Journal of Biomedical Informatics 46 (2013) 200–211

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Methodological Review

Biomedical text mining and its applications in cancer research

Fei Zhu [a,b], Preecha Patumcharoenpol [c,d], Cheng Zhang [a], Yang Yang [a,b], Jonathan Chan [c], Asawin Meechai [e], Wanwipa Vongsangnak [a], Bairong Shen [a,f,*]

[a] Center for Systems Biology, Soochow University, Suzhou 215006, China
[b] School of Computer Science and Technology, Soochow University, Suzhou 215006, China
[c] School of Information Technology, King Mongkut's University of Technology Thonburi, Thailand
[d] School of Bioresources and Technology, King Mongkut's University of Technology Thonburi, Thailand
[e] Department of Chemical Engineering, King Mongkut's University of Technology Thonburi, Thailand
[f] Institute for Translational Bioinformatics and Systems Medicine, School of Biomedical Informatics, Suzhou University of Science and Technology, Jiangsu 215009, China

ARTICLE INFO

Article history:
Received 28 March 2012
Accepted 30 October 2012
Available online 15 November 2012

Keywords:
Biomedical text
Cancer
Systems biology
Text mining

ABSTRACT

Cancer is a malignant disease that has caused millions of human deaths. Its study has a long history of well over 100 years. There have been an enormous number of publications on cancer research. This integrated but unstructured biomedical text is of great value for cancer diagnostics, treatment, and prevention. The immense body and rapid growth of biomedical text on cancer has led to the appearance of a large number of text mining techniques aimed at extracting novel knowledge from scientific text. Biomedical text mining on cancer research is computationally automatic and high-throughput in nature. However, it is error-prone due to the complexity of natural language processing. In this review, we introduce the basic concepts underlying text mining and examine some frequently used algorithms, tools, and data sets, as well as assessing how much these algorithms have been utilized. We then discuss the current state-of-the-art text mining applications in cancer research and we also provide some resources for cancer text mining. With the development of systems biology, researchers tend to understand complex biomedical systems from a systems biology viewpoint. Thus, the full utilization of text mining to facilitate cancer systems biology research is fast becoming a major concern. To address this issue, we describe the general workflow of text mining in cancer systems biology and each phase of the workflow. We hope that this review can (i) provide a useful overview of the current work of this field; (ii) help researchers to choose text mining tools and datasets; and (iii) highlight how to apply text mining to assist cancer systems biology research.

© 2012 Elsevier Inc. All rights reserved.

Contents

1. Introduction . . . 201
2. Biomedical text mining phases and tasks . . . 201
   2.1. Information retrieval . . . 201
   2.2. Named entity recognition and relation extraction . . . 202
   2.3. Knowledge discovery . . . 203
   2.4. Hypothesis generation . . . 203
3. Data sets and tools for biomedical text mining . . . 203
4. Application of biomedical text mining in cancer research . . . 205
5. Cancer systems biology research with text mining approach . . . 207
   5.1. Workflow of text mining based cancer systems biology research . . . 207
   5.2. Examples of integrated biomedical text mining tools . . . 207
6. Future work and challenges . . . 207
7. Conclusions . . . 208
   Acknowledgments . . . 208
   References . . . 209

* Corresponding author at: Center for Systems Biology, Soochow University, Suzhou 215006, China. Fax: +86 512 65110951.
E-mail address: bairong.shen@suda.edu.cn (B. Shen).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.10.007

1. Introduction

The vast numbers of biomedical text provide a rich source of knowledge for biomedical research. Text mining can help us to mine information and knowledge from a mountain of text and it is now widely applied in biomedical research. As shown in Fig. 1A, the number of publications obtained from PubMed using "text mining" as the query word in the title or abstract has grown substantially since 2000. Many researchers have taken advantage of text mining technology to discover novel knowledge to improve the development of biomedical research, especially those pertaining to malignant diseases, such as cancer.

As a notoriously lethal human disease, cancer caused 7.4 million deaths in 2008 [1]. Thus, cancer is one of the most important study areas for biomedical researchers. It has been widely studied for more than 100 years. The huge body and rapid growth of text on cancer research provides a valuable resource. As can be seen in Fig. 1B, there are many publications on cancer research and the number of publications keeps increasing every year. We searched PubMed with "cancer" in the title or abstract and we retrieved more than 847,000 publications. It is almost impossible for people to read all of these publications and discover new knowledge. Text mining is able to help researchers to complete this difficult task. Realizing the advantages of text mining will facilitate cancer research, by helping to find new knowledge for cancer diagnostics, treatment, and prevention.

Text mining employs many computational technologies, such as machine learning, natural language processing, biostatistics, information technology, and pattern recognition, to find new exciting outcomes hidden in unstructured biomedical text. There are many applications of cancer-related text mining, such as identifying malignant tumor related biomedical mentions (genes, proteins, etc.), finding relationships among biomedical entities (protein–protein, gene–disease, etc.), extracting knowledge from text and generating hypotheses, and constructing or improving pathways. Several review articles for biomedical text mining have been published in past years [2–8], in this review, we pay much attention to the application of text mining in cancer research.

In the following, we introduce the fundamental concepts and tasks of text mining. We address some representative algorithms for each major task in text mining and we discuss at great lengths how far these algorithms have been utilized in biomedical text mining. We then present some state-of-the-art text mining applications and datasets, especially those developed for the genomic era. We review work that has applied text mining techniques to cancer research and some resources for cancer text mining. Finally, we highlight the general workflow of text mining during cancer systems biology and talk about each phase in detail.

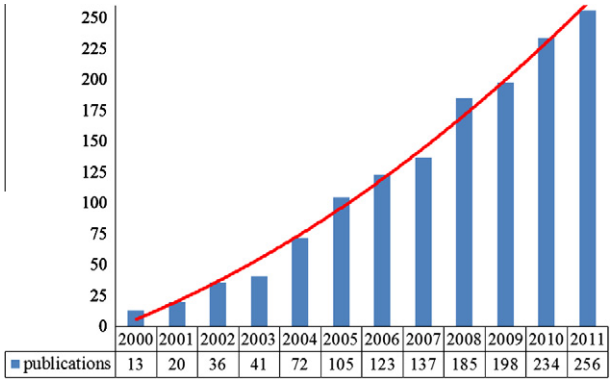

Fig. 1A. The number of publications in PubMed using the query word "text mining" or "literature mining" in the title or abstract. Search detail: text mining [Title/Abstract] OR literature mining [Title/Abstract].

2. Biomedical text mining phases and tasks

The goal of text mining is to derive implicit knowledge that hides in unstructured text and present it in an explicit form. This generally has four phases: information retrieval, information extraction, knowledge discovery, and hypothesis generation. Information retrieval systems aim to get desired text on a certain topic; information extraction systems are used to extract predefined types of information such as relation extraction; knowledge discovery systems help us to extract novel knowledge from text; hypothesis generation systems infer unknown biomedical facts based on text, as shown in Fig. 2. Thus, the general tasks of biomedical text mining include information retrieval, named entity recognition and relation extraction, knowledge discovery and hypothesis generation.

2.1. Information retrieval

Besides conventional information retrieval systems, there are also advanced knowledge information retrieval systems that integrate data from different resources into a single context to enhance our understanding of complex biomedical systems. For example, to access text mining results and other data, Maier et al. [9] generated a chronic obstructive pulmonary disease knowledge base and developed an integrated knowledge management systems. Salivaomics Knowledge Base [10] defined the Saliva Ontology as a terms and relations vocabulary to facilitate data retrieval and integration across multiple fields of research together with data analysis and data mining. QuExT [11], a PubMed-based document retrieval system, followed a concept-oriented query expansion methodology to find documents containing concepts related to

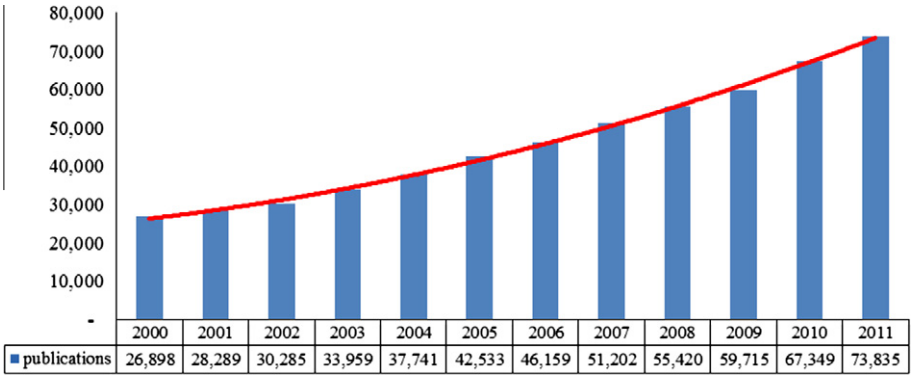

Fig. 1B. The number of publications in PubMed using the query word "cancer" in the title or abstract. Search detail: cancer [Title/Abstract].

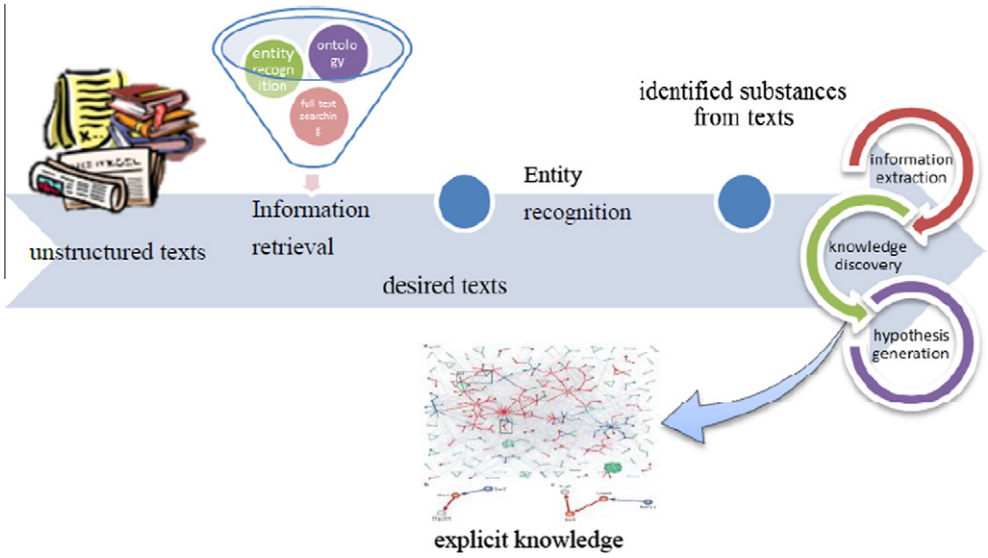

Fig. 2. Conventional phases and tasks involved in biomedical text mining.

query words. In the genome era, with advances in biotechnology and high-throughput methods for gene analysis, there will be a continually growing need for text mining and information retrieval tools to help researchers find relevant articles for their studies.

2.2. Named entity recognition and relation extraction

Named entity recognition is the most important step in the extraction of knowledge [12], which has the overall aim of identifying specific terms, such as gene, protein, disease, and drug. Several technologies in computing have been employed for biomedical term identification. However, in practice, there are still many obstacles for automatically identifying biomedical terms. For example, a biomedical term may have several different written forms, e.g., epilepsy and falling sickness refer to the same disease, which is a disorder of the central nervous system characterized by the loss of consciousness and convulsions [13]. In addition, an entity can be represented differently, e.g., cancer can be represented as a disease as well as an astronomical sign. Moreover, abbreviations of terms can cause ambiguity problems. For example, PC may mean prostate cancer, phosphatidyl choline, or even personal computer. Many biomedical terms also consist of phrases or compound words, or they may have an affix.

Current biomedical named entity recognition technique falls into three major categories: dictionary-based approaches, rule-based approaches and machine learning approaches [2,14]. However, dictionary-based approaches tends to miss undefined terms that are not mentioned in the dictionary [15], rule-based approaches require rules that identify terms from text, and the resulting rules are often not effective in all cases [15]. Machine learning approaches generally require standard annotated training data sets which usually takes tremendous human efforts to build [16]. Moreover, most machine learning approaches tend to be data-driven and application domain-oriented and precision, recall rate, and F_1 rate are often used to evaluate the performance of the recognition, as follows [17]:

$$precision = \frac{number\ of\ true\ positive}{number\ of\ true\ positive + number\ of\ false\ positive} \quad (1)$$

$$recall = \frac{number\ of\ true\ positive}{number\ of\ true\ positive + number\ of\ false\ negative} \quad (2)$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

For example, when we identify a gene term, true positive refers to a gene term is correctly identified as a gene; false positive indicates a non-gene term is incorrectly identified as a gene, and false negative incorrectly identifies a gene term as non-gene term.

Machine learning approaches are now used widely for named entity recognition, e.g., Hidden Markov Models (HMM) [18], Support Vector Machines (SVMs) [19], Conditional Random Fields (CRFs) [20], and Maximum Entropy (ME) [21]. For examples, Zhou et al. [22] used an HMM-based system with biomedical information as domain knowledge to recognize protein, DNA, RNA, cell-type, and cell-line. Kazama et al. [23] used SVMs to identify protein, DNA, cell-type, cell-line, and lipid, with a 73.6% F_1 rate. Tsai et al. [24] developed a CRFs system to extract protein mentions, achieving a 78.4% F_1 rate. Lin et al. [25] used ME to recognize 23 categories of biological terms with a 72% F_1 rate.

Presently, the best F_1 rates for biomedical named entity recognition systems are not as good as the results from general purpose ones [26]. Researchers have tried many methods to improve the performance, by combining different approaches and proposing hybrid approaches [27], conducting post-processing after machine learning, and adding biomedical domain knowledge [28,29]. Some of these applications are discussed in the following section.

A biomedical term may appear in the form of abbreviation and may also have multiple synonyms in text. Abbreviation recognition and synonym recognition are helpful for unifying and normalizing biomedical terms in named entity recognition. There are many such systems. For example, Chang et al. [30] used logistic regression to score abbreviations and obtained an 83% recall rate and 80% precision rate with the Medstract corpus. An abbreviation recognition system was developed in [31], based on a machine learning approach, with a 95.86% precision rate and an 84.64% recall rate with the AB3P corpus. Yu et al. [32] developed a set of pattern-matching rules to map an abbreviation to its full form and achieved a 70% recall rate and a 95% precision rate. Based on collocations, Liu and Friedman's system [33] achieved an 88.5% recall rate and a 96.3% precision rate. McCrae and Collier [34] developed a rule-based synonym recognition system, and the system implemented by Cohen et al. [35] was based on pattern extraction.

More current research is now interested in terms identification and normalization [36]. One of the tasks in BioCreative III is focused on gene normalization, which identifies gene mentions and links these genes to standard identifiers (e.g., database identifiers)

[37]. Such kind of systems are keeping emerging, e.g. the system developed by Liu et al. [38].

Conventional relationship extraction is focused on investigating biomedical relation extraction (e.g., protein–protein interaction and gene–disease relation) from biomedical terms (e.g., genes, proteins, diseases, or drugs) [39]. Many researchers have done much work on relationship extraction. The system developed by Ben Abacha et al. [40] is able to identify the correct semantic relationship between each pair of entities using MetaMap [41] to identify medical substances while a linguistic patterns approach determines the semantic relationship between each pair. The systems developed by Chun et al. [42] could extract gene–disease relations from Medline. They used a machine learning-based named entity recognition system to remove incorrect disease and gene names caused by dictionary matching-based term recognition. They found that improving the terms recognition performance could also improve the relationship extraction precision.

In the current genomic era, many researchers are interested in mining gene–gene interactions, protein–protein interactions, and other interactions in genome-wide associations that provide useful scaffolds for further integrative analysis of gene expression and database annotation [2,43–45], as well as other extensive relationships [16]. Eskin and Agichtein [46] applied text mining technology and combined it with sequence analysis to discover protein subcellular localizations, and the results seemed to be highly accurate. Li et al. [47] took a co-occurrence-based text mining approach to determine interactions from the biomedical literature where they used a naïve Bayesian approach to verify the resulting interactions by integrating heterogeneous types of evidence from genomic and proteomic data sets. The systems developed by Agarwal et al. [17] can be used to determine whether an article is related to protein–protein interactions and to map the interaction to relevant articles. Tsai [48] presented a text mining and visualization framework to find the details of protein–protein interactions and provide a deeper understanding of protein function by identifying the sequence of amino acids at the interface of a protein interaction.

In addition, researchers are focusing on the relationship between genes and other biomedical entities, as well as the relationship between proteins and other biomedical entities, such as gene–disease relationships and protein sub-cellular relationships. For example, the system developed by Krallinger et al. [49] can systematically access information to analyze genetic, cellular, and molecular aspects of the plant *Arabidopsis thaliana*. Srinivasan and Wedemeyer [50] studied the relationship between diseases and disease areas. Shetty and Dalal [51] constructed a statistical document classifier that was based on MEDLINE citations to determine whether a drug had caused adverse effects. Their systems contributed to current drug safety procedure.

2.3. Knowledge discovery

Knowledge including facts, information, or descriptions, implicit or explicit, refers to the theoretical or practical understanding of a domain or a subject [52]. Knowledge discovery is the creation of knowledge from large volumes of structured or unstructured data. The knowledge obtained may become additional data that can be used for further usage and discovery [53]. Knowledge discovery is a very important part of data mining. Text mining, also referred as text data mining, is a branch of data mining that particularly deals with text. Discovering knowledge from biomedical text is a process with the aims to find answers for biomedical questions, such as identifying new drug targets or novel cancer diagnostic biomarkers. The CRAB, a fully integrated text mining tool developed by Korhonen et al. [54], extracted relevant data in literature and assessed cancer risk by utilizing knowledge discovery technologies. Their work demonstrated that text mining pipeline can facilitate complex research tasks in biomedicine. In addition, Nam and Park [55] took advantage of text mining to integrate existed work and discovered two pathways functionally involved in the predictor gene set indicative of susceptibility to early-onset colorectal cancer, overcoming shortage of whole-genome expression studies of colorectal cancer.

Knowledge discovery is able to integrate biomedical text with other multiple sources of data to generate a novel interpretive context [56]. For example, through text mining technology together with microarray data, Urzua et al. [57] found out post-transcriptional control of ovarian processes as possible cause for the observed tumor and reproductive phenotypes. They also inferred that it was repetitive cycling that represented the actual link between ovarian tumorigenesis and reproductive records [57].

2.4. Hypothesis generation

Based on facts or information that cannot be satisfactorily explained with the available knowledge, a scientific hypothesis, which is a trial solution to a problem rather than a theory, can be proposed for suggestion on further research [58]. Experiments may be used to evaluate the proposed hypotheses before solving the problem. Scientific hypothesis is somewhat like a scientific imagination which is based on existing evidence and knowledge. As it says, imagination is so important that it embraces the entire world, and all there ever will be to know and understand. Hypothesis generation is to get unproved inference with clues hidden in the text while knowledge discovery means to extract novel knowledge.

The biomedical literature is a treasure trove of potential information for making biomedical inferences and generating new hypotheses. Hypothesis generation is an important task in text mining, which is very helpful for biomedical researchers who want to infer unknown biomedical facts that can be used to guide the design of experiments or explain existing experimental results. This task is gradually receiving much more attention from researchers. Swanson [59] used a pattern rule to determine a hidden link between fish oil and Raynaud's syndrome in published text. Li et al. [60] built Alzheimer's Disease-specific drug–protein connectivity maps based on protein interaction networks and literature mining. By exploring the Alzheimer's disease connectivity map, they proposed a new hypothesis that diltiazem and quinidine may be investigated as candidate drugs for Alzheimer's disease treatment. Hettne et al. [61,62] used an association-based technique and a natural language processing tool to generate a ranked list of genes associated with diseases and extracted the relations between genes and lipopolysaccharide. Topinka and Shyu [63], taking an biomedical text mining based approach, along with structure-based protein–protein interaction, predicted cancer interaction networks.

3. Data sets and tools for biomedical text mining

In terms of information retrieval systems, PubMed [64,65] is one of the best known biomedical databases and it contains more than 20 million citations on biomedical articles from MEDLINE and life science journals, which provides a convenient web-based search portal for users as well as an application program interface for developers. Textpresso [66,67] uses an ontology, returns searching goals for classes of biological concepts (e.g., gene, allele, cell, or phenotype), classes of relations of objects (e.g., association, regulation), and related descriptions (e.g., biological process). GoPubMed [68,69] classifies literature abstracts according to a Gene Ontology and shows the ontology terms that are related to the query words. In addition, it allows users to explore PubMed search results with an ontology viewer.

Table 1
Some frequently used biomedical named entity recognition systems.

System	Brief introduction
ABNER [130,131]	ABNER is a software tool for molecular biology text analysis. It uses linear-chain conditional random fields approach with orthographic and contextual features
GENIATagger [132,133]	The GENIA tagger is specifically tuned for biomedical text such as MEDLINE abstracts. It is a useful pre-processing tool for information extraction from biomedical documents
LingPipe [134–136]	LingPipe provides three generic, trainable chunkers to carry on named entity recognition. LingPipe can be used to identify biomedical entities such as genes, organisms, malignancies, and chemicals
Yapex [137,138]	Yapex is a rule-based system named entity recognition system that utilizes lexical and syntactic analysis to identify protein names

Table 2
Standard annotated data sets for biomedical named entity recognition.

Corpus name	Brief introduction
Acromine [139,140]	The abbreviation dictionary of Acromine is automatically constructed from the whole MEDLINE. Acromine showed it was quite good then it was applied to the whole MEDLINE
BioLexicon [94]	The BioLexicon brings together terminologies from several large public bioinformatics data resources such as UniProtKb, ChEBI and NCBI. The BioLexicon represents terms in conjunction with lexical and statistical information so as to improve performance of text mining
GENETAG [141,142]	GENETAG is one of the most important standardized standard data sets for biomedical named entity recognition testing. It has 20,000 MEDLINE sentences for gene/protein term identification. 1, 5000 GENETAG sentences were used for the BioCreAtIvE Task 1A Competition
GO [143]	The Gene Ontology (GO) project is a major bioinformatics initiative aiming at standardizing the representation of gene and gene product. GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data

Table 3
Some useful tools for relationship extraction.

System	Brief introduction
BCMS [144,145]	BioCreative MetaServer (BCMS) is a meta-service for information extraction which can generate annotations for PubMed/Medline abstracts, covering gene names, gene IDs, species, and protein–protein interactions
Chilibot [146,147]	Chilibot searches PubMed abstracts about specific relationships between proteins, genes, or keywords. The results are returned as a graph
HPID [148,149]	The Human Protein Interaction Database (HPID) provides human protein interaction information from existing structural and experimental data, and integrated human protein interactions derived from BIND, DIP and HPRD. Users can find potential interaction between with input protein and proteins of the databases. The protein IDs in EMBL, Ensembl, MIM, RefSeq, HPRD and NCBI can be used during interaction search
HPRD [150–152]	The Human Protein Reference Database (HPRD) is a platform for human protein interaction networks and disease association. All the information in HPRD has been manually extracted from the literature by experts. For each in the proteome. HPRD can visually deploy the results
iHOP [70,71,153]	Information Hyperlinked over Proteins (iHOP) can generate a network of concurring genes and proteins from millions of PubMed abstracts. iHOP utilizes genes and proteins as hyperlinks between sentences and abstracts; hence the information can be converted into an integrated navigable resource
IntAct [154,155]	IntAct provides analysis tools for molecular interaction as well as interaction database of which data were derived from literature curation or user submissions
MedScan [156,157]	MedScan collected information and data retrieval from multiple sources of public information, text, journals, and various datasets, and then transformed into biological relationships which could be used for hypothesis generating and verification, disease understanding, drug and patient management
PubGene [158,159]	The retrieve names of gene and protein by PubGene are cross-referenced to each other and to relevant terms with goal of understanding biological function, importance in disease and their relationship
Reactome [160–162]	Reactome is an open-source data analysis tools, as well as a manually curated and peer-reviewed database including interaction, reaction and pathway data. Reactome can be used for interaction, reaction and pathway-based analysis

Table 4
Some standard annotated data sets for relation extraction.

Data set name	Brief introduction
BioInfer [163–165]	BioInfer is a XML-based format corpus protein–protein interaction. The data of BioInfer were from five well-known protein–protein interaction corpora: AIMed, BioInfer, LLL, IEPA, and HPRD50
HIV-1, human PI [166–169]	HIV-1 corpus contains summary of all known interactions of HIV-1 proteins with host cell proteins, other HIV-1 proteins, or proteins from disease organisms associated with HIV/AIDS
LLL 05 [170]	The LLL05 is composed by annotation indicating agent and target of a gene interaction, a dictionary of named entities as well as variants and synonyms, and linguistic information. The LLL05 can be used to evaluate the ability of systems to identify gene/proteins interactions
PICorpus [171,172]	PICorpus is a protein–protein interaction corpus which was originally created at the PDG. PICorpus can be used for a variety of biomedical text mining tasks, such as named entity extraction, relation identification and relation extraction systems
PDZBase [173,174]	PDZBase contains 339 PDZ-domain mediated protein–protein interactions, which have been manually extracted. All the interactions are mediated directly by the PDZ-domain, and identified in vivo or in vitro experiments. The information of the binding-sites of interacting proteins are known.
STRING [175,176]	STRING provides known and predicted protein interactions, including physical and functional associations derived from Genomic context, high-throughput experiments, coexpression and previous knowledge

Table 5
Some commonly used standard annotated data sets for text mining.

Data set name	Brief introduction
BioCreative III [177]	BioCreative III works for evaluating text mining and information extraction systems applied to the biomedical domain. BioCreative III has several data set for three tasks: cross-species gene identification and normalization, protein–protein interactions extraction, and interactive demonstration task for gene indexing and retrieval task
BioInfer [163–165]	BioInfer is a XML-based format corpus protein–protein interaction. The data of BioInfer were from five well-known protein–protein interaction corpora: AIMed, BioInfer, LLL, IEPA, and HPRD50
BioText [178–184]	BioText was initially constructed by 1000 randomly selected MEDLINE abstracts from the results of a query on the term yeast. The dataset was then manually annotated and further verified. BioText has 954 correct pairs, including abbreviation definitions, protein–protein interaction data, and relations between disease treatment entities
GENIA [185,186]	The GENIA data set is one of the most frequently used dataset for evaluation of biomedical and biological information extraction and text mining systems. The data set contains 1999 Medline abstracts, selected using a PubMed query for terms human, blood cells, and transcription factors The GENIA data set has many sub data set, aiming for part-of-Speech annotation, constituency (phrase structure) syntactic annotation, term annotation, event annotation, relation annotation, and coreference annotation
PICorpus [171,172]	PICorpus is a protein–protein interaction corpus which was originally created at the PDG. PICorpus can be used for a variety of biomedical text mining tasks, such as named entity extraction, relation identification and relation extraction systems

For biomedical named entity recognition, there are several powerful systems and data sets. Table 1 provides a list of some useful biomedical named entity recognition systems. Table 2 lists standard data sets that can be used to evaluate the performance of a named entity recognition system or to develop a machine learning-based named entity recognition system. Tables 1 and 2 also have some synonym and abbreviation recognition systems and resources.

There are many useful relationship extraction systems, such as iHOP [70,71], which detects the interactions between genes by using genes or proteins as hyperlinks between sentences and abstracts based on a co-occurrence approach. More relationship extraction systems are shown in Table 3.

To overcome the lack of integration between genomic data and biological literature, Baran et al. [72] developed a tool that linked over 2 million articles in PubMed to nearly 150,000 genes in Ensembl from 50 species. The data set for relationship extraction is also important. Some common data sets are shown in Table 4. Finally, some commonly used standard annotated data sets for text mining purposes are listed in Table 5.

Many hypothesis generation systems are available. BioTextQuest [73] is a biomedical text mining system for concept discovery that provides services such as biomedical named entity recognition, concept association, and hypothesis generation. Arrowsmith [74,75] identified meaningful links between two sets of Medline articles. BITOLA [76–78] can be used to mine new discoveries among biomedical entities or concepts, such as disease candidate gene in the literature.

4. Application of biomedical text mining in cancer research

There is a vast body of work on biomedical text mining in cancer research. In particular, DNA methylation is one of the hottest topics. Methylation profiles have been successfully used for the early detection and personalized treatment of cancer [79,80]. Different databases have been developed for DNA methylation. PubMeth [81] and MeInfoText [79,80] are the two most popular databases in this area. PubMeth [81] is a cancer methylation database with text mining tools and expert annotations. Associations among genes, methylation, and cancers in MeInfoText [79,80] are extracted from a large body of biomedical literature.

As a complex disease, cancer is related to a large number of genes and proteins. Biomedical researchers are interested in mining cancer-related genes and proteins from the literature to study cancer diagnostics, treatment, and prevention. Chun et al. [82] developed maximum entropy-based system that recognizes named entities and the relationships among prostate cancer and relevant genes. Deng et al. [83] employed a text mining approach to identify prostate cancer-related genes as candidate genes and they used the OMIM (Online Mendelian Inheritance in Man) database to verify them. Natarajan et al. [84] also built gene–gene interaction networks for 72 genes using a text mining approach. They discovered novel knowledge in the effect of S1P (sphingosine-1-phosphate) on angiogenesis and the invasion of glioblastoma which contributed to understanding the interaction between invasive glioblastoma and S1P. Krallinger et al. implemented two cancer-related text mining applications [85]. One was used to extract human gene mutations of predefined types of cancer from literatures; the other was particularly used for breast cancer categorization and text-based breast cancer gene ranking. Clancy et al. [86] developed a ranked immunological relevance score for all human genes, which is used to evaluate gene expression profiles and quantify the immunological component of tumors. They applied it to expression profiles in melanomas to find the early activation of the adaptive immune response and the diversity of the immune component during melanoma progression.

Other biomedical interactions are also important areas for cancer researchers. Kolluru et al. [87] used text mining workflows to automatically extract substances from microorganisms and their habitats in free text. They used conditional random fields to extract microorganisms, habitats, and the inter-relationships between organisms and their habitats from the literature. Xu et al. [88] utilized scientific literature from PubMed to extract experimental data on protein phosphorylation. The resulting information proved to be valuable for biomedical researchers studying cellular processes and cancers. PESCADOR, a web-based tool for text mining, can be used to extract network of interactions from PubMed abstracts, and refine the interaction network in accordance with user-defined concepts [89]. It has been applied in the exploring protein aggregation in neurodegenerative disease and in the expansion of pathways associated with colon cancer [89].

It is believed that early detection, evidence-based strategies for prevention and patient management can be used to reduce and control the causes of cancer. Thus, an important part of cancer research is cancer risk assessment, which determines the likelihood of developing cancer by evaluating the available evidence. Korhonen et al. [90] applied biomedical text mining technology to cancer risk assessment. They extracted evidence from the literature as features and developed several classes for risk levels of the causes of cancer, from which researchers can acquire the risk levels. Guo et al. [91] developed classifiers for the automatic identification of schemes from abstracts to help cancer risk assessment.

Clinical records normally include an abundance of information on disease diagnosis and treatment, so it is possible to use them

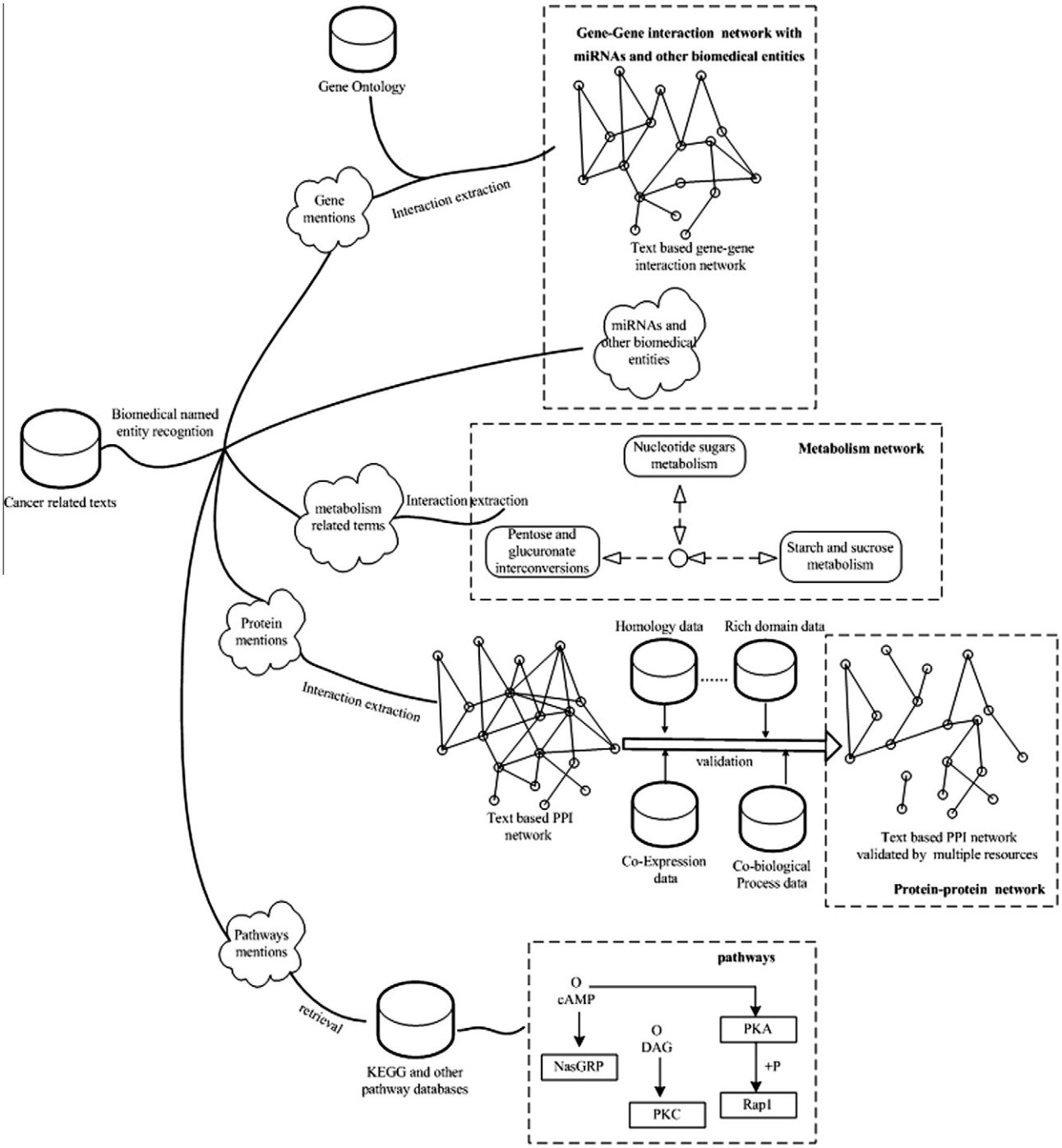

Fig. 3. An illustration of a text mining-assisted cancer study workflow from a systems biology viewpoint.

to facilitate the study of cancer and other potential factors. Ben Abacha et al. [40] used a supervised machine learning approach to extract the relationships among medical problems, treatments, and tests. You et al. [92] proposed a multi-class classifier to analyze diagnostic syndromes in the clinical records to improve experience and techniques. Lee et al. [93] developed a text mining based system to discover the relationships among cancer and potential factors. They mined relationships among diseases and potential factors in clinical medical records using self-organizing maps while they used SVM to evaluate them.

There are several case studies in the context of text mining resources for cancer research, such as the large-scale terminological resources for biomedical text mining provided by Thompson et al. [94] that has over 2.2 million lexical entries and over 1.8 million terminological variants, as well as over 3.3 million semantic relations, including over 2 million synonymy relations. In addition, Thompson et al. [95] also provided an important resource for the training of domain-specific information extraction systems, to facilitate semantic-based searching of documents for interaction extraction. Maqungo et al. [96] developed a database for prostate cancer-associated genes. The database provides pre-compiled biomedical text mining information on prostate cancer and it also integrates data on molecular interactions, pathways, gene ontologies, gene regulation at the molecular level, and predicted transcription factor binding sites in the promoters of prostate cancer implicated genes and transcription factors. The genes and miRNAs

contained in HlungDB, an integrated database of human lung cancer, were collected through text mining [97]. The database also provides lung cancer associated networks for the further investigation of molecular mechanisms of lung cancer.

5. Cancer systems biology research with text mining approach

5.1. Workflow of text mining based cancer systems biology research

Today, researchers tend to understand complex biological systems from a systems biology viewpoint [98]. Systems biology-based networks can be constructed by aggregating previously reported associations from the literature or various databases. For example, Hayasaka et al. [99] constructed a network of genes, genetic diseases, and brain areas based on associations reported in the literature. Sharma et al. [100] collected a set of known disease-related genes and built an interaction network by the mining literature, finding 19 genes that were confirmed to be related to prostate cancer after analysis. Consequently, the full utilization of text mining to facilitate cancer systems biology research is a new hot topic. Generally the conventional flow of text mining based cancer systems biology research is text acquisition, bio-entity terms recognition, complex relation extraction, new knowledge discovery, and hypothesis generation in turn, as showed in Fig. 3. We will discuss this procedure in detail in the following.

In the general phase of text mining of cancer systems biology, we initially obtained related biomedical text from many available sources, such as PubMed. A number of literature databases provide packed data download service. However, although it is convenient, the included text is not timely updated, and text quantity is also limited. Many literature database systems offers application programming interface, by which we can use scripts to download the text automatically by computers. For examples, through E-utility of PubMed [64,101], users can easily get up-to-date text.

Named entity recognition tools can then be used to extract biomedical mentions from the text obtained. The mentions usually include terms such as gene names, protein names, mRNA (message RNA) names, miRNA (micro-RNA) names, metabolism related terms, and cell terms. After finding the biomedical terms, we can build a gene–gene interaction network, metabolism pathways, and other networks. Resources such as Gene Ontology can be used for network building. MicroRNAs are considered to be connected with cancer, so we can investigate how miRNAs work in gene–gene interaction. In the next phase, we can study how components and structures change in dynamic contexts. Certain networks and their variations, such as protein–protein interaction networks [102] and variations in metabolism network, can be built from text. Due to the high false negative rate in text mining-based networks, we can employ some validation and inference algorithms to correct and optimize the network. In each phase, we can use many resources to validate the network, such as homology, co-expression data, rich domain data, and co-biological process data, as well as other information. Through validation, some nodes and interactions with strong evidence will be strengthened, whereas a false one will be removed or updated. Consequently, we can develop a protein–protein interactome based on multiple sources of interaction evidence [47]. Finally, all the networks and components can be used for further studies.

Signaling pathway reconstruction plays a significant role in understand the molecular mechanisms in cancer. Signaling pathway maps are usually obtained from manual literature search, automated text mining, or canonical pathway databases [103]. Pena-Hernandez et al. implemented an extraction tool to find gene relationship and up-to-date pathways from literature [104].

5.2. Examples of integrated biomedical text mining tools

An integrated biomedical text mining systems is supposed to provide the stated functionalities. There are many tools dominated in cancer research. However blindly using the results from text mining tools is not a wise idea because the information and knowledge derived from uncurated text are error prone. Many tools choose to manually curate text by experts. In the following we will briefly introduce the three most popular commercial tools, i.e., Pathway Studio [105], GeneGO [106] and Ingenuity [107].

By Pathway Studio [105], we can analyze pathway, gene regulation networks, protein interaction maps and navigate molecular networks. Its background knowledge database contains more than 100,000 events of regulation, interaction and modification between proteins, cell processes and small molecules. It has a natural language processing module, MedScan, which enables Pathway Studio for entity identification and then applied handcrafted context free grammar (CFG) rules to extract relationships. Pathway Studio can access the entire PubMed database and online resource, full-text journal, literature, experimental and electronic notebooks. Pathways and networks from the extracted facts and interactions extracted from retrieved text. Many algorithms such as Find direct interactions, Find shortest paths, Find common targets or Find common regulators are available.

MetaCore, one of key products of GeneGO [106] is an integrated knowledge database and software suite for pathway analysis of experimental data and gene lists. The knowledge base of MetaCore is manually curated database derived from extensive full-text literature annotation. MetaMiner of GeneGo, mainly including MetaMiner Disease Platforms, MetaMiner Stem Cells, MetaMiner Prostate Cancer, MetaMiner Cystic Fibrosis, offers a knowledge mining and data analysis platforms for oncology. The most important disease reconstruction function is based on three fundamentals, manual annotation of all gene–disease associations, reconstruction of disease pathways and functional data and knowledge mining of OMICs experimental studies published in a disease area. GeneGo also provides API for third party software development.

Ingenuity [107] helps researchers model, analyze, and understand the complex biomedical, biological and chemical systems by integrating data from a variety of experimental platforms. One application example of Ingenuity Systems is analysis of CD44hi breast cancer stem cell-like subpopulations using Ingenuity iReport. The base knowledge of Ingenuity is also extracted by experts from the full text of the scientific literature, including findings about genes, drugs, biomarkers, chemicals, cellular and disease processes, and signaling and metabolic pathways. Researchers can search the scientific literature and find insights most relevant to the desired experimental model or question, build dynamic pathway models, and get confidence in hypotheses and conclusions.

6. Future work and challenges

With the development of the next-generation sequencing technologies, high throughput experimental methods are revolutionizing the life sciences rapidly. The widespread of the cloud computing application is also accelerating the application of text mining technology in the frontier research in life science. We here discuss the work and challenges in the future application of text mining in cancer researches as follows.

The first challenge is to apply biomedical text mining technologies in the personalized medicine development. It is well-known that cancer is a complex disease. Many factors such as race, gender, age and environments may correlate with risk of cancer [108–114]. The personalized medicine is becoming a trend and the therapies will be tailored to individual patients with their biomedical

information collected and analyzed. Ando et al. have applied the text mining technique to qualitatively identify the differences in the focus of life review interviews by patient's age, gender, disease age and stage [115]. Ahmed et al. integrated compound–target relationships related with cancer by text mining and presented the spectrum of research on personalized medicine and compound–target interactions [116]. The personalized medicine in cancer will take in all these important aspects into consideration during text mining [117]. One solution is to categorize data before text mining rather than treat them together without any preprocessing. It is a really tough task to categorize data at individual level features. On the other hand, one of the negative consequence of categorization is making it harder for text mining to find a good biomarker for all cases.

The second challenge is the complex of cancer molecular mechanisms. The same cancer phenotype could be caused by different gene or gene sets from the same pathway or network. To study the complex mechanisms of cancer, we need to mine text from a hierarchical network view rather than from a single level. Systems biomedicine carries on analysis and study from different levels, including motif [118,119], pathway [120–122], module [123–125] and network [126,127]. The resulting hierarchical data provide us valuable materials to conduct text mining on different levels. However, how to correctly categorize text to hierarchical network, and how to integrate text mining results from different levels and discover new knowledge with a systems biomedicine view are really a hard work.

The third challenge is to apply the text mining techniques in translational medicine research. Translational medicine, an emerging field of biomedicine, involves the transformation of laboratory findings into novel diagnosis and treatment of patients [128]. The knowledge of pre-clinical can be used in clinic to improve treatment. Translational medicine facilitates the course of diseases predicting, preventing, diagnosing, and treating. Bioinformatics will be a driver rather than a passenger for translational biomedical research [128], such as the data integration and data mining platform presented by Liekens et al. [129] could retrospectively confirm recently discovered disease genes and identify potential susceptibility genes. It will add tough tasks for text mining, since translation biomedical text mining should consider various stages of information and various sources of evidence, and integrate the Omics and clinical data sets to find out novel knowledge for both biology and medicine domains. There are many this kind of applications, such as the data integration and data mining platform presented by Liekens et al. [129] could retrospectively confirm recently discovered disease genes and identify potential susceptibility genes.

The fourth challenge for text mining will be the integration of the text information at molecule, cell, tissue, organ, individual and even population levels to understand the complex biological systems. Nevertheless, most of the current text mining studies focus on molecular level, and very little text mining work reported at high levels, which in fact has a close relationship with cancer phenotypes. Text mining at high levels and integrate the text information at all these levels will be a big challenge for cancer study and provide also opportunities for successful cancer diagnosis and treatments.

The last challenge will be the de-noising and testing of the text mining results. Text mining results are often obtained with noising information and false positives since natural language text are often inconsistent. It contains ambiguities caused by semantics, slang and syntax. It can be also suffered from noise and error in text. As a result, the mined information cannot be used blindly. Many methods have been developed to solve the problem. The first is to manually read and understand the contexts, analyze them, and then add semantic tags. This pre-processing in fact turns the unstructured text into structured text with semantic tags. Thereby, the developed tools can easily achieve the goal with high precision rate. However, the approach is very restricted as it needs vast human efforts and turns out to be very time consuming. As a result, the data source for mining could be modest in size, only limiting mining ability. The second method is to carry on text mining on vast biomedical text, and then analyze the results and screen out the final results with prior domain experience. During the mining process, domain knowledge is usually employed to improve mining efficiency as well as the quality of the mined knowledge. This approach although the mined results may still contain more errors, is more powerful on knowledge discovering compared with the first approach. These two approaches are distinct on treating the text to be mined. The first one ensures correctness by carefully manual pre-processing, while the second one is to select correct ones by post-processing by experts. The third approach is to take a compromise between pre-processing and post-processing, where some advanced statistical analysis will be used to roughly clean data at first stage and then conduct mining on them.

7. Conclusions

Currently, there is a huge body of biomedical text and their rapid growth makes it impossible for researchers to address the information manually. Researchers can use biomedical text mining to discover new knowledge. We have reviewed the important research issues related to text mining in the biomedical field. We also provided a review of the state-of-the-art applications and datasets used for text mining in cancer research, thereby providing researchers with the necessary resources to apply or develop text mining tools in their research. We introduced the general workflow of text mining to support cancer systems biology and we illustrated each phase in detail. We can see that text mining has been used widely in cancer research. However, to fully utilize text mining, it is still necessary to develop new methods for full text mining and for highly complex text, as well as platforms for integrating other biomedical knowledge bases.

In spite of the huge potential of applying text mining on biomedicine, it still needs further development. Biomedical text mining systems are not as golden standard tools of biomedical researchers as retrieval systems and sequencing tools. The next important mission of text mining for us is to develop applications that are really helpful to biomedical research, so that researchers can get more productive and make more progress in the information rapid growing ear. To achieve the goal, more concerns should be put on helping biological biomedical scientists to remove the obstacles that block the development rather than discussions that are not related with actual demands. One of the hottest topics of text mining is to coordinate and cooperate with multiple subjects. That is, biomedical text mining, coupled with other data and means, should yield consistent, measurable, and testable results.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (91230117, 31170795), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20113201110015). International S&T Cooperation Program of Suzhou (SH201120) and the National High Technology Research and Development Program of China (863 program, Grant No. 2012AA02A601).

References

[1] World Health Organization. World health statistics 2009. Geneva: World Health Organization; 2009.
[2] Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005;6:57–71.

[3] Scherf M, Epple A, Werner T. The next generation of literature analysis: integration of genomic analysis into text mining. Brief Bioinform 2005;6:287–97.
[4] Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform 2005;6:239–51.
[5] Winnenburg R, Wachter T, Plake C, Doms A, Schroeder M. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? Brief Bioinform 2008;9:466–78.
[6] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform 2007;8:358–75.
[7] Ananiadou S, Kell DB, Tsujii J. Text mining and its potential applications in systems biology. Trends Biotechnol 2006;24:571–9.
[8] Ananiadou S, Pyysalo S, Tsujii J, Kell DB. Event extraction for systems biology by text mining the literature. Trends Biotechnol 2010;28:381–90.
[9] Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, et al. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. BMC Syst Biol 2011;5:38.
[10] Ai J, Smith B, Wong DT. Saliva Ontology: an ontology-based framework for a Salivaomics Knowledge Base. BMC Bioinformatics 2010;11:302.
[11] Matos S, Arrais JP, Maia-Rodrigues J, Oliveira JL. Concept-based query expansion for retrieving gene related publications from MEDLINE. BMC Bioinformatics 2010;11:212.
[12] Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform 2005;6:357–69.
[13] Dagar A, Chandra PS, Chaudhary K, Avnish C, Bal CS, Gaikwad S, et al. Epilepsy surgery in a pediatric population: a retrospective study of 129 children from a tertiary care hospital in a developing country along with assessment of quality of life. Pediatr Neurosurg 2011;47(3):186–93.
[14] Li L, Zhou R, Huang D. Two-phase biomedical named entity recognition using CRFs. Comput Biol Chem 2009;33:334–8.
[15] Rebholz-Schuhmann D, Yepes AJ, Li C, Kafkas S, Lewin I, Kang N, et al. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. J Biomed Semantics 2011;2(Suppl. 5):S11.
[16] Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-Aryamontri A, Winter A, et al. The Protein–Protein Interaction tasks of BioCreative III: Classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinformatics 2011;12(Suppl. 8):S3.
[17] Agarwal S, Liu F, Yu H. Simple and efficient machine learning frameworks for identifying protein–protein interaction relevant articles and experimental methods used to study the interactions. BMC Bioinformatics 2011;12(Suppl. 8):S10.
[18] Ephraim YMN. Hidden Markov processes. IEEE Trans Inform Theory 2002;48:1518–69.
[19] Habib MS, Kalita J. Scalable biomedical Named Entity Recognition: investigation of a database-supported SVM approach. Int J Bioinform Res Appl 2010;6:191–208.
[20] He Y, Kayaalp M. Biological entity recognition with conditional random fields. In: AMIA annu symp proc; 2008. p. 293–7.
[21] Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inform 2009;42:905–11.
[22] Zhou GD, Su J. Exploring deep knowledge resources in biomedical name recognition. In: JNLPBA; 2004. p. 96–99
[23] Kazama J, Makino T, Ohta Y, Tsujii J. Tuning support vector machines for biomedical named entity recognition. In: Association for computational linguistics Morristown, NJ, USA; 2002. p. 1–8.
[24] Tsai T, Chou WC, Wu SH, Sung TY, Hsiang J, Hsu WL. Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. Expert Syst Appl 2006;30:117–28.
[25] Lin YF, Tsai TH, Chou WC, Wu KP, Sung TY, Hsu WL. A maximum entropy approach to biomedical named entity recognition. In: The 4th ACM SIGKDD workshop on data mining in bioinformatics; 2004. p. 56–61.
[26] Yen-Ching CR, Tzong-Han Tsai, Wen-Lian Hsu. New challenges for biological text-mining in the next decade. J Comput Sci Technol 2010;25:169–79.
[27] Fei Zhu BS. Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing. PLoS One 2012;7(6):e39230.
[28] Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S. How to make the most of NE dictionaries in statistical NER. BMC Bioinformatics 2008;9(Suppl. 11):S5.
[29] Zhou GDaJS. Exploring deep knowledge resources in biomedical name recognition. In: JNLPBA; 2004.
[30] Chang JT, Schutze H, Altman RB. Creating an online dictionary of abbreviations from MEDLINE. J Am Med Inform Assoc 2002;9:612–20.
[31] Kuo CJ, Ling MH, Lin KT, Hsu CN. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. BMC Bioinformatics 2009;10(Suppl. 15):S7.
[32] Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc 2002;9:262–72.
[33] Liu H, Friedman C. Mining terminological knowledge in large biomedical corpora. Pac Symp Biocomput 2003:415–26.
[34] McCrae J, Collier N. Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC Bioinformatics 2008;9:159.
[35] Cohen AM, Hersh WR, Dubay C, Spackman K. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC Bioinformatics 2005;6:103.
[36] Zhiyong Lu H-YK, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, et al. The gene normalization task in BioCreative III. BMC Bioinformatics 2011;12.
[37] Arighi CN, Roberts PM, Agarwal S, Bhattacharya S, Cesareni G, Chatr-Aryamontri A, et al. BioCreative III interactive task: an overview. BMC Bioinformatics 2011;12(Suppl. 8):S4.
[38] Huang M, Liu J, Zhu X. GeneTUKit: a software for document-level gene normalization. Bioinformatics 2011;27:1032–3.
[39] Arighi CN, Lu Z, Krallinger M, Cohen KB, Wilbur WJ, Valencia A, et al. Overview of the BioCreative III workshop. BMC Bioinformatics 2011;12(Suppl. 8):S1.
[40] Ben Abacha A, Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomed Semantics 2011;2(Suppl. 5):S4.
[41] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17:229–36.
[42] Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, et al. Extraction of gene–disease relations from Medline using domain dictionaries and machine learning. In: Citeseer; 2006. p. 4–15.
[43] Wren JD, Garner HR. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. Bioinformatics 2004;20:191–8.
[44] Raychaudhuri S, Schutze H, Altman RB. Using text analysis to identify functionally coherent gene groups. Genome Res 2002;12:1582–90.
[45] Raychaudhuri S, Altman RB. A literature-based method for assessing the functional coherence of a gene group. Bioinformatics 2003;19:396–401.
[46] Eskin E, Agichtein E. Combining text mining and sequence analysis to discover protein functional regions. Pac Symp Biocomput 2004:288–99.
[47] Li X, Cai H, Xu J, Ying S, Zhang Y. A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. Amino Acids 2010;38:1237–52.
[48] Tsai FS. Text mining and visualisation of Protein–Protein Interactions. Int J Comput Biol Drug Des 2011;4:239–44.
[49] Krallinger M, Rodriguez-Penagos C, Tendulkar A, Valencia A. PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction. Nucleic Acids Res 2009;37:W160–5.
[50] Srinivasan P, Wedemeyer M. Mining concept profiles with the vector model or where on earth are diseases being studied. In: Citeseer; 2003.
[51] Shetty KD, Dalal SR. Using information mining of the medical literature to improve drug safety. J Am Med Inform Assoc 2011;18:668–74.
[52] Frawley William J, Piatetsky-Shapiro G, Matheus Christopher J. Knowledge discovery in databases: an overview. AI Mag 1992;13:57–70.
[53] Fayyad Usama, Piatetsky-Shapiro G, Smyth Padhraic. From data mining to knowledge discovery in databases. AI Mag 1996;17:37–54.
[54] Korhonen A, Seaghdha DO, Silins I, Sun L, Hogberg J, Stenius U. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PLoS One 2012;7:e33427.
[55] Nam S, Park T. Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with epithelial–mesenchymal transition. PLoS One 2012;7:e31685.
[56] Mack R, Hehenberger M. Text-based knowledge discovery: search and mining of life-sciences documents. Drug Discov Today 2002;7:S89–98.
[57] Urzua U, Owens GA, Zhang GM, Cherry JM, Sharp JJ, Munroe DJ. Tumor and reproductive traits are linked by RNA metabolism genes in the mouse ovary: a transcriptome-phenotype association analysis. BMC Genomics 2010;11(Suppl. 5):S1.
[58] Hilborn RM, Marc. The ecological detective: confronting models with data. Princeton University Press; 2011.
[59] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30:7–18.
[60] Li J, Zhu X, Chen JY. Building disease-specific drug–protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol 2009;5:e1000450.
[61] Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics 2005;6(Suppl. 1):S14.
[62] Hettne KM, Weeber M, Laine ML, ten Cate H, Boyer S, Kors JA, et al. Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study. J Clin Periodontol 2007;34:1016–24.
[63] Topinka CM, Shyu CR. Predicting cancer interaction networks using text-mining and structure understanding. In: AMIA annu symp proc; 2006. p. 1123.
[64] McEntyre J, Lipman D. PubMed: bridging the information gap. Can Med Assoc J 2001;164:1317–9.
[65] Pubmed. <http://www.ncbi.nlm.nih.gov/pubmed/>.
[66] Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004;2:e309.
[67] Textpresso. <http://www.textpresso.org/>.
[68] Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 2005;33:W783–6.
[69] GoPubMed. <http://www.gopubmed.org/>.
[70] Hoffmann R, Valencia A. A gene network for navigating the literature. Nat Genet 2004;36:664.
[71] Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 2005;21(Suppl. 2):ii252–8.
[72] Baran J, Gerner M, Haeussler M, Nenadic G, Bergman CM. Pubmed2ensembl: a resource for mining the biological literature on genes. PLoS One 2011;6:e24716.

[73] Papanikolaou N, Pafilis E, Nikolaou S, Ouzounis CA, Iliopoulos I, Promponas VJ. BioTextQuest: a web-based biomedical text mining suite for concept discovery. Bioinformatics 2011;27:3327–8.
[74] Arrowsmith. <http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/start.cgi>.
[75] Smalheiser NR, Torvik VI, Zhou W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. Comput Methods Programs Biomed 2009;94:190–7.
[76] BITOLA. <http://ibmi.mf.uni-lj.si/bitola/>.
[77] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Improving literature based discovery support by genetic knowledge integration. Stud Health Technol Inform 2003;95:68–73.
[78] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005;74:289–98.
[79] Fang YC, Huang HC, Juan HF. MeInfoText: associated gene methylation and cancer information from text mining. BMC Bioinformatics 2008;9:22.
[80] Fang YC, Lai PT, Dai HJ, Hsu WL. MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. BMC Bioinformatics 2011;12:471.
[81] Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W. PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Res 2008;36:D842–6.
[82] Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, et al. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. BMC Bioinformatics 2006;7(Suppl. 3):S4.
[83] Deng X, Geng H, Bastola DR, Ali HH. Link test – a statistical method for finding prostate cancer biomarkers. Comput Biol Chem 2006;30:425–33.
[84] Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, DeSesa C, 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;7:373.
[85] Krallinger M, Leitner F, Valencia A. Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol 2010;593:341–82.
[86] Clancy T, Pedicini M, Castiglione F, Santoni D, Nygaard V, Lavelle TJ, et al. Immunological network signatures of cancer progression and survival. BMC Med Genomics 2011;4:28.
[87] Kolluru B, Nakjang S, Hirt RP, Wipat A, Ananiadou S. Automatic extraction of microorganisms and their habitats from free text using text mining workflows. J Integr Bioinform 2011;8:184.
[88] Xu Y, Teng D, Lei Y. MinePhos: A literature mining system for protein phoshphorylation information extraction. IEEE/ACM Trans Comput Biol Bioinform 2012;9(1):311–5.
[89] Barbosa-Silva A, Fontaine JF, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinformatics 2011;12:435.
[90] Korhonen A, Silins I, Sun L, Stenius U. The first step in the development of Text Mining technology for Cancer Risk Assessment: identifying and organizing scientific evidence in risk assessment literature. BMC Bioinformatics 2009;10:303.
[91] Guo Y, Korhonen A, Liakata M, Silins I, Hogberg J, Stenius U. A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. BMC Bioinformatics 2011;12:69.
[92] You M, Zhao RW, Li GZ, Hu X. MAPLSC: a novel multi-class classifier for medical diagnosis. Int J Data Min Bioinform 2011;5:383–401.
[93] Lee CH, Wu CH, Yang HC. Text mining of clinical records for cancer diagnosis. In: Proceedings of the second international conference on innovative computing, informatio and control: IEEE computer society; 2007.
[94] Thompson P, McNaught J, Montemagni S, Calzolari N, Del Gratta R, Lee V, et al. The BioLexicon: a large-scale terminological resource for biomedical text mining. BMC Bioinformatics 2011;12:397.
[95] Thompson P, Nawaz R, McNaught J, Ananiadou S. Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics 2011;12:393.
[96] Maqungo M, Kaur M, Kwofie SK, Radovanovic A, Schaefer U, Schmeier S, et al. DDPC: dragon database of genes associated with prostate cancer. Nucleic Acids Res 2011;39:29.
[97] Wang L, Xiong Y, Sun Y, Fang Z, Li L, Ji H, et al. HLungDB: an integrated database of human lung cancer research. Nucleic Acids Res 2010;38:D665–9.
[98] Macilwain C. Systems biology: evolving into the mainstream. Cell 2011;144:839–41.
[99] Hayasaka S, Hugenschmidt CE, Laurienti PJ. A network of genes, genetic disorders, and brain areas. PLoS One 2011;6:e20907.
[100] Sharma P, Senthilkumar RD, Brahmachari V, Sundaramoorthy E, Mahajan A, Sharma A, et al. Mining literature for a comprehensive pathway analysis: a case study for retrieval of homocysteine related genes for genetic and epigenetic studies. Lipids Health Dis 2006;5:1.
[101] Palakal M, Bright J, Sebastian T, Hartanto S. A comparative study of cells in inflammation, EAE and MS using biomedical literature data mining. J Biomed Sci 2007;14:67–85.
[102] Papp B, Notebaart RA, Pal C. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet 2011;12:591–602.
[103] Alexopoulos LG, Melas IN, Chairakaki AD, Saez-Rodriguez J, Mitsos A. Construction of signaling pathways and identification of drug effects on the liver cancer cell HepG2. Conf Proc IEEE Eng Med Biol Soc 2010;2010:6717–20.
[104] Pena-Hernandez KE, Mahamaneerat WK, Kobayashi T, Shyu CR, Arthur G, Caldwell CW. Mapping biomedical literature with WNT signaling pathway. In: AMIA annu symp proc; 2008. p. 1089.
[105] Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio – the analysis and navigation of molecular networks. Bioinformatics 2003;19:2155–7.
[106] http://www.genego.com/.
[107] Jimenez-Marin A, Collado-Romero M, Ramirez-Boo M, Arce C, Garrido JJ. Biological pathway analysis by ArrayUnlock and Ingenuity Pathway Analysis. BMC Proc 2009;3(Suppl. 4):S6.
[108] Kountourakis P, Ajani JA, Davila M, Lee JH, Bhutani MS, Izzo JG. Barrett's esophagus: a review of biology and therapeutic approaches. Gastrointest Cancer Res 2012;5:49–57.
[109] Chandolu V, Dass CR. Cell and molecular biology underpinning the effects of PEDF on cancers in general and osteosarcoma in particular. J Biomed Biotechnol 2012;2012:740295.
[110] Chlebowski RT, McTiernan A, Wactawski-Wende J, Manson JE, Aragaki AK, Rohan T, et al. J Clin Oncol 2012;30(23):2844–52.
[111] Foroughi F, Saadat N, Salehian MT. Encapsulated insular carcinoma of the thyroid arising in Graves' disease: report of a case and review of the literature. Int J Surg Pathol 2012;10: 1066896912449688.
[112] Wei MY, Giovannucci EL. Lycopene, tomato products, and prostate cancer incidence: a review and reassessment in the PSA screening era. J Oncol 2012;2012:271063.
[113] Hassanein M, Callison JC, Callaway-Lane C, Aldrich MC, Grogan EL, Massion PP. The state of molecular biomarkers for the early detection of lung cancer. Cancer Prev Res 2012;5:992–1006.
[114] Hoffe S, Balducci L. Cancer and age: general considerations. Clin Geriatr Med 2012;28:1–18.
[115] Ando M, Morita T, O'Connor SJ. Primary concerns of advanced cancer patients identified through the structured life review process: a qualitative study using a text mining technique. Palliat Support Care 2007;5:265–71.
[116] Ahmed J, Meinel T, Dunkel M, Murgueitio MS, Adams R, Blasse C, et al. CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. Nucleic Acids Res 2011;39:D960–7.
[117] Mattila J, Koikkalainen J, Virkki A, van Gils M, Lotjonen J. Alzheimer's Disease Neuroimaging I. Design and application of a generic clinical decision support system for multiscale data. IEEE Trans Biomed Eng 2012;59:234–40.
[118] Wang B. BRCA1 tumor suppressor network: focusing on its tail. Cell Biosci 2012;2:6.
[119] Chatterjee S, Kumar D. Unraveling the design principle for motif organization in signaling networks. PLoS One 2011;6:e28606.
[120] Staiger C, Cadot S, Kooter R, Dittrich M, Muller T, Klau GW, et al. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PLoS One 2012;7:e34796.
[121] Giordano CN, Sinha AA. Cytokine networks in Pemphigus vulgaris: an integrated viewpoint. Autoimmunity 2012;45(6):427–39.
[122] Liu KQ, Liu ZP, Hao JK, Chen L, Zhao XM. Identifying dysregulated pathways in cancers from pathway interaction networks. BMC Bioinformatics 2012;13:126.
[123] Hjermstad MJ, Bergenmar M, Fisher SE, Montel S, Nicolatou-Galitis O, Raber-Durlacher J, et al. The EORTC QLQ-OH17: a supplementary module to the EORTC QLQ-C30 for assessment of oral health and quality of life in cancer patients. Eur J Cancer 2012;48(14):2203–11.
[124] Chaudhry Z, Siddiqui S. Health related quality of life assessment in Pakistani paediatric cancer patients using PedsQLTM 4.0 generic core scale and PedsQLTM cancer module. Health Qual Life Outcomes 2012;10:52.
[125] Khoshnevisan A, Yekaninejad MS, Ardakani SK, Pakpour AH, Mardani A, Aaronson NK. Translation and validation of the EORTC brain cancer module (EORTC QLQ-BN20) for use in Iran. Health Qual Life Outcomes 2012;10:54.
[126] Ramasubbu R, Taylor VH, Samaan Z, Sockalingham S, Li M, Patten S, et al. The Canadian Network for Mood and Anxiety Treatments (CANMAT) task force recommendations for the management of patients with mood disorders and select comorbid medical conditions. Ann Clin Psychiatry 2012;24: 91–109.
[127] Logue JS, Morrison DK. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. Genes Dev 2012;26:641–50.
[128] Azuaje FJ, Heymann M, Ternes AM, Wienecke-Baldacchino A, Struck D, Moes D, et al. Bioinformatics as a driver, not a passenger, of translational biomedical research: perspectives from the 6th Benelux bioinformatics conference. J Clin Bioinformatics 2012;2:7.
[129] Liekens AM, De Knijf J, Daelemans W, Goethals B, De Rijk P, Del-Favero J. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. Genome Biol 2011;12:R57.
[130] Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005;21:3191–2.
[131] ABNER. <http://pages.cs.wisc.edu/~bsettles/abner/>.
[132] Tsuruoka Y, Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: Association for computational linguistics Morristown, NJ, USA; 2005. p. 467–74.
[133] GENIATagger. <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Tagger>.
[134] Carpenter B. LingPipe for 99.99% recall of gene mentions; 2007. p. 307–9.
[135] Carpenter B. Character language models for Chinese word segmentation and named entity recognition; 2006. p. 169–72.
[136] LingPipe. <http://www.alias-i.com/lingpipe/>.

[137] Franzen K, Eriksson G, Olsson F, Asker L, Lidïn P, Cöster J. Protein names and how to find them. Int J Med Inform 2002;67:49–61.
[138] Yapex. <http://www.sics.se/humle/projects/prothalt/>.
[139] Acromine. <http://www.nactem.ac.uk/software/acromine/>.
[140] Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics 2006;22:3089–95.
[141] Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 2005;6(Suppl. 1):S3.
[142] GENETAG. <ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/>.
[143] GO. <http://www.geneontology.org/>.
[144] BCMS. <http://bcms.bioinfo.cnio.es/>.
[145] Leitner F, Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, Kuo CJ, et al. Introducing meta-services for biomedical information extraction. Genome Biol 2008;9(Suppl. 2):S6.
[146] Chilibot. <http://www.chilibot.net/>.
[147] Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 2004;5:147.
[148] HPID. <http://wilab.inha.ac.kr/hpid/>.
[149] Han K, Park B, Kim H, Hong J, Park J. HPID: the human protein interaction database. Bioinformatics 2004;20:2466–70.
[150] HPRD. <http://www.hprd.org/>.
[151] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003;13:2363–71.
[152] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database – 2009 update. Nucleic Acids Res 2009;37:D767–72.
[153] iHOP. <http://www.ihop-net.org/UniPub/iHOP/>.
[154] IntAct. <http://www.ebi.ac.uk/intact/main.xhtml>.
[155] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct – open source resource for molecular interaction data. Nucleic Acids Res 2007;35:D561–5.
[156] MedScan. <http://www.ariadnegenomics.com/technology-research/medscan/>.
[157] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 2003;19:1699–706.
[158] PubGene. <http://www.pubgene.org/>.
[159] Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001;28:21–8.
[160] Reactome. <http://www.reactome.org/>.
[161] Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol 2007;8:R39.
[162] Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, et al. Correction: Reactome: a knowledge base of biologic pathways and processes. Genome Biol 2009;10:402.
[163] Pyysalo S, Ginter F, Heimonen J, Bjorne J, Boberg J, Jarvinen J, et al. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 2007;8:50.
[164] Pyysalo S, Ginter F, Heimonen J, Björne J, Boberg J, Järvinen J, et al. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinform 2007;8(50):1–24.
[165] BioInfer. <http://mars.cs.utu.fi/BioInfer/>.
[166] HIV-1ProteinInteraction. <http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html>.
[167] Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG. Human immunodeficiency virus type 1, human protein interaction database at NCBI. Nucleic Acids Res 2009;37:D417–22.
[168] Ptak RG, Fu W, Sanders-Beer BE, Dickerson JE, Pinney JW, Robertson DL, et al. Cataloguing the HIV type 1 human protein interaction network. AIDS Res Hum Retroviruses 2008;24:1497–502.
[169] Pinney JW, Dickerson JE, Fu W, Sanders-Beer BE, Ptak RG, Robertson DL. HIV–host interactions: a map of viral perturbation of the host system. AIDS 2009;23:549–54.
[170] LLL05. <http://genome.jouy.inra.fr/texte/LLLchallenge/>.
[171] Johnson HL, Baumgartner Jr WA, Krallinger M, Cohen KB, Hunter L. Corpus refactoring: a feasibility study. J Biomed Discov Collab 2007;2:4.
[172] PICorpus. <http://bionlp-corpora.sourceforge.net/picorpus/index.shtml>.
[173] PDZBase. <http://icb.med.cornell.edu/services/pdz/start>.
[174] Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H. PDZBase: a Protein–Protein Interaction database for PDZ-domains. Bioinformatics 2005;21:827–8.
[175] STRING. <http://string.embl.de/>.
[176] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009;37:D412–6.
[177] BioCreAtIvE. <http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/results/>.
[178] BioText. <http://biotext.berkeley.edu/data.html>.
[179] Rosario B, Hearst MA. Multi-way relation classification: application to protein–protein interactions. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP); 2005. p. 732–9.
[180] Rosario B, Hearst MA. Classifying semantic relations in bioscience texts. In: Proceedings of the 42nd annual meeting on association for computational linguistics; 2004. p. 1–8.
[181] A BRaM. Classifying semantic relations in bioscience text. In: proceedings of the 42nd annual meeting of the association for computational linguistics (ACL 2004). Barcelona; 2004.
[182] Hearst BRaM. Multi-way relation classification: application to protein–protein interaction. In: HLT-NAACL'05. Vancouver; 2005.
[183] Hearst BRaM. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: Proceedings of 2001 conference on empirical methods in natural language processing (EMNLP 2001). Pittsburgh, PA; 2001.
[184] Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput 2003:451–62.
[185] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus – semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(Suppl. 1):i180–2.
[186] GENIA. <http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/geniaform.cgi>.