Informatics methods, such as text mining and natural language processing, are always involved in bioinformatics research. In this study, we discuss text mining and natural language processing methods in bioinformatics from two perspectives. First, we aim to search for knowledge on biology, retrieve references using text mining methods, and reconstruct databases. For example, protein-protein interactions and gene-disease relationship can be mined from PubMed. Then, we analyze the applications of text mining and natural language processing techniques in bioinformatics, including predicting protein structure and function, detecting noncoding RNA. Finally, numerous methods and applications, as well as their contributions to bioinformatics, are discussed for future use by text mining and natural language processing researchers.

1. Introduction

Text mining and natural language processing refer to comprehending and analyzing natural language by using computer algorithms and programs. It is an important research direction in the application field of artificial intelligence. Research on natural language processing and text mining has been reported as early as the emergence of computers. With continuous and extensive research on machine learning and data mining algorithms, existing text mining technologies have achieved good results in automatic abstraction, automatic question answering, web relational network analysis, and anaphora resolution [1, 2].

Bioinformatics is an interdiscipline that emerged with the progress and accomplishment of the Human Genome Project. It predicts and solves live science problems related to genetics by using computer and statistical informatics. Data storage, retrieval, and analysis are the key processes in bioinformatics [3–7]. The National Center for Biotechnology Information established various databases for biological data, including sequence databases for storing DNA and protein data (e.g., dbEST and dbSNP) [8, 9], Online Mendelian Inheritance in Man database for storing disease data, Gene Expression Omnibus database for storing gene chip data, and PubMed database for storing biological and medical literature [10].

Text mining and natural language processing techniques are necessary to retrieve user preference knowledge from expanding databases. Therefore, researchers retrieve papers on certain topics of interest, such as determining protein-protein interactions, from PubMed using computer algorithms and programs. With the cracking of genetic codes, researchers have determined that biological sequences, particularly protein sequences, are similar to human language in terms of composition. In addition to using text mining to retrieve bioinformatics articles directly, an increasing number of researchers are regarding protein sequences as a special “text” and analyzing them based on existing text mining technologies. The relationship between bioinformatics and natural language processing is shown in Figure 1. Researchers have also predicted the structures and functions of proteins. Based on these two aspects, we summarize the text mining technologies used in bioinformatics research. We aim to present these technologies to more bioinformatics researchers and hope that the number of researchers who can use good text mining technologies in bioinformatics studies will increase.
2. Mining Bioinformatics Literature

The development of text mining technology plays an important role in retrieving biological literature, particularly in establishing biological information databases. A special workshop on biological literature retrieval problems was conducted during the Annual Meeting of the Association for Computational Linguistics and the Annual International Conference on Intelligent Systems for Molecular Biology in 2005 to discuss literature mining problems related to bioinformatics. Extracting protein-protein interactions and the relationship between gene functions and diseases are two leading application subjects.

2.1. Extracting Protein-Protein Interactions. Extracting the protein interaction network is an important research topic in bioinformatics and systems biology [11–14]. In previous studies, researchers searched for protein-protein interactions manually. However, with the exponential growth of biological literature, a program that can recognize protein-protein interactions automatically from PubMed abstracts is necessary. Nevertheless, no unified naming rule for proteins has been established yet. Many proteins and genes use the same name. Consequently, recognizing protein names from the literature abstracts and further determining their interactions are key problems in the application of text mining in searching for protein-protein interactions.

Initially, researchers extracted protein-protein interactions through statistical and counting methods. They manually created dictionaries of protein names and then searched abstracts that involve elements occurring at least twice. On this basis, researchers determined that associated proteins interact with one another [15]. Some researchers also used dynamic planning to extract and compare protein-protein interactions [16].

Extracting protein-protein interactions has been a research hot spot in bioinformatics for a long time and has attracted an increasing number of researchers in the fields of text mining and natural language processing. First, the grammar of literature abstracts is analyzed more carefully, rather than making a simple statistics of dictionary words. Kim et al. converted a complicated semantic structure analysis into calculating the shortest path in a graph by creating a nucleus [17]. Similar analysis methods of literature abstracts include grammatical analysis [18–21], context-free grammar analysis [22], ontology analysis [23], and other information retrieval methods. Protein-protein interactions are examined using these analysis methods. In addition, many machine learning methods, such as ensemble learning [24] and Bayesian network [25], are applied to recognize protein names and interactions.

2.2. Extracting the Relationship between Gene Functions and Diseases. Extracting protein-protein interactions involves searching for two proteins in the text and determining whether they interact with each other. Similarly, extracting the relationship between gene functions and diseases also involves searching for gene names and disease names simultaneously in the literature and then determining whether a particular gene is related to a certain disease [26].

In general, such extraction process can be divided into three steps. First, the abstracts of associated papers are searched through comparison with a dictionary. Second, the search scope has to be expanded forward and backward sometimes based on the location of the related word or clause.
3.1. Predicting Protein Structure. Protein structure determines function [48]. Hence, it should be analyzed to determine protein function. The structural analysis of protein mainly focuses on certain protein sequences and classifies regions into the α-helix, β-lamella, and protein disordered regions. Predicting the α-helix and β-lamella regions is the same as predicting the secondary protein structure.

If a protein sequence is regarded as a natural language, then analyzing the type of protein in a region is similar to calibrating grammar in natural language processing. First, the secondary protein structure is predicted by combining rules and statistics [49–52]. However, faced with the bottleneck of statistical prediction, some researchers have proposed using machine learning prediction methods, including methods based on artificial neural network (ANN) [53], support vector machine (SVM) [54, 55], random forest [56–58], and maximum entropy [59].

Predicting the protein disordered region is also conducted. This region refers to the area without a stable or unique 3D structure in the protein space structure. Many text mining and machine learning methods, including ANN [60–62], SVM [63–65], conditional random field [66], and random forest [67], have been used to predict the protein disordered region. Common existing server addresses are listed in Table 1.

3.2. Predicting Protein Function. Predicting protein function is one of the most basic research topics in bioinformatics. It involves predicting protein-protein interactions and interaction sites [68, 69], localizing subcellular protein [70–78], predicting and classifying transmembrane protein [79–82], protein remote homology detection [83, 84], classifying protein functions [85–93], recognizing multifunctional enzymes [94–96], and DNA binding protein identification [97, 98].

The protein sequence is easy to determine. Similar to natural language, the protein sequence has many complicated rules. However, summarizing and understanding the rules of protein sequences are difficult. Therefore, analyzing and predicting the "protein language" expressed by amino acid sequences by using computational linguistics and machine learning methods are necessary. Through these procedures, we may be able to understand the functions of protein sequences.

Predicting protein-protein interactions is one of the most basic research topics in protein functions. Many researchers are committed to predicting whether two protein sequences exhibit interactions. To date, many machine learning methods have been applied, including SVM [99], kernel method [100, 101], decision-making tree [102, 103], random forest [104], Bayesian network [105], and the autoregressive model [106]. Several text processing methods, such as ontology annotation and sample weighting [107], are used to detect features and process training data. When predicting protein-protein interactions, researchers also aim to analyze the region of protein-protein interactions, which is used to predict protein-protein interaction sites. Information approaches commonly used in grammatical analyses, such as condition random fields [108] and a hidden Markov model (HMM) [109], have been used to analyze interaction sites and have achieved good results. Moreover, random forest [110], SVM [111], ANN [112], Bayesian network [113], linear regression [114], and other machine learning methods
are used to predict protein-protein interaction sites. Nevertheless, some researchers doubt that determining the protein sequence alone is inadequate to provide sufficient information for predicting interactions [115]. Text mining and machine learning researchers should develop new features and classification methods to solve this problem. The websites of existing common software used to predict protein-protein interactions and interaction sites are provided in Table 2.

### 4. Applying Natural Language Processing Techniques to Noncoding RNA Identification

#### 4.1. Comparative RNA Prediction Methods

Alignment is also an important topic in natural language processing. DNA or RNA sequences can also be viewed as text. Sequence-based multiple sequence alignment methods can be used only at the sequence similarity level. The secondary structures of ncRNAs are usually more conserved than their sequences [116, 117]; for example, miRNA precursors share the common hairpin-like structure and tRNAs form cloverleaf structures [118, 119]. The functions of many ncRNAs are therefore determined by their secondary structure rather than by their sequences. As a result, structure-based multiple sequence alignment methods have been developed to align an input sequence to known ncRNA structures to determine the ncRNA class to which the input sequence belongs.

LocARNA [120] can produce fast and high-quality pairwise and multiple alignments of RNA sequences. It uses a complex RNA energy model for simultaneous folding and sequence/structure alignment of the RNAs. LocARNA performs global and local sequence alignments as well as local structural alignment of RNA molecules. An upgraded version of LocARNA, called LocARNA-P, has been developed recently [121]. The new version incorporates a probabilistic model that can compute accurate multiple alignments based on a probabilistic consistency transformation and reliability profiles for assessing local alignment quality and localizing RNA motifs. These features are based on computing sequence and structure match probabilities based on the LocARNA alignment model.
Table 3: Multiple sequence alignment tools.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Alignment method</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAT</td>
<td>Sequence-based</td>
<td><a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a></td>
</tr>
<tr>
<td>BWA-SW</td>
<td></td>
<td><a href="http://bio-bwa.sourceforge.net">http://bio-bwa.sourceforge.net</a></td>
</tr>
<tr>
<td>Multilign</td>
<td></td>
<td><a href="http://rna.urmc.rochester.edu/">http://rna.urmc.rochester.edu/</a></td>
</tr>
<tr>
<td>FoldalignM</td>
<td></td>
<td><a href="http://foldalign.ku.dk/">http://foldalign.ku.dk/</a></td>
</tr>
<tr>
<td>LocARNAN/LocARNAP</td>
<td></td>
<td><a href="http://www.bioinf.uni-freiburg.de/Software/LocARNAN/">http://www.bioinf.uni-freiburg.de/Software/LocARNAN/</a></td>
</tr>
<tr>
<td>MASTR</td>
<td></td>
<td><a href="http://mastr.binf.ku.dk/">http://mastr.binf.ku.dk/</a></td>
</tr>
<tr>
<td>RAF</td>
<td></td>
<td><a href="http://contra.stanford.edu/contrastfold/">http://contra.stanford.edu/contrastfold/</a></td>
</tr>
<tr>
<td>RNAsampler</td>
<td>Structure-based</td>
<td><a href="http://ural.wustl.edu/software.html">http://ural.wustl.edu/software.html</a></td>
</tr>
<tr>
<td>RNAsshapes</td>
<td></td>
<td><a href="http://bibiserv.techfak.uni-bielefeld.de/rnasshapes/">http://bibiserv.techfak.uni-bielefeld.de/rnasshapes/</a></td>
</tr>
<tr>
<td>RNAlifold</td>
<td></td>
<td><a href="http://www.tbi.univie.ac.at/RNA/">http://www.tbi.univie.ac.at/RNA/</a></td>
</tr>
<tr>
<td>StemLoc</td>
<td></td>
<td>N.A.</td>
</tr>
<tr>
<td>MAFFT</td>
<td></td>
<td><a href="http://mafft.cbrc.jp/alignment/software/index.html">http://mafft.cbrc.jp/alignment/software/index.html</a></td>
</tr>
<tr>
<td>MiRAlign</td>
<td></td>
<td><a href="http://bioinfo.au.tsinghua.edu.cn/miralign/">http://bioinfo.au.tsinghua.edu.cn/miralign/</a></td>
</tr>
</tbody>
</table>

Although comparative methods perform well in most cases, they have three intrinsic limitations: (1) they are highly dependent on the availability of homologous sequences or structures and cannot make predictions when no relevant sequence similarity or structure similarity is available; (2) they cannot correctly identify real ncRNAs that have low homology with known ncRNAs; and (3) they can identify only ncRNAs that are homologous with members of known ncRNA classes but cannot identify members of novel ncRNA classes. Most IncRNAs (long noncoding RNAs) cannot be predicted using comparative methods because they do not have specific structures or sequence similarity. These limitations mean that comparative methods display low specificity for identifying ncRNAs. The multiple sequence alignment tools that are currently available are listed in Table 3.

### 4.2. Noncomparative RNA Prediction Methods

The noncomparative methods are independent of homologous information and can, therefore, detect nonconserved ncRNAs. Most noncomparative methods employ machine learning techniques to make the predictions [122], which are similar to the text mining techniques.

Table 4: miRNA identification methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Online service</th>
<th>Local service</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiPred</td>
<td><a href="http://www.bioinf.seu.edu.cn/miRNA/">http://www.bioinf.seu.edu.cn/miRNA/</a></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>microPred</td>
<td><a href="http://www.cs.ox.ac.uk/people/manohararuksan.batuwita/microPred.htm">http://www.cs.ox.ac.uk/people/manohararuksan.batuwita/microPred.htm</a></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>TripletSVM</td>
<td><a href="http://bioinfo.au.tsinghua.edu.cn/mirnasvm">http://bioinfo.au.tsinghua.edu.cn/mirnasvm</a></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PlantMiRNApred</td>
<td><a href="http://nclab.hit.edu.cn/PlantMiRNApred/">http://nclab.hit.edu.cn/PlantMiRNApred/</a></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>miRNApre</td>
<td><a href="http://121.192.180.205:8080/miRNApreWeb/">http://121.192.180.205:8080/miRNApreWeb/</a></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MIReNA</td>
<td><a href="http://www.ihes.fr/~carbone/data8/">http://www.ihes.fr/~carbone/data8/</a></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>HuntMi</td>
<td><a href="http://adaa.polsl.pl/agudys/huntmi/huntmi.htm">http://adaa.polsl.pl/agudys/huntmi/huntmi.htm</a></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Mirident</td>
<td><a href="http://www.regulatoryrna.org/pub/mirident">http://www.regulatoryrna.org/pub/mirident</a></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>CHMM</td>
<td><a href="http://web.iiit.ac.in/~sumeet/mirna/">http://web.iiit.ac.in/~sumeet/mirna/</a></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>HeteroMirPred</td>
<td><a href="http://ncrna-pred.com/premiRNA.html">http://ncrna-pred.com/premiRNA.html</a></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Because of the importance of RNA structure, several computational RNA folding tools have been developed, such as mfold, RNAfold, vsfold, evofold, and sfold. Generally, these algorithms determine the folded secondary structure from input sequence by optimizing the intermolecular base pairing to minimize the free energy. Some miRNA identification methods are shown in Table 4 and existing RNA secondary prediction tools are listed in Table 5.

5. Conclusion and Future Research

As research on natural language and text mining methods develops, different application fields will be the key to future
studies. Interdisciplines represented by bioinformatics are becoming the focus of an increasing number of information science researchers. The application of text mining technologies and methods in bioinformatics study will become the focus of text mining researchers. Meanwhile, bioinformatics researchers have to learn text mining technologies intensively to solve specific bioinformatics problems.

In retrieving biological literature, apart from the aforementioned prediction of protein-protein interactions and gene-disease relationship, many problems, particularly those that require updating literature retrieval results, such as the relationships between adverse drug reaction and molecule composition as well as among single nucleotide polymorphism sites, diseases, and adverse drug effects, require the use of text mining to search for related knowledge in a literature database.

In bioinformatics, nearly all studies related to proteomics and predicting protein structure according to amino acid sequences can be conducted using text mining and natural language processing technology. Many mature texts mining technologies, such as word frequency statistics, condition random fields, HMM, and context-free grammar, have been successfully applied to predict secondary protein structures, irregular regions, interactions, and interaction sites. However, the latest research results in text mining and natural language processing should be verified by applying them in protein and DNA languages. No effective computation method is available yet for predicting third and fourth protein structures, protein homology remote detection, protein disorder region detection, interaction network establishment, and drug target prediction. Information science researchers should develop and provide more effective algorithms. In addition, new machine learning and text mining methods (e.g., semisupervised learning and active learning) have been proposed and will be applied in biological literature retrieval and bioinformatics. At present, recommending systems based on feedback has become a new hot spot problem in retrieving biological literature. And the Hadoop technique for big data is another hot spot for biology sequences [123].

The development of bioinformatics relies on information science. In particular, text mining and natural language processing researchers should provide a more extensive application space. Researchers of text mining algorithms should develop more effective intelligent algorithms based on the characteristics of biological data. This study does not only summarize text mining methods used in bioinformatics and corresponding problems, but it also provides related websites of successful prediction software. Recently, text mining researchers who are involved in bioinformatics can test and compare different types of software. The authors hope that the number of text mining researchers who can apply their own methods in bioinformatics will increase, which will facilitate the development of bioinformatics and even genetic studies.

**Conflict of Interests**

The authors declare that they have no competing interests.

**Acknowledgments**

This work was supported by Natural Science Foundation of China (Grant no. 31200769), the Natural Science Foundation of Fujian Province of China (Grants no. 2013J05103 and no. 2014J01253), Xiamen Science and Technology Planning Project (Grant no. 3502ZZ20143030), and Scientific Research Plan Project of Fujian Education Department (Grants nos. JB12184 and JB09203).

**References**


