Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 674296, 10 pages
http://dx.doi.org/10.1155/2015/674296

Review Article

Survey of Natural Language Processing Techniques in Bioinformatics

Zhiqiang Zeng,[1] Hua Shi,[1] Yun Wu,[1] and Zhiling Hong[2]

[1] *College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China*
[2] *Software School, Xiamen University, Xiamen 361005, China*

Correspondence should be addressed to Zhiling Hong; hongzl@xmu.edu.cn

Received 10 May 2015; Revised 12 June 2015; Accepted 21 June 2015

Academic Editor: Tao Huang

Copyright © 2015 Zhiqiang Zeng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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.

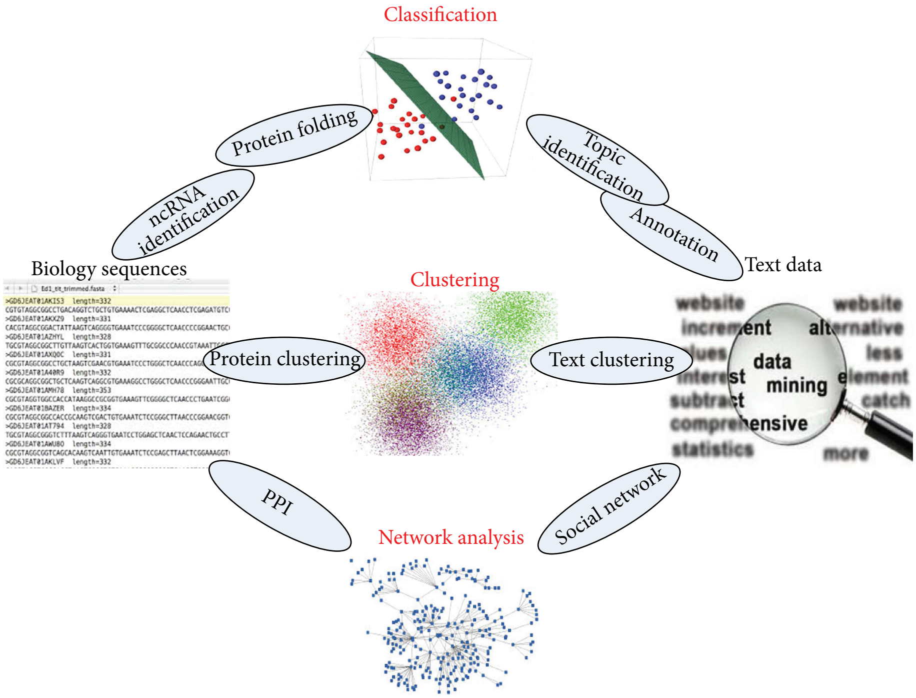

Figure 1: Problems and methodology relationship between NLP and bioinformatics.

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 to ensure accuracy. Finally, facts are evaluated using grammar analysis methods or machine learning methods. Such extraction methods frequently yield good results for special genes and diseases. Bui et al. examined the relationship between drugs and HIV variation in PubMed [27]. Jiang et al. determined the relationship between approximately 3000 microRNAs and different diseases based on the naming rule of microRNA [28]. Cheng et al. developed a text mining system based on the relationship among human diseases, variations, and drug effects [29]. Iossifov et al. focused on investigating malformations of human and mouse encephalon [30]. Jensen et al. made a detailed summary of related document databases, literature mining software, and functions [31].

2.3. Retrieving References. A considerable amount of bioscience literature has been published. Searching for interacting proteins and examining the relationship between genes and diseases are only two application cases. Text mining technology is required to obtain answers to many other bioscience and bioinformatics problems in various databases, such as PubMed.

Biological literature mining and related problem solving have to cope with two major problems, namely, recognizing name entities and extracting relations. These problems are mainly solved by (1) methods based on linguistic analysis [32], (2) methods based on dictionaries [33], (3) machine learning methods [34, 35], and (4) statistical methods [36].

Several important databases are also selected with text mining. STRING [37] and BioGRID [38] are built for protein-protein interaction with literature mining. For predicting gene function, PubTator [39] and GeneCards [40] are important databases using text mining techniques. Related works were reviewed in detail in Huang and Lu's work [41] recently. As the development of crowdsource, artificial text searching and mining can also be helpful for biomedicine literature collection [42].

Moreover, converting PubMed database into an Extensible Markup Language relational database [43] and a fuzzy search of papers and author names through short-term matching are also current research hot spots [44].

3. Applying Text Mining Technologies to Protein Research

DNA and protein sequences are a meaningful genetic language and are regarded as the sealed book of life. Therefore, an increasing number of natural language processing and text mining algorithms are being applied to study bioinformatics. For example, latent semantic analysis was applied to protein remote homology detection [45, 46], and protein spectral analysis originates from word frequency statistics in natural language processing. Furthermore, some grammar rules of protein, DNA, and RNA sequences were discovered, and several web servers were constructed so as to extract these features and rules [47].

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

Table 1: Web server for protein disorder prediction.

Problem	Name	Websites	Input format
Protein disorder prediction	DisProt	http://www.disprot.org/pondr-fit.php http://www.disprot.org/metapredictor.php http://www.dabi.temple.edu/disprot/predictor.php	Fasta or EMBL sequence format
	DisEMBL	http://dis.embl.de/	SwissProt ID
	DRIPPRED	http://www.sbc.su.se/~maccallr/disorder/cgi-bin/submit.cgi	Only plain sequence; one sequence once; slow
	FoldIndex	http://bip.weizmann.ac.il/fldbin/findex	Only plain sequence; one sequence once
	IUPred	http://iupred.enzim.hu/	SwissProt ID or plain sequence
	PONDR	http://www.pondr.com/cgi-bin/PONDR/pondr.cgi	Fasta
	PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1	Raw sequence or fasta format
	SCRATCH	http://scratch.proteomics.ics.uci.edu/	Only plain sequence; one sequence once; slow
	Spritz	http://distill.ucd.ie/spritz/	Raw sequence or fasta format
	RONN	http://www.strubi.ox.ac.uk/RONN/	Fasta, but only one sequence once

Table 2: Web server for protein-protein interaction and sites prediction.

Problem	Name	Websites	Input format
Protein interaction sites prediction	PPISP	http://pipe.scs.fsu.edu/ppisp.html http://pipe.scs.fsu.edu/meta-ppisp.html	PDB file
	Protemot	http://protemot.csbb.ntu.edu.tw/index.html	PDB ID
	SPPIDER	http://sppider.cchmc.org	PDB file or PDB ID
	Whiscy	http://nmr.chem.uu.nl/Software/whiscy/index.html	PDB file
Protein-protein interaction prediction	InterPreTS	http://www.russell.embl.de/cgi-bin/tools/interprets.pl	Fasta, 40 sequences at most
	PIE	http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/	Gene ID or name
	PPI	http://121.192.180.204:8080/PPI/Home.jsp	Fasta
	PredHS	http://www.predhs.org/	PDB files, 10 files at most
	Pred-PPI	http://cic.scu.edu.cn/bioinformatics/predict_ppi/default.html	Two fasta sequences
	Prism	http://cosbi.ku.edu.tr/prism/	Two PDB IDs or PDB files
	Struct2Net	http://groups.csail.mit.edu/cb/struct2net/webserver/	Gene names or keywords

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.

Tool	Alignment method	URL
BLAT	Sequence-based	http://genome.ucsc.edu/
BLAST		http://www.ncbi.nlm.nih.gov/
BWA-SW		http://bio-bwa.sourceforge.net
Multilign	Structure-based	http://rna.urmc.rochester.edu/
FoldalignM		http://foldalign.ku.dk/
LocARNA/LocARNA-P		http://www.bioinf.uni-freiburg.de/Software/LocARNA/
MASTR		http://mastr.binf.ku.dk/
RAF		http://contra.stanford.edu/contrafold/
RNASampler		http://ural.wustl.edu/software.html
RNAshapes		http://bibiserv.techfak.uni-bielefeld.de/rnashapes/
RNAalifold		http://www.tbi.univie.ac.at/RNA/
StemLoc		N.A.
MAFFT		http://mafft.cbrc.jp/alignment/software/index.html
MiRAlign		http://bioinfo.au.tsinghua.edu.cn/miralign/

Table 4: miRNA identification methods.

Method	URL	Online service	Local service
MiPred	http://www.bioinf.seu.edu.cn/miRNA/	√	√
microPred	http://www.cs.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm		√
TripletSVM	http://bioinfo.au.tsinghua.edu.cn/mirnasvm		√
PlantMiRNAPred	http://nclab.hit.edu.cn/PlantMiRNAPred/	√	√
miRNApre	http://121.192.180.205:8080/miRNApreWeb/	√	√
MIReNA	http://www.ihes.fr/~carbone/data8/		√
HuntMi	http://adaa.polsl.pl/agudys/huntmi/huntmi.htm		√
Mirident	http://www.regulatoryrna.org/pub/mirident		√
CSHMM	http://web.iitd.ac.in/~sumeet/mirna/		√
HeteroMirPred	http://ncrna-pred.com/premiRNA.html	√	√

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 lncRNAs (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 5: Secondary prediction tools.

Tool	URL
RNAfold	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
RNAstructure	http://rna.urmc.rochester.edu/rnastructure.html
mfold	http://www.bioinfo.rpi.edu/applications/mfold/
vsfold	http://www.rna.it-chiba.ac.jp/~vsfold/vsfold4/
evofold	http://users.soe.ucsc.edu/~jsp/EvoFold/
sfold	http://sfold.wadsworth.org/cgi-bin/index.pl

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 and 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 disordered 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. 3502Z20143030), and Scientific Research Plan Project of Fujian Education Department (Grants nos. JB12184 and JB09203).

References

[1] C. Lin, Z. Huang, F. Yang, and Q. Zou, "Identify content quality in online social networks," *IET Communications*, vol. 6, no. 12, pp. 1618–1624, 2012.

[2] L. Chen, L. Chun, L. Ziyu, and Z. Quan, "Hybrid pseudo-relevance feedback for microblog retrieval," *Journal of Information Science*, vol. 39, no. 6, pp. 773–788, 2013.

[3] Y. Li, C. Wang, Z. Miao et al., "ViRBase: a resource for virus-host ncRNA-associated interactions," *Nucleic Acids Research*, vol. 43, no. 1, pp. D578–D582, 2015.

[4] L. Wang, K. Qian, Y. Huang et al., "SynBioLGDB: a resource for experimentally validated logic gates in synthetic biology," *Scientific Reports*, vol. 5, article 8090, 2015.

[5] Y. Wang, L. Chen, B. Chen et al., "Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network," *Cell Death & Disease*, vol. 4, no. 8, article e765, 2013.

[6] X. Zhang, D. Wu, L. Chen et al., "RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein) interaction," *RNA*, vol. 20, no. 7, pp. 989–993, 2014.

[7] Y. Li, L. Zhuang, Y. Wang et al., "Connect the dots: a systems level approach for analyzing the miRNA-mediated cell death network," *Autophagy*, vol. 9, no. 3, pp. 436–439, 2013.

[8] J. Wang, Q. Zou, and M. Z. Guo, "Mining SNPs from EST sequences using filters and ensemble classifiers," *Genetics and Molecular Research*, vol. 9, no. 2, pp. 820–834, 2010.

[9] J. Wang, L. Zhang, Q. Zou, J. Tan, X. Chen, and Y. Wu, "Association studies on mtDNA and Parkinson's disease population discrimination using the statistical classification," *Current Bioinformatics*, vol. 9, no. 5, pp. 481–489, 2014.

[10] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed Research International*. In press.

[11] B. Liu, X. Wang, L. Lin, B. Tang, Q. Dong, and X. Wang, "Prediction of protein binding sites in protein structures using hidden Markov support vector machine," *BMC Bioinformatics*, vol. 10, article 381, 2009.

[12] F. Guo, S. C. Li, P. Du, and L. Wang, "Probabilistic models for capturing more physicochemical properties on protein-protein interface," *Journal of Chemical Information and Modeling*, vol. 54, no. 6, pp. 1798–1809, 2014.

[13] F. Guo, S. C. Li, L. Wang, and D. Zhu, "Protein-protein binding site identification by enumerating the configurations," *BMC Bioinformatics*, vol. 13, article 158, 2012.

[14] F. Guo, S. C. Li, and L. Wang, "Protein-protein binding sites prediction by 3D structural similarities," *Journal of Chemical Information and Modeling*, vol. 51, no. 12, pp. 3287–3294, 2011.

[15] M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li, "Discovering patterns to extract protein-protein interactions from full texts," *Bioinformatics*, vol. 20, no. 18, pp. 3604–3612, 2004.

[16] Y. Hao, X. Zhu, M. Huang, and M. Li, “Discovering patterns to extract protein-protein interactions from the literature: part II,” *Bioinformatics*, vol. 21, no. 15, pp. 3294–3300, 2005.

[17] S. Kim, J. Yoon, and J. Yang, “Kernel approaches for genic interaction extraction,” *Bioinformatics*, vol. 24, no. 1, pp. 118–126, 2008.

[18] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi, “Automated extraction of information on protein-protein interactions from the biological literature,” *Bioinformatics*, vol. 17, no. 2, pp. 155–161, 2001.

[19] K. Fundel, R. Küffner, and R. Zimmer, “RelEx—relation extraction using dependency parse trees,” *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[20] J. Šarić, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork, “Extraction of regulatory gene/protein networks from Medline,” *Bioinformatics*, vol. 22, no. 6, pp. 645–650, 2006.

[21] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, “GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles,” *Bioinformatics*, vol. 17, no. 1, pp. S74–S82, 2001.

[22] J. M. Temkin and M. R. Gilder, “Extraction of protein interaction information from unstructured text using a context-free grammar,” *Bioinformatics*, vol. 19, no. 16, pp. 2046–2053, 2003.

[23] A. Skusa, A. Rüegg, and J. Köhler, “Extraction of biological interaction networks from scientific literature,” *Briefings in Bioinformatics*, vol. 6, no. 3, pp. 263–276, 2005.

[24] R. Malik, L. Franke, and A. Siebes, “Combination of text-mining algorithms increases the performance,” *Bioinformatics*, vol. 22, no. 17, pp. 2151–2157, 2006.

[25] R. Chowdhary, J. Zhang, and J. S. Liu, “Bayesian inference of protein–protein interactions from biological literature,” *Bioinformatics*, vol. 25, no. 12, pp. 1536–1542, 2009.

[26] Q. Zou, J. Li, C. Wang, and X. Zeng, “Approaches for recognizing disease genes based on network,” *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

[27] Q.-C. Bui, B. T. Nualláin, C. A. Boucher, and P. M. A. Sloot, “Extracting causal relations on HIV drug resistance from literature,” *BMC Bioinformatics*, vol. 11, article 101, 2010.

[28] Q. Jiang, Y. Wang, Y. Hao et al., “miR2Disease: a manually curated database for microRNA deregulation in human disease,” *Nucleic Acids Research*, vol. 37, no. 1, pp. D98–D104, 2009.

[29] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart, “PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites,” *Nucleic Acids Research*, vol. 36, pp. W399–W405, 2008.

[30] I. Iossifov, R. Rodriguez-Esteban, I. Mayzus, K. J. Millen, and A. Rzhetsky, “Looking at cerebellar malformations through text-mined interactomes of mice and humans,” *PLoS Computational Biology*, vol. 5, no. 11, Article ID e1000559, 2009.

[31] L. J. Jensen, J. Saric, and P. Bork, “Literature mining for the biologist: from information retrieval to biological discovery,” *Nature Reviews Genetics*, vol. 7, no. 2, pp. 119–129, 2006.

[32] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, “Textpresso: an ontology-based information retrieval and extraction system for biological literature,” *PLoS Biology*, vol. 2, no. 11, article e309, 2004.

[33] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, and K. Takeda, “A text-mining system for knowledge discovery from biomedical documents,” *IBM Systems Journal*, vol. 43, no. 3, pp. 516–533, 2004.

[34] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, “Open information extraction from the web,” in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 51, pp. 68–74, New York, NY, USA, 2007.

[35] M. Banko and O. Etzioni, “The tradeoffs between open and traditional relation extraction,” in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 28–36, Columbus, Ohio, USA, June 2008.

[36] M. Abulaish and L. Dey, “Biological relation extraction and query answering from MEDLINE abstracts using ontology-based text mining,” *Data and Knowledge Engineering*, vol. 61, no. 2, pp. 228–262, 2007.

[37] D. Szklarczyk, A. Franceschini, S. Wyder et al., “STRING v10: protein-protein interaction networks, integrated over the tree of life,” *Nucleic Acids Research*, vol. 43, no. 1, pp. D447–D452, 2015.

[38] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred et al., “The BioGRID interaction database: 2015 update,” *Nucleic Acids Research*, vol. 43, no. 1, pp. D470–D478, 2015.

[39] C.-H. Wei, H.-Y. Kao, and Z. Lu, “PubTator: a web-based text mining tool for assisting biocuration,” *Nucleic Acids Research*, vol. 41, no. 1, pp. W518–W522, 2013.

[40] M. Safran, I. Dalah, J. Alexander et al., “GeneCards version 3: the human gene integrator,” *Database*, vol. 2010, Article ID baq020, 16 pages, 2010.

[41] C. C. Huang and Z. Lu, “Community challenges in biomedical text mining over 10 years: success, failure and the future,” *Briefings in Bioinformatics*, 2015.

[42] R. Khare, B. M. Good, R. Leaman, A. I. Su, and Z. Lu, “Crowdsourcing in biomedicine: challenges and opportunities,” *Briefings in Bioinformatics*, 2015.

[43] D. E. Oliver, G. Bhalotia, A. S. Schwartz, R. B. Altman, and M. A. Hearst, “Tools for loading MEDLINE into a local relational database,” *BMC Bioinformatics*, vol. 5, article 146, 2004.

[44] J. Wang, I. Cetindil, S. Ji et al., “Interactive and fuzzy search: a dynamic way to explore MEDLINE,” *Bioinformatics*, vol. 26, no. 18, Article ID btq414, pp. 2321–2327, 2010.

[45] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, “A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis,” *BMC Bioinformatics*, vol. 9, article 510, 2008.

[46] B. Liu, J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, “Using distances between Top-n-gram and residue pairs for protein remote homology detection,” *BMC Bioinformatics*, vol. 15, supplement 2, article S3, 2014.

[47] B. Liu, F. Liu, L. Fang, X. Wang, and K. Chou, “repDNA: a python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects,” *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2015.

[48] B. Liu, D. Zhang, R. Xu et al., “Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection,” *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[49] P. Y. Chou and G. D. Fasman, “Empirical predictions of protein conformation,” *Annual Review of Biochemistry*, vol. 47, pp. 251–276, 1978.

[50] J. Garnier, D. J. Osguthorpe, and B. Robson, “Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins,” *Journal of Molecular Biology*, vol. 120, no. 1, pp. 97–120, 1978.

[51] Q. Dong, X. Wang, L. Lin, and Y. Wang, “Analysis and prediction of protein local structure based on structure alphabets,” *Proteins: Structure, Function and Genetics*, vol. 72, no. 1, pp. 163–172, 2008.

[52] Q. Dong, X. Wang, and L. Lin, “Prediction of protein local structures and folding fragments based on building-block library,” *Proteins: Structure, Function and Genetics*, vol. 72, no. 1, pp. 353–366, 2008.

[53] B. Rost and C. Sander, “Prediction of protein secondary structure at better than 70% accuracy,” *Journal of Molecular Biology*, vol. 232, no. 2, pp. 584–599, 1993.

[54] H. Ding, H. Lin, W. Chen et al., “Prediction of protein structural classes based on feature selection technique,” *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 235–240, 2014.

[55] H. Lin, C. Ding, Q. Song et al., “The prediction of protein structural class using averaged chemical shifts,” *Journal of Biomolecular Structure & Dynamics*, vol. 29, no. 6, pp. 643–649, 2012.

[56] C. Lin, Y. Zou, J. Qin et al., “Hierarchical classification of protein folds using a novel ensemble classifier,” *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[57] W. Chen, X. Liu, Y. Huang, Y. Jiang, Q. Zou, and C. Lin, “Improved method for predicting protein fold patterns with ensemble classifiers,” *Genetics and Molecular Research*, vol. 11, no. 1, pp. 174–181, 2012.

[58] X. Zhao, Q. Zou, B. Liu, and X. Liu, “Exploratory predicting protein folding model with random forest and hybrid features,” *Current Proteomics*, vol. 11, no. 4, pp. 289–299, 2014.

[59] Y. Liu, J. Carbonell, J. Klein-Seetharaman, and V. Gopalakrishnan, “Comparison of probabilistic combination methods for protein secondary structure prediction,” *Bioinformatics*, vol. 20, no. 17, pp. 3099–3107, 2004.

[60] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, “Sequence complexity of disordered protein,” *Proteins: Structure, Function and Genetics*, vol. 42, no. 1, pp. 38–48, 2001.

[61] C.-T. Su, C.-Y. Chen, and Y.-Y. Ou, “Protein disorder prediction by condensed PSSM considering propensity for order or disorder,” *BMC Bioinformatics*, vol. 7, article 319, 2006.

[62] C.-T. Su, C.-Y. Chen, and C.-M. Hsu, “IPDA: integrated protein disorder analyzer,” *Nucleic Acids Research*, vol. 35, no. 2, pp. W465–W472, 2007.

[63] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, “Prediction and functional analysis of native disorder in proteins from the three kingdoms of life,” *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, 2004.

[64] K. Shimizu, S. Hirose, and T. Noguchi, “POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix,” *Bioinformatics*, vol. 23, no. 17, pp. 2337–2338, 2007.

[65] S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi, “POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions,” *Bioinformatics*, vol. 23, no. 16, pp. 2046–2053, 2007.

[66] L. Wang and U. H. Sauer, “OnD-CRF: predicting order and disorder in proteins conditional random fields,” *Bioinformatics*, vol. 24, no. 11, pp. 1401–1402, 2008.

[67] P. Han, X. Zhang, R. S. Norton, and Z.-P. Feng, “Large-scale prediction of long disordered regions in proteins using random forests,” *BMC Bioinformatics*, vol. 10, article 8, 2009.

[68] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, “Exploiting three kinds of interface propensities to identify protein binding sites,” *Computational Biology and Chemistry*, vol. 33, no. 4, pp. 303–311, 2009.

[69] B. Liu, B. Liu, F. Liu, and X. Wang, “Protein binding site prediction by combining hidden markov support vector machine and profile-based propensities,” *The Scientific World Journal*, vol. 2014, Article ID 464093, 6 pages, 2014.

[70] Z. Wang, Q. Zou, Y. Jiang, Y. Ju, and X. Zeng, “Review of protein subcellular localization prediction,” *Current Bioinformatics*, vol. 9, no. 3, pp. 331–342, 2014.

[71] H. Lin, H. Ding, F.-B. Guo, A.-Y. Zhang, and J. Huang, “Predicting subcellular localization of mycobacterial proteins by using Chou’s pseudo amino acid composition,” *Protein & Peptide Letters*, vol. 15, no. 7, pp. 739–744, 2008.

[72] H. Lin, H. Ding, F.-B. Guo, and J. Huang, “Prediction of subcellular location of mycobacterial protein using feature selection techniques,” *Molecular Diversity*, vol. 14, no. 4, pp. 667–671, 2010.

[73] H. Lin, H. Wang, H. Ding, Y.-L. Chen, and Q.-Z. Li, “Prediction of subcellular localization of apoptosis protein using Chou’s pseudo amino acid composition,” *Acta Biotheoretica*, vol. 57, no. 3, pp. 321–330, 2009.

[74] H. Lin, W. Chen, L.-F. Yuan, Z.-Q. Li, and H. Ding, “Using over-represented tetrapeptides to predict protein submitochondria locations,” *Acta Biotheoretica*, vol. 61, no. 2, pp. 259–268, 2013.

[75] H. Ding, S.-H. Guo, E.-Z. Deng et al., “Prediction of Golgi-resident protein types by using feature selection technique,” *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[76] H. Lin, C. Ding, L.-F. Yuan et al., “Predicting subchloroplast locations of proteins based on the general form of Chou’s pseudo amino acid composition: approached from optimal tripeptide composition,” *International Journal of Biomathematics*, vol. 6, no. 2, Article ID 1350003, 2013.

[77] P.-P. Zhu, W.-C. Li, Z.-J. Zhong et al., “Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition,” *Molecular BioSystems*, vol. 11, no. 2, pp. 558–563, 2015.

[78] H. Ding, L. Liu, F.-B. Guo, J. Huang, and H. Lin, “Identify golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition,” *Protein & Peptide Letters*, vol. 18, no. 1, pp. 58–63, 2011.

[79] Q. Zou, X. Li, Y. Jiang, Y. Zhao, and G. Wang, “BinMemPredict: a web server and software for predicting membrane protein types,” *Current Proteomics*, vol. 10, no. 1, pp. 2–9, 2013.

[80] H. Lin, “The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou’s pseudo amino acid composition,” *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 350–356, 2008.

[81] C. Ding, L.-F. Yuan, S.-H. Guo, H. Lin, and W. Chen, “Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions,” *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[82] H. Lin and H. Ding, “Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition,” *Journal of Theoretical Biology*, vol. 269, pp. 64–69, 2011.

[83] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, “Protein remote homology detection by combining Chou’s pseudo amino acid composition and profile-based protein representation,” *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[84] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[85] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction with incomplete annotations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 3, pp. 579–591, 2014.

[86] Q. Zou, Z. Wang, X. Guan, B. Liu, Y. Wu, and Z. Lin, "An approach for identifying cytokines based on a novel ensemble classifier," *BioMed Research International*, vol. 2013, Article ID 686090, 11 pages, 2013.

[87] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction using multi-label ensemble classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 1045–1057, 2013.

[88] H. Ding, E.-Z. Deng, L.-F. Yuan et al., "iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels," *BioMed Research International*, vol. 2014, Article ID 286419, 10 pages, 2014.

[89] W.-X. Liu, E.-Z. Deng, W. Chen, and H. Lin, "Identifying the subfamilies of voltage-gated potassium channels using feature selection technique," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 12940–12951, 2014.

[90] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, 2015.

[91] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Molecular BioSystems*, vol. 10, no. 8, pp. 2229–2235, 2014.

[92] L.-F. Yuan, C. Ding, S.-H. Guo, H. Ding, W. Chen, and H. Lin, "Prediction of the types of ion channel-targeted conotoxins based on radial basis function network," *Toxicology in Vitro*, vol. 27, no. 2, pp. 852–856, 2013.

[93] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 67–70, 2011.

[94] X.-Y. Cheng, W.-J. Huang, S.-C. Hu et al., "A global characterization and identification of multifunctional enzymes," *PLoS ONE*, vol. 7, no. 6, Article ID e38979, 2012.

[95] H. Lin, W. Chen, and H. Ding, "AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes," *PLoS ONE*, vol. 8, no. 10, Article ID e75726, 2013.

[96] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[97] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining chou's PseAAC and Physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[98] B. Liu, J. Xu, X. Lan et al., "IDNA-Prot—dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS ONE*, vol. 9, no. 9, Article ID e106691, 2014.

[99] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[100] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, supplement 1, pp. i38–i46, 2005.

[101] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins: Structure, Function and Genetics*, vol. 63, no. 3, pp. 490–500, 2006.

[102] L. V. Zhang, S. L. Wong, O. D. King, and F. P. Roth, "Predicting co-complexed protein pairs using genomic and proteomic data integration," *BMC Bioinformatics*, vol. 5, article 38, 2004.

[103] S. J. Darnell, D. Page, and J. C. Mitchell, "An automated decision-tree approach to predicting protein interaction hot spots," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 813–823, 2007.

[104] X.-W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.

[105] R. Jansen, H. Yu, D. Greenbaum et al., "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[106] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein-protein interactions from protein sequences," *Bioinformatics*, vol. 19, no. 15, pp. 1875–1881, 2003.

[107] M.-H. Li, X.-L. Wang, L. Lin, and T. Liu, "Effect of example weights on prediction of protein-protein interactions," *Computational Biology and Chemistry*, vol. 30, no. 5, pp. 386–392, 2006.

[108] M.-H. Li, L. Lin, X.-L. Wang, and T. Liu, "Protein-protein interaction site prediction based on conditional random fields," *Bioinformatics*, vol. 23, no. 5, pp. 597–604, 2007.

[109] T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Müller, "Modelling interaction sites in protein domains with interaction profile hidden Markov models," *Bioinformatics*, vol. 22, no. 23, pp. 2851–2857, 2006.

[110] M. Šikić, S. Tomić, and K. Vlahoviček, "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests," *PLoS Computational Biology*, vol. 5, no. 1, Article ID e1000278, 2009.

[111] J. R. Bradford and D. R. Westhead, "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics*, vol. 21, no. 8, pp. 1487–1494, 2005.

[112] P. Fariselli, F. Pazos, A. Valencia, and R. Casadio, "Prediction of protein-protein interaction sites in heterocomplexes with neural networks," *European Journal of Biochemistry*, vol. 269, no. 5, pp. 1356–1361, 2002.

[113] J. R. Bradford, C. J. Needham, A. J. Bulpitt, and D. R. Westhead, "Insights into protein-protein interfaces using a bayesian network prediction method," *Journal of Molecular Biology*, vol. 362, no. 2, pp. 365–386, 2006.

[114] I. Kufareva, L. Budagyan, E. Raush, M. Totrov, and R. Abagyan, "PIER: protein interface recognition for structural proteomics," *Proteins*, vol. 67, no. 2, pp. 400–417, 2007.

[115] J. Yu, M. Guo, C. J. Needham, Y. Huang, L. Cai, and D. R. Westhead, "Simple sequence-based kernels do not predict protein-protein interactions," *Bioinformatics*, vol. 26, no. 20, pp. 2610–2614, 2010.

[116] Q. Zou, T. Zhao, Y. Liu, and M. Guo, "Predicting RNA secondary structure based on the class information and Hopfield network," *Computers in Biology and Medicine*, vol. 39, no. 3, pp. 206–214, 2009.

[117] Q. Zou, C. Lin, X.-Y. Liu, Y.-P. Han, W.-B. Li, and M.-Z. Guo, "Novel representation of RNA secondary structure used to improve prediction algorithms," *Genetics and Molecular Research*, vol. 10, no. 3, pp. 1986–1998, 2011.

[118] B. Liu, L. Fang, F. Liu et al., “Identification of real microRNA precursors with a pseudo structure status composition approach,” *PLoS ONE*, vol. 10, no. 3, Article ID e0121501, 2015.

[119] B. Liu, L. Fang, J. Chen, F. Liu, and X. Wang, “miRNA-dis: microRNA precursor identification based on distance structure status pairs,” *Molecular BioSystems*, vol. 11, no. 4, pp. 1194–1204, 2015.

[120] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen, “Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering,” *PLoS Computational Biology*, vol. 3, no. 4, article e65, 2007.

[121] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen, “LocARNA-P: accurate boundary prediction and improved detection of structural RNAs,” *RNA*, vol. 18, no. 5, pp. 900–914, 2012.

[122] C. Wang, L. Wei, M. Guo, and Q. Zou, “Computational approaches in detecting non-coding RNA,” *Current Genomics*, vol. 14, no. 6, pp. 371–377, 2013.

[123] Q. Zou, X.-B. Li, W.-R. Jiang, Z.-Y. Lin, G.-L. Li, and K. Chen, “Survey of MapReduce frame operation in bioinformatics,” *Briefings in Bioinformatics*, vol. 15, no. 4, Article ID bbs088, pp. 637–647, 2014.