Biological and Medical Big Data Mining

George Tzanis
Aristotle University of Thessaloniki, Greece

Abstract. This paper discusses the concept of big data mining in the domain of biology and medicine. Biological and medical data are increasing at very rapid rates, which in many cases outpace even Moore's law. This is the result of recent technological development, as well as the exploratory attitude of human beings, that prompts scientists to answer more questions by conducting more experiments. Representative examples are the advances in sequencing and medical imaging technologies. Challenges posed by this data deluge, and the emerging opportunities of their efficient management and analysis are also part of the discussion. The major emphasis is given to the most common biological and medical data mining applications.

Introduction

Data collection and data analysis were actually taking place from ancient time, even if they were in a primitive form. Many ancient human civilizations had gained important knowledge by observing the planets and the stars. By analyzing these observations they were able to accurately predict the time of the seasonal changes over a year. These predictions were very valuable, especially for agricultural and habitation purposes, providing the means for the survival and development of these civilizations.

Later on in the age of scientific revolution data collection and analysis became a more mature process that guided to a large number of important scientific discoveries. Worth mentioning is the large number of accurate and comprehensive astronomical observations that were collected by the Danish astronomer Tycho Brahe during the early years of scientific revolution in 16th century. After Brahe's death, Johannes Kepler used those astronomical data, a fact that implies a kind of data sharing, and developed his three laws of planetary motion. Another important example of data collection and data analysis was the one of Charles Darwin's in 19th century. Darwin made a voyage that lasted almost five years. During the voyage he investigated geology of the lands he visited and made a lot of natural history collections. The notes and observations he made during his voyage were determinant for the development of natural selection and evolution theories.

In the 20th century the important discoveries concerning DNA, such as the clarification of the correct double-helix model of DNA structure (Watson & Crick, 1953) established molecular biology as one of the most important research fields of biology. These discoveries attracted much attention and changed the direction of research in biology, as well as in medicine. Although the advances in biology during the 20th century were great, the scientific theories and discoveries of physicists are considered even greater. Therefore 20th century is described as the century of physics. However, as it is widely believed we are now living in the century of biology, which promises important advances that will enlighten the constitutive details and rules that characterize and govern life (Venter & Cohen, 2004).

The acquisition of more data has been proceeding through various inventions and technological advancements. For example, the invention and use of telescope made possible the observation of more objects in the sky, whereas the invention and use of the microscope made possible the discovery and study of microscopic organisms such as bacteria. One of the most important recent technological advancements in biology was the development of the polymerase chain reaction (PCR) by Kary Mullis in 1983. The first scientific publication about PCR presented by Mullis et al. three years later (1986). PCR is a biochemical process that amplifies a single or a small number of copies of a piece of DNA sequence across several orders of magnitude. The great importance of PCR is reflected in the fact that PCR was the cornerstone of developing large-scale experiments and sequencing projects making possible to decipher the genetic code of organisms. The representative example is the Human Genome Project, which was founded in 1990 by the U.S. Department of Energy and the U.S. National Institutes of Health (NIH) and was completed in 2003.

After the recent technological advances that made possible the conduction of many large scale experiments, the collection of biological data has been increasing at explosive rates. An important example to perceive the rapidness of this data growth is to consider that the number of transistors on integrated circuits and consequently the processing speed as well as storage capacity of computing hardware doubles approximately every 18 months. This is a very good estimation made by Gordon Moore (Moore, 1965) and is widely known as Moore’s law. However, nowadays Moore’s law seems reaching its limits. In contrast, new biological data is doubling approximately every 9 months, and this rate seems to increase dramatically (EMBL, 2013).

Basic Molecular Biology Concepts

Diversity is a key property of life and is reflected in the tremendous heterogeneity among living creatures. Surprisingly, the underlying molecular details of organisms are almost universal. All organisms depend on the activities of proteins, a complex family of molecules that comprise the main structural and functional units of cells. The hypothesis of molecular unity of organisms is strengthened by the fact that similar protein sets with similar functions are found in very different organisms. Nucleic acids, namely deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), are another family of molecules found in every organism having the role to carry the code of life. Both unity and diversity of living things have been arisen through the force of evolution (Hunter, 2004).

Both proteins and nucleic acids are linear polymers of smaller molecules called monomers. The term sequence is used to refer to the order of monomers that constitute these molecules. Because their sequence is usually long, they are both called macromolecules. Sequences of proteins and nucleic acids are usually represented by strings of different symbols, one for each monomer. Proteins are constructed by the linear combination of twenty amino acids, whereas nucleic acids are constructed by the linear combination of four nucleotides. Each nucleotide symbol refers to the nitrogen base it contains. DNA may contain adenine (A), cytosine (C), guanine (G), or thymine (T). In RNA molecules thymine is replaced by uracil (U). DNA is usually double-stranded, including two complementary chains, where each A of one chain binds to a T of the opposite and each C of one chain to a G of the opposite. These DNA strands are antiparallel. On the other hand, RNA is usually single-stranded.

DNA is the genetic material of almost every living organism. Instead, RNA is the genetic material for some viruses and carries out a variety of other functions mostly related to transcription and translation, which are described below. There are various types of RNA, such as messenger RNA (mRNA), which carries information from DNA to protein, ribosomal RNA (rRNA), which is part of ribosomes (translate mRNA to proteins), and transfer RNA (tRNA), which carry amino acids to protein synthesis location. Function of proteins is generally determined by their structure. Depending on various molecular forces, proteins are arranged in four levels of conformation (primary, secondary, tertiary, and quaternary structure).

The genetic material of most organisms is organized in long double-stranded DNA molecules called chromosomes. An organism may contain one or more chromosomes. For example human cells contain 23 pairs of chromosomes. Two of them (X and Y) are sex chromosomes. Each copy of a pair of chromosomes is inherited from each parent. The term ploidy refers to the number of sets of chromosomes. For example, human somatic cells are diploids (contain 2×23 chromosomes), whereas the human gamete cells are haploids (contain 23 chromosomes) A gene, that represents a molecular hereditary unit, is a DNA sequence located in a particular chromosome and encodes the information for the synthesis of a protein or RNA molecule. A chromosome usually contains a large number of genes. All the genetic material of a particular organism constitutes its genome. It is estimated that there are about 20,500 genes in the haploid human genome, whereas the length of the haploid human genome is around 3×10^9 base pairs (NHGRI, 2013).

The central dogma of molecular biology, as coined and re-stated by Francis Crick (1958; 1970), describes the flow of the biological information (Figure 1). The general transfers that take place in most organisms are described by the filled arrows. In particular, DNA is transcribed into RNA that is finally translated into protein. The circular arrow around DNA denotes its replicability. Furthermore, there are some special transfers, described by the unfilled arrows. RNA of retroviruses (e.g. HIV) is reverse transcribed into DNA, which is then integrated into the infected cell's DNA. Also, there are viruses that replicate their RNA. Finally, in the laboratory it is possible to directly translate DNA into a protein.

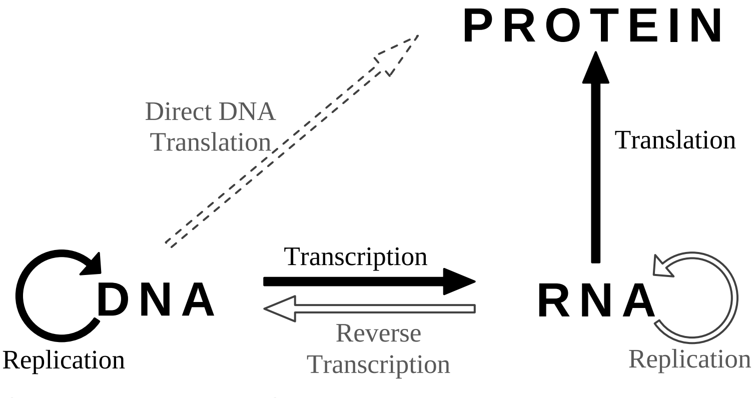

Figure 1: The flow of biological sequence information

There have been proposed a lot of classification systems of organisms. One of the most recent and widely used systems (Woese et al., 1990) divides organisms into three domains, bacteria, archaea, and eukarya. Eukaryotic (comes from the Greek word "karyon" meaning “nut” or “kernel”) cells contain a nucleus, which contains their genetic material, whereas bacteria and archaea (prokaryotic cells) lack such a kind of structure. Eukarya include many organisms, such as animals, plants, fungi, and protists. Most eukarya are multicellular organisms, whereas bacteria and archaea are unicellular.

New Biological Fields and Paradigms

The introduction of new technologies as well as the improvement of existing ones unavoidably led to the emergence of new fields and paradigms either because of the need to manage data masses or because the resulted new knowledge pointed to new directions.

The explosive growth of biological data has made impractical their efficient organization, maintenance, dissemination and analysis without the use of computers. This need led to the evolution of bioinformatics, an inter-disciplinary field at the intersection of biology and informatics. The basic aims of bioinformatics are the organization of data in a way that allows access and update, the development of biological data analysis tools, and the analysis of data in order to gain new biological insights.

Another popular term that is used sometimes in order to refer to the research area that bioinformatics covers is computational biology. However, many scientists consider bioinformatics and computational biology as two distinct fields, although strongly related. As mentioned by Searls (2010), most of those who discriminate bioinformatics and computational biology describe the former as a toolkit and the latter as a science.

A neologism that has become very popular last years is the use of suffix "omics", which usually refers to a field of study in biology, such as genomics and proteomics. The subject of study of these fields, which is usually a collection of objects (e.g. an organism’s genetic material or an organism’s whole set of proteins), is described by using the suffix "ome", such as the genome and proteome respectively. There have been informally proposed a lot of such terms leading to an overuse of this new terminology paradigm. A large list of related terms can be found in (MediaWiki, 2013).

Another new paradigm is systems biology, which is the approach followed in biological and biomedical research in order to understand biology at the system level. It is an inter-disciplinary field of study that focuses on the structure and dynamics of cell and organism function in a more holistic perspective, rather than on the characteristics of isolated parts of cells or organisms in the framework of the traditional reductionism. Properties of systems become of central concern. The understanding of these properties may have a strong impact on the future of medicine. However, many advances in experimental technologies, software, as well as data analysis methods are required, before systems biology will be able to provide all its promising potential (Kitano, 2002).

Epigenetics is the study of changes in gene function that cannot be explained by changes in DNA sequence (Riggs et al., 1996). These changes are heritable through mitosis, the process by which a cell doubles its genetic material in order to be divided into two daughter cells. Epigenetic mechanisms add epigenetic marks, such as methyl groups, to DNA or to other molecules, usually histones (take part in DNA packaging). These mechanisms work together and affect gene expression at many locations throughout the genome. The resulting epigenetic state of the genome, called epigenome, varies by cell type. The diversity of epigenetic marks is huge. In a diploid human genome there are tens of different post-translational histone modifications and more than 50×10^6 sites of potential DNA methylation. This is translated in about $2^{50,000,000}$ possible epigenotypes, and seems that no two human cells would have identical epigenomes (Cortessis et al., 2012). Moreover, epigenetic states change over time depending on normal developmental or pathological processes, as well as environmental exposures or random variation. This huge variability in epigenome dictates that it is a second genome positioned over the original one. That is the reason why the Greek prefix "epi", which means over or above, has been utilized in the term “epigenome”.

Data Mining

Data mining emerged in order to cope with the challenges that traditional data analysis techniques where facing up when dealing with large amounts of often peculiar data. Strictly speaking, data mining is the main step in the process of Knowledge Discovery in Databases or KDD Process (Figure 2). Knowledge Discovery in Databases is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996). However, the term “data mining” is very often used to describe the whole KDD Process. Although the core of the process is the data mining step, where a data mining algorithm is applied in order to extract the patterns from data, the pre-processing and post-processing phases are very important too and contribute sensibly to the quality of the extracted knowledge. The pre-processing phase usually includes the selection of an appropriate portion of data, the cleaning of the selected data, as well as the transformation of data. The post-processing phase deals with the management of the produced patterns and models and focuses on the evaluation and interpretation of data mining results.

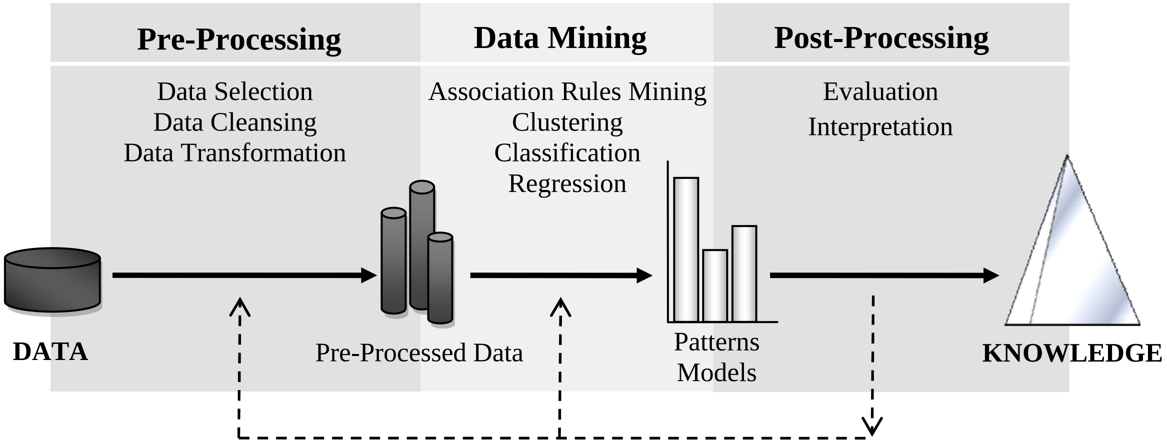

Figure 2: The KDD Process

The most common data mining tasks are classification, regression, clustering, and association rules mining (Table 1). The main goal of the first two data mining tasks is prediction, whereas the goal of the other two is description. In the first case, a predictive model is fitted in known data examples, and is used for predicting new data. In contrast, descriptive algorithms identify patterns or relationships in the data. Depending on the nature of the data and the desired knowledge there is a large number of algorithms for each task.

Table 1: Most common data mining tasks

Predictive	Descriptive
Classification. Maps data into predefined classes, using a model that was constructed on data with known class mappings.	*Association Rules Mining*. Extracts rules that describe relationships among variables of data.
Regression. Maps data into a real valued prediction variable, using a model that was constructed on data with known mappings.	*Clustering*. Groups data, so that data in the same group (cluster) are more similar to each other than to those in other groups.

Big Data Challenges and Opportunities

The term “big data” is very popular nowadays and related research has attracted the attention of a large portion of the scientific community. As mentioned by Diebold (2012) the first appearance of this term was in a Silicon Graphics slide deck with the title of “Big Data and the Next Wave of InfraStress" in 1998. The first book mentioning “big data” was a data mining book by Weiss & Indrukya (1998), whereas the first academic paper including this term in the title appeared two years later (Diebold, 2000).

The popularity of “big data” concept is justified by the fact that big data raise a lot of challenges in terms of management and analysis, as well as by the fact that recent technological advances have made possible the collection of vast amounts of data in short time due to the research efforts conducted in the framework of many scientific disciplines and research fields. Among these domains are biology and medicine, which have recently joined the big data race, encountering data management challenges that were traditionally faced up by astronomers and high-energy physicists. The challenging character of big data is described by 5-V’s. Laney (2001) was the first who talked about the 3-V's in big data management, namely volume, variety, and velocity. However, nowadays there are being proposed even more V’s, but two of them, variability and value, are the most popular (Fan & Bifet, 2012). These 5-V’s are described in the following lines (Fan & Bifet, 2012; Fayyad, 2012):

- *Volume*. The volume of the data becomes larger than ever before, and increases rapidly, providing challenges in loading, processing, and transferring.
- *Velocity*. The velocity of data streams that arrive continuously and need to be analyzed pose challenging real-time constraints.
- *Variety*. There are many different types of data (e.g. text, sensor data, images, audio, video, graph), various degrees of structure in data (structured, semi-structured, unstructured), and often different types and structures of data are mixed.
- *Variability*. The structure of the data as well as the way users want to interpret these data usually changes with time providing extra data and knowledge management challenges.
- *Value*. The value refers to the quality of the extracted knowledge that may give the ability of making better decisions and answering more questions.

The increasing volume, heterogeneity, and complexity of biological and medical data raise difficulties that cannot be overcome by current approaches. Biological and medical big data usually derive from three categories of sources (NIH, 2013):

(1) A small number of groups that produce very large amounts of data (usually large funded projects).
(2) Individual investigators who produce large datasets utilizing new technologies.
(3) A large number of sources that each produces small datasets (e.g. research data or clinical data in electronic medical records).

To illustrate the volume of human genome data, it should be considered that a DNA sequence can be represented in a binary format (0 and 1), which is the alphabet of computers, by using at least two bits for each individual base pairs. Two bits provide 4 different words of information, namely 00, 01, 10, and 11. Those four 2-bit words are adequate to represent the 4 different DNA bases. In that way a byte, which consists of 8 bits, can store 4 DNA bases. So, in order to represent the entire diploid human genome (6×10^9 base pairs) in digital format there are needed 12×10^9 bits or 1.5×10^9 bytes, that is 1.5 Gigabytes (Gb). However, the data files that are provided by current sequencers occupy much more volume than 1.5 Gb, that can be about 140 Gb as Lawrence Hunter, a computational biologist at the University of Colorado Denver, describes in a Nature's technology feature article (Marx, 2013).

Another example of the vast amount of data produced by current sequencing projects is that of BGI, which was formerly known as the Beijing Genomics Institute, in China. BGI is one of the largest producers of genomic data in the world generating 6 terabytes of genomic data every day. Every one of BGI's 157 genome sequencing instruments can decode a human genome per week, an effort taking months or years a few years ago (Marx, 2013).

The existing solutions of the technical problems in analysis of large volumes of distributed biological data are grouped in three broad categories by Huttenhower, & Hofmann (2010):

(1) Web applications aggregating information from multiple sources and providing pre-computed data mining results.
(2) Application programming environments (APIs) allowing more sophisticated queries on individual large data sources.
(3) Do-it yourself solutions relying on manually obtaining and processing bulk data from various public repositories, which is a highly time consuming approach.

The major challenges faced up in the use of biological and medical big data have given rise to new efforts, such as the NIH Big Data to Knowledge (BD2K) initiative (NIH, 2013), which aim to deal with this new paradigm effectively. The mission of these projects includes the localization and access of data and software tools, the standardization of data and metadata, the extension of data and software sharing policies and practices, the organization, management, and processing of biological and medical big data, and the development of new data analysis and integration tools. Finally, the training of researchers for the effective use of the big data resources is of particular importance.

As presented in introduction, biological data are becoming available at a rate that notably outpaces Moore's law. Consequently, individual hardware units are not any more adequate to satisfy the computational requirements for managing and analyzing these data. New algorithmic approaches with increased scalability and efficiency, as well as infrastructures that exploit the computational power of multiple hardware units, such as distributed and cloud computing are deemed necessary.

The prevalent current trend in dealing with the big data storm is the concept of cloud computing. According to NIST definition (Mell & Grance 2011), "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction". A lot of effort in biology and medicine focuses on cloud computing, in order to utilize data and software tools situated in huge, off-site centers. As Hunter describes (Marx, 2013), researchers do not have to buy their own hardware, thus cloud computing becomes extremely attractive in an age of reduced research funding.

Another great initiative supported by CERN, ESA, EMBL, and many other partners is the Helix Nebula - the Science Cloud (Helix Nebula, 2013). It is an effort aiming to support the massive IT requirements of European scientists, based on cloud computing infrastructure.

Finally, there have been developed big data infrastructures with the most popular being Apache Hadoop (The Apache Software Foundation, 2013a). It is a platform for writing applications that efficiently processes large amounts of data in parallel using clusters of computer nodes. It is based on a distributed file system called Hadoop Distributed Filesystem (HDFS), as well as on the MapReduce programming model. There are also open source initiatives, such as Apache Mahout (The Apache Software Foundation, 2013b), which is a scalable machine learning and data mining software based mostly in Hadoop. It implements a lot of machine learning and data mining algorithms.

Biological and Medical Data Mining Applications

This section describes the most common data mining application in biological and medical data that have entered or have the potential to enter the big data realm.

Biological Sequence Mining

Sequence analysis refers to the use of analytical methods to unravel the structure, the function, and various features of biological sequences, such as DNA, RNA, and protein sequences. Genome sequencing projects provide vast amounts of biological sequences, which have to be analyzed and annotated. Gene prediction usually follows sequencing and aims to identify biologically functional stretches of DNA. This is not a trivial task, especially for the more complex eukaryotic genomes. Therefore, the use of advanced techniques including data mining is demanded.

The most important sequence analysis tasks include the comparison of sequences in order to find similarities that may imply structural and functional relation, the identification of sequence differences and variations, such as mutations and single nucleotide polymorphisms (SNPs), and the evolution and genetic diversity of sequences and organisms. Moreover, sequence analysis includes the identification of intrinsic characteristics, such as the prediction of regulatory regions (i.e. promoters and enhancers), which are segments of DNA where regulatory proteins bind preferentially and thus control gene expression and consequently protein abundance. Prediction of the transcription start site, the translation initiation site, the splice sites, and the polyadenylation site are also some important sequence analysis tasks. Finally, it should be mentioned that epigenetic states are also recorded using sequencing methods.

However, it is worth mentioning that sequencing and the subsequent analysis of the sequenced genomes is not only constricted by the strict scientific frame of the elucidation of the rules that govern life. It is also applied in clinical practice and it will be even more frequently applied in the future, following the concurrent scientific advances and knowledge insights in molecular biology and bioinformatics. The analysis of the genome of individuals can reveal the genetic profile of the person, providing the potential of actions that were unreachable up to now. The detection of genetic predisposition for specific diseases and the preparation of the best possible drug according to patient's genetic profile are two representative examples of the value that molecular biology and bioinformatics add to clinical practice, guiding to a more personalized medicine. However, the high genetic and molecular complexity and heterogeneity of many diseases, like cancer, keeps, at least for the moment, genomics away from an everyday clinical practice. This suggests an even harder effort and an even greater network of scientific collaboration to overcome the problems that constantly keep personalized medicine at its infancy (Katsios & Roukos, 2010).

Gene Expression Analysis

Each organism contains a number of genes that code the synthesis of an RNA or protein molecule. Almost each cell contains the same set of chromosomes and genes. However, there are many cell types that have very different properties and functions. These differences are dictated by respective differences in protein abundance. The concentration of a protein mostly depends on the levels of its corresponding mRNA molecule, which in turn are determined by the expression of the related gene. The process of information transferring from a gene to mRNA and then to a protein is called gene expression.

A very popular tool for analyzing gene expression is microarray or gene chip (Schena, et al., 1995), which measures the relative mRNA levels of thousands of genes, providing the ability to compare the expression levels of different biological samples. A microarray consists of a large number of spots. Every spot contains an amount of a specific short DNA sequence (oligonucleotide), called probe. This sequence is usually a short part of a gene or other DNA sequence that can be hybridized with its complementary sequence in the sample. The degree of hybridization indicates the level of gene expression. However, as the number of samples increases, the time and cost of the experiment may increase importantly. These samples may correlate with different time points taken during a biological process or at different developmental stages. They may also correlate with different tissue types, such as normal cells and cancerous cells.

Another method for measuring genome-wide gene expression is Serial Analysis of Gene Expression (SAGE), which allows the quantitative profiling of a large number of mRNA transcripts (Velculescu et al., 1995). It is based on sequencing technology and uses small tags (10-13 base pairs) that correspond to fragments of mRNA transcripts. Although, SAGE had the advantage that the experimenter does not have to know in advance which mRNA sequences will be studied, its diminished reliability didn't allow this method to be as popular as microarrays. However, several improved variants have been developed since then, with the most advanced being SuperSAGE (Matsumura et al., 2005), which uses longer tags (26 base pairs). SuperSAGE provides the benefits of reduced time, cost, and effort for analysis by exploiting the advantages of next-generation sequencing, which reduces complexity and allows better quantification.

A more recently developed technology for genome-wide expression analysis is Whole Transcriptome Shotgun Sequencing (WTSS) or RNA-Seq (Morin et al., 2008). This method uses high throughput sequencing technologies to determine the expression level and exact nucleotide sequence of each mRNA transcript that is expressed in a sample.

The greatest challenge posed by gene expression data is their high dimensionality. They contain a small number of samples, which is in the order of tens or a few hundreds, but a very large number of features, namely genes, that is usually in the order of thousands. The majority of genes are usually irrelevant and uninformative. Feature selection methods aim to reduce the possibility of shadowing the relevant genes' information. Many feature selection approaches have been utilized for reducing the dimensionality of the data by selecting a small number of genes. A thorough study of such methods is provided by Hua et al. (2009).

Clustering is maybe the most used method in gene expression analysis. Clustering methods can be used to cluster genes with similar behavior or samples with similar gene expressions together. A category of clustering algorithms, called subspace clustering, bi-clustering, or two-mode clustering algorithms (Van Mechelen et al., 2004), are used to simultaneously cluster genes and samples. This family of algorithms is very interesting in the domain of gene expression analysis, since a small set of genes is usually expressed in a small group of samples. However, the high dimensionality and the noise of the data, as well as the high computational complexity of bi-clustering algorithms constrict the performance. A survey of gene expression clustering techniques is presented by Jiang et al. (2004).

Data Mining in Structural Bioinformatics

Structural bioinformatics is the subfield of bioinformatics which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules, especially for proteins. The application of data mining in structural bioinformatics is quite challenging, since structural data are not linear. Moreover, the search space for most structural problems is continuous, namely infinite and demands highly efficient and heuristic algorithms (Gu & Bourne, 2009).

Data mining algorithms have been utilized for the prediction of various protein properties, such as active sites of enzymes, modification sites, and protein domains. One of the most interesting applications of data mining in structural bioinformatics is the prediction of secondary structure of proteins given their sequence of amino acids. This problem has been studied for about four decades, with the first published work based on single residue statistics (Chou & Fasman, 1974). Many techniques have been developed since then for dealing with this problem and combined methods have provided improved prediction accuracy.

Other important problems of structural bioinformatics that utilize data mining methods are the RNA secondary structure prediction, the inference of a protein's function from its structure, and the efficient design of drugs, based on structural knowledge of their target.

Biomedical Text Mining

Biomedical text mining, concerns the automatic extraction of information from usually unstructured text documents with content related to biology and medicine. Biomedical text mining has become an attractive research area, since it is critical for researchers in the fields of biology and medicine to access the most recent information concerning the domain of their expertise. Current practice involves on-line search for information, employing the latest technologies in text mining, information retrieval, and semantic web.

From a data miner’s point of view, biomedical text data have special characteristics that raise essential challenges. The most important of these characteristics are the heavy use of domain-specific terminology, the polysemy of words, namely the existence of many words that have multiple meanings (word sense disambiguation), the low frequency words (sparse data), the use of multiple terms having the same meaning (semantics), the frequent creation of new terms, and the different writing styles. Moreover, the huge volume of data makes biomedical text mining an even more challenging big data mining paradigm. Illustrative of the large volume of biomedical text data is the fact that PubMed, which is one of the most popular databases providing free access to references and abstracts on life sciences and biomedical topics, has about 23 million records, as of August 2013.

The tasks of biomedical text mining have been categorized in many studies from different points of view. In the following lines the most common tasks are presented according to Cohen and Hersh (2005) as well as Simpson & Demner-Fushman (2012):

- *Named entity recognition (NER)*. Refers to the automatic identification of the presence of specific biological or medical terms in unstructured text.
- *Synonym and abbreviation extraction*. Provides more advantages to users’ searches, since many biological and medical entities have multiple names or abbreviations.
- *Relationship extraction*. Refers to the identification of relationships among entities.
- *Hypothesis generation*. Involves the extraction of previously unknown possibly valuable relationships among entities.
- *Event extraction*. It is the result of a recent shift in biomedical information extraction from binary relationship identification to the more motivated task of recognizing complex, possibly nested event structures.
- *Text classification*. It is the automatic recognition whether a document or part of it has specific characteristics of interest.
- *Summarization*. Refers to the automatic summarization of biological and medical documents.
- *Question answering*. Involves the process of answering directly and precisely to natural language questions.

Biological Network Analysis

Network analysis can be applied in various problems in biology and medicine, including finding gene function, drug target identification, designing strategies for disease diagnosis and treatment, as well as dealing with epidemics and biosurveillance. The large number of pieces that compose a biological system and the interactions among them are more accurately described as networks. These networks are represented by graphs that may contain many thousands of vertices and edges. The most popular network analysis applications in biology and medicine are described in more detail in the following lines (Pavlopoulos et al., 2011).

- *Protein-protein interaction (PPI) networks*. These networks contain information of functional interactions among various proteins that need to cooperate with each other in order to accomplish the various biological processes. The majority of proteins interac with multiple partners (on average six to eight other proteins) and form complex interaction networks (Panchenko & Przytycka, 2008). The most common representation of these networks is as undirected graphs. The amino acid sequence of most proteins is known. However, the functions of the majority of these proteins have not been accurately recognized. The prediction of proteins functions is a challenging problem. The researchers believe that they will gain important knowledge on how each protein works by studying its interactions with other proteins. Since there are techniques that make possible the detection of protein interactions within an organism, research has been focused on the analysis of protein interaction networks in order to infer protein functions.
- *Gene regulatory networks (GRNs)*. These networks contain information concerning the control of gene expression in cells. They include a set of DNA segments that interact with each other as well as with other substances in the cell. Those interactions are accomplished indirectly through their RNA or protein products. There is a number of factors that regulate and modify those products, such as transcription factors, which regulate the process of transcription and post-translational modifications (e.g. phosphorylation or glycosylation) that alter proteins in order to change their structure and function. Gene regulatory networks are usually represented as directed graphs in order to model the way that proteins and other molecules participate in gene expression.
- *Signal transduction networks*. Signal transduction networks contain interactions among different biological molecules such as proteins or other macromolecules, hormones, and neurotransmitters, representing the transmission of signals either within the cell or from the cell's environment to the inside of the cell. These networks are usually represented by directed graphs with multiple edges.
- *Metabolic networks*. These networks contain information related to metabolic processes of organisms. The series of biochemical reactions that take place in cells are called metabolic pathways. Metabolic networks represent the metabolic pathways along with the interactions that regulate these biochemical reactions. Enzymes, which are the proteins that catalyze biochemical reactions, have a key role in metabolic networks.

A lot of methods have been proposed for dealing with network analysis and graph mining. In particular, most of the research efforts concern analysis of the World Wide Web and social networks. However, biological network analysis benefits from the advances in general network analysis, but also demands special attention due to the specific characteristics of such networks as described above.

Matagenomics Data Mining

Metagenomics is a field that deals with the study of metagenomes, which include the genetic material of microorganisms that is received directly from the environment they inhibit. Metagenomics enables the study of uncultured microorganisms in genomic level. The result of studying only cultured microorganisms was that the vast majority of microbial biodiversity had been missed (Hugenholz et al., 1998). The recent utilization of sequencing methods (Eisen, 2007) in order to sequence the genome of uncultured microorganisms that are sampled directly from their habitats cures significantly this problem. Higher speed and lower cost of sequencing technologies provide the means for thorough and wide study of this microscopic world.

It is obvious that this novel way of genomic data gathering pose new big data challenges that were not met before. First of all, since the data are received from uncultured microorganisms, there is a large genomic heterogeneity, often containing more than 10,000 species. Moreover, the sequential data are typically noisy and partial (Wooley et al., 2010). This recently evolved field seems quite promising to biologists and has attracted a lot of attention. The basic aim of bioinformatics and big data scientists is to provide the tools to deal efficiently and effectively with the large volume, high noise and incompleteness of metagenomic data.

Medical Image Mining

Medical imaging refers to any process that is used in order to create images of the human body or parts of it. The aim of creating these images is to provide clinical diagnostic information or to contribute to research in medicine. The most of the medical imaging data are received in the application framework of radiology, nuclear medicine, endoscopy, and microscopy.

Medical imaging data are collected in an incredibly rapid rate resulting to huge collections of data that pose big data mining challenges to researchers. According to Frost & Sullivan’s recent predictions (Nafziger, 2013), diagnostic imaging alone is going to generate about 1 exabyte (10^{18} bytes) of data in three years (2013-2016). Advances in medical imaging technology have greatly increased the amount of medical information included in a medical image posing challenges to clinicians too. This arises due to the increased spatial resolution, which provides greater anatomical detail, as well as the increased contrast resolution, which permits evaluating even more subtle structures (Rao et al., 2009). For this reason, although these technological advances provide diagnostic benefits, they may result in data overload. Moreover, the increased image acquisition rate increases even more clinician’s workload.

There have been developed a lot of clinically motivated data mining products aiming to extract the key information from the vast amount of imaging data in order to provide to radiologists the means for a more accurate and timely diagnosis with a consequent positive impact on patient’s health. Common methodology of all of the medical image mining applications, which may target on different clinical tasks, is the transformation of raw imaging data into clinically relevant information, through data mining algorithms. Rao et al. (2009) review such applications focusing on the data mining challenges faced up in developing commercial products.

Data Mining for Biosurveillance and Epidemiology

Biosurveillance is dedicated to managing and analyzing health related data for the early detection of threats and hazards, so that the most effective and timely actions can be taken to protect public health (DHHS, 2010). There are many potential threats to human health, including various environmental exposures and disease outbreaks. This new paradigm for public health aims to integrate and efficiently manage health related data contained in many distributed data sources. The large volume and the noisy character of these data demands special treatment and cloud computing infrastructures are considered particularly important for managing this kind of data (Ramanathan, 2012).

Epidemiology is a field of medical science that deals with the incidence, distribution, and control of disease in populations depending on the study of patterns, causes, and effects of health and disease conditions. Classical statistical methods have been immensely used in epidemiology. However, the vast amount of data along with their embedded complexity makes data mining an indispensable tool for analyzing efficiently and effectively these data in a hypothesis generating manner rather than the traditional statistical hypothesis testing style (Fefferman, 2006).

Both biosurveillance and epidemiology deal with the detection of unusual events, such as disease outbreaks. Consequently, anomaly detection and emerging patterns mining are common data mining tasks applied in this domain. Moreover, network analysis and clustering methods are also considered of particular importance in the study of this kind of problems.

Conclusions

For a large period of time research progress in biology and medicine was constricted by the poor yield of data acquisition methods. However, the great technological progress of the last decades has provided the means for collecting huge data masses with lower cost and effort. Examples are the advances in sequencing and medical imaging technologies. After surpassing the data acquisition limits, scientists are now facing up the great challenges of managing and analyzing the collected data. Moreover, new scientific paradigms, such as systems biology, tend to adopt a holistic perspective to their studies and deal with whole systems, rather than isolated parts. Such a paradigm shift widens the field of scientific view demanding an increased amount of data analyses and consequently additional computational resources.

New highly scalable algorithms that can mine efficiently big data, as well as infrastructures that incorporate multiple hardware units, such as distributed and cloud computing, are considered essential in order to deal with the storm of data. A number of initiatives have been arisen to support the requirements of big data mining. Cloud computing seems to prevail at the moment among other research efforts.

However, there is a lot of controversy about big data (Fan & Bifet, 2012). Some believe that there is no need to distinguish big data from traditional data analytics, since data will never be small again. Nevertheless, this is more or less a terminology issue and does not practically affect the direction of research on the topic. The thought that bigger data are not necessarily better data is realistic. Noise, misrepresentation, and search for information that is not included in given data are important aspects that have to be considered, in order to improve quality of big data. Moreover, statistical significance of results obtained from big data analyses has to be a concern, since bigger data may be more prone to randomness. Finally, ethical worries about data accessibility remain an important issue.

Regardless of the controversies about big data, common sense dictates that more data have the potential to embody more knowledge, especially if data acquisition conditions do not reduce acquired data quality. The near future of big data science seems to be predictable and what is believed is that big data are going to become even bigger not only due to the technological advances but also because of the strong intention of scientists to answer more questions. Data deluge as well as scientific and technological progresses have driven for various reasons to the emergence of new interdisciplinary fields that fuse seemingly unrelated fields. Examples are bioinformatics and astrobiology. Analysis of big data has the potential to provide the essential means to bridge the gaps of even more scientific disciplines in the future.

References

Chou, P.Y., Fasman G.D. (1974). Prediction of protein conformation. Biochemistry, 13, 222–245.

Cohen, A.M. & Hersh, W.R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57–71.

Cortessis, V.K., Thomas, D.C., Levine, A.J., Breton, C.V., Mack, T.M., Siegmund, K.D., Haile, R.W., & Laird P.W. (2012). Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. Human Genetics, 131(10):1565-1589.

Crick, F.H.C. (1958). On protein synthesis. Symposium of the Society for Experimental Biology XII, 139-163.

Crick, F.H.C. (1970). Central Dogma of Molecular Biology. Nature, 227, 561-563.

Diebold, F. (2000). "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discussion Read to the Eighth World Congress of the Econometric Society, 2000.

Diebold, F. (2012). On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

DHHS - US Department of Health and Human Services. (2010). National Biosurveillance Strategy for Human Health.

Eisen, J.A. (2007). Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbe". PLoS Biology 5(3): e82.

EMBL (2013), ELIXIR, How fast is life science data growing? Retrieved August 9, 2013, from http://www.elixir-europe.org/news/how-fast-life-science-data-growing.

Fan, W. and Bifet, A. (2012). Mining Big Data: Current Status, and Forecast to the Future. SIGKDD Explorations Volume 14, Issue 2, 1-5.

Fayyad, U.M. (2012). Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. Keynote Talk. BigMine: BigData Mining Workshop KDD-2012, Beijing, China.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.). (1996). Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, Menlo Park, California, USA, pp. 1-34.

Fefferman, N.H. (2006). Selected Problems in Epidemiology. DIMACS Tutorial on Data Mining and Epidemiology.

Gu, J. & Bourne, P.E. (2009). Structural Bioinformatics, 2nd Edition, Wiley-Blackwell, New York.

Helix Nebula. (2013). Helix Nebula - the Science Cloud. Retrieved August 9, 2013, from http://helix-nebula.eu

Hua, J., Tembe, W.D., & Dougherty, E.R. (2009). Performance of feature-selection methods in the classification of high-dimension data, Pattern Recognition, 42(3), 409-424.

Hugenholz, P, Goebel, B.M., & Pace, N.R. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. Journal of Bacteriology, 180(18), 4765–4774.

Hunter, L. (2004). Life and Its Molecules: A Brief Introduction. AI Magazine, 25(1), 9-22.

Huttenhower, C., & Hofmann, O. (2010). A Quick Guide to Large-Scale Genomic Data Mining. PLoS Computational Biology 6(5): e1000779.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering, 16(11), 1370-1386.

Katsios, C., & Roukos, D.H. (2010). Individual genomes and personalized medicine: life diversity and complexity. Personalized Medicine, 7(4), 347-350.

Kitano, H. (2002). Systems Biology: A Brief Overview. Science, 295(5560), 1662-1664.

Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note.

Marx, V. (2013). Biology: The big challenges of big data. Nature, 498, 255–260.

Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Kruger, D.H., & Terauchi, R. (2005). SuperSAGE. Cellular Microbiology 7(1), 11–18.

MediaWiki. (2013). Omics. Retrieved August 9, 2013, from http://omics.org.

Mell, P. & Grance, T. (2011). The NIST Definition of Cloud Computing. Recommendations of the National Institute of Standards and Technology. NIST, U.S. Department of Commerce.

Moore, G.E., (1965). Cramming More Components onto Integrated Circuits, Electronics, 114–117.

Morin, R.D., Bainbridge, M., Fejes, A, Hirst, M., Krzywinski, M., Pugh, T.J., McDonald, H., Varhol, R., Jones, S.J.M., & Marra, M.A. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques, 45(1), 81–94.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. Cold Spring Harbor Symp. Quant. Biol. 51, 263–273.

Nafziger, B. (2013). "Big Data" comes to medical imaging. DOTmed Business News, 42-45.

NHGRI (2013). An Overview of the Human Genome Project. Retrieved August 9, 2013, from http://www.genome.gov/12011238.

NIH. (2013). Big Data to Knowledge (BD2K). Retrieved August 9, 2013, from http://bd2k.nih.gov.

Pavlopoulos, G.A., Secrier, M., Moschopoulos, C.N., Soldatos, T.G., Kossida, S., Aerts, J., Schneider, R., & Bagos, P.G. (2011). Using graph theory to analyze biological networks. BioData Mining, 4:10.

Panchenko, A. and Przytycka, T. (2008). Protein-Protein Interactions and Networks. Identification, Computer Analysis and Prediction, Springer, New York.

Ramanathan, A. (2012). Cloud-Based Computational Bio-surveillance Framework for Discovering Emergent Patterns From Big Data, Invited Talk, NDIA Bio-surveillance Conference, Washington.

Rao, R.B., Fung, G., Krishnapuram, B., Bi, J., Dundar, M., Raykar, V.C., Yu, S., Krishnan, S., Zhou, X., Krishnan, A., Salganicoff, M., Bogoni, L., Wolf, M., Jerebko, A., & Stoeckel, J. (2009). Mining medical images. In Proceedings of the Third Workshop on Data Mining Case Studies and Practice Prize, Fifteenth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining.

Riggs, A.D., Martienssen, R.A., & Russo V.E.A. (1996). Introduction. In Epigenetic mechanisms of gene regulation (ed. V.E.A. Russo et al.), p. 1. Cold Spring Harbor Laboratory Press, New York.

Schena, M, Shalon, D, Davis, R.W., & Brown, P.O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 270, 467-470.

Searls, D.B. (2010). The Roots of Bioinformatics. PLoS Computational Biology, 6(6): e1000809.

Simpson, M.S. & Demner-Fushman, D. (2012). Biomedical Text Mining: A Survey of Recent Progress. In Aggarwal, C.C. & Zhai, C.X. (Eds.). Mining Text Data, 465-517.

The Apache Software Foundation. (2013a). Apache Hadoop. Retrieved August 9, 2013, from http://hadoop.apache.org.

The Apache Software Foundation. (2013b). Apache Mahout. Retrieved August 9, 2013, from http://mahout.apache.org.

Van Mechelen, I., Bock, H.H., & De Boeck, P. (2004). Two-mode clustering methods: a structured overview. Statistical Methods in Medical Research, 13(5), 363–394.

Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. (1995). Serial Analysis of Gene Expression, Science, 270(5235), 484-487.

Venter, C., & Cohen, D. (2004). The Century of Biology. New Perspectives Quarterly, 21: 73–77.

Watson, J.D., & Crick, F.H.C. (1953). A Structure for Deoxyribose Nucleic Acid. Nature 171 (4356): 737–738.

Weiss, S.M. & Indurkhya, N. (1998). Predictive data mining: a practical guide. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences, 87, 4576-4579.

Wooley, J.C., Godzik, A., Friedberg, I. (2010). A Primer on Metagenomics. PLoS Computational Biology, 6(2): e1000667.