WJBC World Journal of Biological Chemistry

Online Submissions: http://www.wjgnet.com/1949-8454office
wjbc@wjgnet.com
doi:10.4331/wjbc.v3.i2.27

World J Biol Chem 2012 February 26; 3(2): 27-33
ISSN 1949-8454 (online)

EDITORIAL

Knowledge acquisition, semantic text mining, and security risks in health and biomedical informatics

Jingshan Huang, Dejing Dou, Jiangbo Dang, J Harold Pardue, Xiao Qin, Jun Huan, William T Gerthoffer, Ming Tan

Jingshan Huang, J Harold Pardue, School of Computer and Information Sciences, University of South Alabama, Mobile, AL 36688, United States
Dejing Dou, Computer and Information Science Department, University of Oregon, Eugene, OR 97403, United States
Jiangbo Dang, Knowledge and Decision Systems Group, Siemens Corporate Research, Princeton, NY 08540, United States
Xiao Qin, Department of Computer Science and Software Engineering, Samuel Ginn College of Engineering, Auburn University, Auburn, AL 36849, United States
Jun Huan, Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66047, United States
William T Gerthoffer, Department of Biochemistry and Molecular Biology, University of South Alabama, Mobile, AL 36688, United States
Ming Tan, Mitchell Cancer Institute; Department of Cell Biology and Neuroscience, University of South Alabama, Mobile, AL 36688, United States
Author contributions: Huang J, Tan M and Gerthoffer WT organize the whole paper; Huang J, Qin X and Huan J contributed to introduction and research motivation, current bio-ontologies, and concluding remarks; Huang J and Qin X contributed to background knowledge in ontologies; Dou D and Qin X contributed to ontological techniques in medical and biological research; Dang J contributed to semantic text mining on clinical and biomedical data section; Pardue JH contributed to security risks to medical data research.
Correspondence to: Jingshan Huang, PhD, Assistant Professor, School of Computer and Information Sciences, University of South Alabama, FCW 20, School of CIS, 307 University Blvd. N., Mobile, AL 36688, United States. huang@usouthal.edu
Telephone: +1-251-4607612 Fax: +1-251-4607274
Received: October 28, 2011 Revised: December 7, 2011
Accepted: December 14, 2011
Published online: February 26, 2012

Abstract

Computational techniques have been adopted in medical and biological systems for a long time. There is no doubt that the development and application of computational methods will render great help in better understanding biomedical and biological functions. Large amounts of datasets have been produced by biomedical and biological experiments and simulations. In order for researchers to gain knowledge from original data, nontrivial transformation is necessary, which is regarded as a critical link in the chain of knowledge acquisition, sharing, and reuse. Challenges that have been encountered include: how to efficiently and effectively represent human knowledge in formal computing models, how to take advantage of semantic text mining techniques rather than traditional syntactic text mining, and how to handle security issues during the knowledge sharing and reuse. This paper summarizes the state-of-the-art in these research directions. We aim to provide readers with an introduction of major computing themes to be applied to the medical and biological research.

Key words: Biomedical informatics; Bioinformatics; Knowledge sharing; Ontology matching; Heterogeneous semantics; Semantic integration; Semantic data mining; Semantic text mining; Security risk

Peer reviewer: Buyong Ma, PhD, Center for Cancer Research Nanobiology Program, SAIC-Frederick, National Cancer Institute at Frederick, NIH, Frederick, MD 21702, United States

Huang J, Dou D, Dang J, Pardue JH, Qin X, Huan J, Gerthoffer WT, Tan M. Knowledge acquisition, semantic text mining, and security risks in health and biomedical informatics. *World J Biol Chem* 2012; 3(2): 27-33 Available from: URL: http://www.wjgnet.com/1949-8454/full/v3/i2/27.htm DOI: http://dx.doi.org/10.4331/wjbc.v3.i2.27

INTRODUCTION AND RESEARCH MOTIVATION

Applying computational techniques to the simulation and analysis of medical and biological systems has a long history dating back to the earliest analog and even mechanical computers. With the explosion of data brought about by various medical and biological techniques such as modern molecular techniques, it is now increasingly clear to many researchers that future progress in better understanding biomedical and biological functions relies inescapably on the development and application of innovative and advanced computational methods. Biomedical and biological experiments and simulations now routinely produce petascale datasets, a prelude to the even larger, extreme-scale datasets that are about to be common in the near future. Unfortunately, most of datasets collected by medical scientists and biologists are not sufficient for analysis by themselves. On the contrary, data must be transformed into knowledge to be of any real value. Transforming data to knowledge is a nontrivial process and is regarded as a critical link in the chain of knowledge acquisition, sharing, and reuse. It is essential for medical scientists and biologists to obtain an enhanced ability (1) to gain knowledge and understanding from data of increasing size and complexity; and (2) to perform hypothesis testing and knowledge discovery in petascale data. Only this way is it possible for us to change fundamentally our understanding about how humans perceive and gain knowledge from large, complex biological datasets resulting from a variety of experiments and from extreme-scale simulations. Fundamental advances in computing are needed during the aforementioned transformation from data to knowledge. In such a transformation process, scientists are facing three main challenges among others: (1) efficient and effective methods to represent human knowledge in formal computing models; (2) semantic instead of traditional text mining techniques; and (3) security issues when sharing and reusing the knowledge obtained from original datasets.

The rest of this paper addresses the aforementioned three challenges and is organized as follows. We first introduce the background knowledge in ontologies, a formal computing model in knowledge representation. Then we survey currently popular bio-ontologies. Next, we summarize the state-of-the-art research efforts in ontological techniques applied to the medical and biological fields, semantic text mining, and security risks, respectively. Finally, we conclude the paper with some remarks.

BACKGROUND KNOWLEDGE IN ONTOLOGIES

Ontology is a computational model of some portion or domain of the world[1]. The model describes the semantics of terms used in some domain of interest. Ontology is often captured in some forms of a semantic network, i.e., a graph whose nodes are concepts or individual objects and whose arcs represent relationships or associations among the concepts. The semantic network is augmented by properties and attributes, constraints, functions, and rules, which govern the behavior of the concepts. In brief, an ontology consists of a finite set of concepts, along with these concepts' properties and relationships. Note that some ontologies also contain instances in addition to the aforementioned graphical structure (also known as "schema").

Ontology heterogeneity is an inherent characteristic of ontologies developed by different parties for the same (or similar) domains. The heterogeneous semantics may occur in two scenarios. (1) Different ontologies could use different terminologies to describe the same conceptual model. That is, different terms could be used for the same concept, or an identical term could be adopted for different concepts; and (2) Even if two ontologies share the same name for a specific concept, that concept's associated properties and relationships with other concepts are most likely to be different.

Ontology matching is a short term for "ontology schema matching", also known as "ontology alignment", or "ontology mapping". It is the process of determining correspondences between concepts from heterogeneous ontologies (often designed by distributed parties). Such correspondences include many relationships, for example, *equivalentWith*, *subClassOf*, *superClassOf*, and *siblings*.

CURRENT BIO-ONTOLOGIES

The value of any kind of data is greatly enhanced when data exist in a form allowing integration with other data. One approach to integration is through the annotation of multiple bodies of data using common controlled vocabularies or ontologies. Therefore, research in this area has led to a proliferation of bio-ontologies. The most successful example is the Gene Ontology (GO) Project[2], which is a major bioinformatics initiative aiming to standardize the representation of gene and gene product attributes across species and databases. Consisting of three sub-ontologies, i.e., *Biological Process*, *Cellular Component*, and *Molecular Function*, GO provides a controlled vocabulary of terms for describing gene product characteristics and annotation data, as well as tools to access and process such data. The focus of GO is to describe how gene products behave in a cellular context. Besides, research has been carried out for ontology-based data integration in bioinformatics. Note that GO itself is part of a larger classification effort, the Open Biomedical Ontologies (OBO) (short for OBO, formerly Open Biological Ontologies), which is an effort to create controlled vocabularies for shared use across different biological and medical domains.

Other existing bio-ontologies that are commonly used by biological and biomedical researchers include, but not limited to, the RiboWeb ontology, the EcoCyc ontology, the Schulze-Kremer ontology for molecular biology

Table 1 Summary of contents, structure, and representation of several bio-ontologies (from Stevens *et al*[3])

Ontology	Application scenario	Modularised?	Domain-oriented component	Task-oriented component	Generic component	Instances	Detail level	KR
GO	Controlled vocabulary for database annotation	Partially	Drosophila, mouse and yeast gene function gene product function, process and cellular location and structure	×	×	√	High	×
EcoCyc	Database schema	√	*Escherichia coli* genes, metabolism, regulation, signal transduction and metabolic pathways	Visualization of biochemical reactions and layout of genes with chromosome	√	√	High	Frames
MBO	Community reference	√	Shallow	Shallow	√	×	Low	×
RiboWeb	Database schema	√	Ribosome Components, covalently bonded molecules, biological macromolecules, regions of molecules	Experimental detail, techniques for analysing data, publication	√	√	High	Frames
TaO	Common access ontology-based search	Partially	Proteins, enzymes, motifs, secondary and tertiary structure, functions and processes, subcellular structure and chemicals, including cofactors. The larger model includes nucleic acid and genes	Bioinformatics search and analysis tasks	√	×	High	DLs

(MBO), and the TAMBIS Ontology (TaO). Table 1 from Stevens *et al*[3] summarizes these bio-ontologies with respect to their organization, structure, purpose, and contents. The column *Domain-oriented Component* includes domain-specific components and domain generalization components; the column *Task-oriented Component* identifies task-specific components and task generalization components; and the column *KR* demonstrates the type of knowledge representation used.

ONTOLOGICAL TECHNIQUES IN MEDICAL AND BIOLOGICAL RESEARCH

Ontological techniques have been widely applied to medical and biological research. All seven systems surveyed in this section have been developed upon some bio-ontologies. We briefly introduce six such systems, followed by the last one where more description of biological aspect is provided to help the reader better understand how biomedical and biological informatics may facilitate domain experts to gain biological insights.

Cantor *et al*[4] have discussed the issue of mapping concepts in GO to the Unified Medical Language System (UMLS). Such a mapping may allow the exploitation of the UMLS semantic network to link disparate genes, through their annotation in GO, to unique clinical outcomes, potentially uncovering biological relationships. This study reveals the inherent difficulties in the integration of vocabularies created in different manners and by specialists in different fields, as well as the strengths of different techniques used to accomplish this integration.

Köhler *et al*[5] have described principles and methods used to implement Semantic Meta Database (SEMEDA). Database owners may use SEMEDA to provide semantically integrated access to their databases; the owners may also collaboratively edit and maintain ontologies and controlled vocabularies. Biologists can use SEMEDA to query the integrated databases in real time without prior knowledge of structures or any technical details of the underlying databases. The authors aim to handle technical problems of database integration and issues related to semantics, e.g., the use of different terms for the same items, different names for equivalent database attributes, and missing links between relevant entries in different databases.

Sulman *et al*[6] have reported a high-resolution integrated map of the region constructed (CompView) to identify all markers in the smallest region of overlapping deletion. A regional somatic cell hybrid panel is used to localize more precisely those markers identified in CompView as within or overlapping the region, and a sequence from clones is used to validate STS content by electronic PCR and to identify transcripts. The authors have concluded that the annotation of a putative tumor suppressor locus provides a resource for further analysis of meningioma candidate genes.

Jakoniene *et al*[7] have argued that during the process of retrieving and information integration from multiple biological data sources, approaches should be enhanced by ontological knowledge. Jakoniene *et al*[7] have identified different types of ontological knowledge that are available on the Internet. In the light of the ontological knowledge, they have proposed an approach to supporting integrated access to multiple biological data sources. Their work also shows that current ontology-based integration approaches only cover parts of their proposed approaches.

Birkland *et al*[8] have presented a system, Biozon, to address the problems encountered in the integration of heterogeneous data types in the biology domain. Biozon offers biologists a new knowledge resource to navigate through and explore by unifying multiple biological da-

tabases that consist of a variety of data types (e.g., DNA sequences, proteins, interactions, and cellular pathways). Biozon is different from previous efforts in the sense that it uses a single extensive and tightly connected graph schema wrapped with hierarchical ontology of documents and relations. Beyond warehousing existing data, Biozon computes and stores novel derived data, similarity relationships and functional predictions, for example. The integration of similarity data allows propagation of knowledge through inference and fuzzy searches.

The value of any kind of data is greatly enhanced when data exist in a form allowing the integration with other data. One approach to integration is through the annotation of multiple bodies of data using common controlled vocabularies or ontologies. Unfortunately, the very success of this approach has led to a proliferation of ontologies, which itself creates obstacles to integration. In order to overcome such problems, Smith *et al*[9] have described a strategy, namely, the OBO Foundry initiative. The long-term goal of the OBO initiative is that the data generated through biomedical research should form a single, consistent, cumulatively expanding, and algorithmically tractable whole. Efforts to realize this goal are still very much in the proving stage. Nevertheless, the initial efforts reflect an attempt to walk the line between the flexibility that is indispensable to scientific advance and the institution of principles that is indispensable to successful coordination.

Huang *et al*[10] have presented a domain-specific knowledge base built upon the Ontology for MicroRNA Targets (OMIT) to facilitate knowledge acquisition in the field of miRNA target gene prediction. The identification and characterization of important roles that miRNAs perform in human cancer has increasingly become an active research area. However, the prediction of miRNA target genes remains a challenging task to cancer researchers. Current prediction processes are time-consuming, error-prone, and subject to biologists' limited prior knowledge. The OMIT system aims to assist biologists in unraveling important roles of miRNAs in human cancer; thus, OMIT can help clinicians make sound decisions when treating cancer patients. To be more specific, it is well known that each miRNA can have hundreds of possible target genes. Currently, there are many different target prediction databases that are geographically distributed worldwide and that have adopted quite different schemas and terminologies. Moreover, in many cases, additional information for target genes is critical for biologists to understand fully these genes' biological functions. More often than not, such additional information is not available in target prediction databases. Instead, other resources such as the GO ontologies are needed for this purpose. Taking *mir-21* as an example, miRDB, TargetScan, and PicTar report 348, 210, and 175 target genes for *mir-21*, respectively. It is very challenging, if not impossible, for biologists to search manually a total of 733 candidate target genes, let alone to further search for useful information on each gene hidden in GO. In fact, the situation could be even worse: biologists usually make use of more than three databases in the miRNA research area. To handle this challenge, the OMIT framework helps biologists discover miRNA candidate target genes in a much more efficient manner: (1) knowledge from various databases is automatically obtained, integrated, and presented to users; and (2) related information from GO is provided for each retrieved target gene. In this manner, biologists can save a large amount of time that would have been spent if a manual search were to be carried out. In addition, due to inference engines (also known as ontology reasoners) specifically designed for OWL ontologies, OMIT is able to identify hidden knowledge that is not explicit in the original data. For example, combining the information of "mir-21 promotes hepatoCellularCarcinoma" obtained from the knowledge base, with the fact that "hepatoCellularCarcinoma" is defined as an instance of the concept *Carcinoma*, which in turn is a subclass of the concept *MalignantNeoplasm*, a new conclusion, "mir-21 promotes MalignantNeoplasm", is acquired by reasoning on the concept hierarchy. Similarly, another conclusion, "mir-21 promotes Tumor", can be readily obtained as well. These extra conclusions will help biologists to generalize their findings to more model systems.

SEMANTIC TEXT MINING ON CLINICAL AND BIOMEDICAL DATA

With ontological knowledge bases, text-mining approaches are evolving and have been applied to a wide range of healthcare applications for clustering clinical data and extracting clinical answers. In general, there are four steps to generate annotated text from raw text for semantic text mining: (1) tokenization; (2) lexical processing; (3) syntactic processing to identify sentence structure; and (4) semantic processing with ontologies.

Spasic *et al*[11] have summarized different approaches in applying ontologies to text-mining applications in biomedicine. They emphasize that ontologies and terminological lexicons are prerequisites for advanced text mining. They have reviewed various approaches for information retrieval (IR), information extraction (IE), and machine learning (ML). Spasic and his team have concluded that ontologies can be of help in all three tasks in text mining. Ontologies help to relax exact matching in IR, and the hierarchical organization of ontologies and relations between described concepts can be used to constrain or relax a search query. In IE, ontologies can be used in both passive and active ways to extract entities. In ML, ontologies can be applied in term classification, term clustering, and term relation extraction.

Although radiology reports contain valuable information, most are just filed and not referred to later. Gong *et al*[12] have proposed a text-mining system to extract and use the information in radiology reports. The text-mining system includes three modules: (1) the Medical Finding Extractor, (2) the Report and Image Retriever, and (3) the Text-Assisted Image Feature Extractor. The first module,

i.e., Medical Finding Extractor, aims to extract medical findings in radiology reports. This module first identifies medical terms from free text radiology reports based on medical lexicons, then extracts findings and modifiers based on semantic rules, and finally it represents findings in the extensible markup language (XML) format; (2) The second module, i.e., Report and Image Retriever, makes report contents searchable. It uses a query analyzer and then uses either exact match or partial match to find field reports; and (3) The last module, i.e., Text-Assisted Image Feature Extractor, uses abnormality detection from text mining to assist feature extraction in medical imaging processing to get favorable results.

Semantic decision support systems can supplement semantic knowledge bases as secondary sources to assist physicians in making sound clinical decisions. Lin *et al*[13] have investigated a clinical evidence retrieval system and have hypothesized that grouping retrieved MEDLINE citations into semantically coherent clusters, based on automatically extracted interventions from the abstract text, represents an effective strategy for presenting results, compared with a traditional ranked list. Based on this hypothesis, they have designed a workflow that: (1) identifies all entities belonging to chemicals and drugs, devices, and procedures from retrieved abstracts; (2) extracts main interventions by assigning each intervention (and the associated abstract) to its own cluster; (3) iteratively merges clusters whose interventions share a common UMLS hypernym, ascending the UMLS hierarchy in the process; and (4) sorts results in the order of the original PubMed results within each formed cluster.

Bundschus *et al*[14] have focused on extracting both the existence of a relation and its type from biomedical text using conditional random fields. Bundschus *et al*[14] have conducted two sets of experiments: (1) disease-treatment relation extraction from PubMed abstracts; and (2) gene-disease relation extraction from GeneRIF. Based on the experiments, they have concluded that the discriminative approaches Conditional Random Fields (CRFs) and Artificial Neural Networks (ANNs) are better than the Generative approach in this task. However, ANNs suffer from feature numbers. Training of CRFs is much faster than Support Vector Machine (SVM) because no feature selection is needed.

SECURITY RISKS TO MEDICAL DATA RESEARCH

In addition to an impact on the quality and rate of discovery and innovation, the abundance of digital data arising from medical informatics promises to transform healthcare through the meaningful use of electronic health records and medical data. Year 2015 is the deadline set by the American Recovery and Reinvestment Act of 2009 for hospitals, clinics, and practices to adopt level 1 meaningful use. This legislation has as its goal the development of a nationwide health and medical information technology infrastructure. This infrastructure is designed to improve healthcare outcomes, reduce healthcare costs, and further innovation and discovery by integrating technology into the flow of clinical practice. Health and medical data must be interoperable, private, and secure[15].

As this health and medical data technology infrastructure emerges and evolves, a host of threats to security and privacy will likewise emerge and evolve. Healthcare and medical systems are a very complex interplay of technologies, people, policy, and legislation. However, these systems operate in an especially challenging security environment because healthcare and medical data must be captured and retrieved at the point-of-care, that is, it must be mobile and wireless.

This section focuses on risks to security and privacy of healthcare and medical data in terms of security and privacy threats. A threat can be defined as the exploitation of a system vulnerability. Vulnerability is defined as "a flaw or weakness in system security procedures, design, implementation, or internal controls that could be exploited to accomplish a security breach or a violation of the system's security policy[16]".

Much work has been done to catalog and classify vulnerabilities to healthcare and medical data[17-21]. Landry *et al*[17] have developed a threat tree to assess and manage risks to healthcare and medical data. Kotz[19] has proposed a framework that organizes a set of 25 threats by identity threats, access threats, and disclosure threats. Samy *et al*[21] have identified 22 categories of health information systems threats. Their research has identified five critical areas, namely, power failure/loss, acts of human error or failure, technological obsolescence, hardware failures or errors, and software failures or errors. The focus of this literature review is vulnerabilities associated with unauthorized manipulation, data loss, and data corruption. Consequences of exploiting security vulnerabilities include exposure to economic harm, mental anguish, social stigma, identity theft, and poor healthcare and medical outcomes. What follows is a selected list of vulnerabilities reported in the literature.

Vandalism of health and medical data can be thought of in the broader context of cyber protest. Cyber protest is an expression of a social movement through the use of information technologies[22]. Cyber protesters have long targeted controversial healthcare such as abortion or medical animal research[23-25].

Samy *et al*[21] have ranked hardware and software failures among the top five out of 22 threats to healthcare and medical data. Hardware failures have long been a security and privacy issue in the application of information technology to the management of healthcare and medical data[26,27]. Hardware failures include "hard" failures (e.g., hard drive crashes and failure of backup technology) as well as "soft" failures (e.g., poor planning of storage requirements and external media).

If inappropriately implemented, information technology can actually lead to healthcare and medical errors that represent an unintended exploitation of a vulnerability. For example, recent studies[28-30] have shown that com-

puter physician order entry (CPOE) systems can facilitate medication and order entry errors. There are two major causes for this type of error. First, healthcare and medical data come from a wide range of sources and in varying formats. If medical data are not correctly integrated in the database or inappropriately juxtaposed on a computer screen, medical errors can occur. Second, the introduction of information technology into the flow of point-of-care processes inevitably changes and affects the normal flow of events. Caregivers operate in a multi-tasked, time-sensitive, and frenetic environment. If the creation and retrieval of healthcare and medical data poorly matches the normal flow of healthcare delivery, time-sensitive therapies such as early resuscitation may be delayed to the detriment of patients' health and chances of survival.

Malicious threats to the healthcare and medical infrastructure are very real. One needs only recall the 1982 Tylenol cyanide contamination case. The information technology equivalent is the threat of malware[31]. Developing "tamper-proof" mechanisms for the information technology that manages healthcare and medical data is a nontrivial challenge. An example of a class of Tylenol-like threats is malware in embedded medical devices[32,33]. Medical devices must be connected to other devices and servers through wireless communications, therefore, the data transmitted *via* wireless networks are vulnerable to both security and privacy threats.

Next-generation information technology infrastructure will have to increase the availability and accuracy of healthcare and medical data. However, unauthorized access is still a serious vulnerability. There are many forms of and reasons for unauthorized access such as acquiring medical treatment with a different person's insurance policy[20]. Such unauthorized access is regarded as a form of identity theft. Note that a common vulnerability arises from the very nature of healthcare delivery. The first directive of caregivers is to do no harm and to save life. So-called "break-the-glass" life-or-death situations require that data access policies are circumvented, thus exposing healthcare and medical data to unauthorized access.

CONCLUDING REMARKS

Innovative computing methodologies built upon the increasingly pervasive cyber infrastructure are required in order for medical scientists and biologists to obtain an enhanced ability to integrate, share, and reuse originally heterogeneous data from distributed laboratories. Researchers are facing many challenges to revolutionize efficiently the traditional medical and biological research, to conceptualize data, and to acquire in-depth knowledge out of original datasets thereafter. In this paper, we have highlighted state-of-the-art research efforts in three related fields: ontological techniques in medical and biological research; semantic text mining on clinical and biomedical data; and security risks to medical data. The goal of this paper is to provide readers with an introduction to major computing themes that can help scientists obtain a better understanding of important biological functions at different levels.

REFERENCES

1 **Singh MP**, Huhns MN. Service-Oriented Computing - Semantics, Processes, Agents. 1st ed. Chichester: John Wiley and Sons Ltd., 2005
2 Gene Ontology Website. September 2011. Available from: URL: http://www.geneontology.org/index.shtml
3 **Stevens R**, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform* 2000; **1**: 398-414
4 **Cantor MN**, Sarkar IN, Gelman R, Hartel F, Bodenreider O, Lussier YA. An evaluation of hybrid methods for matching biomedical terminologies: mapping the gene ontology to the UMLS. *Stud Health Technol Inform* 2003; **95**: 62-67
5 **Köhler J**, Philippi S, Lange M. SEMEDA: ontology based semantic integration of biological databases. *Bioinformatics* 2003; **19**: 2420-2427
6 **Sulman EP**, White PS, Brodeur GM. Genomic annotation of the meningioma tumor suppressor locus on chromosome 1p34. *Oncogene* 2004; **23**: 1014-1020
7 **Jakoniene V,** Lambrix P. Ontology-based Integration for Bioinformatics. Proceedings of the VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems; 2005 Sep 2-3; Trondheim, Norway. New York: Springer, 2007: 55-58
8 **Birkland A**, Yona G. BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics* 2006; **7**: 70
9 **Smith B**, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007; **25**: 1251-1255
10 **Huang J**, Townsend C, Dou D, Liu H, Tan M. OMIT: a domain-specific knowledge base for microRNA target prediction. *Pharm Res* 2011; **28**: 3101-3104
11 **Spasic I**, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform* 2005; **6**: 239-251
12 **Gong T**, Tan CL, Leong TY, Lee CK, Pang BC, Tchoyoson Lim CC, Tian Q, Tang S, Zhang Z. Text mining in radiology reports. 8th IEEE International Conference on Data Mining; 2008 Dec 15-19; Pisa: Data Mining, 2008: 815-820
13 **Lin J**, Demner-Fushman D. Semantic clustering of answers to clinical questions. *AMIA Annu Symp Proc* 2007; 458-462
14 **Bundschus M**, Dejori M, Stetter M, Tresp V, Kriegel HP. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics* 2008; **9**: 207
15 **President's Council of Advisors on Science and Technology**. Report to the President Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward. 2010
16 **Stoneburner G**, Goguen A, Feringa A. Risk management guide for information technology systems: recommendations of the National Institute of Standards and Technology. Gaithersburg, MD: US Dept. of Commerce, National Institute of Standards and Technology, 2002
17 **Landry JP**, Pardue JH, Johnsten T, Campbell M, Patidar P. A Threat Tree for Health Information Security and Privacy. Detroit, MI: Americas Conference on Information Systems, 2011
18 **Appari A**, Johnson ME. Information security and privacy in healthcare: current state of research. *IJIEM* 2010; **6**: 279-314
19 **Kotz D.** A threat taxonomy for mHealth privacy. In: Workshop on Networked Healthcare Technology. Bangalore: Communication Systems and Networks, 2011

20 **Nematzadeh A**, Camp LJ. Threat analysis of online health information system. In: Makedon F, Maglogiannis I, Kapidakis S, editors. Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments. New York: Association for Computing Machinery, 2010: 1-7
21 **Samy GN**, Ahmad R, Ismail Z. Security threats categories in healthcare information systems. *Health Informatics J* 2010; **16**: 201-209
22 **Zimbra D**, Abbasi A, Chen H. A cyber-archaeology approach to social movement research: Framework and case study. *JCMC* 2010; **16**: 48-70
23 **Curran WJ**, Stearns B, Kaplan H. Privacy, confidentiality and other legal considerations in the establishment of a centralized health-data system. *N Engl J Med* 1969; **281**: 241-248
24 **Forrest JD**, Henshaw SK. Providing controversial health care: abortion services since 1973. *Womens Health Issues* 1993; **3**: 152-157
25 **Jackson T**. Website of the week: Animal research. *BMJ* 2001; **322**: 244
26 **Hersh W**. Health care information technology: progress and barriers. *JAMA* 2004; **292**: 2273-2274
27 **Kilbridge P**. Computer crash--lessons from a system failure. *N Engl J Med* 2003; **348**: 881-882
28 **Ash JS**, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assoc* **2004; 11**: 104-112
29 **Han YY**, Carcillo JA, Venkataraman ST, Clark RS, Watson RS, Nguyen TC, Bayir H, Orr RA. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. *Pediatrics* 2005; **116**: 1506-1512
30 **Koppel R**, Metlay JP, Cohen A, Abaluck B, Localio AR, Kimmel SE, Strom BL. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005; **293**: 1197-1203
31 **Keese J**, Motzo L. Pro-active approach to malware for healthcare information and imaging systems. In: CARS 2005: Computer Assisted Radiology and Surgery. Salt Lake City: Elsevier, 2005: 943-947
32 **Fu K**. Inside risks: Reducing risks of implantable medical devices. *CACM* 2009; **52**: 25-27
33 **Maisel WH**, Kohno T. Improving the security and privacy of implantable medical devices. *N Engl J Med* 2010; **362**: 1164-1166

S- Editor Cheng JX **L- Editor** Kerr C **E- Editor** Zheng XM