In silico discoveries for biomedical sciences

Haagen, H. van

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.

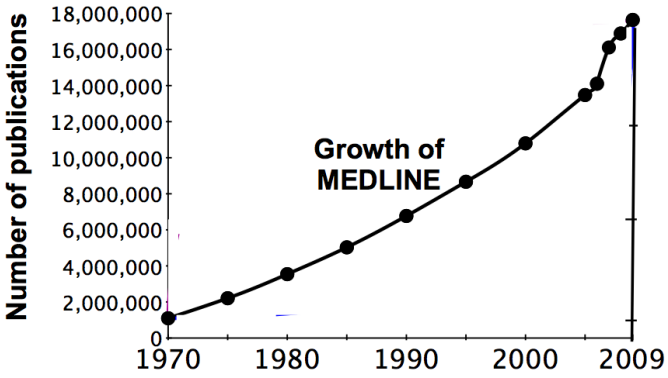


Figure 1. The growth of scientific literature over the last 40 years.

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.

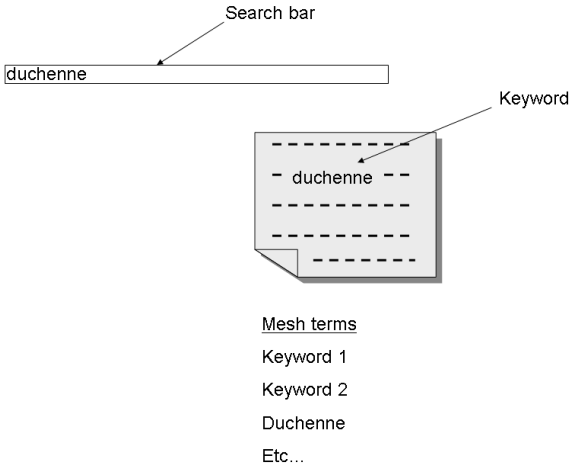


Figure 2. Schematic drawing of a simple text-mining system

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.

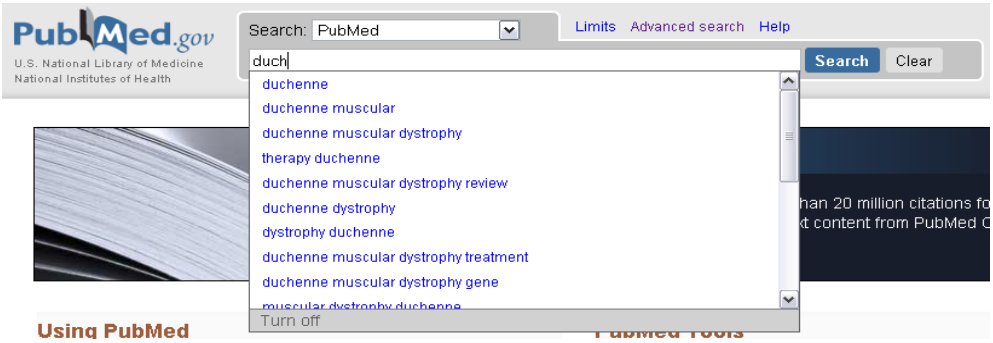


Figure 3. Example for a search query when only the first four letters in a search bar are entered.

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**

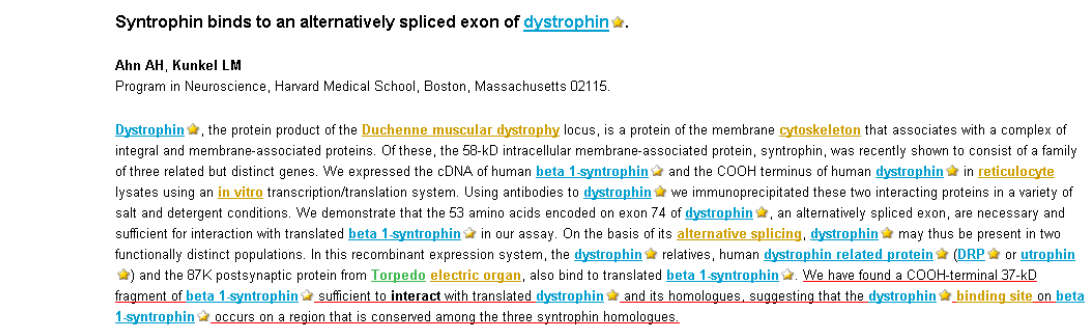
Figure 4. Example of an entry in UMLS for DMD. The field contains descriptions, synonyms and a unique identifier 13264 (last field).

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).

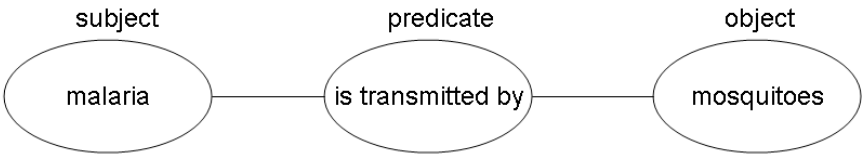


Figure 6. Schematic drawing of a relationship in triplet format.

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.

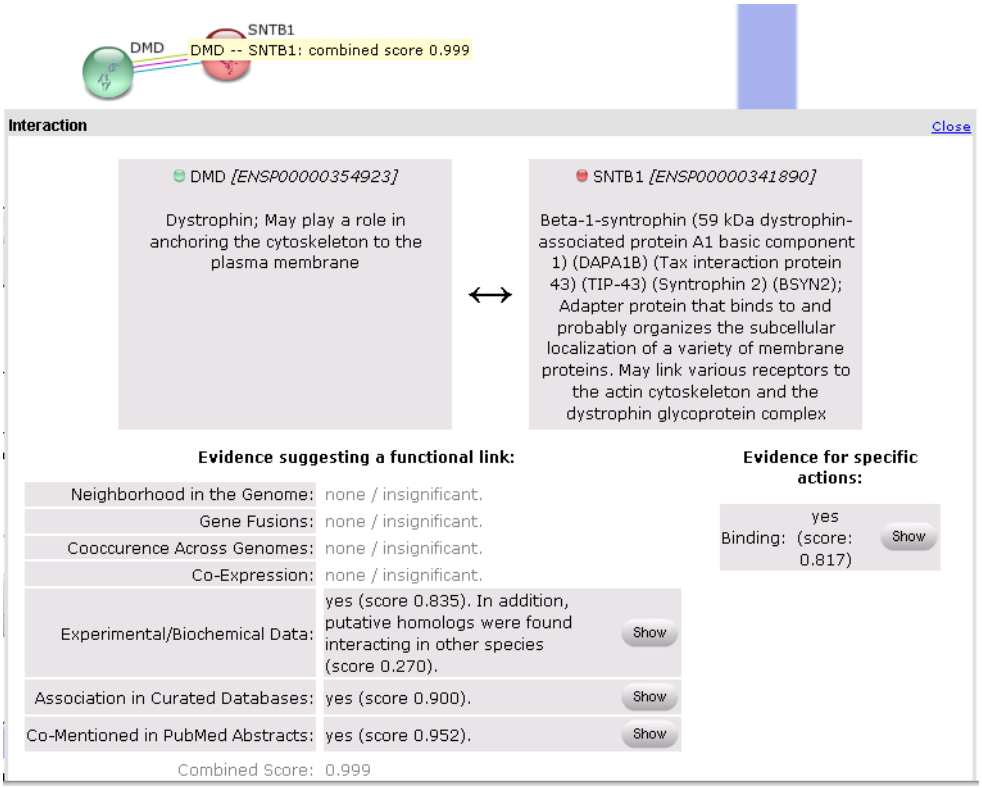


Figure 7. Screenshot of the STRING website. Here the evidence for DMD and STBN1 is mostly found in PubMed abstracts and curated databases.

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.

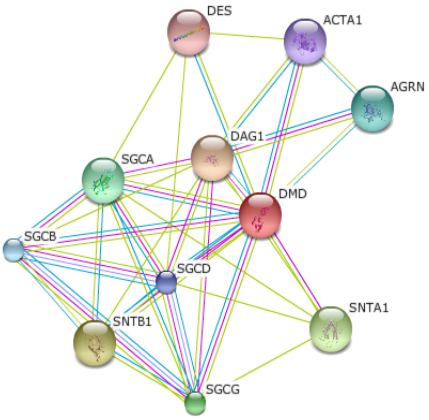


Figure 8. Example of a protein network surrounded around the dystrophin protein (DMD)

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).

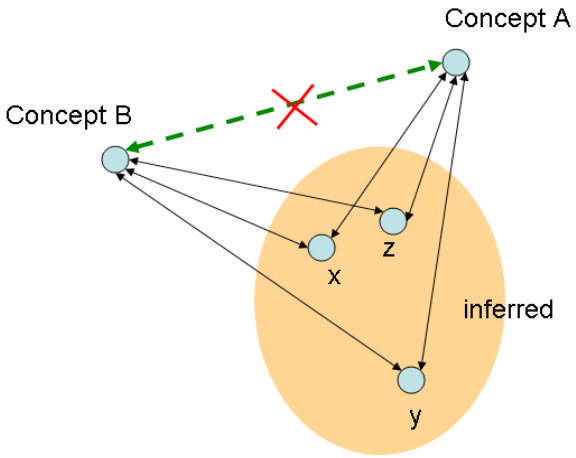


Figure 9. Principle of inferred relationship extraction. Concept A and B are never co-mentioned together in a document. Therefore they do not have a direct relationship. However, via the intermediate concepts X, Y, and Z an indirect relationship can be inferred.

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.