Chapter 1
KNOWLEDGE MANAGEMENT, DATA MINING, AND TEXT MINING IN MEDICAL INFORMATICS

Hsinchun Chen[1], Sherrilynne S. Fuller[2], Carol Friedman[3], and William Hersh[4]

[1]Management Information Systems Department, Eller College of Management, University of Arizona, Tucson, Arizona 85721; [2]University of Washington, Biomedical and Health Informatics, Seattle, Washington 98195-7155; [3]Columbia University, Department of Biomedical Informatics New York, New York 10032; [4]Oregon Health and Science University, Medical Informatics and Clinical Epidemiology, Portland, Oregon 97239-3098

Chapter Overview

In this chapter we provide a broad overview of selected knowledge management, data mining, and text mining techniques and their use in various emerging biomedical applications. It aims to set the context for subsequent chapters. We first introduce five major paradigms for machine learning and data analysis including: probabilistic and statistical models, symbolic learning and rule induction, neural networks, evolution-based algorithms, and analytic learning and fuzzy logic. We also discuss their relevance and potential for biomedical research. Example applications of relevant knowledge management, data mining, and text mining research are then reviewed in order including: ontologies; knowledge management for health care, biomedical literature, heterogeneous databases, information visualization, and multimedia databases; and data and text mining for health care, literature, and biological data. We conclude the paper with discussions of privacy and confidentiality issues of relevance to biomedical data mining.

Keywords

knowledge management; data mining; text mining

1. INTRODUCTION

The field of biomedical informatics has drawn increasing popularity and attention, and has been growing rapidly over the past two decades. Due to the advances in new molecular, genomic, and biomedical techniques and applications such as genome sequencing, protein identification, medical imaging, and patient medical records, tremendous amounts of biomedical research data are generated every day. Originating from individual research efforts and clinical practices, these biomedical data are available in hundreds of public and private databases, which have been made possible by new database technologies and the Internet. The digitization of critical medical information such as lab reports, patient records, research papers, and anatomic images has also resulted in large amounts of patient care data. Biomedical researchers and practitioners are now facing the "info-glut" problem. Currently, the rate of data accumulation is much faster than the rate of data interpretation. These data need to be effectively organized and analyzed in order to be useful.

New computational techniques and information technologies are needed to manage these large repositories of biomedical data and to discover useful patterns and knowledge from them. In particular, knowledge management, data mining, and text mining techniques have been adopted in various successful biomedical applications in recent years. *Knowledge management* techniques and methodologies have been used to support the storing, retrieving, sharing, and management of multimedia and mission-critical tacit and explicit biomedical knowledge. *Data mining* techniques have been used to discover various biological, drug discovery, and patient care knowledge and patterns using selected statistical analyses, machine learning, and neural networks methods. *Text mining* techniques have been used to analyze research publications as well as electronic patient records. Biomedical entities such as drug names, proteins, genes, and diseases can be automatically extracted from published documents and used to construct gene pathways or to provide mapping into existing medical ontologies.

In the following sections, we first survey the background of knowledge management, data mining, and text mining research. We then discuss the use of these techniques in emerging biomedical applications.

2. KNOWLEDGE MANAGEMENT, DATA MINING, AND TEXT MINING: AN OVERVIEW

Knowledge management, data mining, and text mining techniques have been widely used in many important applications in both scientific and business domains in recent years.

Knowledge management is the system and managerial approach to the gathering, management, use, analysis, sharing, and discovery of knowledge in an organization or a community in order to maximize performance (Chen, 2001). Although there is no universal definition of what constitutes knowledge, it is generally agreed there is a continuum of data, information, and knowledge. Data are mostly structured, factual, and oftentimes numeric, and reside in database management systems. Information is factual, but unstructured, and in many cases textual. Knowledge is inferential, abstract, and is needed to support decision making or hypothesis generation. The concept of knowledge has become prevalent in many disciplines and business practices. For example, information scientists consider taxonomies, subject headings, and classification schemes as representations of knowledge. Consulting firms also have been actively promoting practices and methodologies to capture corporate knowledge assets and organizational memory. In the biomedical context, knowledge management practices often need to leverage existing clinical decision support, information retrieval, and digital library techniques to capture and deliver tacit and explicit biomedical knowledge.

Data mining is often used during the knowledge discovery process and is one of the most important subfields in knowledge management. *Data mining aims to analyze a set of given data or information in order to identify novel and potentially useful patterns* (Fayyad et al., 1996). These techniques, such as Bayesian models, decision trees, artificial neural networks, associate rule mining, and genetic algorithms, are often used to discover patterns or knowledge that are previously unknown to the system and the users (Dunham, 2002; Chen and Chau, 2004). Data mining has been used in many applications such as marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, among others.

Text mining aims to extract useful knowledge from textual data or documents (Hearst, 1999; Chen, 2001). Although text mining is often considered a subfield of data mining, some text mining techniques have originated from other disciplines, such as information retrieval, information visualization, computational linguistics, and information science. Examples of text mining applications include document classification, document clustering, entity extraction, information extraction, and summarization.

Most knowledge management, data mining, and text mining techniques involve learning patterns from existing data or information, and are therefore built upon the foundation of *machine learning* and *artificial intelligence*. In the following, we review several major paradigms in machine learning, important evaluation methodologies, and their applicability in biomedicine.

2.1 Machine Learning and Data Analysis Paradigms

Since the invention of the first computer in the 1940's, researchers have been attempting to create knowledgeable, learnable, and intelligent computers. Many knowledge-based systems have been built for various applications such as medical diagnosis, engineering troubleshooting, and business decision-making (Hayes-Roth and Jacobstein, 1994). However, most of these systems have been designed to acquire knowledge manually from human experts, which can be a very time-consuming and labor-intensive process. To address this problem, machine learning algorithms have been developed to acquire knowledge automatically from examples or source data. Simon (1983) defined machine learning as "*any process by which a system improves its performance*." Mitchell (1997) gives a similar definition, which considers machine learning to be "*the study of computer algorithms that improve automatically through experience*." Although the "machine learning" term has been widely adopted in the computer science community, in the context of medical informatics, "data analysis" is more commonly used to represent "*the study of computer algorithms that improve automatically through the analysis of data*." Statistical data analysis has long been adopted in biomedical research.

In general, machine learning algorithms can be classified as supervised learning or unsupervised learning. In supervised learning, training examples consist of input/output pair patterns. Learning algorithms aim to predict output values of new examples based on their input values. In unsupervised learning, training examples contain only the input patterns and no explicit target output is associated with each input. The unsupervised learning algorithms need to use the input values to discover meaningful associations or patterns.

Many successful machine learning systems have been developed over the past three decades in the computer science and statistics communities. Chen and Chau (2004) categorized five major paradigms of machine learning research, namely probabilistic and statistical models, symbolic learning and rule induction, neural networks, evolution-based models, and analytic learning and fuzzy logic. We will briefly review research in each of these areas and discuss their applicability in biomedicine.

2.1.1 Probabilistic and Statistical Models

Probabilistic and statistical analysis techniques and models have the longest history and strongest theoretical foundation for data analysis. Although it is not rooted in artificial intelligence research, statistical analysis achieves data analysis and knowledge discovery objectives similar to machine learning. Popular statistical techniques, such as regression analysis, discriminant analysis, time series analysis, principal component analysis, and multi-dimensional scaling, are widely used in biomedical data analysis and are often considered benchmarks for comparison with other newer machine learning techniques.

One of the more advanced and popular probabilistic models in biomedicine is the *Bayesian model*. Originating in pattern recognition research (Duda and Hart, 1973), this method was often used to classify different objects into predefined classes based on a set of features. A Bayesian model stores the probability of each class, the probability of each feature, and the probability of each feature given each class, based on the training data. When a new instance is encountered, it can be classified according to these probabilities (Langley et al., 1992). A variation of the Bayesian model, called the *Naïve Bayesian model*, assumes that all features are mutually independent within each class. Because of its simplicity, the Naïve Bayesian model has been adopted in different domains (Fisher, 1987; Kononenko, 1993). Due to its mathematical rigor and modeling elegance, Bayesian learning has been widely used in biomedical data mining research, in particular, genomic and microarray analysis.

A machine learning technique gaining increasing recognition and popularity in recent years is the *support vector machines* (SVMs). SVM is based on statistical learning theory that tries to find a hyperplane to best separate two or multiple classes (Vapnik, 1998). This statistical learning model has been applied in different applications and the results have been encouraging. For example, it has been shown that SVM achieved the best performance among several learning methods in document classification (Joachims, 1998; Yang and Liu, 1999). SVM is also suitable for various biomedical classification problems, such as disease state classification based on genetic variables or medical diagnosis based on patient indicators.

2.1.2 Symbolic Learning and Rule Induction

Symbolic learning can be classified according to its underlying learning strategy such as rote learning, learning by being told, learning by analogy, learning from examples, and learning from discovery (Cohen and Feigenbaum, 1982; Carbonell et al., 1983). Among these, *learning from* *examples* appears to be the most promising symbolic learning approach for knowledge discovery and data mining. It is implemented by applying an algorithm that attempts to induce a general concept description that best describes the different classes of the training examples. Numerous algorithms have been developed, each using one or more different techniques to identify patterns that are useful in generating a concept description. Quinlan's ID3 decision-tree building algorithm (Quinlan, 1983) and its variations such as C4.5 (Quinlan, 1993) have become one of the most widely used symbolic learning techniques. Given a set of objects, ID3 produces a decision tree that attempts to classify all the given objects correctly. At each step, the algorithm finds the attribute that best divides the objects into the different classes by minimizing entropy (information uncertainty). After all objects have been classified or all attributes have been used, the results can be represented by a decision tree or a set of production rules.

Although not as powerful as SVM or neural networks (in terms of classification accuracy), symbolic learning techniques are computationally efficient and their results are easy to interpret. For many biomedical applications, the ability to interpret the data mining results in a way understandable to patients, physicians, and biologists is invaluable. Powerful machine learning techniques such as SVM and neural networks often suffer because they are treated as a "black-box."

2.1.3 Neural Networks

Artificial neural networks attempt to achieve human-like performance by modeling the human nervous system. A neural network is a graph of many active nodes (neurons) that are connected with each other by weighted links (synapses). While knowledge is represented by symbolic descriptions such as decision trees and production rules in symbolic learning, knowledge is learned and remembered by a network of interconnected neurons, weighted synapses, and threshold logic units (Rumelhart et al., 1986a; Lippmann, 1987). Based on training examples, learning algorithms can be used to adjust the connection weights in the network such that it can predict or classify unknown examples correctly. Activation algorithms over the nodes can then be used to retrieve concepts and knowledge from the network (Belew, 1989; Kwok, 1989; Chen and Ng, 1995).

Many different types of neural networks have been developed, among which the *feedforward/backpropagation model* is the most widely used. Backpropagation networks are fully connected, layered, feed-forward networks in which activations flow from the input layer through the hidden layer and then to the output layer (Rumelhart et al., 1986b). The network usually starts with a set of random weights and adjusts its weights according to each learning example. Each learning example is passed through the network to activate the nodes. The network's actual output is then compared with the target output and the error estimates are then propagated back to the hidden and input layers. The network updates its weights incrementally according to these error estimates until the network stabilizes. Other popular neural network models include Kohonen's *self-organizing map* and the *Hopfield network*. Self-organizing maps have been widely used in unsupervised learning, clustering, and pattern recognition (Kohonen, 1995); Hopfield networks have been used mostly in search and optimization applications (Hopfield, 1982). Due to their performances (in terms of predictive power and classification accuracy), neural networks have been widely used in experiments and adopted for critical biomedical classification and clustering problems.

2.1.4 Evolution-based Algorithms

Evolution-based algorithms rely on analogies to natural processes and Darwinian *survival of the fittest*. Fogel (1994) identifies three categories of evolution-based algorithms: *genetic algorithms*, *evolution strategies*, and *evolutionary programming*. Among these, genetic algorithms are the most popular and have been successfully applied to various optimization problems. Genetic algorithms were developed based on the principle of genetics (Holland, 1975; Goldberg, 1989; Michalewicz, 1992). A population of individuals in which each individual represents a potential solution is first initiated. This population undergoes a set of genetic operations known as *crossover* and *mutation*. Crossover is a high-level process that aims at exploitation while mutation is a unary process that aims at exploration. Individuals strive for survival based on a selection scheme that is biased toward selecting fitter individuals (individuals that represent better solutions). The selected individuals form the next generation and the process continues. After some number of generations the program converges and the optimum solution is represented by the best individual. In medical informatics research, genetic algorithms are among the most robust techniques for feature selection problems (e.g., identifying a subset of genes that are most relevant to a disease state) due to their stochastic, global-search capability.

2.1.5 Analytic Learning and Fuzzy Logic

Analytic learning represents knowledge as logical rules and performs reasoning on such rules to search for proofs. Proofs can be compiled into more complex rules to solve similar problems with a smaller number of searches required. For example, Samuelson and Rayner (1991) used analytic learning to represent grammatical rules that improve the speed of a parsing system.

While traditional analytic learning systems depend on hard computing rules, there is usually no clear distinction between values and classes in the real world. To address this problem, *fuzzy systems* and *fuzzy logic* have been proposed. Fuzzy systems allow the values of False or True to operate over the range of real numbers from 0 to 1 (Zedah, 1965). Fuzziness has been applied to allow for imprecision and approximate reasoning. In general, we see little adoptation of such approaches in biomedicine.

2.1.6 Hybrid Approach

As Langley and Simon (1995) pointed out, the reasons for differentiating the paradigms are "more historical than scientific." The boundaries between the different paradigms are usually unclear and many systems have been built to combine different approaches. For example, fuzzy logic has been applied to rule induction and genetic algorithms (e.g., Mendes et al., 2001), genetic algorithms have been combined with neural network (e.g., Maniezzo, 1994; Chen and Kim, 1994), and because neural network has a close resemblance to probabilistic model and fuzzy logic they can be easily mixed (e.g., Paass, 1990). It is not surprising to find that many practical biomedical knowledge management, data mining, and text mining systems adopt such a hybrid approach.

2.2 Evaluation Methodologies

The accuracy of a learning system needs to be evaluated before it can become useful. Limited availability of data often makes estimating accuracy a difficult task (Kohavi, 1995). Choosing a good evaluation methodology is very important for machine learning systems development.

There are several popular methods used for such evaluation, including *holdout sampling*, *cross validation*, *leave-one-out*, and *bootstrap sampling* (Stone, 1974; Efron and Tibshirani, 1993). In the holdout method, data are divided into a training set and a testing set. Usually 2/3 of the data are assigned to the training set and 1/3 to the testing set. After the system is trained by the training set data, the system predicts the output value of each instance in the testing set. These values are then compared with the real output values to determine accuracy.

In cross-validation, a data set is randomly divided into a number of subsets of roughly equal size. Ten-fold cross validation, in which the data set is divided into 10 subsets, is most commonly used. The system is trained and tested for 10 iterations. In each iteration, 9 subsets of data are used as training data and the remaining set is used as testing data. In rotation, each subset of data serves as the testing set in exactly one iteration. The accuracy of the system is the average accuracy over the 10 iterations. Leave-one-out is the extreme case of cross-validation, where the original data are split into *n* subsets, where *n* is the size of the original data. The system is trained and tested for *n* iterations, in each of which *n*–1 instances are used for training and the remaining instance is used for testing.

In the bootstrap method, *n* independent random samples are taken from the original data set of size *n*. Because the samples are taken with replacement, the number of unique instances will be less than *n*. These samples are then used as the training set for the learning system, and the remaining data that have not been sampled are used to test the system (Efron and Tibshirani, 1993).

Each of these methods has its strengths and weaknesses. Several studies have compared them in terms of their accuracies. Hold-out sampling is the easiest to implement, but a major problem is that the training set and the testing set are not independent. This method also does not make efficient use of data since as much as 1/3 of the data are not used to train the system (Kohavi, 1995). Leave-one-out provides the most unbiased estimate, but it is computationally expensive and its estimations have very high variances, especially for small data sets (Efron, 1983; Jain et al., 1987). Breiman and Spector (1992) and Kohavi (1995) conducted independent experiments to compare the performance of several different methods, and the results of both experiments showed ten-fold cross validation to be the best method for model selection.

In light of the significant medical and patient consequences associated with many biomedical data mining applications, it is critical that a systematic validation method be adopted. In addition, a detailed, qualitative validation of the data mining or text mining results needs to be conducted with the help of domain experts (e.g., physicians and biologists), and therefore this is generally a time-consuming and costly process.

3. KNOWLEDGE MANAGEMENT, DATA MINING, AND TEXT MINING APPLICATIONS IN BIOMEDICINE

Knowledge management, data mining, and text mining techniques have been applied to different areas of biomedicine, ranging from patient record management to clinical diagnosis, from hypothesis generation to gene clustering, and from spike signal detection to protein structure prediction. In this section, we briefly survey some of the relevant research in the field, covering the applications of learning techniques in knowledge management, and data mining and text mining in biomedicine. More exhaustive and detailed reviews and discussions of selected knowledge management, data mining, and text mining techniques and applications in biomedicine can be found in the subsequent chapters in this book.

3.1 Ontologies

Before we examine different biomedical applications, it is important to understand the role of ontologies in knowledge management and knowledge discovery, especially for text mining applications. *An ontology is a specification of conceptualization. It describes the concepts and relationships that can exist and formalizes the terminology in a domain* (Gruninger and Lee, 2002). Ontologies are often used to facilitate knowledge sharing between people, information processing, data mining, communication between software agents, or other knowledge processing applications.

Many ontologies have been developed in the biomedical field. The *Unified Medical Language System* (UMLS), supported by the National Library of Medicine (NLM), is a major resource for facilitating computer programs to process and manage biomedical documents (McCray et al., 1993; Humphreys et al., 1993; Campbell et al., 1998; Humphreys et al., 1998). The UMLS offers three knowledge sources: the Metathesaurus, the Semantic Network, and the Specialist Lexicon. The Metathesaurus is a large multilingual controlled vocabulary database for biomedicine that allows users to map biomedical names and textual terms to concepts (i.e., controlled vocabulary terms), or to identify a set of different terms that are associated with a single concept. The Metathesaurus is formed by integrating about 100 different controlled vocabularies including the Medical Subject Headings (MeSH), a controlled vocabulary, and SNOMED-CT, a controlled clinical vocabulary established by the College of American Pathologists. The Semantic Network provides the categorization of the concepts in the Metathesaurus and also the relationships among the concepts. The Specialist Lexicon, designed to facilitate natural language processing for biomedical text, is a lexicon containing syntactic definitions for both biomedical terms and general English terms. These resources provide a framework and ontology for knowledge representation in biomedicine. UMLS resources have been widely used in biomedical language processing (Baclawski et al., 2000; Bodenreider and McCray, 2003; Perl and Geller, 2003; Rosse and Mejino, 2003; Zhang et al., 2003; Caviedes and Cimino, 2004). Several studies have investigated the mapping of concepts from the Metathesaurus to the Semantic Network (Cimino et al., 2003; Rindflesch and Fiszman, 2003).

Besides biomedical documents, it is also important for researchers and computers to understand the different terminologies for genes and proteins. The *Gene Ontology* (GO) project is an effort to address the need for consistent descriptions of gene products in different databases (The Gene Ontology Consortium, 2000). Aiming to produce a dynamic, controlled vocabulary of genes that can be applied to all eukaryotes, the project includes many databases, including FlyBase (Drosophila), the Saccharomyces Genome Database (SGD), the Mouse Genome Database (MGD), and several other major genome databases. GO consists of three structured ontologies that describe genes and gene products. GO terms are also cross-referenced with indexes from other databases. Similarly, the *Human Genome Nomenclature* (HUGO) specifies the standard, approved names and symbols for human genes (Wain et al., 2002). Most of this data can be searched on the Web as text files. There are numerous public databases specifying gene and gene products that are associated with multiple organisms as well as with specific model organisms.

3.2 Knowledge Management

Artificial intelligence techniques have been used in knowledge management in biomedicine as early as the 1970s, when the *MYCIN* program was developed to support consultation and decision making (Shortliffe, 1976). In MYCIN, the knowledge obtained from experts was represented as a set of IF-THEN production rules. Systems of this type would be later known as *expert systems* and become very popular in the 1980s. Expert systems relied on expert knowledge that was *engineered* into it, which was a time-consuming and labor-intensive process.

The performance of MYCIN was encouraging and it even outperformed human experts in some cases (Yu et al., 1979). Despite its early success, it was never used in actual clinical settings. Other medical diagnostic systems were also seldom used clinically. The reasons were two-fold. First, people were skeptical about computer technologies and system performances. Computers were not popular at that time, and many physicians did not believe that computers could perform better than humans. Second, computers were big, expensive machines in the 1970s. It was not feasible to support complex programs like MYCIN on an affordable computer to provide fast responses (Shortliffe, 1987). However, with the improved performance and lower cost of modern computers and medical knowledge-based systems, we believe there is a great opportunity for adopting selected knowledge management systems and technologies in the biomedical context, in particularly, not as a human replacement (i.e., expert systems) but as a biomedical decision making aide.

3.2.1 Knowledge Management in Health Care

It has been generally recognized that patient record management systems is highly desired in clinical settings (Heathfield and Louw, 1999; Jackson, 2000; Abidi, 2001). The major reasons include physicians' significant information needs (Dawes and Sampson, 2003) and clinical information overload. Hersh (1996) classified textual health information into two main categories: patient-specific clinical information and knowledge-based information, which includes research reported in academic journals, books, technical reports, and other sources. Both types of information are growing at an overwhelming pace.

Although early clinical systems were mostly simple data storage systems, knowledge management capabilities have been incorporated in many of them since the 1980s. For example, the *HELP* system, developed at the Latter Day Saints Hospital in Utah, provides a monitoring program on top of a traditional medical record system. Decision logic was stored in the system to allow it to respond to new data entered (Kuperman et al., 1991). The *SAPHIRE* system performs automatic indexing of radiology reports by utilizing the UMLS Metathesaurus (Hersh et al., 2002). The clinical data repository at Columbia-Presbyterian Medical Center (Friedman et al., 1990) is another example of a database that is used for decision support (Hripcsak, 1993) as well as well as physician review. The clinical data repository at the University of Virginia Health System is another example (Schubart and Einbinder, 2000). In their data warehouse system, clinical, administrative, and other patient data are available to users through a Web browser. Case-based reasoning also has been proposed to allow physicians to access both operative knowledge and medical literature based on their medical information needs (Montani and Bellazzi, 2002). Janetzki et al. (2004) use a natural language processing approach to link electronic health records to online information resources. Other advanced text mining techniques also have been applied to knowledge management in health care and will be discussed in more detail later in the chapter.

3.2.2 Knowledge Management for Biomedical Literature

Besides clinical information, knowledge management has been applied to research articles and reports, mostly via selected information retrieval and digital library techniques. The National Library of Medicine (NLM) offers the PubMed service, which includes over 13 million citations for biomedical articles from MEDLINE and other relevant journals. Many search systems have been built to help users retrieve relevant biomedical research papers and reports in database systems and over the Web. Automatic indexing and retrieval techniques are often applied. For example, the *Telemakus* system offers researchers a framework for information retrieval, visualization, and knowledge discovery (Fuller et al., 2002; Fuller et al., 2004; Revere et al., 2004). Using information extraction and visualization techniques, the system allows researchers to search the database of research articles for a statistically significant finding. The *HelpfulMed* system allows users to search for biomedical documents from several databases including MEDLINE, CancerLit, PDQ, and other evidence-based medicine databases (Chen et al., 2003). The HelpfulMed database includes high-quality health care-related Web pages collected from reputable sites using a neural-network-based spreading activation algorithm (Chau and Chen, 2003). The system also provides a term-suggestion tool called *Concept Mapper*, which allows users to consult a system-generated thesaurus and the NLM's UMLS to refine their search queries (Houston et al., 1999; Leroy and Chen, 2001).

MARVIN is an example of medical information retrieval systems that applied selected machine learning techniques (Baujard et al., 1998). Built on a multi-agent architecture, the system filters relevant documents from a set of Web pages and follows links to retrieve new documents. While MARVIN's filtering was based on simple document similarity metrics, other algorithms such as maximum-distance, artificial neural networks, and support vector machines have been applied to filtering medical Web pages (Palakal et al., 2001; Chau and Chen, 2004). A Bayesian model based on term strength analysis also has been used in biomedical document retrieval (Wilbur and Yang, 1996). Shatkay et al. (2000; 2002) use a probabilistic similarity-based search to retrieve biomedical documents that share similar themes.

Other text mining techniques also have been used to facilitate the management and understanding of biomedical literature. For example, natural language processing and noun phrasing techniques have been applied to extract noun phrases from medical documents (Tolle and Chen, 2000). Noun phrases often convey more precise meanings than single terms and are often more useful in further analysis. Named-entity extraction also has been widely applied to automatically identify from text documents the names of entities of interest (Chau et al., 2002). While mostly tested on general entities such as people names, locations, organizations, dates, times, number expressions, and email addresses (Chinchor, 1998), named-entity extraction has been used to extract specific biomedical entities such as gene names, protein names, diseases, and symptoms with promising results (Fukuda et al., 1998; Leroy et al., 2003). The extracted entities and relations are useful for information retrieval and knowledge management purposes. Both entity and relation extraction techniques will be discussed in more detail in our review of text mining later in the article.

3.2.3 Accessing Heterogeneous Databases

In the post-genome era, biomedical data are now being generated at a speed much faster than researchers can handle using traditional methods (National Research Council, 2000). The abundance of genomic and biomedical data has created great potential for research and applications in biomedicine, but the data are often distributed in diverse databases. As biological phenomena are often complex, researchers are faced with the challenge of information integration from heterogeneous data sources (Barrera et al., 2004). Many techniques have been proposed to allow researchers and the general public to share their data more effectively. For example, Sujansky (2001) proposes a framework to integrate heterogeneous databases in biomedicine by providing a uniform conceptual schema and using selected query-translation techniques. The *BLAST* programs are widely used to search protein and DNA databases for sequence similarities (Altschul et al., 1997). The MedBlast system, making use of BLAST, allows researchers to search for articles related to a given sequence (Tu et al., 2004). Sun (2004) uses automated algorithms to identify equivalent concepts available in different databases in order to support information retrieval. A software agent architecture also has been proposed to help users retrieve data from distributed databases (Karasavvas et al., 2004).

3.2.4 Information Visualization and Multimedia Information Access

Information (and knowledge) visualization for biomedical informatics is critical for understanding and sharing knowledge. With the rapid increase in computer speed and reduction in cost, graphical visualization has become increasingly popular in biomedical applications. Visualization techniques support display of more meaningful information and facilitate user understanding. Maps, trees, and networks are among some of the most popular information visualization representations. In the HelpfulMed system discussed earlier, documents retrieved from different databases are clustered using a self-organizing map algorithm (Kohonen, 1995) and a two-dimensional map is generated to display the document clusters (Chen et al., 2003). Bodenreider and McCray (2003) apply radial diagrams and correspondence analysis techniques to visualize semantic groups in the UMLS semantic network. Han and Byun (2004) use a three-dimensional display to visualize protein interaction networks. Virtual reality also has been applied in visualizing metabolic networks (Rojdestvenski, 2003).

Three-dimensional displays, interactive visualization, multimedia displays, and other advanced visualization techniques have been applied successfully in many biomedical applications. The most prominent example is the NLM's Visible Human Project (Ackerman, 1991), which produces three-dimensional representations of the normal male and female human bodies by obtaining transverse CT, MR, and cryosection images of representative male and female cadavers. The data is complete and anatomically detailed as the male was sectioned at one millimeter intervals and the female at one-third of a millimeter intervals. The data provides a good testbed for medical imaging and multimedia processing algorithms and has been applied to various diagnostic, educational, and research uses.

Because text processing algorithms cannot be applied to multimedia data directly, image processing and indexing techniques are often needed for selected biomedical applications. These techniques enable users to visualize, retrieve, and manage multimedia data such as X-ray and CAT-scan images more effectively and efficiently. For example, Yoo and Chen (1994) developed a system to provide a natural navigation of patient data using three-dimensional images and surface rendering techniques. Antani et al. (2004) study different shape representation methods to measure the similarity between X-ray images in order to enable users to manage and organize these images. Their system allows users to retrieve vertebra shapes significant to the pathology indicated in the query. Due to the increasing popularity and maturity of medical imaging systems, we foresee a pressing need for advanced multimedia processing and knowledge management capabilities in biomedicine.

3.3 Data Mining and Text Mining

Data mining techniques have been widely used to find new patterns and knowledge from biomedical data. While Bayesian models were widely used in the early days, more advanced machine learning methods, such as artificial neural networks and support vector machines, have been applied in recent years. These techniques are used in different areas of biomedicine, including genomics, proteomics, and medical diagnosis, among others. In the following, we review some of the major applications of data mining and knowledge discovery techniques in the field.

3.3.1 Data Mining for Health care

Because of their predictive power, data mining techniques have been widely used in diagnostic and health care applications. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion.

Classification is the most widely used technique in medical data mining. Dreiseitl et al. (2001) compare five classification algorithms for the diagnosis of pigmented skin lesions. Their results show that logistic regression, artificial neural networks, and support vector machines performed comparably, while *k*-nearest neighbors and decision trees performed worse. This is more or less consistent with the performances of these classification algorithms in other applications (e.g., Yang and Liu, 1999). Classification techniques are also applied to analyze various *signals* and their relationships with particular diseases or symptoms. For example, Acir and Guzelis (2004) apply support vector machines in automatic spike signal detection in ElectroEncephaloGrams (EEG), which can be used in diagnosing neurological disorders related to epilepsy. Kandaswamy et al. (2004) use artificial neural network to classify lung sound signals into six different categories (e.g., *normal*, *wheeze*, and *rhonchus*) to assist diagnosis.

Data mining is also used to extract rules from health care data. For example, it has been used to extract diagnostic rules from breast cancer data (Kovalerchuk et al., 2001). The rules generated are similar to those created manually in expert systems and therefore can be easily validated by domain experts. Data mining has also been applied to clinical databases to identify new medical knowledge (Prather et al., 1997; Hripcsak et al., 2002).

3.3.2 Data Mining for Molecular Biology

New sequencing technologies and low computation cost have resulted in an overwhelming abundance of biological data that can be accessed easily by researchers. It is not feasible to analyze these data manually, and the gap between the amount of submitted sequence data and related annotations, structures, or expression profiles is rapidly growing.

Data mining has begun to play an important role in addressing this problem. Clustering is probably the most widely used data mining technique for biological data. For example, clustering analysis is often applied to microarray gene expression data to identify groups of genes sharing similar expression profiles. Eisen et al. (1998) applied hierarchical clustering on the *Saccharomyces cerevisiae* gene expression data and achieved promising results. Various other clustering algorithms also have been tested on gene expression data, including *k*-means clustering (Herwig et al., 1999), backpropagation neural networks (Sawa and Ohno-Machado, 2003), self-organizing maps (Tamayo et al., 1999; Herrero et al., 2001), fuzzy clustering (Belacel et al., 2004), expectation maximization (Qu and Xu, 2004), and support vector machines (Brown et al., 2000). Qin et al. (2003) used the idea of kernel (as in support vector machines) and combined it with hierarchical clustering. Gene expression analysis also has been applied in cancer class discovery and prediction (Golub et al., 1999; Hsu et al., 2003).

Besides clustering, other predictive data mining techniques also have been applied to biomedical data. For example, neural network models have been widely used in predicting protein secondary structure (Qian and Sejnowski, 1988; Hirst and Sternberg, 1992). Increasingly, data mining algorithms also have been used for prediction in various biomedical applications including protein backbone angle prediction (Kuang et al., 2004), protein domains (Nagarajan and Yona, 2004), biological effects (Krishnan and Westhead, 2004), and DNA binding (Ahmad et al., 2004). These predictive methods are often based on classification (supervised learning) algorithms such as neural networks or support vector machines.

3.3.3 Text Mining for Literature and Clinical Records

Text mining has been widely used to analyze biomedical literature. Because of the large amount of research articles in public databases and the diversity of biomedical research, it is not uncommon that researchers encounter some sequences or new genes that they have no knowledge about. It is quite likely that some important relationships between biological entities remain unnoticed because relevant data are scattered and no researcher has linked them together (Swanson, 1986; Smalheiser and Swanson, 1998). Given the large amount of published literature and that many researchers only specialize in a small sub-domain (e.g., several particular genes), text mining techniques could be invaluable in discovering new knowledge patterns or hypotheses from the large amount of existing and new literature in biomedicine (Yandell and Majoros, 2002).

Text mining for biomedical literature often involves two major steps. First, it must identify biomedical entities and concepts of interests from free text using natural language processing techniques. For instance, if we want to study the relationship between a gene (e.g., *p53*) and a disease (e.g., *brain tumors*), the names of both entities need to be correctly identified from the relevant textual documents. Many text mining algorithms have been applied to this problem. For example, Fukuda et al. (1998) use simple morphological clues to recognize the names of proteins and other materials with high accuracy. Support vector machines have been used in entity extraction by classifying words into the 24 entity classes in the *GENIA* corpus (Kazama et al., 2002). Tanabe and Wilbur (2002) use part-of-speech tagging and a Bayesian model to identify genes and proteins in text. Hatzivassiloglou et al. (2001) compared three machine learning techniques, namely Naïve Bayesian model, decision trees, and inductive rule learning, to resolve the classification of a biological entity (e.g., protein, gene, and RNA) after it was identified. Their results showed that the three learning models had comparable performance. Other studies have investigated the mapping between abbreviations and full names such that these names will not be considered by the system as different entities (Yu et al., 2002).

After the entity names have been identified, further analyses are performed to see whether these entities have any relationships, such as gene regulations, metabolic pathways, or protein-protein interactions (Blaschke et al., 1999; Dickerson et al., 2003). *Shallow parsing* is often used to focus on specific parts of the text to analyze predefined words such as verbs and nouns (Leroy et al., 2003). Sekimizu et al. (1998) identified the set of most frequently used verbs in a collection of abstracts and developed a set of rules to identify the subjects and objects of the verbs. Pustejovsky et al. (2002) used relational parsing and finite state automata to identify *inhibit* relationships from biomedical text. The *GENIES* system, based on the *MedLEE* parser (Friedman and Hripcsak, 1998), also has been used to extract molecular pathways from texts (Friedman et al., 2001). The *Telemakus* system extracts information by analyzing the headings and surrounding texts of tables and figures (Fuller et al., 2002; Revere et al., 2004). The *Genescene* system utilizes an ontology-based approach to relation extraction by integrating the *Gene Ontology*, the *Human Genome Nomenclature*, and the *UMLS* (Leroy and Chen, forthcoming). The system combines natural language processing and co-occurrence analysis techniques to identify terms and gene pathway relations from biomedical abstracts. The *EDGAR* system extracts drugs, genes, and relationships from text (Rindflesch et al., 2000). Wren et al. (2004) developed a system that uses a random network model to rank the relationships identified from text. Machine learning techniques also have been used to automate the process of annotation. For instance, Kretschmann et al. (2001) used a *C4.5* algorithm to generate rules for keyword annotation in the *SWISS-PROT* database.

Text mining also has been applied to patient records and other clinical documents to facilitate knowledge management. It adopts a process similar to that of text mining from literature. For example, the system reported by Harris et al. (2003) extracts terms from clinical texts. Using natural language processing techniques, the *MedLEE* system (Friedman and Hripcsak, 1998) has been applied to free-text patient records. It extracts useful entities in order to identify patients having tuberculosis or breast cancer based on their admission chest radiographs and mammogram reports, respectively (Knirsch et al., 1999; Jain and Friedman, 1997). Chapman et al. (2004) use a similar text mining approach for automated fever detection from clinical records to detect possible infectious disease outbreaks.

3.4 Ethical and Legal Issues for Data Mining

Medical records and biological data generated from human subjects contain private and confidential information. Patients' and human subjects' data must be handled with great caution in order to protect their privacy and confidentiality. Researchers do not automatically acquire the rights to use patient or subject data for data mining purposes unless they obtain the patients' or subjects' consent (Berman, 2002). In the US, the 1996 Health Insurance Portability and Accountability Act (HIPAA) set the standards for using and handling patient data in electronic format. The "Common Rule" also specifies how to protect human subjects in federally-funded research. In Europe, the EU Data Protection Directive specifies rules on handling and processing any information about individuals. Violations of these standards could result in legal responsibilities and penalties including fine and imprisonment. Data mining results that are relevant to patients and subjects need to be interpreted in the proper medical context and with the help of the biomedical professionals.

In biomedical data mining, under most conditions patient data should not be *individually identifiable,* i.e., no record should provide sufficient data to identify the individual related to the record. These include anonymous data (data collected without patient-identification information), anonymized data (data collected with patient-identification information which is removed later), or de-identified data (data with patient-identification information encoded or encrypted) (Cios and Moore, 2002).

4. SUMMARY

In this chapter we provide a broad overview of selected knowledge management, data mining, and text mining techniques and their use in various emerging biomedical applications. However powerful they may be, these techniques need to be used with great care in the biomedical applications. One concern, as discussed earlier, is that medical data are often sensitive and involve private and confidential information. It is important that patients' confidentiality and privacy are not compromised due to the introduction of advanced knowledge management, data mining, and text mining technologies. Another caveat is that findings generated from selected machine learning techniques need to be interpreted carefully. Knowledge and patterns discovered by computers need to be experimentally or clinically validated in order to be considered rigorous, just like any knowledge generated by human. Errors and incorrect associations could propagate quickly through electronic media, especially when large databases and powerful computational techniques are involved.

Nonetheless, these new knowledge management, data mining, and text mining techniques are changing the way new knowledge is discovered, organized, applied, and disseminated. With the increasing speed of computers, the connectivity of the Internet, the abundance of biomedical data, and the advances in medical informatics research, we believe we will continue to generate, manage, and harvest biomedical knowledge effectively and efficiently, allowing us to better understand the complex biological processes of life and assist in addressing the well-being of human kind.

REFERENCES

Abidi, S. S. R. (2001). “Knowledge Management in Healthcare: Towards ‘Knowledge-driven’ Decision-support Services,” *International Journal of Medical Informatics*, 63, 5-18.

Acir, N. and Guzelis, C. (2004). “Automatic Spike Detection in EEG by a Two-stage Procedure Based on Support Vector Machines,” *Computers in Biology and Medicine*, 34(7), 561-575.

Ackerman, M. J. (1991). “The Visible Human Project,” *Journal of Biocommunication*, 18(2), 14.

Ahmad, S., Gromiha, M. M., and Sarai, A. (2004). “Analysis and Prediction of DNA-binding Proteins and Their Binding Residues Based on Composition, Sequence, and Structural Information,” *Bioinformatics*, 20(4), 477-486.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs,” *Nucleic Acids Research*, 25(17), 3389-3402.

Antani, S., Lee, D. J., Long, L. R., and Thoma, G. R. (2004). “Evaluation of Shape Similarity Measurement Methods for Spine X-ray Images,” *Journal of Visual Communication and Image Representation*, 15, 285-302.

Baclawski, K., Cigna, J., Kokar, M. W., Mager, P., and Indurkhya, B. (2000). “Knowledge Representation and Indexing Using the Unified Medical Language System,” in *Proceedings of the Pacific Symposium on Biocomputing*, 493-504.

Barrera, J., Cesar-Jr, R. M., Ferreira, J. E., and Gubitoso, M. D. (2004). “An Environment for Knowledge Discovery in Biology,” *Computers in Biology and Medicine*, 34, 427-447.

Baujard, O., Baujard, V., Aurel, S., Boyer, C., and Appel, R. D. (1998). “Trends in Medical Information Retrieval on the Internet,” *Computers in Biology and Medicine*, 28, 589-601.

Belacel, B., Cuperlovic-Culf, M., Laflamme, M., and Ouellette, R. (2004). “Fuzzy J-Means and VNS Methods for Clustering Genes from Microarray Data,” *Bioinformatics*, 20(11), 1690-1701.

Belew, R. K. (1989). "Adaptive Information Retrieval: Using a Connectionist representation to Retrieve and Learn about Documents," in *Proceedings of the 12th ACM-SIGIR Conference*, Cambridge, MA, June 1989.

Berman, J. J. (2002). "Confidentiality Issues for Medical Data Miners," *Artificial Intelligence in Medicine*, 26(1-2), 25-36.

Blaschke, C., Andrade, M. A., Ouzounis, C. and Valencia, A. (1999). "Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, 60-67.

Bodenreider, O. and McCray, A. T. (2003). "Exploring Semantic Groups through Visual Approaches," *Journal of Biomedical Informatics*, 36, 414-432.

Breiman, L. and Spector, P. (1992). "Submodel Selection and Evaluation in Regression: The X-random Case," *International Statistical Review*, 60(3), 291-319.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). "Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," in *Proceedings of the National Academy of Sciences*, 97, 262-267.

Campbell, K. E., Oliver, D. E., and Shortliffe, E. H. (1998). "The Unified Medical Language System: Toward a Collaborative Approach for Solving Terminologic Problems," *Journal of the American Medical Informatics Association*, 5(1), 12-16.

Carbonell, J. G. Michalski, R. S., Mitchell, T. M. (1983). "An Overview of Machine Learning," in R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.), *Machine Learning, An Artificial Intelligence Approach*, Palo Alto, CA: Tioga.

Cavideds, J. E. and Cimino, J. J. (2004). "Towards the Development of a Conceptual Distance Metric for the UMLS," *Journal of Biomedical Informatics*, 37, 77-85.

Chapman, W. W., Dowling, J. N., and Wagner, M. M. (2004). "Fever Detection from Free-text Clinical Records for Biosurveillance," *Journal of Biomedical Informatics*, 37, 120-127.

Chau, M. and Chen, H. (2003). "Comparison of Three Vertical Search Spiders," *IEEE Computer*, 36(5), 56-62.

Chau, M. and Chen, H. (2004). "Using Content-based and Link-based Analysis in Building Vertical Search Engines," in *Proceedings of the International Conference on Asian Digital Libraries*, Shanghai, China, December 13-17, 2004.

Chau, M., Xu, J. J., and Chen, H. (2002). "Extracting Meaningful Entities from Police Narrative Reports," in *Proceedings of the National Conference for Digital Government Research*, Los Angeles, California, USA, May 19-22, 2002, 271-275.

Chen, H. (2001). *Knowledge Management Systems: A Text Mining Perspective*, Tucson, AZ: The University of Arizona.

Chen, H. and Chau, M. (2004). "Web Mining: Machine Learning for Web Applications," *Annual Review of Information Science and Technology*, 38, 289-329.

Chen, H. and Kim, J. (1995). "GANNET: A Machine Learning Approach to Document Retrieval," *Journal of Management Information Systems*, 11(3), 9-43.

Chen, H., Lally, A. M., Zhu, B., and Chau, M. (2003). "HelpfulMed: Intelligent Searching for Medical Information over the Internet," *Journal of the American Society for Information Science and Technology*, 54(7), 683-694, 2003.

Chen, H. and Ng, T. (1995). "An Algorithmic Approach to Concept Exploration in a Large Knowledge Network (Automatic Thesaurus Consultation): Symbolic Branch and Bound Search vs. Connectionist Hopfield Net Activation," *Journal of the American Society for Information Science*, 46(5), pp. 348-369.

Chinchor, N. A. (1998). “Overview of MUC-7/MET-2,” in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Virginia, USA, April 29 - May 1, 1998.

Cimino, J. J., Min, H., and Perl, Y. (2003) “Consistency across the Hierarchies of the UMLS Semantic Network and Metathesaurus,” *Journal of Biomedical Informatics*, 36, 450-461.

Cios, K. J. and Moore, G. W. (2002). “Uniqueness of Medical Data Mining,” *Artificial Intelligence in Medicine*, 26(1-2), 25-36.

Cohen, P. R. and Feigenbaum, E. A. (1982). *The Handbook of Artificial Intelligence: Volume III*, Reading, MA: Addison-Wesley.

Dawes, M. and Sampson, U. (2003). “Knowledge Management in Clinical Practice: A Systematic Review of Information Seeking Behavior in Physicians,” *International Journal of Medical Informatics*, 71, 9-15.

Dickerson, J. A., Berleant, D., Cox, Z., Fulmer, A. W., and Wurtele, E. (2003). “Creating and Modeling Metabolic and Regulatory Networks Using Text Mining and Fuzzy Expert Systems,” in J. T. L. Wang, C. H. Wu, and P. P. Wang (Eds.), *Computational Biology and Genome Informatics*, World Scientific.

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., Binder, M. (2001). “A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions,” *Journal of Biomedical Informatics*, 34, 28-36.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*, New York: John Wiley and Sons.

Dunham, M. H. (2002). *Data Mining: Introductory and Advanced Topics*, New Jersey, USA: Prentice Hall.

Efron, B. (1983). “Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation,” *Journal of the American Statistical Association*, 78(382), 316-330.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). “Cluster Analysis and Display of Genome-wide Expression Patterns,” in *Proceedings of the National Academy of Sciences*, 95, 14863-14868.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, 17(3), 37-54.

Fisher, D. H. (1987). “Knowledge Acquisition via Incremental Conceptual Clustering,” *Machine Learning*, 2, 139-172.

Fogel, D. B. (1994). “An Introduction to Simulated Evolutionary Optimization,” *IEEE Transactions on Neural Networks*, 5, 3-14.

Friedman, C. Hripcsak, G., Johnson, S. B., Cimino, J. J., Clayton, P. D. (1990). “A Generalized Relational Schema for an Integrated Clinical Patient Database,” in *Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care*, 335-339.

Friedman, C. and Hripcsak, G. (1998). “Evaluating Natural Language Processors in the Clinical Domain,” *Methods of Information in Medicine*, 37, 334-344.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). “GENIES: A Natural-language Processing System for the Extraction of Molecular Pathways from Journal Articles,” *Bioinformatics*, 17(Supp. 1), S74-S82.

Fukuda K., Tamura A., Tsunoda T., and Takagi T. (1998). “Toward Information Extraction: Identifying Protein Names from Biological Papers,” in *Proceedings of the Pacific Symposium on Biocomputing*, 707-718.

Fuller, S., Revere, D., Soderland, S., Bugni, P., Kadiyska, Y., Reber, L., Fuller, H., and Martin, G. (2002). “Modeling a Concept-Based Information System to Promote Scientific Discovery: The Telemakus System,” in *Proceedings of the AMIA 2002 Annual Symposium*, 1023.

Fuller, S., Revere, D., Bugni, P., Fuller, H., and Martin, G. (2004). “A Knowledgebase System to Enhance Scientific Discovery: Telemakus,” *Biomedical Digital Libraries,* 1(2-15).

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading, MA: Addison-Wesley.

Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S. (1999). “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” *Science*, 286(5439), 531-537.

Gruninger, M. and Lee, J. (2002). “Ontology: Applications and Design,” *Communications of the ACM*, 45(2), 39-41

Han, K., and Byun, Y. (2004). “Three-dimensional Visualization of Protein Interaction Networks,” *Computers in Biology and Medicine*, 34, 127-139.

Harris, M. R., Savova, G. K., Johnson, T. M., and Chute, C. G. (2003). “A Term Extraction Tool for Expanding Content in the Domain of Functioning, Disability, and Health: Proof of Concept,” *Journal of Biomedical Informatics*, 36, 250-259.

Hatzivassiloglou, V., Duboue, P. A., and Rzhetsky, A. (2001). “Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach,” *Bioinformatics*, 17(Supp. 1), S96-S106.

Hayes-Roth, F. and Jacobstein, N. (1994). “The State of Knowledge-based Systems,” *Communications of the ACM*, 37, 27-39.

Hearst, M. A. (1999). “Untangling Text Data Mining,” in *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, June 20-26.

Heathfield, H. and Louw, G. (1999). “New Challenges for Clinical Informatics: Knowledge Management Tools,” *Health Informatics Journal*, 5(2), 67-73.

Herrero, J., Valencia, A., and Dopazo, J. (2001). “A Hierarchical Unsupervised Growing Neural Network for Clustering Gene Expression Patterns,” *Bioinformatics*, 17, 126-136.

Hersh, W. (1996). *Information Retrieval: A Health Care Perspective*. Berlin, Germany: Springer-Verlag.

Hersh, W., Mailhot, M., Arnott-Smith, C., and Lowe, H. (2002). “Selective Automated Indexing of Findings and Diagnoses in Radiology Reports,” *Journal of Biomedical Informatics*, 34, 262-273.

Herwig, R., Poustka, A., Müller, C., Bull, C., Lehrach, H., and O'Brien, J. (1999). “Large-scale Clustering of cDNA Fingerprinting Data,” *Genome Research*, 9, 1093-1105.

Hirst, J. D. and Sternberg, M. J. E. (1992). “Prediction of Structural and Functional Features of Protein and Nucleic Acid Sequences by Artificial Neural Networks,” *Biochemistry*, 31, 7211-7218.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: University of Michigan Press.

Hopfield, J. J. (1982). “Neural Network and Physical Systems with Collective Computational Abilities,” in *Proceedings of the National Academy of Science,* USA, 1982, 79(4), pp. 2554-2558.

Houston, A. L., Chen, H., Hubbard, S. M., Schatz, B. R., Ng, T. D., Sewell, R. R. and Tolle, K. M. (1999). “Medical Data Mining on the Internet: Research on a Cancer Information System,” *Artificial Intelligence Review*, 13, 437-466.

Hsu, A. L., Tang, S., and Halgamuge, S. K. (2003). “An Unsupervised Hierarchical Dynamic Self-organizing Approach to Cancer Class Discovery and Market Gene Identification in Microarray Data,” *Bioinformatics*, 19(16), 2131-2140.

Hripcsak, G. (1993). “Monitoring the Monitor: Automated Statistical Tracking of a Clinical Event Monitor,” *Computers and Biomedical Research, 26(5), 449-466.*

Hripcsak, G., Austin, J. H., Alderson, P. O., and Friedman, C. (2002). “Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports,” *Radiology*, 224(1), 157-163.

Humphreys, B. L., Lindberg, D. A. B., and McCray, A. (1993). “The Unified Medical Language System,” *Methods of Information in Medicine*, 32(4), 281.

Humphreys, B. L., Lindberg, D. A. B., Schoolman, H. M., and Barnett, G. O. (1998). “The Unified Medical Language System: An Informatics Research Collaboration,” *Journal of the American Medical Informatics Association*, 5(1), 1-11.

Jackson, J. R. (2000). “The Urgent Call for Knowledge Management in Medicine,” *The Physician Executive*, 26(1), 28-31.

Jain, A. K., Dubes, R. C. and Chen, C. (1987). “Bootstrap Techniques for Error Estimation,” *IEEE Transactions on Pattern Analysis and Machine Learning*, 9(5), 628-633.

Knirsch, C.A., Jain, N. L., Pablos-Mendez, A., Friedman, C., and Hripcsak, G. (1996). “Respiratory Isolation of Tuberculosis Patients Using Clinical Guidelines and an Automated Clinical Decision Support System,” *Infection Control and Hospital Epidemiology*, 19(2), 94-100.

Jain, N. L. and Friedman, C. (1997). “Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports.” in *Proceedings of the Fall 1997 AMIA Conference*, Philadelphia, USA, 829-833.

Janetzki, V., Allen, M., and Cimino, J. J. (2004). “Using Natural Language Processing to Link from Medical Text to On-line Information Resources,” *Proceedings of Medinfo*, 2004, 1665.

Joachims, T. (1998). “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” in *Proceedings of the European Conference on Machine Learning*, Berlin, 1998, pp. 137-142.

Kandaswamy, A., Kumar, C. S., Ramanathan, R. P. Jayaraman, R., and Malmurugan, N. (2004). “Neural Classification of Lung Sounds Using Wavelet Coefficients,” *Computers in Biology and Medicine*, 34, 523-537.

Karasavvas, K. A., Baldock, R., and Burger, A. (2004). “Bioinformatics Integration and Agent Technology,” *Journal of Biomedical Informatics*, 37, 205-219.

Kazama, J., Maino, T., Ohta, Y., and Tsujii, J. (2002). “Tuning Support Vector Machines for Biomedical Named Entity Recognition,” in *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, USA, July 2002, 1-8.

Kohavi, R. (1995). “A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, San Francisco, CA, 1995, Morgan Kaufmann, pp. 1137-1143.

Kohonen, T. (1995). *Self-organizing Maps*, Springer-Verlag, Berlin.

Kononenko, I. (1993). “Inductive and Bayesian Learning in Medical Diagnosis,” *Applied Artificial Intelligence*, 7, 317-337, 1993.

Kovalerchuk, B., Vityaev, E., and Ruiz, J. F. (2001). “Consistent and Complete Data and ‘Expert’ Mining in Medicine,” in Cios, K. J. (Ed.), *Medical Data Mining and Knowledge Discovery*, New York, USA: Physica-Verlag.

Kretschmann, E., Fleischmann, W., and Apweiler, R. (2001). “Automatic Rule Generation for Protein Annotation with the C4.5 Data Mining Algorithm Applied on SWISS-PROT,” *Bioinformatics*, 17(10), 920-926.

Krishnan, V. G. and Westhead, D. R. (2003). “A Comparative Study of Machine-Learning Methods to Predict the Effects of Single Nucleotide Polymorphisms on Protein Function,” *Bioinformatics*, 19(17), 2199-2209.

Kuperman, G. J., Gardner, R.M., Pryor, T.A. (1991). *The HELP System*, New York: Springer-Verlag.

Kwok, K. L. (1989). “A Neural Network for Probabilistic Information Retrieval,” in *Proceedings of the 12th ACM-SIGIR Conference on Research and Development in Information Retrieval*, Cambridge, Massachusetts, June 1989, pp.21-30.

Langley, P. and Simon, H. (1995). “Applications of Machine Learning and Rule Induction,” *Communications of the ACM*, 38(11), 55-64.

Leroy, G. and Chen, H. (2001). “Meeting Medical Terminology Needs – The Ontology-Enhanced Medical Concept Mapper,” *IEEE Transactions on Information Technology in Biomedicine*, 5(4), 261-270.

Leroy, G. and Chen, H. (forthcoming). “Genescene: An Ontology-enhanced Integration of Linguistic and Co-occurrence-based Relations in Biomedical Texts” *Journal of the American Society for Information Science and Technology*, forthcoming.

Leroy, G., Chen, H., and Martinez, J. D. (2003). “A Shallow Parser Based on Closed-class Words to Capture Relations in Biomedical Text,” *Journal of Biomedical Informatics*, 36, 145-158.

Lippmann, R. P. (1987). An Introduction to Computing with Neural Networks, *IEEE Acoustics Speech and Signal Processing Magazine*, 4, 4-22.

Maniezzo V. (1994). “Genetic Evolution of the Topology and Weight Distribution of Neural Networks,” *IEEE Transactions on Neural Networks*, 5(1), 39-53.

Mendes, R. R. F., Voznika, F. B., Freitas, A. A. and Nievola, J. C. (2001). “Discovering Fuzzy Classification Rules with Genetic Programming and Co-evolution,” *Principles of Data Mining and Knowledge Discovery*, *Lecture Notes in Artificial Intelligence*, 2168, pp. 314-325. Springer-Verlag, 2001.

Michalewicz, Z. (1992). Genetic Algorithms + Data Structures =Evolution Programs. Berlin: Springer-Verlag.

Mitchell, T. (1997). *Machine Learning*, McGraw Hill, 1997.

Montani, S. and Bellazzi, R. (2002). “Supporting Decisions in Medical Applications: The Knowledge Management Perspective,” *International Journal of Medical Informatics*, 68, 79-90.

Nagarajan, N. and Yona, G. (2004). “Automatic Prediction of Protein Domains from Sequence Information Using a Hybrid Learning System,” *Bioinformatics*, 20(9), 1335-1360.

National Research Council (2000). *Bioinformatics: Converting Data to Knowledge: Workshop Summary*, Washington, D.C.: National Academies Press.

Paass, G. (1990), “Probabilistic Reasoning and Probabilistic Neural Networks,” in *Proceedings of the 3rd International Conference on Information Processing and Management of Uncertainty*, pp.6-8.

Palakal, M., Mukhopadhyay, S., Mostafa, J., Raje, R., N’Cho, M., and Mishra, S. (2001). “An Intelligent Biological Information Management System,” *Bioinformatics*, 18(10), 1283-1288.

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., and Hammond, W. E. (1997). “Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse,” in *Proceedings of the AMIA Annual Symposium Fall 1997*, 101-105.

Perl, Y. and Geller, J. (2003). “Research on Structural Issues of the UMLS – Past, Present, and Future,” *Journal of Biomedical Informatics*, 36, 409-413.

Pustejovsky J., Castano J., Zhang J., Kotecki M., and Cochran B. (2002). “Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations,” *Pacific Symposium on Biocomputing*, 362-373.

Qian, N. and Sejnowski, T. J. (1988). “Predicting the Secondary Structure of Globular Proteins Using Neural Network Models,” *Journal of Molecular Biology*, 202, 865-884.

Qin, J., Lewis, D. P., and Noble, W. S. (2003). "Kernel Hierarchical Gene Clustering from Microarray Expression Data," *Bioinformatics*, 19(16), 2097-2104.

Qu Y. and Xu., S. (2004). "Supervised Cluster Analysis for Microarray Data Based on Multivariate Gaussian Mixture," *Bioinformatics*, 20(12), 1905-1913.

Quinlan, J. R. (1983). "Learning Efficient Classification Procedures and Their Application to Chess End Games," in R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Palo Alto, CA: Tioga.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Los Altos, CA: Morgan Kaufmann.

Revere, D., Fuller, S. S, Bugni, P. F., and Martin, G. M. (2004). "An Information Extraction and Representation System for Rapid Review of the Biomedical Literature," in *Proceedings of Medinfo*, 2004.

Rindflesch, T. C., Tanabe, L., and Weinstein, J. N., and Hunter, L. (2000). "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature," in *Proceedings of the Pacific Symposium on Biocomputing* , 514-525.

Rindflesch, T. C. and Fiszman, M. (2003) "The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text," *Journal of Biomedical Informatics*, 36, 462-477.

Rojdestvenski, I. (2003). "VRML Metabolic Network Visualizer," *Computers in Biology and Medicine*, 33, 169-182.

Rosse, C. and Mejino, J. L. V. (2003). "A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy," *Journal of Biomedical Informatics*, 36, 478-500.

Rumelhart, D. E., Hinton, G. E., and McClelland, J. L. (1986a). "A General Framework for Parallel Distributed Processing," in D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing*, pp. 45-76, Cambridge, MA: The MIT Press.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). "Learning Internal Representations by Error Propagation," in D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing*, pp. 318-362, Cambridge, MA: The MIT Press.

Samuelson, C. and Rayner, M. (1991). "Quantitative Evaluation of Explanation-based Learning as an Optimization Tool for a Large-scale Natural Language System," in *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991, pp. 609-615.

Sawa, T. and Ohno-Machado, L. (2003). "A Neural Network-based Similarity Index for Clustering DNA Microarray Data," *Computers in Biology and Medicine*, 33, 1-15.

Schubart, J. R. and Einbinder, J. S. (2000). "Evaluation of a Data Warehouse in an Academic Health Sciences Center," *International Journal of Medical Informatics*, 60, 319-333.

Sekimisu, T., Park, H. S., and Tsujii, J. (1998). "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in MEDLINE Abstracts," *Genome Informatics*, 9, 62-71.

Shatkay, H., Edwards, S., and Boguski, M. (2002). "Information Retrieval Meets Gene Analysis," *IEEE Intelligent Systems*, 17(2), 45-53.

Shatkay, H., Edwards, S., Wilbur, W. J., and Boguski, M. (2000). "Genes, Themes, and Microarrays: Using Information Retrieval for Large-scale Gene Analysis," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, 317-328.

Shortliffe, E. (1976). *Computer-based Medical Consultations: MYCIN*, New York: Elsevier/North Holland.

Shortliffe, E. (1987). "Computer Programs to Support Clinical Decision Making," *Journal of the American Medical Association*, 258, 61-66.

Simon, H. A. (1983). “Why Should Machines Learn?” In R. S. Michalski, J. Carbonell, and T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Palo Alto, CA: Tioga Press.

Smalheiser, N. R. and Swanson, D. R. (1998). “Using ARROWSMITH: A Computer-assisted Approach to Formulating and Assessing Scientific Hypotheses,” *Computer Methods and Programs in Biomedicine*, 57, 149-153.

Stone, M. (1974). “Cross-validation Choices and Assessment of Statistical Predictions,” *Journal of the Royal Statistical Society*, 36, 111-147.

Sujansky, W. (2001). “Heterogeneous Database Integration in Biomedicine,” *Journal of Biomedical Informatics*, 34, 285-298.

Sun, Y. (2004). “Methods for Automated Concept Mapping between Medical Databases,” *Journal of Biomedical Informatics*, 37, 162-178.

Swanson, D. R. (1986). “Fish Oil, Raynaud’s Syndrome, and Undiscovered Public Knowledge,” *Perspectives in Biology and Medicine*, 30(1), 7-18.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub T. R. (1999). “Interpreting Patterns of Gene Expression with Self-organizing Maps: Methods and Application to Hematopoietic Differentiation,” in *Proceedings of the National Academy of Sciences*, 96, 2907-2912.

Tanabe, L. and Wilbur, W. J. (2002). “Tagging Gene and Protein Names in Biomedical Text,” *Bioinformatics*, 18(8), 1124-1132.

The Gene Ontology Consortium (2000). “Gene Ontology: Tool for the Unification of Biology,” *Nature Genetics*, 25(1), 25-29.

Tolle, K. and Chen, H. (2000) “Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools,” *Journal of the American Society for Information Science*, 51(4), 352-370.

Tu, Q., Tang, H., and Ding, D. (2004). “MedBlast: Searching Articles Related to a Biological Sequence,” *Bioinformatics*, 20(1), 75-77.

Vapnik, V. (1998). *Statistical Learning Theory*, Wiley, Chichester, GB, 1998.

Wain, H. M., Lush, M., Ducluzeau, F., Povey, S. (2002). “Genew: The Human Gene Nomenclature Database,” *Nucleic Acids Research*, 30(1), 169-171.

Wilbur, W. J. and Yang, Y. (1996). “An Analysis of Statistical Term Strength and Its Use in the Indexing and Retrieval of Molecular Biology Texts,” *Computers in Biology and Medicine*, 26(3), 209-222.

Yandell, M. D. and Majoros, W. H. (2002). “Genomics and Natural Language Processing,” *Nature Reviews Genetics*, 3(8), 601-610.

Yang, Y. and Liu, X. (1999). “A Re-examination of Text Categorization Methods, in *Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR’99), 1999, pp. 42-49.

Yoo, T. S., and Chen, D. T. (1994). “Interactive 3D Medical Visualization: A Parallel Approach to Surface Rendering 3D Medical Data,” in *Proceedings of the Symposium for Computer Assisted Radiology*, North Carolina, USA, June 12-15, 1994, 100-105.

Yu, H., Hatzivassiloglou, V., Rzhetsky, A., and Wilbur, W. J. (2002). “Automatically Identifying Gene/Protein Terms in MEDLINE Abstracts,” *Journal of Biomedical Informatics*, 35, 322-330.

Yu, V. L., Fagan, L. M., Wraith, S. M., Clancey, W. J., Scott, A. C., Hannigan, J. Blum, R. L., Buchanan, B. G., and Cohen, S. N. (1979). “Antimicrobial Selection by a Computer: A Blinded Evaluation by Infectious Disease Experts,” *Journal of the American Medical Association*, 242(12), 1279-1282.

Zadeh, L. A. (1965). “Fuzzy sets,” *Information and Control*, 8, 338-353.

Zhang, L., Perl, Y., Halper, M., and Geller, J. (2003). "Designing Metaschemas for the UMLS Enriched Semantic Network," *Journal of Biomedical Informatics*, 36, 433-449.

SUGGESTED READINGS

Shortliffe, E. H. and Perreault, L. E. (2002). Medical Informatics: Computer Applications in Health Care and Biomedicine, Springer.
This excellent introductory book provides a comprehensive overview of the applications of computer and information technologies in health care and biomedicine.

Baldi, P. and Brunak, S. (2000). *Bioinformatics: The Machine Learning Approach*, The MIT Press.
The book describes bioinformatics from a technical perspective and explains in detail the application of data mining algorithms for biomedical sequence and structure analysis.

Mitchell, T. (1997). *Machine Learning*, McGraw Hill, 1997.
This introductory book includes useful reviews of various machine learning techniques and their applications.

Chen, H., Lally, A. M., Zhu, B., and Chau, M. (2003). "HelpfulMed: Intelligent Searching for Medical Information over the Internet," *Journal of the American Society for Information Science and Technology*, 54(7), 683-694, 2003.
This article provides an overview of medical information retrieval techniques on the Internet, including Web crawling, co-occurrence analysis, and document visualization.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). "Cluster Analysis and Display of Genome-wide Expression Patterns," in *Proceedings of the National Academy of Sciences*, 95, 14863-14868.
This article presents a study on performing clustering techniques on gene expression data.

Swanson, D. R. (1986). "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge," *Perspectives in Biology and Medicine*, 30(1), 7-18.
This article describes the interesting story of how public knowledge could remain "undiscovered" as there were no researchers linking the literature in two separate fields, and how the computer was used to discover such knowledge.

Yandell, M. D. and Majoros, W. H. (2002). "Genomics and Natural Language Processing," *Nature Reviews Genetics*, 3(8), 601-610.
This article reviews research studies that apply natural language processing and text mining techniques in genomics.

ONLINE RESOURCES

National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/
NCBI, a division of the National Library of Medicine, provides access to many excellent molecular biology resources, including GenBank (an annotated collection of all publicly available DNA sequences), Entrez (a cross-database search engine), and BLAST (a sequence similarity search engine).

Unified Medical Language Systems (UMLS)
http://www.nlm.nih.gov/research/umls/

Developed by the Lister Hill Center of the NLM, UMLS provides a large-scale and widely-used medical ontology for information retrieval and text mining applications in biomedicine. The three major components include the Metathesaurus, the Semantic Network, and the Specialist Lexicon.

ExPASy Proteomics Server
http://us.expasy.org/
The ExPASy (Expert Protein Analysis System) proteomics server is hosted by the Swiss Institute of Bioinformatics (SIB). It focuses on the analysis of protein sequences and structures. It provides access to Swiss-PROP, TrEMBL, and other proteomics and sequence analysis tools and resources.

Protein Data Bank
http://www.rcsb.org/pdb/
The Protein Data Bank is the single worldwide repository for 3-D biological macromolecular structure data.

European Bioinformatics Institute (EBI)
http://www.ebi.ac.uk/
EBI is the European equivalent of NCBI and is part of the European Molecular Biology Laboratory (EMBL). It manages several biological databases including: nucleic acid, protein sequences, and macromolecular structures.

GenomeNet
http://www.genome.jp/
Developed in Japan, GenomeNet includes several databases for genome research and molecular and cellular biology. Its services include the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the DBGET Integrated Database Retrieval System, among others.

GenomeWeb
http://www.rfcgr.mrc.ac.uk/GenomeWeb/
This site provides a comprehensive directory of genome-related Web sites and information.

Saccharomyces Genome Database (SGM)
http://www.yeastgenome.org
This database contains information about the molecular biology and genetics of the yeast Saccharomyces cerevisiae. Commonly known as the baker's or budding yeast, its genome has been widely studied in bioinformatics.

The Visible Human Project
http://www.nlm.nih.gov/research/visible/
This site includes a detailed description of NLM’s Visible Human Project, instructions on how to obtain the data, and some other related resources and conference information.

The UCI Machine Learning Repository
http://www.ics.uci.edu/~mlearn/MLRepository.html
This repository at the University of California, Irvine, contains data in many different domains (including biomedicine) that have been widely used to test and compare machine learning techniques.

WEKA
http://www.cs.waikato.ac.nz/ml/weka/
Developed at the University of Waikato in New Zealand, WEKA is an open-source machine learning software written in Java, containing a wide range of useful algorithms.

QUESTIONS FOR DISCUSSION

1. What are the similarities and differences between bioinformatics and medical informatics? How can research in the two areas be beneficial to each other?

2. What is an intelligent system? Can an intelligent system be more intelligent than humans? What are the important characteristics of an intelligent system in biomedicine?

3. Discuss the characteristics of major machine learning paradigms and their applicability in biomedicine.

4. Explain what knowledge management is and why it is useful for medical informatics. What are some of the good examples of biomedical knowledge management systems? How can a knowledge management system be created and used in industry?

5. Please compare the knowledge discovery process by computers with that in humans. Do you think that data mining and text mining techniques have begun to change the way that research is done in biomedicine?

6. What are the social, ethical, and legal concerns for future biomedical knowledge management, data mining, and text mining applications?