Biomedical text mining and its applications in cancer research

Fei Zhu\textsuperscript{a, b}, Preecha Patumcharoenpol\textsuperscript{c,d}, Cheng Zhang\textsuperscript{a}, Yang Yang\textsuperscript{a,b}, Jonathan Chan\textsuperscript{c}, Asawin Meechai\textsuperscript{e}, Wanwipa Vongsangnak\textsuperscript{a}, Bairong Shen\textsuperscript{a, f, *}

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

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.

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...
The conventional phases and tasks involved in biomedical text mining.

For example, when we identify a gene term, true positive refers to a gene term that 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 CRF 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).
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 naive 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. Tsi [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 existing 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. Hettete 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 a 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.

<table>
<thead>
<tr>
<th>System</th>
<th>Brief introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABNER [130,131]</td>
<td>ABNER is a software tool for molecular biology text analysis. It uses linear-chain conditional random fields approach with orthographic and contextual features</td>
</tr>
<tr>
<td>GENIA Tagger [132,133]</td>
<td>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</td>
</tr>
<tr>
<td>LingPipe [134–136]</td>
<td>LingPipe provides three generic, trainablechunkers to carry on named entity recognition. LingPipe can be used to identify biomedical entities such as genes, organisms, malignancies, and chemicals</td>
</tr>
<tr>
<td>Yapex [137,138]</td>
<td>Yapex is a rule-based system named entity recognition system that utilizes lexical and syntactic analysis to identify protein names</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Corpus name</th>
<th>Brief introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acromine [139,140]</td>
<td>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</td>
</tr>
<tr>
<td>BioLexicon [94]</td>
<td>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</td>
</tr>
<tr>
<td>GENETAG [141,142]</td>
<td>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</td>
</tr>
<tr>
<td>GO [143]</td>
<td>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</td>
</tr>
</tbody>
</table>

Table 3
Some useful tools for relationship extraction.

<table>
<thead>
<tr>
<th>System</th>
<th>Brief introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCMS [144,145]</td>
<td>BioCreative MetaServer (BCMS) is a meta-service for information extraction which can generate annotations for PubMed/Medline abstracts, covering genes, proteins, gene IDs, species, and protein–protein interactions</td>
</tr>
<tr>
<td>Chilibot [146,147]</td>
<td>Chilibot searches PubMed abstracts about specific relationships between proteins, genes, or keywords. The results are returned as a graph</td>
</tr>
<tr>
<td>HPID [148,149]</td>
<td>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</td>
</tr>
<tr>
<td>HPRD [150–152]</td>
<td>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 display the results</td>
</tr>
<tr>
<td>iHOP [70,71,153]</td>
<td>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</td>
</tr>
<tr>
<td>IntAct [154,155]</td>
<td>IntAct provides analysis tools for molecular interaction as well as interaction database of which data were derived from literature curation or user submissions</td>
</tr>
<tr>
<td>MedScan [156,157]</td>
<td>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</td>
</tr>
<tr>
<td>PubGene [158,159]</td>
<td>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</td>
</tr>
<tr>
<td>Reactome [160–162]</td>
<td>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</td>
</tr>
</tbody>
</table>

Table 4
Some standard annotated data sets for relation extraction.

<table>
<thead>
<tr>
<th>Data set name</th>
<th>Brief introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioInf [163–165]</td>
<td>BioInf is a XML-based format corpus protein–protein interaction. The data of BioInf were from five well-known protein–protein interaction corpora: AImed, BioInf, LLL, IEPA, and HPRD50</td>
</tr>
<tr>
<td>HIV-1, human Pr [166–169]</td>
<td>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</td>
</tr>
<tr>
<td>LLL 05 [170]</td>
<td>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</td>
</tr>
<tr>
<td>PICorpus [171,172]</td>
<td>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</td>
</tr>
<tr>
<td>PDZBase [173,174]</td>
<td>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</td>
</tr>
<tr>
<td>STRING [175,176]</td>
<td>STRING provides known and predicted protein interactions, including physical and functional associations derived from Genomic context, high-throughput experiments, coexpression and previous knowledge</td>
</tr>
</tbody>
</table>
Table 5

Some commonly used standard annotated data sets for text mining.

<table>
<thead>
<tr>
<th>Data set name</th>
<th>Brief introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioCreative III [177]</td>
<td>BioCreative III works for evaluating text mining and information extraction systems applied to the biomedical domain. BioCreative III has several data sets for three tasks: cross-species gene identification and normalization, protein–protein interactions extraction, and interactive demonstration task for gene indexing and retrieval task.</td>
</tr>
<tr>
<td>BioInfer [163–165]</td>
<td>BioInfer is a XML-based format corpus protein–protein interaction. The data of BioInfer were from five well-known protein–protein interaction corpora: ALLmed, BioInfer, ILE, IEPA, and HPRD950.</td>
</tr>
<tr>
<td>BioText [178–184]</td>
<td>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.</td>
</tr>
<tr>
<td>GENIA [185,186]</td>
<td>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.</td>
</tr>
<tr>
<td>PICorpus [171,172]</td>
<td>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.</td>
</tr>
</tbody>
</table>

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 MedInfoText [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 MedInfoText [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
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

Fig. 3. An illustration of a text mining-assisted cancer study workflow from a systems biology viewpoint.
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 shown 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 offer application programming interface, by which we can use scripts to download the text automatically by computers. For examples, through E-utils 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 OMCs 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 pre-processing. 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 noise 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 bio-medicine, 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


Arrowsmith D. http://arrowsmith.psych.ucc.ie/cgi-bin/arrowsmith_ui/start.cgi.


BITOLA. http://bitomi.mi.unis-j.si/bitola/.


