Supp Info Table 4). Thus, for each p-value, we know the fraction of the predicted pairs that are due to implicit information (Fig 3). For p=0.0 only gene-disease pairs with minimally one co-occurrence are found. But even for extremely significant p-values (p=0.00002) we already find some gene-disease pairs for which their association is only due to implicit information (*i.e.*, no co-occurrences found in MEDLINE). For p=0.003, still a highly significant gene-disease p value, the amount of implicit information is already 47%. For commonly accepted p-values around p=0.05, 88% of the gene-disease pairs are due to implicit information. Conclusion: The vast majority of

useful information in text is implicit, and this information is accessible with concept profiles.

There are 5330 gene-disease predictions that are better than p = 0.0002. To facilitate the expert evaluation of these predicted gene-disease pairs, the shared concepts from the concept profiles have been posted online along with the related PubMed IDs. Experts can search this data on gene or disease or any other related concept, and can provide their estimation of the quality of the prediction and leave commentary regarding possible biological mechanisms.

References

1. Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., et al., *Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.* Am J Hum Genet, 2006. **78**(6): p. 1011-25.
2. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
3. Oti, M., Snel, B., Huynen, M.A., and Brunner, H.G., *Predicting disease genes using protein-protein interactions.* J Med Genet, 2006. **43**(8): p. 691-8.
4. Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., et al., *A human phenome-interactome network of protein complexes implicated in genetic disorders.* Nat Biotechnol, 2007. **25**(3): p. 309-16.
5. Jelier, R., Schuemie, M.J., Roes, P.J., van Mulligen, E.M., and Kors, J.A., *Literature-based concept profiles for gene annotation: the issue of weighting.* Int J Med Inform, 2008. **77**(5): p. 354-62.
6. Jelier, R., Schuemie, M.J., Veldhoven, A., Dorssers, L.C., Jenster, G., et al., *Anni 2.0: a multipurpose text-mining tool for the life sciences.* Genome Biol, 2008. **9**(6): p. R96.
7. Perez-Iratxeta, C., Bork, P., and Andrade, M.A., *Association of genes to genetically inherited diseases using data mining.* Nat Genet, 2002. **31**(3): p. 316-9.
8. van Driel, M.A., Cuelenaere, K., Kemmeren, P.P., Leunissen, J.A., and Brunner, H.G., *A new web-based data mining tool for the identification of candidate genes for human genetic disorders.* Eur J Hum Genet, 2003. **11**(1): p. 57-63.

9. Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B., et al., *Integration of text- and data-mining using ontologies successfully selects disease gene candidates.* Nucleic Acids Res, 2005. **33**(5): p. 1544-52.
10. Kohler, S., Bauer, S., Horn, D., and Robinson, P.N., *Walking the interactome for prioritization of candidate disease genes.* Am J Hum Genet, 2008. **82**(4): p. 949-58.
11. Radivojac, P., Peng, K., Clark, W.T., Peters, B.J., Mohan, A., et al., *An integrated approach to inferring gene-disease associations in humans.* Proteins, 2008. **72**(3): p. 1030-7.
12. Linghu, B., Snitkin, E.S., Hu, Z., Xia, Y., and Delisi, C., *Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network.* Genome Biol, 2009. **10**(9): p. R91.
13. Day, A., Dong, J., Funari, V.A., Harry, B., Strom, S.P., et al., *Disease gene characterization through large-scale co-expression analysis.* PLoS One, 2009. **4**(12): p. e8491.
14. Ala, U., Piro, R.M., Grassi, E., Damasco, C., Silengo, L., et al., *Prediction of human disease genes by human-mouse conserved coexpression analysis.* PLoS Comput Biol, 2008. **4**(3): p. e1000043.
15. Schuemie, M.J., Jelier, R., and Kors, J.A. *Peregrine: Lightweight gene name normalization by dictionary lookup*. in *Biocrative 2 workshop*. 2007. Madrid.
16. Tuason, O., Chen, L., Liu, H., Blake, J.A., and Friedman, C., *Biological nomenclatures: a source of lexical knowledge and ambiguity.* Pac Symp Biocomput, 2004: p. 238-49.
17. van Haagen, H.H.H.B.M., t Hoen, P.A.C., Botelho Bovo, A., de MorrÃ©e, A., van Mulligen, E.M., et al., *Novel Protein-Protein Interactions Inferred from Literature Context.* PLoS ONE, 2009. **4**(11): p. e7894.
18. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.

Supplementary information

Seckel syndrome is known as a rare autosomal recessive disorder characterized by growth retardation, microcephaly with mental retardation, and a characteristic 'bird-headed' facial appearance. Presently, only one gene in OMIM, ataxia-telangiectasia, is related to Seckel Syndrome via its mutated from RAD3-related protein (ATR)[1]. Recently a second gene that encodes for Centromere protein J (CENPJ) has been identified by Al-Dosari et. al.[2], but this gene had yet to be entered in OMIM at the time this analysis was completed. Of the top 20 proteins (out of the 12,391 proteins that had sufficient information to build a concept profile) having the highest match score to Seckel Syndrome, CENPJ ranks number 14, although CENPJ has no co-occurrences with Seckel syndrome in PubMed abstracts (Table 1). The concept microcephaly contributes the most to the match score and is the strongest implicit (or indirect) link between Seckel syndrome and CENPJ. Other candidate genes appear in this list that have been co-mentioned with Seckel syndrome before. For instance the protein pericentrin (PCNT, ranks 2) has three articles. The article with PMID: 18157127 describes that PCNT is another gene that causes Seckel syndrome. The article with PMID:19546241 gives a nice overview of related diseases with similar phenotype such as 'Primary microcephaly' and 'microcephalic osteodysplastic primordial dwarfism type II' (MOPD II). This article also lists microcephalin (MCPH1, ranks 5) as another disease-causing candidate. The last article (PMID: 16278902) also mentions the concept MOPD II. These results prompted us to further investigated whether CENPJ might be associated with PCNT and ATR. We generated a prioritized list for ATR and PCNT and checked the rank of CENPJ. Surprisingly CENPJ also showed no co-occurrences with ATR and it ranked 706. However, CENPJ is co-mentioned once with PCNT (PMID: 18174396). In this article PCNT is given as the cause for primordial dwarfism, and other candidate genes for Seckel syndrome are postulated (CDK5RAP2 ranks 32 not in table 1, and ASPM ranks 15). We performed a new search where related concepts for Seckel syndrome were used as PubMed input query and the results aggregated into a single rank using (Table 2). In this case, CENPJ ranks 4th and although ATR is a known gene for Seckel syndrome (recorded in OMIM), its information content is poor compared to the other related Seckel syndrome concepts.

From the PubMed abstracts we selected candidate genes that have been co-mentioned with Seckel syndrome related concepts and used them as a search query in the STRING database of known and predicted protein-protein interactions to see if there are any relations between these candidate genes (Fig 1). Again, ATM has many links with ATR, while all other links are mainly in the network of PCNT. CENPJ has a known physical interaction with PCNT. From a biological view, it would be highly interesting to identify the missing link between the ATR and the

PCNT pathway. In seeking potential relations between Seckel syndrome and CENPJ, it is clearly much easier to inspecting the concept profile overlap between (Table 3) than performing multiple PubMed search queries manually reading up to 17 articles. Lastly, this case demonstrates that using conventional co-occurrence approaches to predicting gene-disease relations could have negative performance results. Here, ATR, although the first choice to use as seed gene when looking for additional genes related to Seckel syndrome, would lead to false negative conclusions.

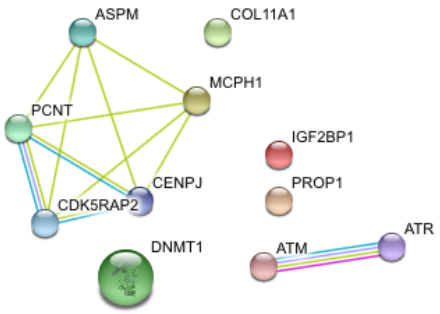


Figure 1. Candidate genes for Seckel syndrome in a network graph generated by STRING.

Supp Info Table 1. Prioritized list of proteins match with the profile of Seckel Syndrome. The column 'main concept' gives information which concept contributes the most to the score and is the strongest implicit (or indirect) link between Seckel syndrome and CENPJ. NUP85 is a homonym for PCNT and retrieves the same articles for PCNT.

rank	Co-occurrences	genename	Main concept	Contribution (%)	OMIM gene	PMIDs
1	7	ATR	ATR	74.43	x	[12640452, 14571270, 15309689, 15496423, 15616588, 16015581, 19504344]
2	3	PCNT	Seckel syndrome	54.38	x	[16278902, 18157127, 19546241]
3	2	NUP85	Seckel syndrome	74.54		[18157127, 19546241]
4	2	ANTXR1	ANTXR1	66.28		[12640452, 19504344]
5	3	MCPH1	MCPH1	61.84	x	[16217032, 17102619,

						19546241]
6	2	NBN	NBN	64.58	x	[15616588, 18664457]
7	0	MCPH2	Primary microcephaly	32.01		
8	2	FANCD2	FANCD2	47.9	x	[15314022, 15616588]
9	0	ATRIP	ATR	92.91		
10	1	DNMT1	DNMT1	98.49		[17015478]
11	1	MDC1	MDC1	19.13		[18664457]
12	1	FANCC	Fanconi's Anemia	31.02	x	[10232749]
13	6	PALB2	Fanconi's Anemia	36.35		[3115102, 6465473, 7686032, 10232749, 15314022, 17224058]
14	0	CENPJ	Microcephaly	17.77	x	
15	0	ASPM	Microcephaly	48.12	x	
16	5	CHEK1	CHEK1	28.75		[15616588, 16217032, 17015478, 18077418, 19504344]
17	0	PROP1	dwarfism	98.96	x	
18	2	MMAB	MMAB	60.9	x	[15314022, 19504344]
19	0	TOPBP1	ATR	64.53		
20	1	FOXL2	FOXL2	59.56	x	[16015581]

Table 2. Rank of proteins in prioritized lists for different concepts that are associated with Seckel syndrome, including Seckel syndrome itself

Gene name	Rank	Seckel syndrome	PCNT	MOPD II	Primary microcephaly	Microcephaly	ATR
PCNT	1	2	1	1	11	70	845
MCPH2	2	7	53	5	1	4	540
ASPM	3	15	38	6	2	2	622
CENPJ	4	14	18	4	3	5	706
ATR	5	1	385	53	29	49	1
MCPH1	6	5	64	7	4	12	271
NUP85	7	3	9	2	15	84	831
NBN	8	6	660	125	24	1	10
MDC1	9	11	233	15	6	63	15
TOPBP1	10	19	257	17	10	237	5
CDK5RAP2	11	32	36	10	5	14	1233
CHEK1	12	16	254	36	20	223	3
TP53BP1	13	27	500	26	13	170	16
RHO	14	41	373	21	8	261	28
CHEK2	15	26	428	87	36	177	4
ERCC2	16	23	1054	48	42	174	9
MRE11A	17	22	1118	584	63	9	11
GCP3	18	39	4	3	31	1077	7133

RAD50	19	25	881	478	67	10	17
SEH1L	20	104	54	12	89	67	753

Table 3. Overlapping concepts between Seckel syndrome and CENPJ

Top	Overlapping concept	Contribution (%)
1	Microcephaly	17.77
2	Primary microcephaly	17.31
3	MCPH1	11.86
4	Mcph1	11.44
5	MCPH1	11.44
6	MCPH1	11.41
7	MCPH1	7.54
8	PCNT	4.38
9	osteodysplastic primordial dwarfism	1.94
10	NUP85	1.11
11	MOPD II	0.93
12	pericentrin	0.77
13	dwarfism	0.72
14	Centrosome	0.55
15	Genes, Recessive	0.32

pvalue	matchscore	implicit	explicit	cum imp	cum exp	% imp	% exp
0	-4	0	29	0	29	0.00	1.00
0	-5	0	287	0	316	0.00	1.00
0.00002	-6	5	1139	5	1455	0.00	1.00
9.01E-05	-7	173	2341	178	3796	0.04	0.96
0.00023	-8	997	4333	1175	8129	0.13	0.87
0.00056	-9	3994	8561	5169	16690	0.24	0.76
0.00127	-10	12863	16612	18032	33302	0.35	0.65
0.00308	-11	38259	30653	56291	63955	0.47	0.53
0.00724	-12	109383	55429	165674	119384	0.58	0.42
0.01756	-13	327773	113012	493447	232396	0.68	0.32
0.04645	-14	1027450	142817	1520897	375213	0.80	0.20
0.11323	-15	2610325	78691	4131222	453904	0.90	0.10
0.23868	-16	5050536	26734	9181758	480638	0.95	0.05
0.41924	-17	7252607	6287	16434365	486925	0.97	0.03
0.60682	-18	7619189	906	24053554	487831	0.98	0.02
0.75026	-19	5790456	40	29844010	487871	0.98	0.02
0.84564	-20	3816141	0	33660151	487871	0.99	0.01
0.90624	-21	2460978	0	36121129	487871	0.99	0.01
0.94567	-22	1586840	0	37707969	487871	0.99	0.01
0.97008	-23	995043	0	38703012	487871	0.99	0.01
0.98434	-24	577399	0	39280411	487871	0.99	0.01
0.99236	-25	322237	0	39602648	487871	0.99	0.01
0.99673	-26	169158	0	39771806	487871	0.99	0.01
0.99855	-27	83531	0	39855337	487871	0.99	0.01
0.99948	-28	36376	0	39891713	487871	0.99	0.01
0.99972	-29	15168	0	39906881	487871	0.99	0.01
0.99982	-30	6123	0	39913004	487871	0.99	0.01
0.99989	-31	2528	0	39915532	487871	0.99	0.01
0.99993	-32	971	0	39916503	487871	0.99	0.01
0.99993	-33	38	0	39916541	487871	0.99	0.01

Table 4. Estimation of Implicit and Explicit Information

1. O'Driscoll, M., Ruiz-Perez, V.L., Woods, C.G., Jeggo, P.A., and Goodship, J.A., *A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome.* Nat Genet, 2003. **33**(4): p. 497-501.
2. Al-Dosari, M.S., Shaheen, R., Colak, D., and Alkuraya, F.S., *Novel CENPJ mutation causes Seckel syndrome.* J Med Genet. **47**(6): p. 411-4.

Chapter 6

General discussion

This thesis presents *in silico* text- and data-mining techniques for the prediction of biologically related concepts. The methods were evaluated on protein-protein interaction data and genes associated with certain diseases. The main part of the research was the evaluation of the text-mining method called concept profiles. Later on we extended concept profiles with other non-textual information. The many hurdles and findings are discussed below. We conclude with the future directions where text-mining and data-mining can be improved.

1. Evaluating set creation

During this research a large part of the effort was needed to collect training and test data.

Collecting good data for the evaluation of a data-mining system is hard. Here we describe the problems we encountered.

1.1 Nature of biological data

The data used in this study has several characteristics that make the application of existing data and text-mining methods difficult. The world of biology is far more complex than a computer system can model. It is no simple 'black and white' or the use of TRUE and FALSE labels.

First, biological data is sometimes not reliable, and highly dependent on the context it appears in. For instance protein-protein interactions (PPIs) are recorded in protein databases and each database has a level of curation. Some protein interactions are very well described in databases like DIP. These PPIs are confirmed with several independent wetlab experiments or have a lot of literature evidence. Other protein interactions come from high throughput experiments and are recorded in a database like IntAct. High throughput experiments normally contain more false positives. The same holds for instance for the annotation of gene functions in the Gene ontology (GO). In an old release of the GO a gene is assigned a GO term describing a molecular function. In later releases sometimes the GO term becomes obsolete because it was wrongly annotated or the GO term is merged with another term.

Second, the current knowledge is limited and incomplete. Only a small fraction of the total interaction space (e.g. all protein-protein interactions in the human body) is described. This results in overestimation of the prediction performance because the performance is biased towards well studied proteins, i.e. biased towards only this small subset of protein-protein interactions.

Third, biological data change over time. For instance when two proteins are not known to interact, a system would label this protein pair as TRUE NEGATIVE. However in a wetlab experiment the two proteins were confirmed to interact. After this discovery the protein pair would be labeled as TRUE POSITIVE.

In an evaluation process, biological data should be used keeping these characteristics in mind.

1.2 Biological nomenclature

The nomenclature of biological names is not standardized. For genes or proteins there exist multiple accession numbers (e.g. Uniprot, Entrez Gene, or HUGO Gene Nomenclature Committee), synonyms, and abbreviations that all need to be mapped to single unique identifiers. To disambiguate genes in text is difficult because many genes share the same synonym, resulting in homonym problems.
For gene-disease relationships it is even harder. Many of the genes are assigned the name of the disease they are associated with. These samples cannot be used as a test sample. In addition the disease name as it is recorded in databases is hard to recognize in text. For instance Alzheimer disease had over 15 variants recorded in OMIM (e.g. Alzheimer type 2). In text normally this will be described as that they found a new type of Alzheimer disease. Hence not the concept Alzheimer type 2 is recognized but the generic concept Alzheimer disease.
Furthermore, concepts are related to each other in a hierarchical ontology. For instance the concept Duchenne muscular dystrophy (DMD) in the ontology is part of the concept muscular dystrophies. Once DMD is recognized in text as a concept, one could argue if it is informative that in the same text the higher level concept muscular dystrophy is recognized.

1.3 Minimum information requirements for text-mining

In chapter 5 we introduced the roll back analysis. This is a way to simulate a prediction over time. We imposed the constraint for gene-disease relationships in our test set that the two concepts should not be co-mentioned together before the relationship was discovered, to prove that they could have been predicted using the implicit information. However, before that first co-occurrence there should be enough information available which is sometimes also not the case. In order to build a concept profile for a concept we maintained a threshold of at least 5 abstracts where that concept is mentioned.
This limitation resulted in a set of only 18 gene disease pairs described in chapter 4 where the original list started out with roughly 5,000 gene disease pairs in HPRD. The same problem probably occurred in the article by Aerts et al. [1] where they obtained a small set of 10 monogenic and 6 polygenic diseases.

1.4 Curation and confirmation of biological data

One aspect of bioinformatics is that it is important to validate (or verify) every *in silico* prediction with a wet lab experiment. The results of the experiments described in this thesis required interpretation by expert biologists. This introduced a dependency on experts, who had to make time in their busy schedules. Luckily,

the biologists at our department were very helpful, but still the amount of work that could be asked of them was limited. Every bioinformatician would love to have his own private biologist.

1.5 Circular reasoning

Another problem is that many databases have a certain level of redundancy. In machine learning a key step to evaluate a prediction system is to divide the data into a training set and a test set. The training data is used to train all the parameters that are used in the model of the prediction system. The test data is used to evaluate how well the system is able to correctly predict the labels of the test data. Training and test data should be independent. That is, no data that is used for training should be used for evaluation, else this could lead to an over estimation of the performance.

However for biological data it is sometimes not possible to divide the data into an independent training and test set. For instance when a wet lab experiment is done for investigating a PPI, the results will be described in an article and published, and the same result are stored in a database like DIP. To separate the database and article information is difficult. Therefore we introduced in chapter 3 and 5 the retrospective study (or roll back analysis) and do a prediction simulation over time to eliminate the bias. To do this, it is necessary to get access to old releases of databases. Most databases do not store previous releases for download. For bioinformatics purposes this would be extremely helpful to keep track of old releases.

2. Findings

2.1 Implicit information extraction and content

Chapter three and five showed that implicit information extraction works. The information, or the indirect links, that connects two concepts can be derived from the concept profile overlap. It seems that for PPI prediction the dominating concept is normally another protein already associated with one of the two proteins. For instance in chapter three CAPN3 was linked with PARVB via the intermediate protein DYSF. For gene disease relationships it is normally an associated phenotype, or another gene also known to cause another disease. For instance in chapter five RECQ4L was also associated with Rothmund-Thomson syndrome and therefore seems to be a good candidate for Baller-Gerald syndrome because these two syndromes show clinical phenotypic overlap. These two examples are indicative that the implicit information is meaningful and well explains the association between two concepts. We further observe that when an implicit link is found between concepts the link is normally one dominating concept. To verify this, more samples should to be evaluated.

2.2 Added value of other data sources.

In chapter four we investigated if concept profiles can be improved by adding other non-textual data sources. We found that some of the problems encountered with text-mining could be solved with the other data sources. We conclude that this may work but the performance is dependent on the sample you are looking at. It was shown for DYSF that the amount of information in additional databases besides the literature was poor. In the PKD1 case study the disambiguation problem was solved by microarray expression data. The nature and the amount of information from every source has its pros and cons dependent on the sample. Figure 1 shows an example of the AuC output for each database for DMD and HTT to illustrate that for each protein another data source is dominating. In many pattern recognition approaches it is usual to do feature selection for dimensionality reduction resulting in the most informative features. For instance in a microarray experiment the goal may be to look for differentially expressed genes. The number of genes checked start with 30,000 and after filtering (feature selection) the number of genes will vary from 10 till 100. However for combining data sources the number of available data sources, suitable for processing, is already limited. Making databases inter-operable is very important. As stated earlier the data source that is most informative changes with each sample (figure 1). A generic feature selection approach therefore seems not appropriate for biological data. A scientist should be able to select the data source he is interested in. Also on the basis of known knowledge and ROC curve analysis a scientist can get a feeling if the data source is informative for his samples (e.g. a protein). Added value of data sources and feature selection should be considered for each question separately.

2.3 Types of relationships

We did the collection of data for the relationship types 'protein interacts with other protein' and 'a gene when mutated causes a certain disease'. As discussed preciously the prediction performance is dependent on each sample. The same holds for types of relationships. This can be very well explained. For protein interactions 70% of the known PPIs recorded in databases cannot be traced back in PubMed abstracts because normally the interactions are stored in a table in full text. When a gene is found for a disease, the landmark paper will always co-mentioned the gene and the disease in the abstract, if not even in the title. After the landmark paper multiple occurrences happen in articles published after the landmark paper. We did an evaluation of the gene disease relationships in OMIM and found that ~83% of the known pairs have a co-occurrence in MedLine abstracts. The distribution is given in figure 3. We checked another relationship type, that of 'gene has function X' taken from the Gene Ontology. For this relationship type the distributions are given in figure 2. These figures clearly show

that the known relationships are very different. Since gene/disease relationships almost always occur in PubMed abstracts, the association score is in general high.
The null distributions (or random distribution) tend to look the same. For knowledge discovery scientists are interested in new concept pairs (e.g. PPI, gene/disease) previously not recorded in any database but found by our text-mining system. Those are all the pairs from the null distribution. In chapter three we generated a null distribution of random protein pairs. In figure 2 and 3 the null distributions for gene/disease and protein/function are given respectively. The distributions look alike. To investigate if null distributions from different semantic types (e.g. protein pairs or gene-disease pair) as the same and can be treated universal we calculated match scores for 100,000 random protein pairs and 100,000 random gene-disease pairs. The results are plotted in figure 4. This plot clearly shows that the gene-disease pairs (blue) are different from the protein pairs (red) even though the two distributions both have a Gaussian characteristic. An explanation of this difference could lie in the fact that proteins or genes in general are more intensively described that diseases (all diseases besides OMIM are taken into account). Therefore concept profiles for protein/genes are more enriched which results in on average higher match scores. This result means that any pairs of two semantic types cannot be treated universally. For instance, when the match score for a protein pair is significant (e.g. $p<0.01$) calculated under the null hypothesis that any concept pair (regardless the semantic type) is not related, this same protein pair could not be significant (or at least is different) under the null hypothesis that protein pairs are not related.

3 Limitations of text-mining

Concept profiles show a better performance in predicting associations between concepts than the direct relationship approach (described in chapter two and three). However there are still limitations in prediction performance even for concept profiles. First, finding a new relationship between two concepts goes as far as there is information. This means that there must be sufficient information for both concepts. For concept profiles we formulated this that there should be at least 5 articles available for both concepts. Many diseases or proteins are rare that they have not been published about. In this case text-mining fails not because of technical shortcomings but just due to the lack of information.
Second, the lack of information can also be within the implicit information. If two concepts are related to each other is does not mean that they will always be linked with each other via intermediate concepts. If they do not, this is also not due to text-mining shortcomings but that there is no implicit links available.
Third, the biggest limitation is the accuracy of the disambiguation process. This is dependent on the style of writing of the author, i.e. which nomenclature he uses for words and if words are abbreviated. The problem of disambiguation lies in that

humans are more adapted to give names to entities that are easy to recognize and easy to remember on how they are named. Normally this is done using an acronym. In addition biologists do not make a standard convention about the word nomenclature for proteins. As Michael Ashburner [2] once said 'Biologists would rather share their toothbrush than share a gene name'.
It is shown in competitive conferences [3] where state of the art text-mining systems compete with each other that there is a maximum performance reachable (e.g. 0.88 and 0.50 recall and precision respectively). No computer system is ever able to retrieve a 100% score.

4 Future directions

We believe that text-mining and in particular concept profiles are indispensable in biological research. We foresee that text-mining will become a core technology in the so called semantic web. The semantic web is a name giving for a trend going on the Internet. The first trend in Internet development was called web 1.0. It was the collection of all static HTML pages with only plain text. The second generation is called web 2.0 where the web became interactive. Think of user input like credits cards, online bookstores or Wikipedia. The third generation is called web 3.0 or the semantic web. Here the plain text on web pages, blogs and published literature will be linked with each other in a web of concepts, where the links between concepts can be facts generated by information extraction (IE) or can be hypothesis being a novel relationship using text-mining techniques.
There is still a lot to gain in research and the development of text-mining and remarkable some of them are not computer oriented.

4.1 Community annotation

The first development is that of community annotation [4]. With the common technology the way that computers can read text have their limitations in terms of accuracy. Disambiguation remains a key aspect and hard to solve for many concepts. However with the future version of the semantic web and the millions of people on the internet every day this can be solved. A person on the web, a so called community annotator, can screen an article of interest that has been tagged by a text-mining system and correct the words that have been misclassified. With misclassification we mean that a word was too ambiguous to resolve or not recognized because the word does not appear as a concept in the ontology. For a human reader the disambiguation can be done manually even so the ontology can be updated with new concepts or synonyms for existing concepts. Or in the case of the Alzheimer example, a new type can be corrected for in an article years after publication. Since the internet contains millions of users every day, this annotation process increases the accuracy of tagged text over time.

4.2 Making standard nomenclature

The second improvement is for standardizing databases, identifiers and names for concepts. In the past many attempts have been made to come to a common ontology that is accepted world wide by all biology scientists. However thus far these attempts have failed for many reasons, some of which are unexplainable. An example is that in the past years companies developed their own databases for data-mining purposes. Once they published about their database (in online website form) it was used frequently over the coming month. After a period of time the database became old and not maintained. In the end the database 'dies' and is buried on the 'database graveyard'. For world wide co-operation we suggest that biologists get inspired by ICT companies and organizations like IEEE for whom standardizing is a well known principle (http://standards.ieee.org/). For instance, with the digital revolution many electronic devices came available for home users that need to be universal. A compact disk that can be played with any CD player that is bought in Germany or Japan. Or a personal computer where a soundcard works and fits in any motherboard. The universal exchangeability works in this field, hence, it may work in other fields like biology.

4.3 Publish everything in blogs

A last improvement would be in the publication of negative results. In data-mining systems there is often the need to compare groups of data. For instance for a microarray this could be a treatment group of affected patients and the control group (reference group) of healthy people. In chapter one we compared the group of PPIs with the group of random protein pairs. There is no database available that explicitly describes that some proteins do not interact. Publication of any experiment ever done would be valuable for a computer scientist (and even for biologists so they do not reinvent the wheel). The publication can now be done via online blogs, which are generally publicly accessible.

4.4 Multidisciplinary environment

A complete non technical aspect of improvement is the communication between different disciplines. The background of today's bioinformatician in most cases is computer science with very little background in biology. In the same way today's biologist lacks the knowledge in the use of computers. The gap in communication between the computer scientist and biologists hampers the further development in bioinformatics research. For instance, biologist and most other disciplines, not engineering related, are sometimes not aware of what is already possible with today's technology. This results in reinventing the wheel or working with old school technology (e.g. massive storage in excel sheets that better could be stored in professional database systems like Oracle and MySQL). Engineers and computer scientists on the other hand have no idea that people are in need of their computer

and engineering skills. When the two worlds never meet they cannot benefit from each others knowledge. We would like to encourage organizations to strive to let biologists meet with bioinformaticians in order to learn from each other. For instance now there are conferences dedicated to bioinformatics research and mostly visited by bioinformaticians. Same holds for conferences mostly oriented for biology. It would be great if a conference was dedicated to present bioinformatics tools and ideas purely to biologists, and that biologists present their ongoing project to bioinformaticians and want feedback or a bioinformatic solution. Such a conference stimulates the increase of collaborations between biologist and bioinformatician.

1. Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
2. Pearson, H., *Biology's name game.* Nature, 2001. **411**(6838): p. 631-2.
3. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., et al., *Overview of BioCreative II gene normalization.* Genome Biol, 2008. **9 Suppl 2**: p. S3.
4. Mons, B., Ashburner, M., Chichester, C., van Mulligen, E., Weeber, M., et al., *Calling on a million minds for community annotation in WikiProteins.* Genome Biol, 2008. **9**(5): p. R89.

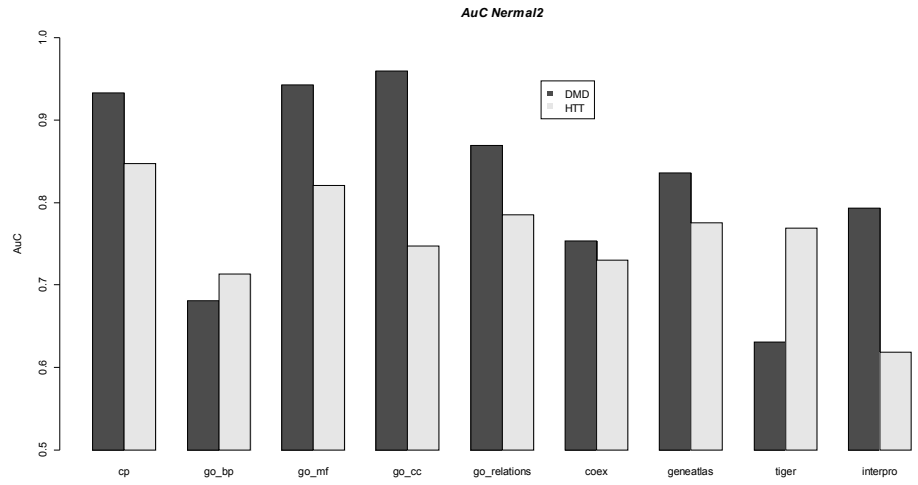


Figure 1. AuC result for DMD and HTT for different databases. The performance is dependent on the protein of interest. Tiger shows opposite behavior then InterPro.

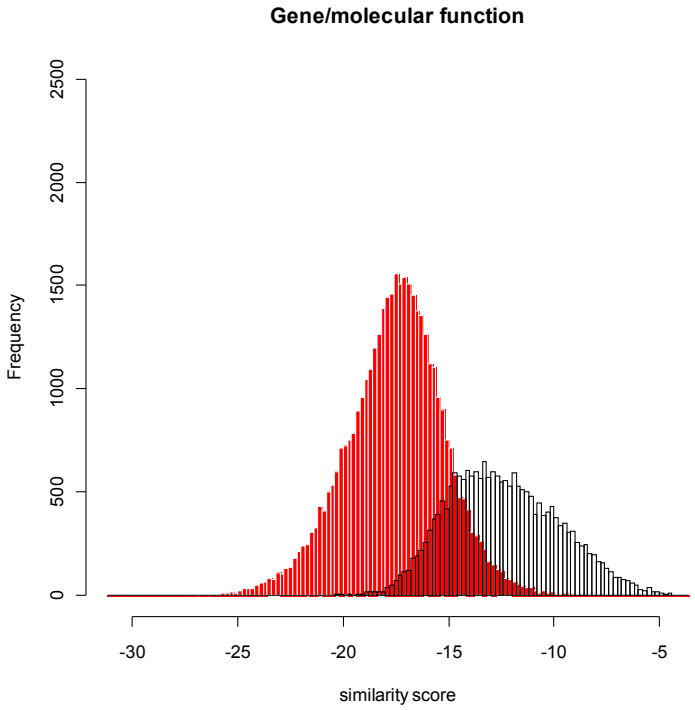


Figure 2. Gene/function distributions

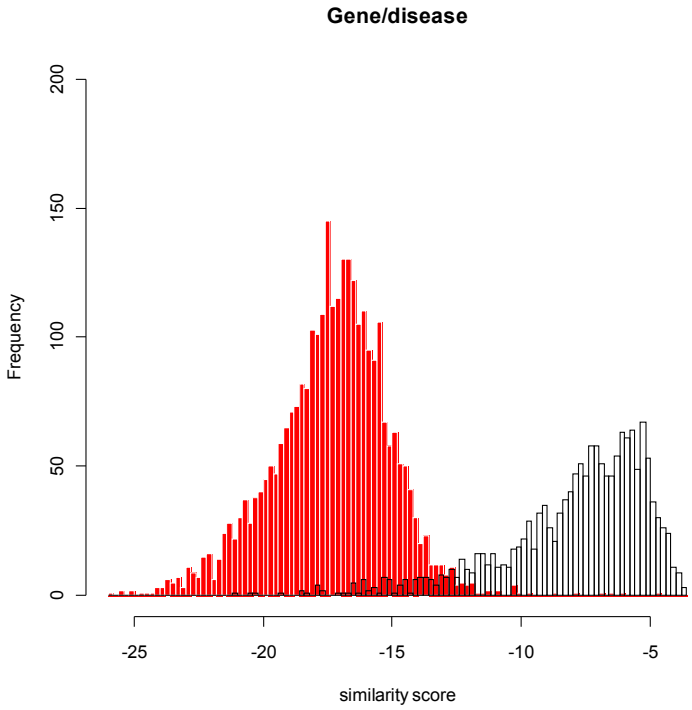


Figure 3. Gene disease distributions

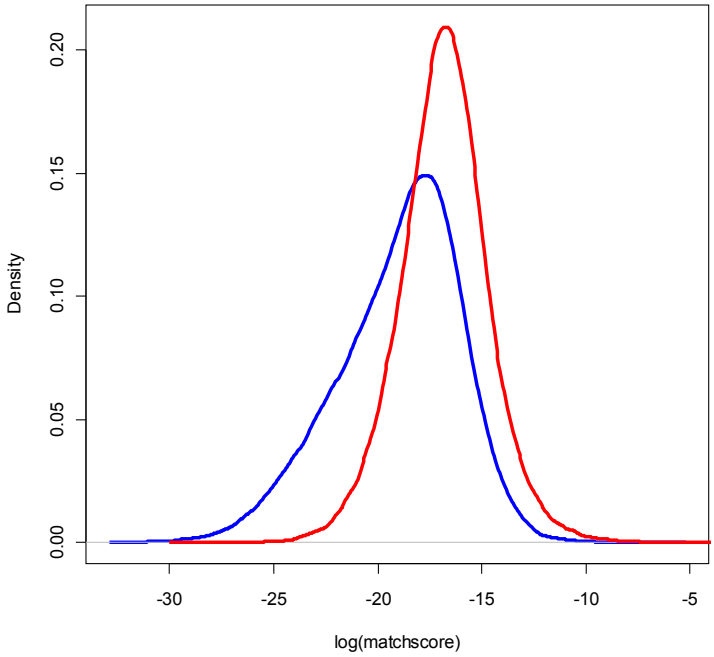


Figure 4. Density plot of random protein pairs (red) and random gene-disease pairs (blue)

Summary

Text-mining is a challenging field of research initially meant for reading large text collections with a computer. Text-mining is useful in summarizing text, searching for the informative documents, and most important to do knowledge discovery. Knowledge discovery is the main subject of this thesis. The hypothesis that knowledge discovery is possible started with the work done by Swanson. He made, as a first finding, links between Raynaud's disease and fish oil using intermediate medical terms to relate them to each other. This principle was formalized in the A-B-C concept. A and C are not directly related to each other but via an intermediate concept B that needs to be discovered.
Tex data can be extended by adding other non textual data such as microarray experiments. Then we are in the field of data-mining. The final goal is to do all kinds of discoveries with computer (in silico) using data sources in order to assist biology research to save time and discover more.

In chapter two we introduced the techniques that are mainly used for the rest of the research. We explained what we mean by concept based text-mining. A concept is an unambiguous unit of thought, meaning that we all agree talking about the same thing. We described how we can define associations between two concepts and the strength of association using statistics. We continued how the direct associations were used to form these A-B-C triplets and used in our text-mining approach called concept profiles. A concept profile is a summary of all concepts related to the main concept stored in a vector with weights. With vector algebra we calculate multiple indirect, A-B-C, links between two concepts. These indirect links we call implicit links. We introduced the statistics used to evaluate our methods such as ROC curves, retrospective analyses, and prioritizers. We concluded what would be the further of text-mining and how it fits into the World Wide Web.

In chapter three we did a large scale analysis on the prediction of protein-protein interactions (PPIs) taken from several protein databases. We compared our concept based text-mining system and concept profiles with MEDLINE, which is word based, and the direct relationship method used by STRING. By direct relationship method we mean that only information is used for two concepts if they co-occur in abstracts, e.g. A-C links, and no C. This is the classical way of doing text-mining. The direct relationship method only detected around 30% of the known PPIs in MEDLINE, concluding that 70% should be detected with indirect or implicit links. The concept profiles outperformed any other method that was based on direct

relations only (Area under ROC curve of 0.90 for concept profiles compared to 0.69 for other methods).
Subsequently we did a retrospective analyses to see if PPIs, added in databases between 2005 and 2007, could be predicted before 2005. Concept profiles showed much better prediction results.
The most interesting result from this analysis was to confirm one prediction in the lab. We made a prioritized list for the protein CAPN3 and predicted PARVB as top candidate with no direct link with CAPN3. It was confirmed with three independent lab experiments that these to proteins physically interact.

We continued on the prediction of PPIs in chapter four. We added five non-textual databases to the text-mining part. This should increase the prediction accuracy and lower the number of false positives. Hence we shifted our analysis from text-mining to data-mining. In this analysis again we used STRING as benchmark. We evaluated different ways to combine data sources. The best method appeared to be Fisher's method for combining single sided p-values. We examined how well our data-mining system was able to predict meaningful protein pairs using three case studies. The first case study was on Dysferlin. This protein showed little information in additional databases and had its information mostly within text. The second case study was on the huntingtin protein. A previous published study of 60 up to 120 putative interaction partners with HTT was used as test data. The prediction of these test samples outperformed that of the STRING method. The last case study on PKD1 showed that adding other databases is also useful for solving homonym problems occurring in text-mining.

In chapter five we switched to another semantic type combination that of the gene-disease and we evaluated how well text-mining is able to predict these kind of relationships. In contrast to the PPI study where we had a large positive set of PPIs, we only evaluated small sets of gene-diseases. This was because it was hard to collect good samples. We generated a new set of 18 known gene-disease pairs known and we used two sets used to evaluate the gene prioritize Endeavour. One contains 10 monogenic diseases and the other six polygenic diseases. We only did roll back analysis to simulate if we could predict these gene-disease pairs. We were able to rank the test gene 2-fold higher than Endeavour on the polygenic diseases.
In this study we delved more into the implicit or indirect links between gene and disease and reasoned if the link was logical. One case study predicted the gene CENPJ when mutated causes Seckel syndrome. Our system was able to rank this gene on position 14 out of more than 12,000 genes before the landmark paper about CENPJ-Seckel syndrome was published.
From these examples and the examples from chapter three is became a burning question how much knowledge can be extracted from text using implicit links, *i.e.*

how much information in the whole of MEDLINE is implicit. We did an analysis on all gene disease pairs and calculated match scores and p-values for each match score. We plotted the p-value against the fraction of explicit links and the fraction of implicit links. We were stunned that for significant scores p-value<0.05 the amount of implicit information already succeeded 80%, concluding that the vast majority of information is implicit.

In chapter six we concluded that implicit information extraction really pays of and that there is far more information in text that we could imagine. However text-mining and data-mining still have their limitations. The best way to solve the shortcomings of the methods is by community annotation. The accuracy of a text-mining system can be increased or even pushed to 100% by manual curation on the internet by millions of users. The ironic thing is that every analysis started *in silico* but ends with the refinement using manual annotation, although it is done by millions of users.

Samenvatting

Text-mining is een uitdagend vakgebied dat oorspronkelijk is bedoeld om grote verzamelingen van tekst documenten te lezen met een computer. Text-mining is nutting voor het samenvatten van tekst, het zoeken naar informatieve documenten, en het meest belangrijke om nieuwe relaties te voorspellen. Het voorspellen van nieuwe relaties is het hoofdonderwerp van dit proefschrift. Deze hypothese van nieuwe kennis extraheren uit tekst begon met het onderzoek verricht door Swanson. Hij heeft als een van de eerste een link (relatie gevonden) gelegd tussen de ziekte van Raynaud en vis olie. De relatie werd gevonden door tussenliggende relaties te gebruiken waardoor ze indirect met elkaar gekoppeld worden. Dit principe werd gemodeleerd als een A-B-C model. A en C zijn niet direct met elkaar in verband gebracht. Maar via het concept B wel (A-B is een relatie en B-C ook). De truuk is om het concept B te vinden en te bepalen of de A-C relatie waar is.

Tekst informatie van worden aangevuldt met ander soort data zoals bijvoorbeeld microarray experimenten. Het uitbreiden van tekst informatie met andere databronnen en dit analyseren op relevante informatie heet data-mining. Text-mining is dus een onderdeel van data-mining. Het hoofdoel van data-mining is om nieuwe relaties te vinden met het gebuik van alle aanwezige databronnen en een computer opdat biologisch onderzoek versneld wordt en ook nieuwe informatie toereikend is.

In hoofdstuk 2 hebben we de technieken uitgelegd die in dit proefschrift zijn gebruikt. We hebben uitgelegd wat wordt bedoeld met concept based text-mining. Een concept is een universieel eenduidige gedachte over een fysiek (tastbaar zoals bijvoorbeeld een fiets) of abstract ding (niet tastbaar zoals bijvoorbeeld de liefde). We hebben besproken hoe we associaties leggen tussen twee concepten en de sterkte van de associatie bepalen met gebruik van statistiek. Vervolgens hebben we uitgelegd hoe deze associaties worden gebruikt in the A-B-C triplets en hoe deze de basis zijn van de *concept profiles* techniek. Een concept profiel is een verzameling van concepten die relateerd zijn met het hoofdconcept van dat profiel en voor elk concept de mate van associatie met het hoofdconcept. Mathematisch gezien is dit een vector met weeggetallen. Met behulp van vector algebra berekenen we meerdere A-B-C relaties tussen twee concepten. Deze indirecte relaties (via concept B) noemen we impliciete relaties.
We hebben statistische methodes omschreven die we gebruiken bij het evalueren van onze text-mining technieken, zoals ROC curves, retrospectieve analyses, en prioritizers. We besloten met wat de toekomst zou zijn van text-mining en hoe dit wordt toegepast binnen het World Wide Web.

In hoofdstuk 3 hebben we een analyse gedaan op grote schaal voor het voorspellen van eiwit-eiwit interacties (PPIs). Deze eiwit-eiwit interacties hebben we genomen uit eiwit databases zoals UniProt. We hebben ons concept based text-mining systeem en de concept profielen vergeleken met het systeem van MEDLINE, dat *word based* is, en vergeleken met de directe relatie methode van STRING.
Met directe relatie methode wordt bedoeld dat alleen informatie wordt gebruikt als twee concepten in hetzelfde abstract voorkomen, bijvoorbeeld A-C zonder een B. Dit is de klassieke manier van text-mining bedrijven. Uit onze resultaten bleek dat de klassieke directe relatie methode slechts 30% van alle bekende PPIs uit databases in MEDLINE kon detecteren. Dit houdt in dat 70% implicit of indirect relateerd zijn. The concept profielen presteerden beter dan welke vorm van directe relatie methode dan ook. Onze methode had een Area under de ROC curve van 0.90 vergeleken met 0.69 van andere methodes.
Vervolgens hebben we een retrospectieve analyse gedaan om te zien of PPIs die aan een database zijn toegevoegd over de periode 2005-2007 konden worden voorspeld met kennis voor 2005. Concept profielen lieten wederom betere resultaten zien dan de klassieke methode.
Het meest interessante resultaat was dat we een van de resultaten gevalideerd hebben in het lab. We hebben een gerangschikte lijst (gerangschikt op hoogste naar laagst associatie) gemaakt voor het eiwit CAPN3. We voorspelde het eiwit PARVB als een top kandidaat die een moleculaire interactie aan zou kunnen gaan met CAPN3. PARVB en CAPN3 hadden tot die tijd geen directe relatie in MEDLINE (noch in STRING). Drie onafhankelijke lab experimenten hebben uiteindelijk bevestigd dat deze twee eiwitten daadwerkelijk een interactie aangaan.

In hoofdstuk 4 zijn we verder gegaan met het voorspellen van eiwit-eiwit interacties. We hebben 5 niet tekstuele databronnen toegevoegd aan het text-mining systeem. De hypothese is dat dit de predictie nauwkeurigheid moet verhogen en het aantal fout positieven moet terugdringen. Deze analyse is dus een data-mining analyse geworden. In deze analyse hebben we wederom STRING als benchmark gebruikt.
We hebben verschillende methodes getest om databronnen met elkaar te combineren. De beste methode die naar voren kwam is Fisher's methode for het combineren van enkelzijdige p-waarden. Om te evalueren hoe goed ons data-mining systeem het deed op het voorspellen van nieuwe relaties hebben we drie casus studies gedaan.
De eerste casus ging over Dysferlin. Dit eiwit liet weining informatie zien in andere databronnen buiten tekst. De tweede casus was voor het huntingtine eiwit. Een eerder gepubliceerde studie naar potentiele eiwit interactie (60 tot 120 stuks) partners werd gebruikt als test set. De voorspellingen van deze testset deed het

velen malen beter dan STRING. De laatste casus was voor het eiwit PKD1. Deze casus liet zien dat het toevoegen van andere databronnen ook het homonym probleem binnen text-mining kunnen oplossen

In hoofdstuk 5 hebben we een ander semantisch type relatie onderzocht. Namelijk de gen-ziekte relatie en we hebben onderzocht hoe goed concept profielen dit type relatie kan voorspellen.
In tegenstelling tot de PPI studie, waar we een grote test dataset hadden, hebben we in deze analyse een zeer kleine set gebruikt van bekende gen-ziekte relaties. De oorzaak hiervan is dat het erg moeilijk is om goede samples te extraheren die geschikt zijn voor testen.
We hebben een testset gegenereerd van 18 bekende gen-ziekte paren en twee testsets die gebruikt zijn bij de studie van de gen prioritizer *Endeavour*. Een van de sets waren monogenetische ziektes en de andere polygenetische ziektes. Voor de analyse deden we alleen een 'terug in de tijd' analyse om te simuleren dat we de gen-ziekte relatie konden voorspellen voordat het expliciet werd in een landmark artikel. Opmerkelijk is dat we het testgen voor de polygenetische ziekten tot 2 keer hoger konden rangschikken dan Endeavour.
In deze studie zijn we ook dieper ingegaan op de impliciete (of indirecte) links en hebben we beredeneerd of deze links een logisch verband vormen met het gen en de ziekte. In een van de case studies hebben we het gen CENPJ voorspeld dat Seckel syndroom veroorzaakt als het gemuteerd is. Dit gen rangschikte op positie 14 van de 12.000 genen. Dit gebeurde al voordat het landmark paper werd gepubliceerd over CENPJ-Seckel syndroom.

Van al deze casussen en de casussen uit hoofdstuk 3 en 4 ontstond er een brandende vraag hoeveel impliciete relaties nog ontdekt kunnen worden uit tekst. Of anders gezegd hoeveel informatie in heel MEDLINE is impliciet? We hebben een analyse gedaan waarbij we voor alle mogelijke gen-ziekte paren een score en p-waarde hebben berekend. We hebben de p-waarde geplot versus de fractie van expliciete relaties met die p-waarde. Zeer opmerkelijk was dat voor significant scorende relaties met eeen p-waarde < 0.05 al bijna 80% van al die relaties impliciet zijn. Dit betekent dat het merendeel van alle informatie in tekst impliciet is.

In hoofdstuk 6 concludeerde we dat door middel van impliciete relaties het mogelijk is om nieuwe relaties te voorspellen en dat er dus veel meer informatie in tekst zit dan men voor mogelijk hield. Text-mining en data-mining hebben echter ook hun limitaties. De beste manier om deze limitaties op te lossen is door 'community annotation'. De nauwkeurigheid van een text-mining systeem kan dan zelfs richting de 100% gaan door het manueel cureren van relaties door miljoenen

gebruikers op het internet. De ironie van dit alles is dat een analyse begon met in silico (alles met de computer) methodes, maar dat aan het einde dit wordt verfijnd met manuele annotatie, daargelaten dat het door miljoenen gebruikers gebeurt.

Curriculum vitae

Herman van Haagen (July 26, 1979) was born and raised in Breda, The Netherlands. In 1998 he moved to Delft to study Electrical engineering at the Delft University of Technology. In 2005 he got his master degree in information and communication theory (ICT). During his master thesis he was introduced in the field of bioinformatics. After his graduation he worked for half a year at the Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital (NKI-AVL). How engineering, and computer science can be applied to biomedical data inspired him to continue working in research. In August 2006 he started his PhD in the field of text- and data-mining at the Leiden University Medical Centre (LUMC). There he was a member of the human genetics department. During his PhD career he collaborated with the biosemantics department at the Erasmus Medical Center in Rotterdam.

Besides the fascination for research Herman has great interest in playing the drums which he started playing at the age of 14.