Hindawi
International Journal of Genomics
Volume 2017, Article ID 6213474, 10 pages
https://doi.org/10.1155/2017/6213474

Review Article

A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries

Kalpana Raja,[1] Matthew Patrick,[1] Yilin Gao,[1] Desmond Madu,[1] Yuyang Yang,[1] and Lam C. Tsoi[1,2,3]

[1] *Department of Dermatology, University of Michigan Medical School, Ann Arbor, MI, USA*
[2] *Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA*
[3] *Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA*

Correspondence should be addressed to Lam C. Tsoi; alextsoi@umich.edu

Received 7 October 2016; Accepted 9 February 2017; Published 26 February 2017

Academic Editor: Margarita Hadzopoulou-Cladaras

In the past decade, the volume of "omics" data generated by the different high-throughput technologies has expanded exponentially. The managing, storing, and analyzing of this big data have been a great challenge for the researchers, especially when moving towards the goal of generating testable data-driven hypotheses, which has been the promise of the high-throughput experimental techniques. Different bioinformatics approaches have been developed to streamline the downstream analyzes by providing independent information to interpret and provide biological inference. Text mining (also known as literature mining) is one of the commonly used approaches for automated generation of biological knowledge from the huge number of published articles. In this review paper, we discuss the recent advancement in approaches that integrate results from omics data and information generated from text mining approaches to uncover novel biomedical information.

1. Introduction

The advances in biotechnology have allowed biomedical research to answer efficiently important biological questions in the different omics scales: genetics, genomics, transcriptomics, epigenomics, proteomics, and metabolomics [1–4]. The omics data can characterize the behaviors of cells, tissues, and organs at the molecular level and allow the comprehensive understanding for the etiology of human diseases. Among the various omics studies, genetic and genomic studies are widely adopted in biomedical research to discover new genes or susceptibility loci associated with different human traits or diseases [5, 6]. Proteomic study is concerned with the structure, function, and modification of proteins expressed in a biological system, specifically the posttranscriptional modifications such as phosphorylation, methylation, and acetylation, which lead to transcription and translation of the same genome into various types of proteomes [7, 8]. Epigenomic study has attracted great attention in the last 5 years. It characterizes the epigenetic modifications of the genome and aims to understand the regulations of the gene expression. Transcriptomic study, in turn, enables the genome-wide assessment of gene expression patterns in cells and tissues by studying the complete set of RNA transcriptomes [9]. Finally, metabolomic study characterizes the metabolites present in cell, tissue, and body fluid and identifies the fluctuation of these metabolites in various disease conditions [10]. The different types of omics studies accumulate a huge volume of data through high-throughput sequencing experiments and provide insights towards the cellular and metabolic processes related to disease diagnoses, treatment, and prevention.

According to the PubMed, over 36,000 research articles have been published in the past ten years and annotated by at least one of the above "omics" experiments (by using the following search phrase: "(genomics [MeSH] OR proteomics [MeSH] OR metabolomics [MeSH] OR transcriptomics [MeSH]) AND humans [MeSH]"). The interest in omics studies has not declined and their applications are evident from the publications in recent years, when compared to only over 10,000 research articles published prior to 2006 by using the same search phrase. However, the acquired data raises various significant challenges: (i) the interpretation of high-throughput results; (ii) the translation of biological data to clinical application; (iii) the data handling, storage, and sharing issues; and (iv) the reproducibility when comparing between different experiments [11, 12]. Among these, the last challenge has been a long-lasting issue, most likely due to the potential discrepancies in processing and interpreting the high-throughput data or due to "cherry-picking" approach to subjectively focus on the components that are indeed false positives. The traditional strategies to overcome these challenges are to conduct extensive literature search and seek professional opinions from domain experts to decipher the mechanism and then conduct downstream experiments to verify the findings. However, this has proven to be time consuming and subjective and has not been a common practice when researchers publish their results from high-throughput experiments. On the other hand, automated approaches have gained much interest in recent years to annotate gene functions [13], to identify biomarkers [14], and to explore genetic mutations [15]. Text mining (also known as literature mining) is a technique that has been used to retrieve and process research articles from PubMed database and can summarize biomedical information present across articles. In molecular biology, text mining is typically used to retrieve relevant documents, prioritize the documents, extract the biomedical concepts (e.g., genes, proteins, cell, tissue, and cell-type), and extract the causal relationships between concepts [16, 17]. Text mining can significantly decrease the time and effort required, compared with traditional labor-intensive approaches.

In this review, we first discuss the various omics techniques used in healthcare and summarize the recent advances in utilizing text mining approaches to facilitate the interpretation and translation of these omics data. We then focus on biomedical literature mining and clinical text mining and further describe the challenges involved in integrating the knowledge from different resources to enhance the biomedical research. Finally, we explain the recent methods to integrate omics and biomedical literature mining data in order to uncover novel biomedical information.

2. The Study of "Omics"

Traditionally, "omics" corresponds to the study of four major biomolecules: genes, proteins, transcriptomes, and metabolites [4]. Since the discovery of DNA [31], much interest has been gained towards understanding the roles of genes and proteins in cellular functions and transduction. Healthcare is considered to vary from one individual to another based on his genome, proteome, transcriptome, and metabolome. The digital revolution has paved the way for integrating patient omics data with the findings in literature for the discovery of novel biomarkers and drug targets [32–34]. Therefore, the study of omics has expanded beyond these four major omics studies, and Table 1 summarizes the various types of omics data applied to biomedical discoveries. The study of omics has introduced the realm of big data to biomedicine [35, 36]. While the first human genome project took more than a decade to complete and involved $3 billion dollars, the entire genome can be sequenced and analyzed within hours for ~$1000 now. Thus, biomedical projects are now possible to generate information at the petabyte (i.e., 1,012 bytes) scale. Nevertheless, the greatest challenge is the large-scale data analysis and its integration with clinical data available in patient electronic health records (EHR) [37].

Cloud [38] and parallel computing [39] are currently used in omics research to handle the huge volume of data. Cloud computing is described as a network of computers connected together through the Internet for effective processing. It is available remotely, through cloud computing providers (e.g., Microsoft, Google, and Amazon), and researchers have an option to make use of it at an affordable cost. Parallel computing speeds up the processing time using the same hardware and Internet setup. The combined approach of using cloud computing and parallel computing together is capable of processing omics data in a feasible time [40, 41]. Other high performance computing platforms include clusters [42], grid computing [43], and graphical processing units [44]. Processing omics data and applying bioinformatics models to the data require expertise to integrate computational, biological, mathematical, and statistical knowledge.

3. Text Mining

PubMed database is a main repository for biomedical literature and contains over 26 million articles. The number of articles being published and indexed by PubMed is increasing exponentially, and therefore text mining has become an attractive (and standard) approach in mining literature data when comparing with the traditional labor-intensive strategies. Researchers use the text mining approach to tackle information overload, both in biomedical and in general areas of big data collection, because it automates data retrieval and information extraction from the unstructured biomedical texts to reveal novel information [45, 46]. While information extraction examines the relationships between specific kinds of information contained within or between documents, information retrieval focuses on summarizing data from the larger units of documents [47]. Another automated approach to deal with unstructured data is Natural Language Processing (NLP). While text mining concentrates on solving a specific problem in a particular domain, NLP attempts to understand the text as a whole [48]. Recently, text mining and NLP have been used to address different biological questions in omics research [49].

3.1. Biomedical Literature Mining. The era of applying text mining approaches to biology and biomedical fields came into existence in 1999. It was first applied to the biomedical domain for gene expression profiling [50], as well as the extraction and visualization of protein-protein interaction [51]. It emerged as a hybrid discipline from the edges of three major fields, namely, bioinformatics, information science, and computational linguistics. Biomedical literature

Table 1: Omics and biomedical applications.

	Omics	Study topic	Biomedical applications†
Genetics/molecular genetics	Genomics	Genes	Gencode, Entrez Gene
	Epigenomics	Epigenetics modifications	Gene Express Omnibus
	Exposomics	Disease-causing environmental factors	Comparative Toxicogenomics Database
	Exomics	Exons in a genome	ICE—a human splice sites database
	ORFeomics	Open Reading Frame (ORF)	—
	Phenomics	Phenotypes	Human Phenotype Ontology
	Pharmacogenomics	Impact of genes on individual's response to drugs	PharmGKB
	Pharmacogenetics	SNPs and their impact on pharmacodynamics and pharmacokinetics	PharmGKB
	Toxicogenomics	Genes response to toxic substances	Comparative Toxicogenomics Database
Molecular biology	Proteomics	Proteins and amino acids	Proteomics Identifications Database (PRIDE)
	Metabolomics	Metabolites	HMDB: Human Metabolome Database
	Transcriptomics	Transcripts (i.e., rRNA, mRNA, tRNA, and microRNA)	Human Transcriptome Map
	Ionomics	Inorganic biomolecules	—
	Kinomics	Protein kinases	KinBase database and KinWeb database
	Metagenomics	Genetic material from multiple organisms	MG-RAST
	Regulomics	Transcription factors and other biomolecules involved in the regulation of gene expression	miRegulome
	Toponomics	Cell and tissue structure	—
Medicine	Trialomics	Human interventional trials from clinical trials	—
	Connectomics	Structural and functional connectivity in brain	—
	Interactomics	Interferons	CREDO

†The list shows example applications.

mining is concerned with the identification and extraction of biomedical concepts (e.g., genes, proteins, DNA/RNA, cells, and cell types) and their functional relationships [17]. The major tasks include (i) document retrieval and prioritization (gathering and prioritizing the relevant documents); (ii) information extraction (extracting information of interest from the retrieved document); (iii) knowledge discovery (discovering new biological event or relationship among the biomedical concepts); and (iv) knowledge summarization (summarizing the knowledge available across the documents). A brief description of the biomedical literature mining tasks is listed as follows.

Biomedical Text Mining Tasks

Document Retrieval. The process of extracting relevant documents from a large collection is called document retrieval or information retrieval [52]. The two basic strategies applied are query-based and document-based retrieval. In query-based retrieval, documents matching with the user specified query are retrieved. In document-based retrieval, a ranked list of documents similar to a document of interest is retrieved.

Document Prioritization. The retrieved documents are usually prioritized to get the most relevant document. Many biomedical document retrieval systems achieve prioritization based on certain parameters including journal-related metrics (e.g., impact factor, citation count) [53] and MeSH index [54, 55] for biomedical articles. The similarity between the documents is estimated with various similarity measurements (e.g., Jaccard similarity, cosine similarity) [56].

Information Extraction. This task aims to extract and present the information in a structured format. Concept extraction and relation/event extraction are the two major components of information extraction [57, 58]. While concept extraction automatically identifies the biomedical concepts present in the articles, relation/event extraction is used to predict the relationship or biological event (e.g., phosphorylation) between the concepts [59, 60].

Knowledge Discovery. It is a nontrivial process to discover novel and potentially useful biological information from the structured text obtained from information extraction. Knowledge discovery uses techniques from a wide range of disciplines such as artificial intelligence, machine learning, pattern recognition, data mining, and statistics [61]. Both information extraction and knowledge discovery find their application in database curation [62, 63] and pathway construction [64, 65].

Knowledge Summarization. The purpose of knowledge summarization is to generate information for a given topic from one or multiple documents. The approach aims to reduce the source text to express the most important key points through content reduction selection and/or generalization [66]. Although knowledge summarization helps to manage the information overload, the state of the art is still open to research to develop more sophisticated approaches that increase the likelihood of identifying the information.

Hypothesis Generation. An important task of text mining is hypothesis generation to predict unknown biomedical facts from biomedical articles. These hypotheses are useful in designing experiments or explaining existing experimental results [67].

Conventional text mining approaches process PubMed abstracts rather than the full-text articles and fail to mine the information not in abstracts. Recently, text mining from the full-text articles is gaining more interest [59]. However, it involves many challenges: (1) the availability of full-text articles is limited (4 million full-text articles in PubMed Central versus 26 million abstracts in PubMed); (2) text mining within tables, figures, and equations is complicated; and (3) information redundancy within the articles. An automated text mining system is generally evaluated using a standard corpus (Table 2). However, the availability of standard corpora in biomedical domain is limited because its generation is expensive, time consuming, and requires domain experts. In general, a gold standard is developed within the research groups when the standard corpora are not available, but mostly not available to other researchers. The text mining systems are commonly evaluated using precision, recall, and f-score. Precision is defined as the relevance accuracy, recall is defined as the retrieval accuracy, and f-score is defined as the harmonic mean of precision and recall [56].

3.2. Clinical Text Mining.

Electronic health records, discharge summaries, and clinical narratives of patients are rich in information that could be useful for improving the healthcare. In addition, the information is also available from the transcription of dictations, direct entry by clinicians/physicians, or speech recognition software. The encoding of structural information from the clinical resources is useful to clinicians and researchers. For example, automated high-throughput clinical applications can be developed to support clinicians' information needs [68]. However, manual encoding is expensive and limited to primary and secondary diagnoses. Clinical text mining, also known as clinical NLP or Medical Language Processing (or simply MLP), is suggested as a potential technology by Institute of Medicine for mining clinical resources. The tasks described above in biomedical literature mining are applicable to clinical text mining and include additional subtasks [69]: (i) negation recognition (e.g., "patient denies on developing rashes"), (ii) temporal extraction (e.g., "small bumps noticed last year"), and (iii) patient-event relationship (e.g., "patient mother had arthritis").

The modern healthcare relies on big data analytics for integrating, organizing, and utilizing different pharmacological or clinical information. A hybrid approach to combine patient genomic data and electronic health record information is expanding as the future vision of healthcare. The omics data has become an emerging tool for diagnosis/clinical investigations of common and rare diseases and helps in clinical decision making (i.e., selecting the best possible treatments for patients). Genome-Wide Association

Table 2: Standard corpora for omics domain.

Corpus	Text mining evaluation task	Brief introduction
JNLPBA (Joint Workshop on NLP in Biomedicine and Its Applications) [18]	Gene/protein concept extraction	The corpus consists of 2,000 PubMed abstracts as training data and 404 PubMed abstracts as test data.
BioCreAtivE 2004 Task 1A dataset [19]	Gene/protein concept extraction	The corpus consists of 15,000 PubMed sentences as training data and 5,000 PubMed sentences as test data.
BioCreAtivE 2 Gene Mention (GM) dataset [20]	Gene/protein concept extraction	The corpus consists of 15,000 PubMed sentences as training data and 5,000 PubMed sentences as test data.
AIMED [21]	Protein-protein interaction	The corpus consists of 225 PubMed abstracts that contain 1,987 sentences with 4,075 protein mentions.
HPRD50 (Human Protein Reference Database) [22]	Protein-protein interaction	The corpus consists of sentences with protein-protein interaction from 50 PubMed abstracts.
BioInfer (Bio Information Extraction Resource) [23]	Protein, gene, and RNA relationships	The corpus consists of 1100 sentences annotated with concept names, relationships, and syntactic dependencies.
IEPA (Interaction Extraction Performance Assessment) [24]	Protein-protein interaction	The corpus consists of more than 200 PubMed sentences annotated with protein-protein interaction.
BioCreAtivE 2.5 Elsevier Corpus [25]	Protein-protein interaction	The corpus consists of 61 PubMed articles as training data and 62 PubMed articles as test data.
BC4GO Corpus [26]	Gene ontology	The corpus consists of 1356 distinct GO terms from 200 PubMed articles.
GREC Corpus [27]	Gene regulation and gene expression events	The corpus consists of 240 PubMed abstracts with annotations on gene regulation and gene expression events.
GETM [28]	Gene expression events	The corpus consists of 150 PubMed abstracts with annotation for gene expression events.
AnEM [29]	Tissue, cell, developing anatomical structure, cellular component	The corpus consists of 500 PubMed sentences with annotations on variety of biomedical concepts.
CellFinder Corpus [30]	Anatomical parts, cell lines, cell types, species, and cell components	The corpus consists of annotations from 10 full-text PubMed articles.

Study (GWAS), also known as Whole Genome Association Study (WGAS), is a relatively new approach for identifying genes (i.e., loci associated with human traits) through rapid scanning of markers across whole DNA or genome [70]. GWAS has been applied also to cancer research for drug repositioning [71], prioritizing susceptible genes in Crohn's disease [72], and analyzing the human variants in the area of precision medicine [73]. As an example, the Michigan Genomics Initiatives (MGI) at the University of Michigan has developed an institutional based DNA and genetics repository combined with patient phenotype. The project aims to bring awareness to each patient/participant about the disease development and response to treatments for better health and wellness. The current studies at MGI include analgesics outcome study (AOS), understanding opioid use in chronic pain patients, a pivotal study on high-frequency nerve block for postamputation pain, Michigan body map (MBM), and positive piggy bag (https://www.michigangenomics.org/).

Clinical text mining faces the following specific challenges: (1) access to patient EHR requires permission from Institutional Review Board (IRB); (2) personal details of the patients should be deidentified; (3) mining approaches depend on the types of clinical documents (e.g., EHR, discharge summary, medical billing, and clinical narratives); (4) mining of dosage information, different types of formulations, and temporal information is demanded; and (5) spelling mistakes and grammatical errors are common in clinical text [69]. The state of the art for both biomedical literature mining and clinical text mining is still open with many challenges and requires more sophisticated and robust approaches.

4. Role of Text Mining in Omics Study

Relationship between concepts of the same kind (e.g., gene-gene) or different kind (e.g., gene-disease) is commonly known as "event" [74]. The events are useful to identify many clinical facts such as disease onset and response to drug treatment. Overwhelming of biomedical articles from omics research has accumulated abundance of information and requires advanced event extraction systems to support the complexity of available information and coverage of varieties of biomedical subdomains [16]. Text mining approaches do not replace the manual curation of biomedical information but support speeding up the process by several-fold [75, 76]. In this section we describe the various text mining approaches developed for mining omics related information.

4.1. Genomics and Text Mining. In the current era of genomics, text mining plays an important role in mining gene-gene interactions [77, 78] and other gene involved interactions (e.g., gene-chemical, gene-disease) [79, 80] to support integrative analysis of gene expression [81, 82], pathway construction [83, 84], ontology development [85], and database annotation [62, 86, 87].

Genes encode proteins and proteins enroll in various biological functions by interacting with other proteins. This encoding process is defined in two steps: transcription (i.e., DNA to RNA) and translation (RNA to protein). Many cellular processes are regulated by microRNA through mRNA degradation and suppression of gene expression such that the protein synthesis is interrupted. This is the fundamental of genomics. In genomics, gene function is assessed from the involvement of genes/proteins in biochemical pathways. The functional genomics is a revolutionary area in text mining where the gene/protein mentions in the biomedical articles and their relationship are considered to be important. Furthermore, gene and protein names are highly complex and text mining has contributed to their recognition in the unstructured text [57, 58].

Different text mining implementations for exploring the finding of genome research have been developed in the past decade. miRTex is a text mining system developed for mining experimentally validated microRNA gene targets from PubMed articles. The system has been successfully implemented to identify the Triple Negative Breast Cancer related genes that are regulated by microRNAs [81]. More sophisticated approaches integrate gene expressions from microarray experiments, biomedical data extracted by text mining, and gene interaction data to predict gene-based drug indications [82]. A similar approach [87] attempts to support manual curation of links between biological databases such as Gene Expression Omnibus (GEO) and PubMed database. Another approach [88] combines text mining data with microarray data for discovering disease-gene association by using unsupervised clustering. The gene-drug interaction information extracted by text mining is used to predict the drug-drug interaction [89]. Above all, the researchers have attempted to use text mining for annotating genome function with gene ontology [90]. Thus, text mining and genomics together uncover much biomedical information that was previously unknown.

4.2. Proteomics and Text Mining. Protein-protein interaction is important to explore the mechanism involved in biological processes and onset of diseases [91]. Intact [92], BIND [93], MIND [94], and DIP [95] are the major databases available for protein-protein interaction. These databases are manually curated by the domain experts, but a larger portion of information is still available only in the biomedical literature. Text mining provides a bridge to cover the gap existing between the manual curation and information hidden in the literature. The approaches to extract protein-protein interaction range from simple rule-based systems and cooccurrence systems to more sophisticated NLP methods [60] and machine learning systems [96]. Apart from protein-protein interaction extraction systems, text mining also provides automated approaches for extracting posttranslational modification of proteins such as protein phosphorylation [59].

4.3. Transcriptomics, Metabolomics, and Text Mining. Text mining approaches for transcriptomics and metabolomics are limited. One major fact is that these two areas of genomics are comparatively new when compared to genomics and proteomics. A recent study compares the metagenome characteristics of healthy individuals with autism patients to analyze the enzymes involved [97]. The computational approach uses text mining for genomics and metabolomics information extraction. A web-based tool called 3Omics is available for integrating, comparing, analyzing, and visualizing data from transcriptomics, metabolomics, and proteomics [98]. Another tool called Babelomics integrates transcriptomics, proteomics, and genomics data to uncover the underlying function profiles [99]. Thus, a wide variety of hidden biomedical information within the omics data are extracted and predicted through text mining.

5. Conclusion

In this review, we summarized the current state of the art in omics research and contribution of text mining approaches to uncover the omics related biomedical information hidden within the published articles. We discussed the core concepts of omics and the challenges involved in storing and analyzing the huge volume of omics data generated from high-throughput experiments. We also highlighted the use of computer techniques such as parallel processing and cloud computing to manage omics data and elaborated on text mining approaches for biomedical literature and clinical text with emphasis on omics. While the omics approach is emerging to be commonly used practice for basic science or clinical diagnosis technique, it is imminent to note that data interpretation and translation is the bottleneck. The advances in text mining can be useful to resolve the challenges with the omics data and further support in novel biomedical discoveries.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors acknowledge the support from the Undergraduate Research Opportunity Program (UROP) from the University of Michigan, the Dermatology Foundation, the Arthritis National Research Foundation, and the National Psoriasis Foundation.

References

[1] E. Morra, M. Lazzarino, E. P. Alessandrino et al., "Central nervous system (CNS) leukemia: the role of high dose cytarabine (HDAra-C)," *Bone Marrow Transplantation*, vol. 4, supplement 1, pp. 101–103, 1989.

[2] D. B. Kell, "The virtual human: towards a global systems biology of multiscale, distributed biochemical network models," *IUBMB Life*, vol. 59, no. 11, pp. 689–695, 2007.

[3] H. V. Westerhoff and B. O. Palsson, "The evolution of molecular biology into systems biology," *Nature Biotechnology*, vol. 22, no. 10, pp. 1249–1252, 2004.

[4] R. P. Horgan and L. C. Kenny, "'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics," *The Obstetrician & Gynaecologist*, vol. 13, no. 3, pp. 189–195, 2011.

[5] L. C. Tsoi, S. L. Spain, E. Ellinghaus et al., "Enhanced meta-analysis and replication studies identify five new psoriasis susceptibility loci," *Nature Communications*, vol. 6, Article ID 7001, 2015.

[6] D. Bertrand, K. R. E. Chng, F. G. H. Sherbaf et al., "Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles," *Nucleic acids research*, vol. 43, no. 7, p. e44, 2015.

[7] P. James, "Protein identification in the post-genome era: the rapid rise of proteomics," *Quarterly Reviews of Biophysics*, vol. 30, no. 4, pp. 279–331, 1997.

[8] G. A. Khoury, R. C. Baliban, and C. A. Floudas, "Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database," *Scientific Reports*, vol. 1, article 90, 2011.

[9] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[10] J. K. Nicholson, J. C. Lindon, and E. Holmes, ""Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," *Xenobiotica*, vol. 29, no. 11, pp. 1181–1189, 1999.

[11] K. Shoenbill, N. Fost, U. Tachinardi, and E. A. Mendonca, "Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations," *Journal of the American Medical Informatics Association*, vol. 21, no. 1, pp. 171–180, 2014.

[12] G. Poste, "Bring on the biomarkers," *Nature*, vol. 469, no. 7329, pp. 156–157, 2011.

[13] Q. Lu, R. L. Powles, Q. Wang, B. J. He, and H. Zhao, "Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies," *PLoS Genetics*, vol. 12, no. 4, Article ID e1005947, 2016.

[14] L. Puchades-Carrasco, M. Palomino-Schätzlein, C. Pérez-Rambla, and A. Pineda-Lucena, "Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers," *Briefings in Bioinformatics*, vol. 17, no. 3, pp. 541–552, 2016.

[15] S. A. Forbes, D. Beare, P. Gunasekaran et al., "COSMIC: exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Research*, vol. 43, pp. D805–D811, 2015.

[16] S. Ananiadou, P. Thompson, R. Nawaz, J. McNaught, and D. B. Kell, "Event-based text mining for biology and functional genomics," *Briefings in Functional Genomics*, vol. 14, no. 3, pp. 213–230, 2015.

[17] M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology," *Genome Biology*, vol. 6, no. 7, article no. 224, 2005.

[18] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets," *Genome Research*, vol. 14, no. 6, pp. 1085–1094, 2004.

[19] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman, "BioCreAtIvE task 1A: gene mention finding evaluation," *BMC Bioinformatics*, vol. 6, no. 1, article no. S2, 2005.

[20] A. Vlachos, "Tackling the BioCreative2 gene mention task with conditional random fields and syntactic parsing," in *Proceedings of the 2nd BioCreative Challenge Evaluation Workshop*, Madrid, Spain, April 2007.

[21] R. Bunescu, R. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

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

[23] S. Pyysalo, F. Ginter, J. Heimonen et al., "BioInfer: a corpus for information extraction in the biomedical domain," *BMC Bioinformatics*, vol. 8, article 50, 2007.

[24] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele, "Mining MEDLINE: abstracts, sentences, or phrases?" *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 7, pp. 326–337, 2002.

[25] F. Leitner, M. Krallinger, G. Cesareni, and A. Valencia, "The FEBS letters SDA corpus: a collection of protein interaction articles with high quality annotations for the BioCreative II.5 online challenge and the text mining community," *FEBS Letters*, vol. 584, no. 19, pp. 4129–4130, 2010.

[26] K. Van Auken, M. L. Schaeffer, P. McQuilton et al., "BC4GO: a full-text corpus for the BioCreative IV GO task," *Database*, vol. 2014, Article ID bau074, 2014.

[27] P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction," *BMC Bioinformatics*, vol. 10, article 349, 2009.

[28] M. Gerner, G. Nenadic, and C. M. Bergman, "An exploration of mining gene expression mentions and their anatomical locations from biomedical text," in *Proceedings of the Workshop on Biomedical Natural Language Processing*, pp. 72–80, Association for Computational Linguistics, Uppsala, Sweden, July 2010.

[29] T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou, "Open-domain anatomical entity mention detection," in *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, Association for Computational Linguistics, Jeju, Korea, July 2012.

[30] M. Neves, E. Damaschun, A. Kurtz, and U. Leser, "Annotating and evaluating text for stem cell research," in *Proceedings of the 3rd Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM '12) at Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012.

[31] L. A. Pray, "Discovery of DNA structure and function: Watson and Crick," *Nature Education*, vol. 1, no. 1, article 100, 2008.

[32] N. T. Issa, S. W. Byers, and S. Dakshanamurthy, "Big data: the next frontier for innovation in therapeutics and healthcare," *Expert Review of Clinical Pharmacology*, vol. 7, no. 3, pp. 293–298, 2014.

[33] S. Jiang, T. E. Hinchliffe, and T. Wu, "Biomarkers of an autoimmune skin disease-psoriasis," *Genomics, Proteomics and Bioinformatics*, vol. 13, no. 4, pp. 224–233, 2015.

[34] A. Tebani, C. Afonso, S. Marret, and S. Bekri, "Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations," *International Journal of Molecular Sciences*, vol. 17, no. 9, p. 1555, 2016.

[35] J. M. Rothberg, W. Hinz, T. M. Rearick et al., "An integrated semiconductor device enabling non-optical genome sequencing," *Nature*, vol. 475, no. 7356, pp. 348–352, 2011.

[36] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, "Continuous base identification for single-molecule nanopore DNA sequencing," *Nature Nanotechnology*, vol. 4, no. 4, pp. 265–270, 2009.

[37] V. Canuel, B. Rance, P. Avillach, P. Degoulet, and A. Burgun, "Translational research platforms integrating clinical and omics data: a review of publicly available solutions," *Briefings in Bioinformatics*, vol. 16, no. 2, pp. 280–290, 2015.

[38] L. Griebel, H. Prokosch, F. Köpcke et al., "A scoping review of cloud computing in healthcare," *BMC Medical Informatics and Decision Making*, vol. 15, article 17, 2015.

[39] K. Ocaña and D. De Oliveira, "Parallel computing in genomic research: advances and applications," *Advances and Applications in Bioinformatics and Chemistry*, vol. 8, pp. 23–35, 2015.

[40] D. P. Wall, P. Kudtarkar, V. A. Fusaro, R. Pivovarov, P. Patil, and P. J. Tonellato, "Cloud computing for comparative genomics," *BMC Bioinformatics*, vol. 11, article no. 259, 2010.

[41] M. Armbrust, A. Fox, R. Griffith et al., "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[42] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster ComSpark: cluster computing with working sets," in *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, p. 10, Boston, Mass, USA, June 2010.

[43] M. Baker, R. Buyya, and D. Laforenza, "Grids and grid technologies for wide-area distributed computing," *Software—Practice & Experience*, vol. 32, no. 15, pp. 1437–1466, 2002.

[44] I. S. Ufimtsev and T. J. Martinez, "Quantum chemistry on graphical processing units. 2. Direct self-consistent-field implementation," *Journal of Chemical Theory and Computation*, vol. 5, no. 4, pp. 1004–1015, 2009.

[45] M. A. Hearst, "Untangling text data mining," in *Proceedings of the the 37th annual meeting of the Association for Computational Linguistics (ACL '99)*, pp. 3–10, College Park, Maryland, June 1999.

[46] K. B. Cohen and L. Hunter, "Natural language processing and systems biology," in *Artificial Intelligence Methods and Tools for Systems Biology*, vol. 5 of *Computational Biology*, pp. 147–173, Springer, Dordrecht, The Netherlands, 2004.

[47] M. Weeber, H. Klein, A. R. Aronson, J. G. Mork, L. T. de Jong-van den Berg, and R. Vos, "Text-based discovery in biomedicine: the architecture of the DAD-system," *Proceedings of the AMIA Symposium*, pp. 903–907, 2000.

[48] A. S. Yeh, L. Hirschman, and A. A. Morgan, "Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup," *Bioinformatics*, vol. 19, supplement 1, pp. i331–i339, 2003.

[49] Y. Liu, Y. Liang, and D. Wishart, "PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more," *Nucleic Acids Research*, vol. 43, no. 1, pp. W535–W542, 2015.

[50] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein, "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling," *BioTechniques*, vol. 27, no. 6, pp. 1210–1217, 1999.

[51] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia, "Automatic extraction of biological information from scientific text: protein-protein interactions," in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pp. 60–67, AAAI, Heidelberg, Germany, August 1999.

[52] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, ACM Press, 2nd edition, 2011.

[53] Y. Lin, W. Li, K. Chen, and Y. Liu, "A document clustering and ranking system for exploring MEDLINE citations," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 651–661, 2007.

[54] S. J. Darmoni, L. F. Soualmia, C. Letord et al., "Improving information retrieval using medical subject headings concepts: a test case on rare and chronic diseases," *Journal of the Medical Library Association*, vol. 100, no. 3, pp. 176–183, 2012.

[55] M. Petrova, P. Sutcliffe, K. W. M. Fulford, and J. Dale, "Search terms and a validated brief search filter to retrieve publications on health-related values in Medline: a word frequency analysis study," *Journal of the American Medical Informatics Association*, vol. 19, no. 3, pp. 479–488, 2012.

[56] C. D. Manning, P. Raghavan, and H. Schuetze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[57] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," in *Proceedings of the 13th Pacific Symposium on Biocomputing (PSB '08)*, pp. 652–663, Kohala Coast, Hawaii, USA, January 2008.

[58] K. Raja, S. Subramani, and J. Natarajan, "A hybrid named entity tagger for tagging human proteins/genes," *International Journal of Data Mining and Bioinformatics*, vol. 10, no. 3, pp. 315–328, 2014.

[59] M. Torii, C. N. Arighi, G. Li, Q. Wang, C. H. Wu, and K. Vijay-Shanker, "RLIMS-P 2.0: a generalizable rule-based information extraction system for literature mining of protein phosphorylation information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 1, pp. 17–29, 2015.

[60] K. Raja, S. Subramani, and J. Natarajan, "PPInterFinder—a mining tool for extracting causal relations on human proteins from literature," *Database (Oxford)*, vol. 2013, Article ID bas052, 2013.

[61] J. Natarajan, D. Berrar, C. J. Hack, and W. Dubitzky, "Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications," *Critical Reviews in Biotechnology*, vol. 25, no. 1-2, pp. 31–52, 2005.

[62] K. E. Ravikumar, K. B. Wagholikar, D. Li, J.-P. Kocher, and H. Liu, "Text mining facilitates database curation—extraction of mutation-disease associations from Bio-medical literature," *BMC Bioinformatics*, vol. 16, no. 1, article 185, 2015.

[63] S. Matos, D. Campos, R. Pinho et al., "Mining clinical attributes of genomic variants through assisted literature curation in Egas," *Database (Oxford)*, vol. 2016, Article ID baw096, 2016.

[64] S. Subramani, R. Kalpana, P. M. Monickaraj, and J. Natarajan, "HPIminer: a text mining system for building and visualizing human protein interaction networks and pathways," *Journal of Biomedical Informatics*, vol. 54, pp. 121–131, 2015.

[65] J. Czarnecki, I. Nobeli, A. M. Smith, and A. J. Shepherd, "A text-mining system for extracting metabolic reactions from full-text articles," *BMC Bioinformatics*, vol. 13, no. 1, article 172, 2012.

[66] R. Mishra, J. Bian, M. Fiszman et al., "Text summarization in the biomedical domain: a systematic review of recent research," *Journal of Biomedical Informatics*, vol. 52, pp. 457–467, 2014.

[67] F. Zhu, P. Patumcharoenpol, C. Zhang et al., "Biomedical text mining and its applications in cancer research," *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 200–211, 2013.

[68] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research," *Yearbook of medical informatics*, pp. 128–144, 2008.

[69] K. Raja and S. R. Jonnalagadda, "Natural language processing and data mining for clinical text," in *Healthcare Data Analytics*, C. K. Reddy and C. C. Aggarwal, Eds., pp. 219–250, CRC Press, 2015.

[70] D. Welter, J. MacArthur, J. Morales et al., "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1001–D1006, 2014.

[71] J. Zhang, K. Jiang, L. Lv et al., "Use of genome-wide association studies for cancer research and drug repositioning," *PLoS ONE*, vol. 10, no. 3, Article ID e0116477, 2015.

[72] D. Muraro, D. A. Lauffenburger, and A. Simmons, "Prioritisation and network analysis of Crohn's disease susceptibility genes," *PLoS ONE*, vol. 9, no. 9, Article ID e108624, 2014.

[73] T. A. Peterson, E. Doughty, and M. G. Kann, "Towards precision medicine: advances in computational approaches for the analysis of human variants," *Journal of Molecular Biology*, vol. 425, no. 21, pp. 4047–4063, 2013.

[74] J.-D. Kim, N. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa, "The genia event and protein coreference tasks of the BioNLP shared task 2011," *BMC bioinformatics*, vol. 13, supplement 11, p. S1, 2012.

[75] T. C. Wiegers, A. P. Davis, K. B. Cohen, L. Hirschman, and C. J. Mattingly, "Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD)," *BMC Bioinformatics*, vol. 10, article 1471, p. 326, 2009.

[76] L. Hirschman, G. A. P. C. Burns, M. Krallinger et al., "Text mining for the biocuration workflow," *Database*, vol. 2012, Article ID bas020, 2012.

[77] E. K. Mallory, C. Zhang, C. Ré, and R. B. Altman, "Large-scale extraction of gene interactions from full-text literature using DeepDive," *Bioinformatics*, vol. 32, no. 1, pp. 106–113, 2015.

[78] J. Hur, A. Özgür, Z. Xiang, and Y. He, "Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions," *Journal of Biomedical Semantics*, vol. 6, no. 1, article no. 2, 2015.

[79] A. P. Davis, C. J. Grondin, R. J. Johnson et al., "The comparative toxicogenomics database: update 2017," *Nucleic Acids Research*, vol. 45, 2017.

[80] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "DISEASES: text mining and data integration of disease-gene associations," *Methods*, vol. 74, pp. 83–89, 2015.

[81] G. Li, K. E. Ross, C. N. Arighi, Y. Peng, C. H. Wu, and K. Vijay-Shanker, "miRTex: a text mining system for mirna-gene relation extraction," *PLoS Computational Biology*, vol. 11, no. 9, Article ID e1004391, 2015.

[82] A. Qabaja, T. Jarada, A. Elsheikh, and R. Alhajj, "Prediction of gene-based drug indications using compendia of public gene expression data and PubMed abstracts," *Journal of Bioinformatics and Computational Biology*, vol. 12, no. 3, Article ID 14500073, 2014.

[83] E. Donnard, A. Barbosa-Silva, R. L. M. Guedes et al., "Preimplantation development regulatory pathway construction through a text-mining approach," *BMC Genomics*, vol. 12, no. 4, article S3, 2011.

[84] R. Lehmann, L. Childs, P. Thomas et al., "Assembly of a comprehensive regulatory network for the mammalian circadian clock: a bioinformatics approach," *PLoS ONE*, vol. 10, no. 5, Article ID e0126283, 2015.

[85] H. Chen, D. Han, Y. Dai, and L. Zhao, "Design of automatic extraction algorithm of knowledge points for MOOCs," *Computational Intelligence and Neuroscience*, vol. 2015, Article ID 123028, 10 pages, 2015.

[86] R. Weikard, F. Hadlich, and C. Kuehn, "Identification of novel transcripts and noncoding RNAs in bovine skin by deep next

generation sequencing," *BMC Genomics*, vol. 14, no. 1, article no. 789, 2013.

[87] A. Neveol, W. J. Wilbur, and Z. Lu, "Improving links between literature and biological data with text mining: a case study with GEO, PDB and MEDLINE," *Database*, vol. 2012, Article ID bas026, 2012.

[88] A. Faro, D. Giordano, and C. Spampinato, "Combining literature text mining with microarray data: advances for system biology modeling," *Briefings in Bioinformatics*, vol. 13, no. 1, Article ID bbr018, pp. 61–82, 2012.

[89] B. Percha, Y. Garten, and R. B. Altman, "Discovery and explanation of drug-drug interactions via text mining," in *Proceedings of the 17th Pacific Symposium on Biocomputing (PSB '12)*, pp. 410–421, Kohala Coast, Hawaii, USA, January 2012.

[90] J. M. Daley, H. Niu, A. S. Miller, and P. Sung, "Biochemical mechanism of DSB end resection and its regulation," *DNA Repair*, vol. 32, pp. 66–74, 2015.

[91] M. G. Kann, "Protein interactions and disease: computational approaches to uncover the etiology of diseases," *Briefings in Bioinformatics*, vol. 8, no. 5, pp. 333–346, 2007.

[92] S. Kerrien, Y. Alam-Faruque, B. Aranda et al., "IntAct—open source resource for molecular interaction data," *Nucleic Acids Research*, vol. 35, no. 1, pp. D561–D565, 2007.

[93] G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, and C. W. V. Hogue, "BIND—The Biomolecular Interaction Network Database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 242–245, 2001.

[94] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "MINT: a molecular INTeraction database," *FEBS Letters*, vol. 513, no. 1, pp. 135–140, 2002.

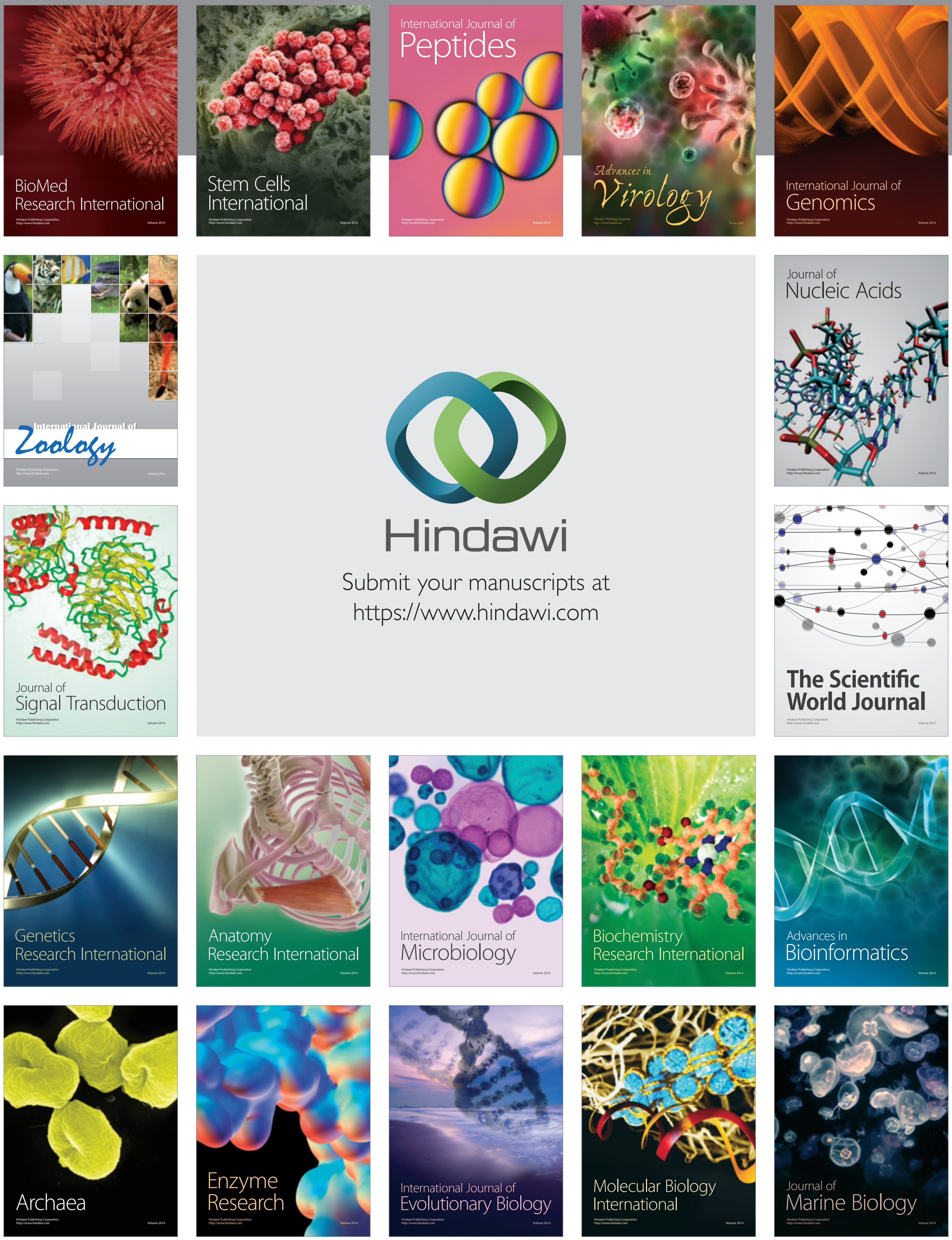
[95] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D449–D451, 2004.

[96] Q.-C. Bui, S. Katrenko, and P. M. A. Sloot, "A hybrid approach to extract protein-protein interactions," *Bioinformatics*, vol. 27, no. 2, pp. 259–265, 2011.

[97] C. Heberling and P. Dhurjati, "Novel systems modeling methodology in comparative microbial metabolomics: identifying key enzymes and metabolites implicated in autism spectrum disorders," *International Journal of Molecular Sciences*, vol. 16, no. 4, pp. 8949–8967, 2015.

[98] T.-C. Kuo, T.-F. Tian, and Y. J. Tseng, "3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data," *BMC Systems Biology*, vol. 7, article 64, 2013.

[99] I. Medina, J. Carbonell, L. Pulido et al., "Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling," *Nucleic Acids Research*, vol. 38, no. 2, pp. W210–W213, 2010.


BioMed Research International
Stem Cells International
International Journal of Peptides
Advances in Virology
International Journal of Genomics
International Journal of Zoology
Journal of Nucleic Acids
Journal of Signal Transduction
The Scientific World Journal
Genetics Research International
Anatomy Research International
International Journal of Microbiology
Biochemistry Research International
Advances in Bioinformatics
Archaea
Enzyme Research
International Journal of Evolutionary Biology
Molecular Biology International
Journal of Marine Biology