Methods 74 (2015) 97–106

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Application of text mining in the biomedical domain

Wilco W.M. Fleuren [a,b], Wynand Alkema [a,c,*]

[a] Computational Discovery & Design Group CMBI, Radboudumc, Nijmegen, The Netherlands
[b] Netherlands eScience Center, The Netherlands
[c] NIZO Food Research BV, Ede, The Netherlands

ARTICLE INFO

Article history:
Received 8 May 2014
Received in revised form 21 January 2015
Accepted 23 January 2015
Available online 30 January 2015

Keywords:
Text mining
Ontology
Natural language processing
Drug discovery
Biomedical research
Automatic information extraction

ABSTRACT

In recent years the amount of experimental data that is produced in biomedical research and the number of papers that are being published in this field have grown rapidly. In order to keep up to date with developments in their field of interest and to interpret the outcome of experiments in light of all available literature, researchers turn more and more to the use of automated literature mining. As a consequence, text mining tools have evolved considerably in number and quality and nowadays can be used to address a variety of research questions ranging from *de novo* drug target discovery to enhanced biological interpretation of the results from high throughput experiments. In this paper we introduce the most important techniques that are used for a text mining and give an overview of the text mining tools that are currently being used and the type of problems they are typically applied for.

1. Introduction

The scientific literature provides a wealth of information to researchers. It may serve as a starting point for assessing the state of the art in a particular field, or as a source of information that can be used for building research hypotheses that subsequently can be experimentally validated. Additionally this knowledgebase may serve as a source for interpretation of experimental results.

A large number of bibliographic databases is available in the life sciences domain, and have been reviewed by Masic and Milinovic [1]. One of the most important entry points to scientific literature sources for biomedical research is PubMed which gives access to more than 24 million scientific literature citations from MEDLINE, life science journals, and online books [2].

The number of articles that are added to the literature databases is growing fast. Fig. 1 shows the results of a PubMed search using terms that describe diseases, drugs and model organisms. In all cases, the number of papers that have been published on these subjects has increased exponentially. In addition to the exponential growth of the literature databases, the rate at which experimental data are produced has increased as well. For example in high throughput gene expression profiling or proteomics experiments, regulation of hundreds or thousands of genes and proteins is measured under multiple experimental conditions.

Retrieval of relevant information from literature databases and combining this information with experimental output is time consuming and requires careful selection of keywords and drafting of queries. This is often a biased and time consuming process, resulting in incomplete search results, preventing the realization of the full potential that these databases can offer [3].

Automated processing and analysis of text (referred to as text mining (TM)) can assist researchers in evaluating the scientific literature. Nowadays TM is applied to answer many different research questions, ranging from the discovery of drug targets and biomarkers from high throughput experiments [4–9] to drug repositioning, the creation of a state-of-the art overview of a certain disease or therapeutic area and for the creation of domain specific databases [10–15].

Due to the heterogeneous nature of written resources, the automated extraction of relevant biological knowledge is not trivial. As a consequence TM has evolved into a sophisticated and specialized field in the biomedical sciences where text processing and machine learning techniques are combined with mining of biological pathways and gene expression databases.

A number of reviews exists about TM in the biomedical domain that often emphasize the technical aspects of TM and the available tools or focus on gene and protein oriented information and less on the applications and real life research questions that even go beyond gene and protein research [16–19].

Here we give a state of the art overview of the use of TM for the biomedical domain and drug discovery. First we give a general

* Corresponding author at: NIZO Food Research BV, Ede, The Netherlands.
E-mail address: wynand.alkema@nizo.com (W. Alkema).

http://dx.doi.org/10.1016/j.ymeth.2015.01.015
1046-2023/

description of TM, the different steps involved and the types of techniques that are used and describe some publicly available systems for TM. Subsequently we discuss a number of examples in which TM approaches have been applied to solve actual research questions. Finally we present an outlook in which we highlight the opportunities that TM can offer in the near future and the challenges that need to be addressed.

2. Text mining

A widely accepted definition of text mining has been provided by Marti Hearst, as "the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different written resources, to reveal otherwise 'hidden' meanings" [20]. New hypothesis or facts that are the results of TM can subsequently be validated by experiments.

TM analysis typically involves a number of distinct phases, reviewed among others in [17,18,21,22], which are shown in Fig. 2 and described in detail below:

In the last decade a large number of applications have been developed (Table 1) that perform TM at various levels, and implement one or more steps from the scheme shown in Fig. 2. Each application has its own flavor of implementation, often driven by the exact question and the type of answer that is required and the intended user group.

2.1. Information retrieval

The first step in TM is to retrieve relevant textual resources for a given subject of interest. This process is referred to as information retrieval (IR) and is typically done by querying bibliographic databases with a set of keywords. The most used IR system by researchers in the biomedical domain is PubMed [9] that gives access to open source full text articles and abstracts of the MEDLINE database. Besides scientific literature, other literature resources such as patents, medical records, FDAs Medwatch reports, EudraVigilance reports, biomedical related blogs and websites are relevant text resources for biomedical research [1,49–51].

Notably, a lot of TM applications are built on the MEDLINE database, because it is freely available, features a rich applied programming interface and provides annotated abstracts with Medical Subject Heading (MeSH) Terms.

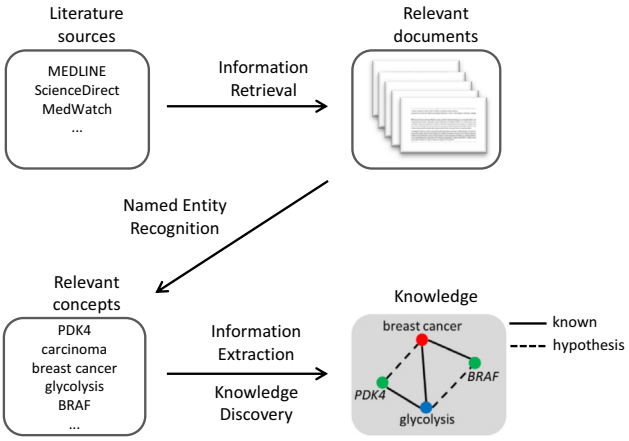

Fig. 2. Overview of a typical TM workflow. A typical TM workflow starts with information retrieval (IR) to get relevant documents for a given subject of interest. Using named entity recognition (NER) these documents will be analyzed for the occurrence of specific keywords. Information extraction (IE) is about detecting links between the found keywords and can be done in a number of ways (see text for more details). During knowledge discovery (KD) links between keywords can be used to infer new relations, so called hidden relations that can be seen as 'true' new knowledge.

A number of TM solutions offer enhanced IR by expanding the queries of the user by organizing similar keywords such as synonyms and alternative names into one concept based on a controlled vocabulary and subsequently incorporating all keywords of the same concept into the query. These tools are marked in Table 1 with the input type 'Concepts'. The use of these extended queries may yield more comprehensive and more specific results. Next to enhancing the IR process, a number of TM tools analyze the results of the query and classify the retrieved documents based on their content or the occurrence of specific keywords in the documents. Sentence extraction, in which only those parts of the document are shown in which specific keywords occur assists the user in focusing on the relevant part of the IR output.

2.2. Named entity recognition

After IR, the resulting document set can be analyzed by search algorithms for the occurrence of specific keywords of interest

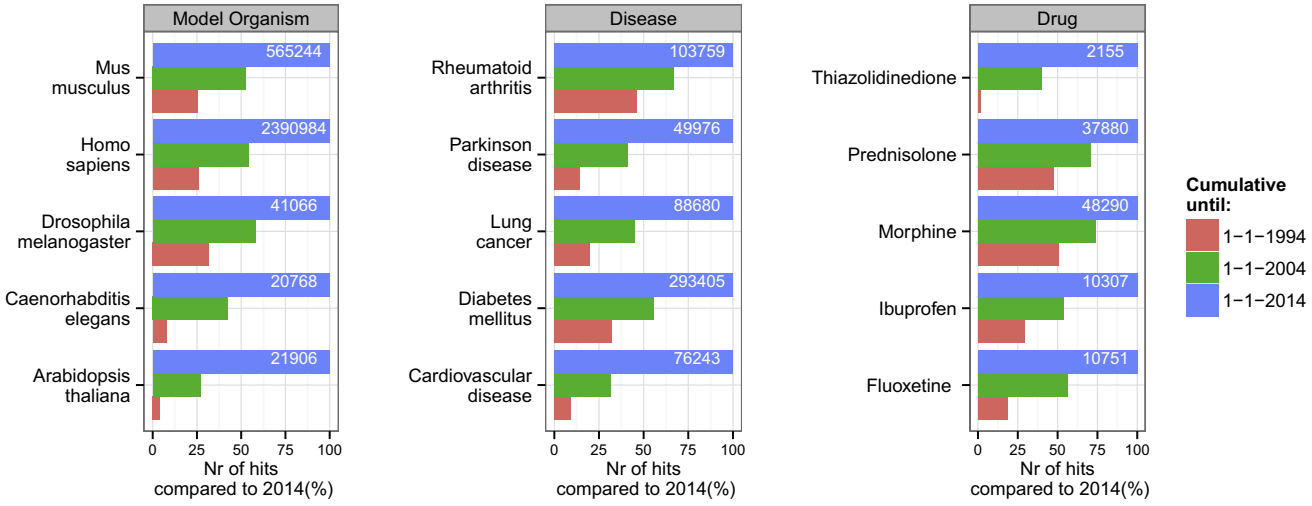

Fig. 1. Growth of the PubMed database. A number of queries for diseases, drugs and model organisms, indicated on the *y*-axis of each graph, were executed. The queries were run with varying cut-off values for the publication date indicated in the legend. The relative number of abstracts retrieved relative to the query with 01-01-2014 as the cut-off date is shown on the *x*-axis. The absolute number of abstracts for the query with 01-01-2014 as the cut-off date is shown on the right side of each group of bars.

Table 1
Overview of text mining applications for the biomedical domain. The table lists the type of information that the tool accepts as input and the type of output that is returned. A 'query' input refers to a standard query with AND and OR keywords that the user can enter. 'Terms' input indicates a list of keywords that can be supplied by the user without the possibility to define AND or OR relations. When a 'concept' input is required, the system directly translates the keyword to an internal set of predefined biological concepts. In that case searches with keywords that are not linked to a concept cannot be done.

Name	Input	Output	Refs.	Tasks	Web link	Description
askMEDLINE	Query	Abstracts	[23,24]	IR	askmedline.nlm.nih.gov/ask/ask.php	Free-text, natural language (English only) query for MEDLINE/PubMed
XplorMed	Query	Ranked abstracts	[24]	IR	xplormed.ogic.ca/	The system provides the main associations between the words in groups of abstracts
eTBLAST	Tekst	Similar abstracts	[25]	IR	etest.vbi.vt.edu/etblast3/	A text-similarity based search engine, using all words in a paragraph to match similar documents
Medline Ranker	Query	Ranked abstracts	[26]	IR	cbdm.mdc-berlin.de/~medlineranker/cms/medline-ranker	Ranks MEDLINE abstracts based on user defined queries
MiSearch	Query	Ranked abstracts	[27]	IR	portal.ncibi.org/gateway/misearch.html	Ranks retrieved articles from PubMed based on a customized personal profile
PICO	Query	Abstracts	[28]	IR	pubmedhh.nlm.nih.gov/nlmd/pico/piconew.php	Search engine for MEDLINE/PubMed with an integrated spelling checker
PubCrawler	Query	Abstracts	[29]	IR	pubcrawler.gen.tcd.ie/	Free alerting service that scans daily updates of the NCBI MEDLINE/PubMed and GenBank databases
PubFocus	Query	Abstracts and statistics	[30]	IR	www.pubfocus.com/	Statistical analysis of the MEDLINE/PubMed search queries enriched with additional information from journal rank database and forward referencing database
PubGet	Query	Abstracts + link to PDFs		IR	pubget.com/	Retrieves PDFs directly based on a user defined query in PubMed
PubMatrix	List of terms	Co-occurrence matrix	[31]	IR	pubmatrix.grc.nia.nih.gov/secure-bin/index.pl	Allows simple text based mining of the NCBI literature search service PubMed using any two lists of keywords terms, resulting in a frequency matrix of term co-occurrence
PubNet	List of terms	Graph	[32]	IR	pubnet.gersteinlab.org/	A flexible system for visualizing literature-derived networks
GeneValorization	Gene list	Abstracts and linked concepts	[33]	IR, NER	bioguide-project.net/gv/start_geneval.php	GeneValorization gives a very clear and handful overview of the bibliography corresponding to user uploaded gene lists
DPWP	Gene list	Linked concepts	[34]	IR, NER	dpwebpage.nia.nih.gov/	Disease/phenotype PAGE is a disease focused gene set analysis web tool to analyze microarray gene expression data with predefined groups of disease related genes
Anne O'Tate	Query	Abstracts and concept statistics	[35]	IR, NER	arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi	Gives an overview of the set of articles retrieved by a PubMed query
Chillibot	Gene list and term list	Graph	[36]	IR, NER	www.chilibot.net/	Chilibot searches PubMed literature database for specific relationships between proteins, genes, or keywords. The results are returned as a graph
ProteinCorral	Protein name	Linked proteins	[37]	IR, NER	www.ebi.ac.uk/Rebholz-srv/pcorral/	Combines information retrieval and extraction from MEDLINE
MEDIE	Tripple	Highlighted phrases		IR, NER	www.nactem.ac.uk/medie/	Retrieve biomedical correlations from MEDLINE, based on indexing by natural language processing and text mining techniques
PubReMiner	Query	Overrepresented concepts	[38]	IR, NER	hgserver2.amc.nl/cgi-bin/miner/miner2.cgi	Breaks down a results of a Pubmed query into categories
PubViz	Concepts	Hyperlinked graphs	[39]	IR, NER	brainarray.mbni.med.umich.edu/	An interactive MEDLINE search engine utilizing external knowledge
Quertle	Query + concepts	Abstracts + overrepresented concepts	[40]	IR, NER, IE	www.quertle.info/	Using a combination of linguistic methods, Quertle finds facts defined within documents, creating its own database of about 300 million relationships, and is able to report the ones that are relevant to your query
CoPub	Concept and gene lists	Abstracts and linked concepts	[9]	IR, NER, IE	www.copub.org	Web application with gene focussed retrieval of co-occurring keywords from MEDLINE

(*continued on next page*)

Table 1 (*continued*)

Name	Input	Output	Refs.	Tasks	Web link	Description
COREMINE medical	Concept	Linked concepts		IR, NER, IE	www.coremine.com/	Presents results about health, medicine and biology in a dashboard format comprised of panels containing various categories of information ranging from introductory sources to the latest scientific articles
FACTA+	Tekst	Abstracts and linked concepts	[41]	IR, NER, IE	www.nactem.ac.uk/facta/	Finds associated concepts with text analysis based on a user query
gopubmed	Concept	Abstracts + overrepresented concepts		IR, NER, IE	www.gopubmed.org/web/gopubmed/	Finds associated concepts with text analysis based on a user query and breaks down query results based on these concepts
iHOP	Protein name	Abstracts and highlighted sentences	[42]	IR, NER, IE	www.ihop-net.org	Information hyperlinked over proteins. Highlights and hyperlinks protein names in text based on user selected protein
PPInterFinder	Abstracts	Linked proteins	[43]	NER	biomining-bu.in/ppinterfinder/html/action.pl	PPInterFinder uses relation keyword co-occurrences with protein names to extract information on protein–protein Interactions from MEDLINE abstracts
Reflect	Tekst	Highlighted text	[44]	NER	reflect.embl.de	Reflect highlights protein and small molecule names, such as IL-5 and rapamycin in text
Whatizit	Tekst	Highlighted text	[45]	NER	www.ebi.ac.uk/webservices/whatizit/info.jsf	Tags concepts in texts based on a number of thesauri
AliBaba	Protein name	Linked terms	[46]	IR	alibaba.informatik.hu-berlin.de/	Parses PubMed abstracts for biological objects and their relations
Martini	Gene lists	Enhanced keywords	[47]	IR	martini.embl.de	Martini uses literature keywords to compare gene sets
STRING	Protein name	Linked genes	[48]	IR	string-db.org/	STRING is a database of known and predicted protein interactions

and statements on the relations between those keywords. An essential step herein is named entity recognition (NER). A named entity is a keyword or a set of keywords that clearly identifies an item or concept. During NER, keywords that are found in the text need to be linked to the specific concept that is being referred to in the document. Concepts can thus be defined as a biological entity that can be referred to by multiple keywords.

This means that for example, a specific gene should be recognized in the text not only by its gene symbol, but also by the synonyms and previous names. Similarly, specific drugs should be recognized in text by reference to the generic substances, the drug trade names and synonyms thereof. In many TM resources, the results of NER are visualized by highlighting the recognized terms and providing links to the originals concepts these terms are referring to. The IHOP server (Table 1) identifies and highlights protein names and MeSH terms in sentences extracted from text and also provides a warning in case of ambiguous protein mappings.

Indeed, one of the biggest challenges of NER in the biomedical domain is the recognition of genes and protein names in scientific text. These are often described using different names and symbols and multiple genes share symbols and names. Results from the gene normalization task of the BioCreative II contest underline this challenge, since none of the participating systems was able to correctly extract all human genes from a set of expert-curated MEDLINE abstracts [52] although it should be noted that also experts only agreed in 90–95% of the cases, setting an upper limit to what can be expected from automated systems [52].

Many TM systems rely on controlled keyword vocabularies, in which keywords that belong together based on for instance the same subject or category, are grouped together for keyword matching in documents. For example, the tool Whatizit (Table 1) allows the user to perform NER in any text with predefined collections of terms for among others genes, proteins, pathways, drugs and diseases [45] that are derived from a variety a research databases.

The text mining application CoPub includes a dedicated vocabulary of liver pathology terms, suitable for the annotation of drug induced gene expression experiments in toxicology studies [9,50]. Next to the controlled vocabularies, a lot of TM tools use the ontologies for NER. Within ontologies, concepts and the keywords that describe these concepts are more formally defined and include relationships and rules that specify the dependencies between concepts. Ontologies are used to formally structure and categorize domain specific information such as information about biological pathways or diseases so it can be for instance used in data mining approaches. Younesi et al. developed a dedicated biomarker ontology for the retrieval of biomarker knowledge from literature with concept classes related to all aspect of biomarker research such as clinical management, diagnosis and prognosis as well as statistics [53]. They used this ontology to retrieve biomarkers for non-small cell lung carcinoma and for neurodegenerative diseases. Ontologies enable query expansion, reformulating a query to improve keyword retrieval performance for instance by exploiting the hierarchical structure of the MeSH descriptors to achieve a significant improvement in image retrieval systems [54].

A number of dedicated ontologies are available in the public domain. The BioPortal for Biomedical Ontologies [55] hosts a large number of these ontologies that can be downloaded or browsed. Many TM applications have implemented ontologies into their workflows to structure their search strategy and visualize and categorize the search results.

Another category of TM systems relies on machine learning based algorithms for NER [56]. Typically, TM system that are based on algorithms such as hidden Markov models (HMM) [57,58], maximum entropy Markov models (MEMM) [59], conditional random fields (CRF) [60,61], and support vector machines (SVM) [62,63],

need to be trained on a carefully constructed annotated training data set that is representative for the real life data set before the actual NER task. Machine learning based TM systems are used for instance to identify chemical entities in text [64], or are used in combination with rule-based and lexical methods to identify organism names in text [65] or used for extraction of cancer staging information from health records to improve clinical decision making [66].

2.3. Information extraction: detection of relationships

After IR and NER, specialized algorithms can be used to detect links between concepts in the text. By linking concepts together, additional context is given to the concepts, which results in valuable knowledge that can be used for downstream analysis. Currently, the most used approaches to extract this knowledge from text are co-occurrence-based methods and natural language processing (NLP) based methods [67].

2.3.1. Co-occurrence-based methods

Co-occurrence based methods are built on the assumption that two concepts that often occur together in the same text are functionally related. For example the co-occurrence of *retinol-binding protein 4 (RBP4)* and *insulin resistance* in MEDLINE abstracts suggests a functional relationship between gene and disease [68,69]. To correct for the fact that many keywords co-occur in text without having a functional relationship, co-occurrence based methods use scoring algorithms based on frequency of occurrence of both keywords, to add a degree of significance to the relationship between the keywords to detect a functional association [70]. Co-occurrence based methods are easy to implement together with NER. They do not, however, provide information with regard to the type of relationship between the keywords and therefore result generally in a higher recall but in a lower precision in comparison with NLP-based methods.

2.3.2. NLP-based methods

NLP is defined as the processing of natural language, i.e. human languages, by computers. NLP methods are based on prior knowledge on how language is structured and on specific knowledge on how biological information is mentioned in the literature. NLP-based methods are phrase based and are able to detect triples in text *e.g. gene A inhibits gene B* or *gene C is involved in disease G* and do, in contrast to co-occurrence based methods, provide information about the type of relationship between two concepts. Therefore phrase-based NLP based methods often have a higher precision in comparison with co-occurrence based methods but are limited to the extraction of concept information for which pre-defined relationships exists. NLP based systems are in general more computationally intensive than co-occurrence based methods. Another disadvantage of some NLP based IE methods is that they these methods are trained for detection of specific relationships on a training set and are thus limited by the availability and quality of the training data and do not scale well when a large number of relations needs to be detected. Open information extraction methods on the other hand do not rely on specific verbs and nouns but focus on how relationships are generically defined in text and can thus in principle extract an unbound number of relations.

In practice, hybrid methods are used in which co-occurrence based methods are used to detect relations between concepts followed by NLP to establish the nature of the relation.

MEDIE (Table 1) is a tool that deploys NLP to detect relationships between concepts. In the interface the user can enter a source and/or target concept and define an interaction of interest, e.g. blocks, binds, causes, regulates. The system then returns statements from MEDLINE abstracts where such a relation has been found. This system can be used to assess whether interactions exists between drugs (sources) and drug targets (targets) in questions such as "what binds to IP-10?". But also more generic terms can be used as targets in questions like "what causes rheumatoid arthritis?".

2.4. Knowledge discovery

If we strictly follow the definition of TM by Hearst then most TM systems can be qualified as information extraction (IE) systems, extracting and ranking relations between concepts that already have been published. Nevertheless, when used in an iterative fashion in which known facts are systematically collected, these IE systems may be suited to infer new relations between keywords based on known facts that are derived from the literature.

2.4.1. The ABC principle

Swanson initiated literature-based knowledge discovery by introducing the ABC-principle. This principle states that keywords A en C, which are never mentioned together in the same text, but are always mentioned with the same set of keywords B, are indirectly related. Swanson applied the ABC-principle in a number of studies, i.e. to infer a beneficial effect of fish-oil on patients suffering from Raynaud's disease, to link magnesium deficiency to migraine and to find out that arginine intake has an effect on levels of somatomedin in blood [71–73].

Since the introduction of the ABC principle by Swanson, a number of TM tools have implemented this principle into their systems. Although the implementation of the ABC principle is conceptually simple, it is not trivial to implement, because rigorous statistics need to be employed to validate new predictions and to estimate the false discovery rate of the predictions.

CoPub discovery is a tool that allows the user to search for hidden relations in a 'open discovery mode' in which the user specifies a target concept (A) and possible intermediate concepts (B) and the system returns all possible relations with other concepts (C). Alternatively the 'closed discovery mode' in which the user specifies keyword A and keyword C can be used to find out whether there are supporting B concepts that associates keyword A to keyword C.

ChemoText is a large repository linking chemicals to diseases based on MEDLINE abstracts. Baker and Hemming used this repository to identify novel compounds for migraine using the ABC concept [74,75].

An extension to the ABC principle is keyword concept profiling. For example a gene concept profile is constructed by listing all its keywords for which associations have been found in the scientific literature, for example based on co-occurrence. By clustering concept profiles using standard multivariate clustering techniques, clusters of related concepts can be found. For example, a drug and a gene that have similar keyword profiles are likely to have a biological relation with each other. Related concepts that cluster together that have never been mentioned together in the same text are especially interesting because they represent new knowledge. The concept profiling technique has been implemented in the tool Anni and was shown to correctly predict cell types and signaling pathways from microarray data [62,76–78].

2.5. Visualization

An important step in TM is representation and visualization of the extracted knowledge to enable fast and correct interpretation of the results and to guide researchers in formulating new hypotheses and initiating follow-up experiments.

In most IR systems, the results of NER are shown by highlighting the recognized concepts in the text and where possible, linking the concepts to data sources where additional information on the concept can be found.

Relationships between keywords can be summarized in tables that can be sorted and filtered based on scoring schemas yielding ranked lists of genes, drugs or diseases which can be used for guiding follow up experiments [51,79,80]. It is important to link the keywords and concepts in these results back to the original text source where they can be highlighted in the text, providing the context in which they are mentioned [9,81].

2.5.1. Literature networks

Most TM applications aim to reveal links between concepts found in the text. Literature networks are therefore an intuitive representation of the results. String, CoPub, PubViz and PubNet (Table 1) are examples of tools that generate hyperlinked graphs that can be used to navigate through literature around a set of concepts. Typically the edges are linked to the underlying literature that connects two concepts, whereas the nodes are linked to databases with additional information about the concepts. A particular useful technology that is used for enrichment of literature network with additional scientific data is the semantic web technology [82]. OpenPhacts is an initiative focused at accelerating drug discovery by connecting clinical, biological and chemical data to pharmacological entities [83]. Leach and coworkers described a system, Hanalyzer that used predefined biological network of more than 8000 mouse genes on which literature findings were mapped using an ontological term definition and connected to multiple gene expression databases [84]. They used this network to analyze a transcriptome data set from mouse craniofacial development. Literature networks allow mapping of links between two concepts into a space in which multiple links between concepts can be visualized. This has the advantage that also indirect links between concepts become apparent, which can give insight into for instance new relations between genes or previously unknown gene disease associations. Moreover, network based clustering tools and graph queries can be used to detect clusters of biologically related genes in these literature networks.

2.5.2. Word clouds

Word clouds are visualizations that display the words that frequently occur in a text. These visualizations are particularly useful when one has no preconceived idea of which concepts should occur in a text. In word clouds, words that appear more frequently in a text are printed in larger fonts than words that occur less often. The principle is illustrated in Fig. 3. To create the word clouds in Fig. 3, all MEDLINE abstracts were obtained for three micro-organisms, *Streptococcus thermophilus*, *Chlamydia trachomatis* and *Mycobacterium tuberculosis*. From these abstracts, common words were removed and the remaining informative words were counted and the most frequent words were displayed in a word cloud [85].

The word clouds provide an insight into the biological properties of these microorganisms and the processes and applications in which they are used. For instance the word cloud of *S. thermophilus* (Fig. 3A) shows occurrences of 'yogurt', 'mozzarella' and 'eps-producing', indicating correctly that *S. thermophilus* plays an important role in the production of dairy products. The keywords in the word cloud in Fig. 3B immediately give insight into the involvement of *C. trachomatis* in chlamydia infection that can lead to 'infertility' and 'salpingitis' (infection and inflammation of the fallopian tubes), and provide clues on how chlamydia is transmitted, e.g. 'sexually', 'intercourse' and how it can be prevented, e.g. 'condom'.

The advantage of word clouds is that this visualization is not biased by the use of a predefined set of concepts or an ontology, but is driven by the raw content of the text. As such they can provide new ideas and insights on a particular concept and can function as a starting point for more specific searches.

3. Application of TM to biomedical problems

The above section describes the steps in a typical TM workflow. Below we discuss a number of examples in which one or more steps from TM workflows have been used to address biomedical questions.

3.1. Genome and gene expression annotation

Genomics, transcriptomics and proteomics have evolved into routinely used techniques in biomedical research. These techniques typically produce lists of genes and proteins that are found in genomes of interest, and expressed under specific experimental conditions. A number of studies describe the use of TM methods for interpretation of these lists of typically hundreds of genes and to assess their functions in pathways, cell types and diseases.

Soldatos et al. [47] used the tool Martini (Table 1) to compare a set of genes associated with a primary cancer versus those associated with the metastatic form of the same cancer. For each list, the associated abstracts were retrieved from MEDLINE and analyzed for biomedical keywords related to among other drugs, chemicals and diseases. The keywords that were overrepresented in the gene list for metastatic cancer were analyzed and indicated that genes

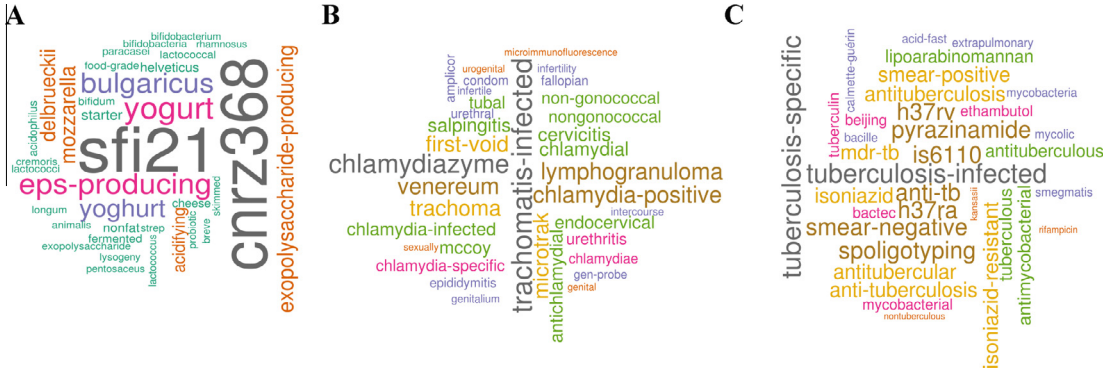

Fig. 3. Word clouds as visualization tools. Word clouds were produced from highly occurring keywords in MEDLINE abstracts about *Streptococcus thermophilus* (A), *Chlamydia trachomatis* (B) and *Mycobacterium tuberculosis* (C). In this visualization, keywords in a larger font have a higher frequency of occurrence than keywords printed in a smaller font.

involved in meiotic cell division in developing sperm might also be involved in metastatic melanoma [86].

Toonen et al. [79] used CoPub (Table 1) to analyze a set of genes that was differentially expressed in blood of healthy volunteers treated with prednisolone. They matched all genes against three keyword categories, i.e. keywords describing glucocorticoid drugs, keywords of metabolic side effects and keywords about inflammatory processes. A number of genes were identified that were not related to inflammation, but showed a clear link to glucocorticoids and metabolic side effects. They proposed that expression of these genes might be a related to the induction of metabolic side effects by glucocorticoids and could be used as biomarkers for these side effects.

Using a similar approach, Hakvoort et al. mapped the murine response to fasting with a combination gene expression and keyword profiling [87]. They used a heat-map representation of over-represented concepts to highlight the biological processes that were regulated during a time course of 72 h and identified neuronal signaling and steroid metabolism as important pathways in this experiment.

The concept of next-generation text-mining was introduced by Hettne et al. They used TM to construct compound–gene associations from the literature, in analogy with the next generation sequencing experiments that experimentally produce potential compound–gene interactions. The resulting gene lists (one gene list per chemical) were then used as input in Gene Set Enrichment Analysis (GSEA) methods for analysis of gene expression data. They showed that the next-generation TM lists were able to correctly predict which chemicals were used in the gene expression profiling experiments and that the next-generation TM lists performed equally well as manually curated gene lists [76].

In general the TM methods for annotating and interpretation work well and give a good overview of the biological processes that are associated with sets of genes and proteins. Most of these methods present direct links to the underlying literature, allowing for easy expert interpretation of the results. Any false positive correlations, arising from e.g. gene name disambiguities are easily noticed. Moreover, since these methods often present aggregate results for an entire gene protein set, false positive relations of some of the members of these sets, do not influence the overall result.

3.2. Drug–target discovery

Text mining has been used extensively in the search for new drug targets and drug candidates. Whereas in drug repositioning information about a known drug and its effects are used to search for new applications, in target discovery the emphasis is on understanding the role of a gene in the onset or progression of a disease. Genes that play a pivotal role in these processes are possible drug target candidates for therapeutic intervention. TM is an established technique in drug discovery to find automatically information in the scientific literature about how these genes are related to the disease and how they are involved in the drug induced effects [50,79,88–90].

Zaravinos et al. used TM to create a literature network of angiogenic components *VEGFA*, *FGF2*, *OPN* and *RHOC* to study their role in urothelial cell carcinoma [91]. Natarajan et al. [92] used text mining on full articles to create networks for genes that were significantly up regulated after treatment of glioblastoma with sphingosine-1-phosphate. Inspection of the networks yielded a novel S1P regulated pathway, involving MMP-9, NRG-1, VEGF and uPA, that could play a role in invasive glioblastoma's.

Novel drug targets can also be obtained by analyzing non-scientific text. A case example was provided by Campillos et al. who used side effect descriptions for a number of drugs to infer sets of drugs with common targets [88]. To this end they classified side effects from drug package inserts using the UMLS ontology. Drugs with similar side effect profiles were then grouped together with their known molecular targets. Novel drug–target relations were then experimentally verified.

To study putative drug targets for coronary heart disease (CHD) and their corresponding interactions, Wu et al. created the knowledgebase CHD@ZJU that records CHD-related information (genes, pathways, drugs) collected from different resources and through text mining followed by manual confirmation [93].

A promising drug target, however, needs to fulfill many more criteria than being disease related. It is important to connect the results of text mining to other information such as druggability of the target, expression in the relevant tissue, knowledge on known side effects, in order to make a balanced judgment on the suitability of a particular protein as a drug target [94].

3.3. Drug repositioning

Drug repositioning or repurposing is the process of identifying and developing new applications for existing drugs, different from the applications for which the drugs originally were developed [95–97]. Drug repositioning can in potential lead to shorter development times, reduce risks and costs of the development. A number of studies have described the use of TM for drug repositioning either with biomedical text as a single source of information [74,75,98,99] or in an integrated approach in which TM was combined with data sets from high throughput experiments [100–103].

Frijters et al. mined MEDLINE abstracts for hidden relations and predicted new relationships between compounds and cell proliferation using genes as intermediate. Two of the found compounds, *dephostatin* and *damnacanthal* were validated and confirmed experimentally in an *in vitro* assay to inhibit cell proliferation [74].

Chiang et al. captured FDA-approved and off-label drug data from the DRUGDEX system and entered them into a Drug–Disease Knowledge Base. They mined this base for diseases with similar therapies using the guilt-by-association method and focused on drugs that were only used for one of these diseases as novel drug suggestions for the other disease. A subset of suggestions was successfully validated by mining clinical trial data to see if they were under evaluation in clinical trials. Using this approach they suggested *rituximab* that is approved for treatment of non-Hodgkin's lymphoma and rheumatoid arthritis, as additional treatment for cataract, gastric ulcer and stomach cancer. Similarly *atorvastatin*, a drug for lowering blood cholesterol, was suggested as an additional treatment for breast cancer, osteosarcoma and Hodgkin's lymphoma [98].

Daminelli et al. introduced an integrated network approach by creating a complex drug–target–disease network based on different types of data and used network motifs to mine this network for drug repositioning. Literature data mining was used to create the network by retrieval of diseases from MeSH terms headers and for validation of the network using the TM application GoPubMed.org [104,105]. Based on this method they suggested repositioning of cardiovascular drugs to parasitic diseases and predicted the cancer-related kinase PIK3CG as a novel target for resveratrol. For an in-depth overview of the use of TM in drug repurposing, the reader is directed to a review of Andronis et al. [106].

The above examples describe successful applications of the ABC principle for the discovery of hidden relations to indeed find truly novel relations. These approaches are prone to produce a large number of potential leads because the number of potential novel relations can quickly explode. Conservative cut-offs in these methods to reduce the false positive rate at the expensive of losing some sensitivity are needed. Moreover, these methods ideally should be

coupled to easily accessible follow up screens to further reduce the number of potential false positives.

3.4. Adverse events

Whereas for drug and drug–target discovery one is mainly interested in establishing relations between the drug and its target and drug efficacy in a disease, it is equally important to explore the possible adverse events that a drug may have.

Hahn et al. reviewed TM efforts on extracting information from pharmacogenomics literature to better understand how human genetic variation impacts drug response [107].

Cheng et al. used TM and data integration to retrieve adverse drug event information from three different sources, i.e. SIDER, CTD, and OFFSIDES, and used analysis with Medical Subject Headings (MeSH) to annotate all compounds and diseases systematically into a database called the MetaADEDB. The database is a useful source for drug discovery to search for known side effects or predict potential side effects for a given drug or compound [108].

Fleuren et al. used keyword co-occurrence methods to automatically identify gene-disease associations in MEDLINE abstracts and created a literature network of genes related to insulin resistance and evaluated the importance of the genes in this network for glucocorticoid induced metabolic side effects and anti-inflammatory processes. Besides known biomarker candidates for glucocorticoid induced insulin resistance they found genes involved in steroid synthesis that were previously not recognized as mediators of GC induced side effects.

Frijters et al. matched gene expression profiles of rat livers treated with various known toxic compounds with keywords from a toxicity thesaurus. Using this method they were able to construct a keyword fingerprint for each compound, i.e. a number of keywords that are most associated with the compound of interest. Analysis of these keywords that gave insight into the biological events and the mode of action of the compounds. A more detailed analysis of the keyword fingerprints of diethylhexylphthalate, dimethylnitrosamine and methapyrilene showed that the differences in the keyword fingerprints of these three compounds are based upon known distinct modes of action [50].

Gurulingappa et al. published a series of papers where they developed TM tools to mine adverse effects from medical case reports. First they created the Adverse Drug Effect (ADE) benchmark corpus, a set of nearly 3000 case reports, manually annotated with 5063 drugs and 5776 conditions [109]. Using this benchmark corpus and an ontology of adverse events, they developed a method to automatically extract adverse events from medical case reports and publicly available data sets on drug adverse effects. With these methods, a number of drug label changes for the drugs rituximab, efalizumab, and natalizumab could successfully be predicted [110,111].

3.5. Electronic health records

Electronic health records (EHRs), electronic patient records (EPRs) or electronic medical records (EMRs) primarily serve as a way to store health information about patients or populations in a sustainable way that enables easy access and exchange of this data by healthcare personnel [112]. However the availability of electronic health information enables secondary use of EHR data. EHRs are a useful resource for the retrieval of information about side-effects and drug interactions of post-market drugs, to potentially establish new patient-stratification principles or for revealing unknown disease correlations. Integrating EHR data with genetic data will also give a better understanding of genotypeã phenotype relationships [113,114].

However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of these data in EHRs and their mining [113]. One of the earliest and most successful research databases that integrated diverse data sources with EHRs is the already mentioned Utah population database (UPDB) [115]. The UPDB earliest success was in the discovery of the breast cancer genes *BRCA1* and *BRCA2*. Because of its linked genealogical and birth certificate information, the UPDB appeared to be a unique resource to search for descendants that have been tested for carrying mutations in cancer genes *BRCA1* or *BRCA2* [116–120].

Lyalina et al. demonstrated that TM and statistical analysis of EMRs of patients suffering from neuropsychiatric illnesses-autism, bipolar disorder, and schizophrenia can be used to extract relevant drugs and phenotypes associated with these neuropsychiatric disorders and characteristic patterns of associations among them. Patient-level analyses suggested a clear separation between autism and the other disorders, while revealing significant overlap between schizophrenia and bipolar disorder. They also enable localization of individual patients within the phenotypic 'landscape' of each disorder [121].

Michelson et al. mined EMRs to detect surgical site infections (SSI) in unstructured clinical notes to improve SSI detection. 22 SSIs detected by traditional hospital-based surveillance were found using TM, along with an additional 37 SSIs not detected by traditional surveillance [122].

Iyer et al. developed a method to mine EHRs for adverse drug reactions caused by drug–drug interactions (DDIs). They used adjusted disproportionality ratios to identify significant drug–drug–event associations among 1165 drugs and 14 adverse events in over 50 million clinical notes [123].

NLP was used to depict the experience of pain in patients with metastatic prostate cancer, from medical records. NLP identified 6387 pain and 13 827 drug mentions in the text of 4409 clinical encounters. Graphical displays revealed the pain 'landscape' described in the textual records and confirmed dramatically increasing levels of pain in the last years of life in all but two patients, all of whom died from metastatic cancer. Heintzelman et al. concluded that tracking longitudinal patterns of pain by TM of free text clinical records may be useful for monitoring pain management and identifying novel cancer phenotypes. [124].

For a more in depth overview of automatic mining of EHRs, the reader is referred to the review of Shivade et al. who reviewed literature describing approaches aimed at automatically identifying patients with a common phenotype, ranging from NLP methods to rule-based system, data mining and statistical analysis [125].

3.6. Domain specific databases

Besides application in drug discovery and genome annotation, TM is used for the creation of domain specific databases. Domain specific databases aggregate and link relevant, domain specific information from a number of general databases and may serve as portals for researchers in these domains. A number of these databases use TM to retrieve and aggregate data from for example MEDLINE [126].

TM can significantly improve the efficiency of construction and maintenance of these databases. In order to successfully apply text mining for building these databases, a good balance needs to be found between the articles that are included and which ones are rejected. A combination of repeated IR steps with manual evaluation of the results by subject matter experts is an essential step in order to fine tune the system such that the desired balance between specificity and sensitivity is achieved. The desired balance is in turn dependent of the aim of the database.

For instance, MeInfoText was created by IE of comprehensive association information about gene methylation and cancer from

large amounts of literature [127], whereas Rodríguez-Penagos et al. used NLP and manual curation of automatic processing of text to validate annotated results and to discover facts and information that might have been overlooked at the triage of curation stages. They successfully applied this technique to generate networks from different sets of documents dealing with regulation in *Escherichia coli* K-12 [128].

The Sjögren's Syndrome Knowledge Base (SSKB) was created by application of TM on PubMed to identify over 7700 abstracts listing approximately 500 potential genes/proteins related to Sjögren's syndrome, a tissue-specific autoimmune disease that affects exocrine tissues [129].

G2Cdb is a neuroscience database aiming to present a global view of the role of synapse proteins in physiology and nervous system related diseases. Here automated TM and expert (human) curation was used to systematically extract information from published neurobiological studies in the fields of synaptic signaling electrophysiology and behavior in knockout and other transgenic mice to annotate experimentally elucidated proteins and genes [130].

Collier et al. used a different application of TM. They developed BioCaster, an ontology-based text mining system for detecting and tracking the distribution of infectious disease outbreaks from linguistic signals on the Web by analyzing documents reported from over 1700 RSS feeds and by plotting the extracted information onto a Google map using geocoded information [131].

The Utah population database (UPDB) that merges demographic information of Utah family histories from over 7 million individuals with diagnostic records about cancer, cause of death, medical details associated with births, claims from statewide inpatient hospital discharge records as well as ambulatory surgery records from hospital outpatient departments and ambulatory surgery centers [115].

4. Outlook

TM pipelines have evolved to a stage where they can be used to efficiently retrieve and analyze the increasing number of research papers in the biomedical field in order to annotate results from high-throughput ∼omics technologies and to support the development of new research hypotheses. However, a number of challenges remain.

4.1. Construction of ontologies

Several ontologies exist in the biomedical domain that have been proven useful in TM pipelines and new ontologies can be constructed by combination of collaborative automated methods and expert curation together with powerful visualization of the results [132–134]. However, more specific ontologies are needed that describe a single disease on various levels. The more refined the ontologies are, the better the literature related to for example a disease can be dissected into the individual pathways and genes, interactions between the genes and their relation to disease progression and phenotypic manifestations.

4.2. From hypotheses to answers

TM workflows are very useful for aiding researchers in hypothesis generation. However, by their nature they potentially produce a lot of data and hypotheses. The added value of the application of (parts of) a TM workflow in research projects or bioinformatics pipelines thus strongly depends the quality and output of the tools. When TM is part of a larger workflow in which predictions by TM are further refined by other bioinformatics methods, TM predictions with a high recall and low specificity (i.e. with a lot of false positive predictions) are allowed and arguably even preferred. If experimental validation experiments on leads derived with TM are planned, TM tools with a high specificity at the expense of recall are needed and direct links to the underlying literature should be available to and researchers in designing their experiments.

If TM tools are used in an IE setting, for example for a first pass interpretation of an ∼omics data set, some false negative results are acceptable. In this case the quality of the tool is dependent on how intuitive the user can browse through the results presented in the form of list, literature networks, document highlighting or word clouds and whether the user has the ability to drill down to the underlying, highlighted, documents to manually evaluate the results.

Given the multitude of TM tools and the wide variety of research questions, the future of TM lies, in our opinion, in "TM on demand" in which TM is used to perform dedicated literature searches, answering specific questions and solving specific problems. This should be done in an iterative process in which biologists, domain experts and data scientists work together as a team, evaluating intermediate TM outcomes and intervening where needed, in order to keep the focus on the required information that can be combined with experimental data to give insight into the biology and that can result in testable hypotheses.

4.3. Additional sources

Given the fact that more data generally beats better algorithms, TM pipelines should be fortified where possible with additional sources of information.

Many open source TM pipelines as well as commercial solutions have been built on scientific peer reviewed papers, and predominantly on abstracts of these papers. With the new initiatives for open access [135] the use of full text into the search results becomes feasible, potentially increasing the scope and level of detail that can be obtained by TM. However, with incorporation of these sources, new methods for evaluating the value of these sources should be developed. Whereas scientific abstracts are information dense and relatively rich in new, experimentally proven facts, full text papers contain introductory statements repeating knowledge in the field and also may contain more speculative statements in the discussion part of the papers. Extraction of new facts from full text papers may thus prove to be challenging.

Similar challenges arise when incorporating text from other sources such as patient records and patient blogs. Also for these sources, techniques for NER and IE should be revalidated and appropriate weighing should be given to information derived from these sources.

Next to integrating additional text sources, TM data should also be amenable for the integration with other biomedical databases to provide a second line of evidence for the findings obtained with TM. Gottlieb et al. presented a nice example of such an approach with DrugRouter, a system for generating drug-specific pathways by combining gene literature data, gene pathway data, drug–response pathways and genome-wide association studies to asses drug responses. These computed pathways suggest novel drug-repositioning opportunities gene-side effect associations, and gene–drug interactions [136].

The exact databases that should be used depend on the question. For example to validate gene–gene relations obtained from the literature, comparison with gene expression, protein–protein interaction or pathway databases can be made. For the discovery of new drug targets, TM results could be augmented with structural information, to select druggable targets.

In interacting with external databases, care should be taken that the input and output is such that the TM tools can easily be incorporated in standard bioinformatics workflows.

If TM tools can meet the above challenges they will continue to be an indispensable asset for researchers in the biomedical domain.

References

[1] I. Masic, K. Milinovic, Acta Inf. Med. 20 (2) (2012) 72–84.
[2] PubMed. Available from: http://www.ncbi.nlm.nih.gov/pubmed.
[3] L.J. Jensen, J. Saric, P. Bork, Nat. Rev. Genet. 7 (2) (2006) 119–129.
[4] C. Plake et al., Nucleic Acids Res. 37 (Web Server issue) (2009) W300–W304.
[5] Z.X. Huang et al., BMC Bioinf. 9 (2008) 308.
[6] A. Kentsis et al., Proteomics Clin. Appl. 3 (9) (2009) 1052–1061.
[7] F. Al-Shahrour et al., Nucleic Acids Res. 35 (Web Server issue) (2007) W91–W96.
[8] A.S. Haqqani et al., J. Proteome Res. 6 (1) (2007) 226–239.
[9] W.W. Fleuren et al., Nucleic Acids Res. 39 (Web Server issue) (2011) W450–W454.
[10] Y. Pan et al., J. Chem. Inf. Model. (2014).
[11] R.A. Abul Seoud, M.S. Mabrouk, Comput. Methods Programs Biomed. 112 (3) (2013) 640–648.
[12] H. Li, C. Liu, Comput. Math. Methods Med. 2012 (2012) 135780.
[13] K. Jensen, G. Panagiotou, I. Kouskoumvekaki, PLoS Comput. Biol. 10 (1) (2014) e1003432.
[14] D. Rebholz-Schuhmann et al., Drug Discov. Today (2013).
[15] D.G. Jamieson et al., Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database, Database (Oxford) 2012 (2012) bas023.
[16] J.J. Kim, D. Rebholz-Schuhmann, Brief. Bioinf. 9 (6) (2008) 452–465.
[17] P. Zweigenbaum et al., Brief. Bioinf. 8 (5) (2007) 358–375.
[18] M. Krallinger, A. Valencia, Genome Biol. 6 (7) (2005) 224.
[19] H. Shatkay, R. Feldman, J. Comput. Biol. 10 (6) (2003) 821–855.
[20] M. Hearst, Proc. Assoc. Comput. Linguist. 37 (1999) 3–10.
[21] S. Ananiadou, D.B. Kell, J. Tsujii, Trends Biotechnol. 24 (12) (2006) 571–579.
[22] L. Hirschman et al., Database (Oxford) 2012 (2012) bas020.
[23] P. Fontelo, F. Liu, M. Ackerman, BMC Med. Inf. Decis. Mak. 5 (2005) 5.
[24] C. Perez-Iratxeta, P. Bork, M.A. Andrade, Trends Biochem. Sci. 26 (9) (2001) 573–575.
[25] J. Lewis et al., Bioinformatics 22 (18) (2006) 2298–2304.
[26] J.F. Fontaine et al., Nucleic Acids Res. 37 (Web Server issue) (2009) W141–W146.
[27] D.J. States et al., Bioinformatics 25 (7) (2009) 974–976.
[28] K.C. Huang et al., J. Biomed. Inf. 46 (5) (2013) 940–946.
[29] K. Hokamp, K.H. Wolfe, Nucleic Acids Res 32 (Web Server issue) (2004) W16–W19.
[30] M.V. Plikus, Z. Zhang, C.M. Chuong, BMC Bioinf. 7 (2006) 424.
[31] K.G. Becker et al., BMC Bioinf. 4 (2003) 61.
[32] S.M. Douglas, G.T. Montelione, M. Gerstein, Genome Biol. 6 (9) (2005) R80.
[33] B. Brancotte et al., Bioinformatics 27 (8) (2011) 1187–1189.
[34] S. De et al., Physiol. Genomics 42A (2) (2010) 162–167.
[35] N.R. Smalheiser, W. Zhou, V.I. Torvik, J. Biomed. Discov. Collab. 3 (2008) 2.
[36] H. Chen, B.M. Sharp, BMC Bioinf. 5 (2004) 147.
[37] C. Li et al., Database (oxford) 2013 (2013) bat030.
[38] R.W. Glynn, M.J. Kerin, K.J. Sweeney, Br. J. Surg. 97 (8) (2010) 1304–1308.
[39] W. Xuan et al., Comput. Syst. Bioinf. Conf. 6 (2007) 359–369.
[40] E. Giglia, Eur. J. Phys. Rehabil. Med. 47 (4) (2011) 687–690.
[41] Y. Tsuruoka et al., Bioinformatics 27 (13) (2011) i111–i119.
[42] J.M. Fernandez, R. Hoffmann, A. Valencia, Nucleic Acids Res. 35 (Web Server issue) (2007) W21–W26.
[43] K. Raja, S. Subramani, J. Natarajan, Database (Oxford) 2013 (2013) bas052.
[44] E. Pafilis et al., Nat. Biotechnol. 27 (6) (2009) 508–510.
[45] D. Rebholz-Schuhmann et al., Bioinformatics 24 (2) (2008) 296–298.
[46] C. Plake et al., Bioinformatics 22 (19) (2006) 2444–2445.
[47] T.G. Soldatos et al., Nucleic Acids Res. 38 (1) (2010) 26–38.
[48] A. Franceschini et al., Nucleic Acids Res. 41 (Database issue) (2013) D808–D815.
[49] M.E. Falagas et al., Arch. Intern. Med. 167 (11) (2007) 1204–1206.
[50] R. Frijters et al., Pharmacogenomics 8 (11) (2007) 1521–1534.
[51] W.W. Fleuren et al., BioData Min. 6 (1) (2013) 2.
[52] A.A. Morgan et al., Genome Biol. 9 (Suppl. 2) (2008) S3.
[53] E. Younesi et al., BMC Med. Inf. Decis. Mak. 12 (2012) 148.
[54] M. Crespo Azcarate, J. Mata Vazquez, M. Mata Lopez, J. Am. Med. Inf. Assoc. 20 (6) (2013) 1014–1020.
[55] P.L. Whetzel, J. Biomed. Semant. 4 (Suppl. 1) (2013) S8.
[56] B. Tang et al., Biomed. Res. Int. 2014 (2014) 240403.
[57] J. Zhang et al., J. Biomed. Inf. 37 (6) (2004) 411–422.
[58] L. Yeganova, L. Smith, W.J. Wilbur, Comput. Biol. Chem. 28 (2) (2004) 97–107.
[59] P. Corbett, A. Copestake, BMC Bioinf. 9 (Suppl. 11) (2008) S4.
[60] M. Skeppstedt et al., J. Biomed. Inf. 49 (2014) 148–158.
[61] L. Li, R. Zhou, D. Huang, Comput. Biol. Chem. 33 (4) (2009) 334–338.
[62] R. Patra, S.K. Saha, Sci. World J. 2013 (2013) 950796.
[63] M.S. Habib, J. Kalita, Int. J. Bioinf. Res. Appl. 6 (2) (2010) 191–208.
[64] S. Eltyeb, N. Salim, J. Cheminf. 6 (2014) 17.
[65] N. Naderi et al., Bioinformatics 27 (19) (2011) 2721–2729.
[66] D. Martinez et al., Artif. Intell. Med. (2014).
[67] K.B. Cohen, L. Hunter, PLoS Comput. Biol. 4 (1) (2008) e20.
[68] N. Mody et al., Am. J. Physiol. Endocrinol. Metab. 294 (4) (2008) E785–E793.
[69] T. Reinehr, B. Stoffel-Wagner, C.L. Roth, J. Clin. Endocrinol. Metab. 93 (6) (2008) 2287–2293.
[70] B.T. Alako et al., BMC Bioinf. 6 (2005) 51.
[71] D.R. Swanson, Perspect. Biol. Med. 30 (1) (1986) 7–18.
[72] D.R. Swanson, Perspect. Biol. Med. 31 (4) (1988) 526–557.
[73] D.R. Swanson, Perspect. Biol. Med. 33 (2) (1990) 157–186.
[74] R. Frijters et al., PLoS Comput. Biol. 6 (9) (2010).
[75] N.C. Baker, B.M. Hemminger, J. Biomed. Inf. 43 (4) (2010) 510–519.
[76] K.M. Hettne et al., BMC Med. Genomics 6 (2013) 2.
[77] R. Jelier et al., Int. J. Med. Inf. 77 (5) (2008) 354–362.
[78] R. Jelier et al., Genome Biol. 9 (6) (2008) R96.
[79] E.J. Toonen et al., Pharmacogenomics 12 (7) (2011) 985–998.
[80] W.W. Fleuren et al., Arch. Physiol. Biochem. 119 (2) (2013) 52–64.
[81] R. Hoffmann, A. Valencia, Bioinformatics 21 (Suppl. 2) (2005) ii252–ii258.
[82] I. Harrow et al., Drug Discov. Today 18 (9–10) (2013) 428–434.
[83] A.J. Williams et al., Drug Discov. Today 17 (21–22) (2012) 1188–1198.
[84] S.M. Leach et al., PLoS Comput. Biol. 5 (3) (2009) e1000215.
[85] W.S.M. Fleuren, J. Boekhorst, J.de Vlieg, W. Alkema, Thesis of Wilco Fleuren: Text mining and information extraction for the lifesciences: an enhanced science approach, 2013.
[86] A.I. Riker et al., BMC Med. Genomics 1 (2008) 13.
[87] T.B. Hakvoort et al., J. Biol. Chem. 286 (18) (2011) 16332–16343.
[88] M. Campillos et al., Science 321 (5886) (2008) 263–266.
[89] D. Hristovski et al., Int. J. Med. Inf. 74 (2–4) (2005) 289–298.
[90] C.A. Trugenberger et al., BMC Bioinf. 14 (2013) 51.
[91] A. Zaravinos et al., Oncol. Rep. 28 (4) (2012) 1159–1166.
[92] J. Natarajan et al., BMC Bioinf. 7 (2006) 373.
[93] L. Wu et al., Database (Oxford) 2013 (2013) bat047.
[94] A.L. Hopkins, C.R. Groom, Nat. Rev. Drug Discov. 1 (9) (2002) 727–730.
[95] T.T. Ashburn, K.B. Thor, Nat. Rev. Drug Discov. 3 (8) (2004) 673–683.
[96] S.L. Kinnings et al., PLoS Comput. Biol. 5 (7) (2009) e1000423.
[97] C. Campas, Drug News Perspect. 22 (2) (2009) 126–128.
[98] A.P. Chiang, A.J. Butte, Clin. Pharmacol. Ther. 86 (5) (2009) 507–510.
[99] J. Li, X. Zhu, J.Y. Chen, Int. J. Data Min. Bioinf. 4 (3) (2010) 241–255.
[100] F. Iorio et al., Proc. Natl. Acad. Sci. U.S.A. 107 (33) (2010) 14621–14626.
[101] J. Scheiber et al., J. Chem. Inf. Model. 49 (2) (2009) 308–317.
[102] G. Hu, P. Agarwal, PLoS ONE 4 (8) (2009) e6536.
[103] E. Kotelnikova et al., J. Bioinf. Comput. Biol. 8 (3) (2010) 593–606.
[104] S. Daminelli et al., Integr. Biol. (Camb.) 4 (7) (2012) 778–788.
[105] A. Doms, M. Schroeder, Nucleic Acids Res. 33 (1) (2005) W783–W786.
[106] C. Andronis et al., Brief. Bioinf. 12 (4) (2011) 357–368.
[107] U. Hahn et al., Brief. Bioinf. 13 (4) (2012) 460–494.
[108] F. Cheng et al., J. Chem. Inf. Model. 53 (4) (2013) 744–752.
[109] H. Gurulingappa et al., J. Biomed. Inf. 45 (5) (2012) 885–892.
[110] H. Gurulingappa et al., Pharmacoepidemiol. Drug Saf. 22 (11) (2013) 1189–1194.
[111] H. Gurulingappa, A. Mateen-Rajput, L. Toldo, J. Biomed. Semant. 3 (1) (2012) 15.
[112] N. Menachemi, T.H. Collum, Risk Manage. Healthc. Policy 4 (2011) 47–55.
[113] P.B. Jensen, L.J. Jensen, S. Brunak, Nat. Rev. Genet. 13 (6) (2012) 395–405.
[114] W.S. Bush et al., Pac. Symp. Biocomput. (2013) 373–384.
[115] Utah Population database (UPDB). Available from: http://healthcare.utah.edu/huntsmancancerinstitute/research/updb/.
[116] D.E. Goldgar et al., Am. J. Hum. Genet. 52 (4) (1993) 743–748.
[117] S.L. Neuhausen et al., Am. J. Hum. Genet. 58 (2) (1996) 271–280.
[118] D.E. Goldgar et al., J. Natl. Cancer Inst. 86 (3) (1994) 200–209.
[119] S.V. Tavtigian et al., Nat. Genet. 12 (3) (1996) 333–337.
[120] K.R. Smith et al., Proc. Biol. Sci. 279 (1732) (2012) 1389–1395.
[121] S. Lyalina et al., J. Am. Med. Inf. Assoc. 20 (e2) (2013) e297–e305.
[122] J.D. Michelson, J.S. Pariseau, W.C. Paganelli, Am. J. Infect. Control (2014).
[123] S.V. Iyer et al., J. Am. Med. Inf. Assoc. (2013).
[124] N.H. Heintzelman et al., J. Am. Med. Inf. Assoc. 20 (5) (2013) 898–905.
[125] C. Shivade et al., J. Am. Med. Inf. Assoc. (2013).
[126] M. Craven, J. Kumlien, Proc. Int. Conf. Intell. Syst. Mol. Biol. (1999) 77–86.
[127] Y.C. Fang, H.C. Huang, H.F. Juan, BMC Bioinf. 9 (2008) 22.
[128] C. Rodriguez-Penagos et al., BMC Bioinf. 8 (2007) 293.
[129] S.U. Gorr et al., BMC Musculoskelet. Disord. 13 (2012) 119.
[130] M.D. Croning et al., Nucleic Acids Res. 37 (Database issue) (2009) D846–D851.
[131] N. Collier et al., Bioinformatics 24 (24) (2008) 2940–2941.
[132] S. Vercruysse, A. Venkatesan, M. Kuiper, BMC Bioinf. 13 (2012) 116.
[133] N.F. Noy et al., Nucleic Acids Res. 37 (Web Server issue) (2009) W170–W173.
[134] C.G. Chute, Yearb. Med. Inf. (2010) 58–63.
[135] R. Van Noorden, Nature 506 (7486) (2014) 17.
[136] A. Gottlieb, R.B. Altman, Clin. Pharmacol. Ther. (2014).