General Framework for Biomedical Knowledge With Data Mining Techniques

B.Madasamy[#1], Dr.J.Jebamalar Tamilselvi[*2]

[#]Assistant Professor & Research Scholar, Dept of MCA,
Agni College of Technology, Anna University, Chennai, India.
[*]Director & Professor, Dept of MCA,
Jaya Engineering College, Anna University, Chennai, India

***Abstract*— Data mining is the process which automates the extraction of predictive information discovers the interesting knowledge from large amounts of data stored in information repositories. Biomedical informatics (BMI) is the science underlying acquisition, maintenance, retrieval, collecting, manipulating, and analysing the biomedical knowledge and information to improve medical data analysis, problem solving, and decision making, inspired by efforts toward progress in medical domain. In this research work a comprehensive framework will be generated which comprises of various data mining techniques and evaluate meaningful information from biomedical data. Data mining field will be applied to biomedical data to analyze the characteristics, identify patterns of interest, for diagnosing and predicting patients' health. These proposed biomedical data mining framework useful to the scholars who are interested in the related researches of data mining and medical domain.**

***Keywords*— Data mining, Biomedical, Framework, Knowledge Discovery.**

I. Introduction

Data mining is a replacement for another popularly used term “Knowledge Discovery in Databases”. This technique widely used in many domains including education, finance, commerce, human resource, geological surveys, weather pattern prediction and telecommunications. KDD process brings together extract knowledge from data in the context of large databases which interest to researchers in machine learning, biomedicine, pattern recognition, statistical analysis, artificial intelligence, knowledge acquisition and data visualization. There are various phases involved in mining data as following. Data Integration: The data are collected and integrated from all the different sources. Data Selection: It may not all the data has collected in integration. To select relevant data used for data mining. Data Cleaning: The data have collected are not clean and may encompass errors, inconsistent, missing values, duplicate, noisy or irrelevant data. Data Transformation: The data even after cleaning are not ready for mining as need to transform them into unique forms for mining. Data mining methods used to attain this smoothing, aggregation, normalization. Data Mining: The data to discover the interesting patterns and techniques like clustering and association analysis. Pattern Evaluation and Knowledge Presentation: It includes visualization, conception, transformation, renovation removing redundant patterns from the produced patterns. Discovered Knowledge: It helps to make use of the knowledge acquired to take better decisions.

Biomedical informatics derives knowledge from biomedical data repositories using computer analysis. These can consist of biomedical information stored in the genetic code, experimental results from various sources, patient statistics, clinical studies and scientific literature. Bioinformatics research includes an improvement of repositories, retrieval, and analysis of the biomedical data. Bioinformatics is a rapidly developing branch of biology and it’s highly interdisciplinary with many practical applications in different areas of biology and medicine. The stored data essential to be accessed in a significant way, and often contents of some databanks or databases have to be accessed simultaneously and correlated with each other. Special techniques and algorithms have been developed to facilitate this task biomedical data mining.

In biomedical research datasets are often collected at multiple locations geographically distributed sources of medical and biological data. Data must be stored, processed easily available to different biomedical contributors, researchers, surgeons and healthcare centers. The integration of biomedical information has become an essential task for biology, biotechnology and health care professionals. The application of information mining techniques to the medical domain is helpful in extracting medical knowledge for diagnosis, decision-making, screening, monitoring, and therapy support medical data management. Biomedical data pose certain challenges to bioinformatics because of their inherent natures of high dimensionality, huge volume, complex, heterogeneous, hierarchical, time-varying, and demand for extremely high accuracy. More recent biomedical applications in epidemiology, bioinformatics, and biosurveillance have received increasing attention and challenges to researchers have tried to incorporate data mining in the medical domain. Biomedical informatics can be classified as follows in Figure.1. This paper proposes a framework that facilitates knowledge building using analysis of biomedical data characteristics.

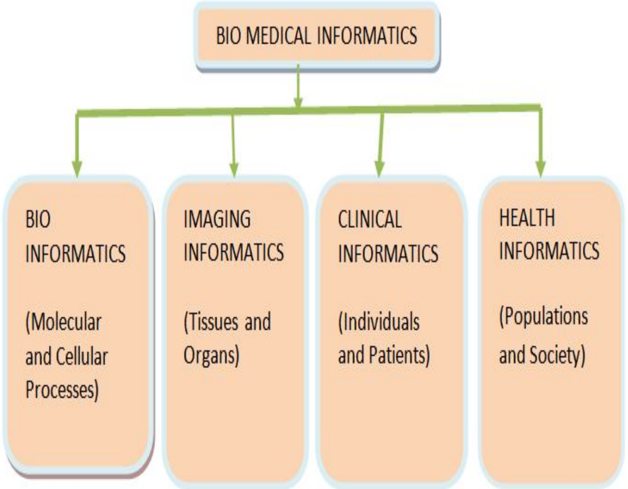


Fig. 1 Biomedical Classification

Biomedical system includes human anatomy and physiology which refers to the structure, control function of human body. Cell biology is the study of the structure and function of cells. Biochemistry is the study of chemical processes which support life. It requires knowledge of key chemical principles which are relevant to biological systems and includes the structure and function of biological molecules and cellular metabolism and its control. Genetics is the study of the structure and function of genes and inheritance. Molecular biology is that branch of biology that deals with the manipulation of nucleic acids such as DNA and RNA. Immunology is the study of components of the immune system, their structure, function and mechanisms of action. It includes innate and acquired immunity. Microbiology is the study of the structure, physiology, biochemistry, classification and control of micro-organisms.

II. LITERATURE REVIEW

This framework deals with public health research data. It is a conceptual framework as well as prototype system for heterogeneous health information including both structured and unstructured data. It developed framework in the cancer patient data warehouse. Integration of the developed techniques into the Cancer Biomedical Informatics Grid will also be carried out. It does not handle all types of biomedical facts. [1] In this framework that facilitates the automatic and systematic construction of systems that detect healthcare fraud and abuse. It does not concentrate about sequences and image data. [2] In this paper propose a Text mining framework to extract the biomedical documents from the biomedical literature. Thus the above framework successfully retrieves the biomedical documents. It concentrates about only text and document based analysis does not analyze clinical data, molecular data and sequence data sets. [3] The goal of the present paper is to devise a new framework addressing microarray data establish superiority of cohesion-based technique both in terms of amount and quality of discovered knowledge. A major achievement is that the framework can produce rules with high accuracy and predictability for molecular biology and biomedical studies. This framework does not handle all types of biomedical data set such as sequencing, clinical data and molecular data. [4] In this article present a genetic algorithm for doing both the tasks of mining and feature selection simultaneously evolving the rules to a real world data mining problem provides the best performance in terms of prediction accuracy and computational efficiency. It does not mine genes, sequences, molecules, text, and clinical facts. [5] This framework is used for electronic detection and medical events to select target events, assessing what information is available electronically, querying the transformed data, verifying the accuracy of event detection, characterizing the events using systems and cognitive approaches, and using what is learned to improve detection. It does not handle complex format of data. But this framework does not deal with offline data repository, unstructured format of the data. [6]

III. FRAMEWORK

The proposed data mining framework Fig. 2 supports biomedical facts, functions, detailed description and design implementation of analysing biomedical data characteristics. It is a conceptual structure intended to support the building of biomedical data mining approaches, explains about the value of data, describe the principles of data mining methods, recognize selected data mining algorithms, describe the advantages of data warehousing, recognize the need for data farming, recognize importance of data and knowledge visualization. There are numerous applications in biomedical technology such as machine learning and statistical methods have been used to discover the potential knowledge from biomedical data. To analyze the characteristics of medical data, to propose a general framework for biomedical data mining, which consists biomedical data selection, biomedical data integration, biomedical data preprocessing, biomedical data transformation, biomedical data mining, biomedical data presentation and biomedical knowledge evaluation.

A. *Biomedical Data Selection*

The main objective of this paper is to propose a biomedical data mining framework with various biomedical UCI datasets and molecular data, protein data, sequence data sets represent a biomedical data analysis. Data mining techniques are needed to support the knowledge discovery process on biomedical data. The selection of appropriate data sources and the choice of suitable data mining methods can lead to a substantial knowledge gain as demonstrated on mining biomedical facts. Data selection is the process of selecting the right data from the database on which the tools in data mining can be used to extract information, gathering knowledge, well structured data and pattern from the following biomedical data repositories under the requirements for high quality, security and robustness. The collection of biomedical information increases the interest in advanced technologies. The multiple data sets collected from population samples, healthcare and patients with specific diseases.

There are famous centralized indexes of distributed biomedical data sources. BIRN project (Marx 2002), GenBank (National Institutes of Health), European Computerized Human Brain Database (ECHBD), fMRI Data Center (Grethe, Van Horn et al. 2001), and the Open Archives Initiative, Brain Map (Fox and Lancaster 2002), Bio Image Database Project (Carazo and Stelzer 1999), National Center for Biotechnology Information (NCBI). The NCBI maintains published medical documents (Pub Med), gene listings (Entrez Gene), protein listings (Entrez Protein), and DNA sequence information (Entrez Sequence). Biological raw data are stored in public databanks such as GenBank or EMBL for primary DNA sequences. The data can be accessed via the www and internet. The most likely translation of all coding sequences in the EMBL databank contains protein sequence databanks like trEMBL. Also databases of scientific literature MEDLINE provide additional functionality, to search for similar articles based on word usage analysis.

B. *Biomedical Data Integration*

Biomedical data pose certain challenges to bioinformatics because of their inherent natures of high dimensionality, huge volume, and demand for extremely high accuracy. Data sources and data types' integration faced many challenges. Biomedical databases cover a growing part of information ranging from clinical findings to genetic structures, including social, behavioural, societal and environmental data. This data is generated and integrated by diverse sources of biomedical history of the subjects and their first degree relatives, laboratory data, data from physical examinations, emerging biochemical markers, instrumental findings, and genomic proteomic data from microarray chips.

The main goal is discovering the biomedical knowledge sources and biomedical data base repository integration with data mining technology. Data Integration relates to the introduction and incorporation of heterogeneous data into the Multi-Knowledge raised area. These needs are mostly related to the integration of main sources of information, namely clinical data, patient-specific genomic proteomic data, the Sequence Retrieval System (SRS), sequences, annotations, proteins, SWISSPROT, Entrez system, Protein 3D structure data bank, access to sequence homology searches and links to other databases are integrated. Structural Bioinformatics, Organism specific databases, Elegans Database, elegans genome, melanogaster are integrated with data mining technology. A major problem is errors in databanks and databases to combining all those data repositories.

C. *Biomedical Data preprocessing and Data cleaning*

Data cleaning and pre-processing approaches data mining techniques helps to systemize the data and store this data in the database or data warehouse for further use of mining and analysis. One of the main tasks related to data analysis is to compare various sequences of the data and search for the similarities among this data which contains disease factors, gene correlation, find the noisy data, statistical data, supervised and unsupervised data, gene information's, redundant data, insignificant and inconsistent data. It tries to eliminate noisy, error and missing biomedical data in the selected biomedical data set repositories. Noise can be defined as some form of error within the data. Some of the tools can be used for filling missing values and elimination of duplicates in the biomedical databases. Due to growth in biomedical research, the large scale of patterns and functions has to be studied. Data cleaning tools helps greatly to study analysis to remove various noisy, error patterns and functions.

D. *Biomedical Data Transformation*

The challenge is to enhance the data collection process so that it combines collection, integration, and analysis effectively by integrating biomedical data, workflow management, tracking data origins, and analysis tools. To automate the extraction of multidimensional biomedical data facilitates stream fusion and pattern detection, which enables researchers to identify the progress how to integrate data sharing in the data collection process type of data selection, integration, and analysis. Individual's biomedical data across various domains and sources are vital component for research decisions achieved by data transformation. Data transformation is the process of changing one data from different data formats. Translating data sources into a unified format with consistent description and logical organization. As a consequence of biomedical source data are typically heterogeneous, inconsistent, fragmented, dirty and difficult to process. Valuable information are embedded within the data cannot be consumed until the data are cleaned, standardized, unified, and integrated.

The aim is to reach an understanding of how researchers are supposed to use available online biomedical data repositories towards new discoveries, and how complex datasets can be transforming the appropriate values. Machine learning algorithms such as Support Vector Machines and Random Forests have been used to transform high dimensional genomic and proteomic data due to their robustness to the dimensionality of the data. Biomedical data transformation can help to improve significantly the classification performance of these algorithms like Naïve Bayes. It can be defined as transformation of the data that is sent for data mining. Biomedical data transformation can increase the data reduction, reduce the dimensionality and increase the efficiency of the data mining with respect to the accuracy and time utilization.

E. *Biomedical Data mining*

Data mining has been widely used in the area of medical science such as biomedical, DNA, genetics and medicine etc. A biomedical knowledge source is a structured database of focused biomedical knowledge. The size and complexity of medical data sets makes it gradually problematic to understand, compare, examine and communicate the data. Data mining helps to extract biomedical data analysis and on-line statistical processing. The data is generally heterogeneous, highly distributed in various areas of research in medical science. It is the major step in KDD when the cleaned and pre-processed data is sent into the intelligent classification, clustering, similarity search algorithms for mining the data. Data mining requires consistent data, meaning that large amounts of biomedical data are needed to convert compatible representations. Knowledge management, data mining, and text mining techniques have been applied to different areas of biomedicine, ranging from biomedical management to biomedical diagnosis, from hypothesis generation to gene Knowledge Management.

A data analysis step the experiment's Tool (MK-DA) supports many data mining processes, such as classification, clustering, class discovery, Gene analysis, and sequence finding. The MK-DA is integrated with the Visualization Tool (MKVIZ) and Report Generator and Manager (MK-REP).

F. *Biomedical Data Interpretation and Presentation*

The individual research efforts biomedical practices, these biomedical data are interpreted with thousands of open and reserved databases, which have been made possible by new database technologies and the web in the form of knowledge management techniques and methodologies have been used to support storing, retrieving, sharing, and management of critical tacit and explicit biomedical knowledge. Knowledge Discovery Data mining emphases on the process of extracting, presenting meaningful patterns from biomedical data discovery using mechanized computational and statistical tools and techniques on large datasets. It can contain methods of data preparation, selection, cleaning, and use of suitable prior knowledge development and application of data mining algorithms, and proper biomedical data analysis.

Data mining is the final presentation of useful knowledge to the user. Data warehouse is a new database model that has been developed to better support data mining techniques. Here the mined biomedical data is presented to the end user in a human viewable format. This involves data presentation, which the user interprets and understands the discovered knowledge obtained by the algorithms.

G. *Knowledge Discovery and Data visualization*

Visualization also plays an important role in biomedical Data mining. The visualization tools helps to present complex structures and sequencing patterns of genes and proteins are most effectively presented in graphs, trees, cubes, and chains. Such visually appealing structures, complex biomedical data structures and facilitate pattern to knowledge discovery data exploration. The developments in biomedical techniques such as molecular, genomic, genome sequencing, protein identification, medical imaging, and patient clinical medical records, these incredible amounts of biomedical research data are visualized every day routine activities. It is an attempt to simplify biomedical data discovery tasks. The results evaluation and interpretation by the expert can lead to iterative refinements of induced model patterns include baseline correction, normalization, scaling and data alignment. Visualization and visual data mining shows a vital part in bionic evolution in biology, medical sciences and DNA technology has led to the accumulation of tremendous amounts of biomedical data.

IV. SUMMARY

Biomedical knowledge discovery is the process of discovering useful knowledge from biomedical data which includes biomedical data analysis and integration of large volumes of data. Extracting knowledge from large volumes of databases, large amounts of data, ability to extract the right information of interest remains the subject of the growing field of knowledge discovery in databases. However in the medical domain, data mining is relatively new field due to certain challenges that are associated with this domain. This paper analysed how data mining may help biomedical data analysis and motivate the further developments of data mining framework for biomedical data analysis and biomedical data mining. Biomedical data are becoming increasingly complex and heterogeneous in nature which is stored in distributed information systems using a variety of data models.

To summarize this paper proposes a framework to address the key questions of how to take advantage of biomedical data based capabilities to improve the efficiency and reliability of accessing biomedical data selection, pre-processing, integration solution development with the help of data mining tasks and algorithms. It analysed the characteristics of biomedical data with the help of data mining techniques. It also addressed how the data mining processing steps are implementing in biomedical data set analysis in a feasible manner. The major involved techniques are biomedical facts selection, biomedical data preprocessing, biomedical data mining besides biomedical knowledge evaluation. Along with the emergence and development of biomedicine, more and more biological information has been discovered by biomarkers.

V. CONCLUSION & FUTURE WORK

The proposed novel biomedical data mining framework yield an underlying acquisition, maintenance, retrieval, collecting, manipulating, analysing, and application of biomedical knowledge and information to improve biomedical data analysis, problem solving and decision making inspired by determinations to improve human health and biomedical researchers. This comprehensive framework comprised of various data mining techniques and evaluates significant information from biomedical data. In this research work, data mining techniques will be applied to biomedical data routinely collected during their day-by-day activity to identify patterns of interest for diagnosing, analyze the characteristics of biomedical data, and predicting patients' health which comprises of biomedical data selection, biomedical data preprocessing, biomedical data mining and biomedical knowledge evaluation.

This framework has been discovered by more biological information which is useful to biomarkers. In near future this framework can be used to analyze specific domain about biomedical research field. To apply various classifications and clustering algorithms to classify and cluster the appropriate biomedical data set. This framework can enhance to develop a new algorithm which is used to classify accurately in biomedical data analysis and domain oriented problem analysis.

REFERENCES

[1] James Gardner, Li Xing "An integrated framework for de-identifying unstructured medical data" Data & Knowledge Engineering www.elsevier.com/locate/datak

[2] Wan-Shiou Yang, San-Yih HwangW.-S. Yang, S.-Y. Hwang "A process-mining framework for the detection of healthcare fraud and abuse" Expert Systems with Applications 31 (2006) 56–68

[3] Latha .K1 Kalimuthu.S2 Dr.Rajaram.R3 "Information Extraction from Biomedical Literature using Text Mining Framework" International Journal of Imaging Science and Engineering (IJISE)

[4] Ramkishore Bhattacharyya "Cohesion: A concept and framework for confident association discovery with potential application in microarray mining" Applied Soft Computing 11 (2011) 592–604 journal homepage: www.elsevier.com/locate/asoc

[5] Riyaz Sikora, Selwyn Piramuthu "Computing, Artificial Intelligence and Information Management Framework for efficient feature selection in genetic algorithm based data mining" European Journal of Operational Research 180 (2007) 723– 737

[6] George Hripcsak, Suzanne Bakken, Peter D. Stetson, and mla L. Patel, "Mining complex clinical data for patient safety research: a framework for event discovery" Journal of Biomedical Informatics 36 (2003) 120–13

[7] Pearson WR (2000) "Flexible sequence similarity searching with the FASTA3 program package" Methods Mol. Biol. 132:185–219

[8] The Genome International Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

[9] Ramos-Pollan, R., et al., "Exploiting eInfrastructures for medical image storage and analysis: A Grid application for mammography CAD," in The Seventh IASTED International Conference on Biomedical Engineering. Austria: Innsbruck, 2010.

[10] Drakos, J., et al., "A perspective for biomedical data integration: Design of databases for flow cytometry" BMC Bioinform. 9:99, 2008.

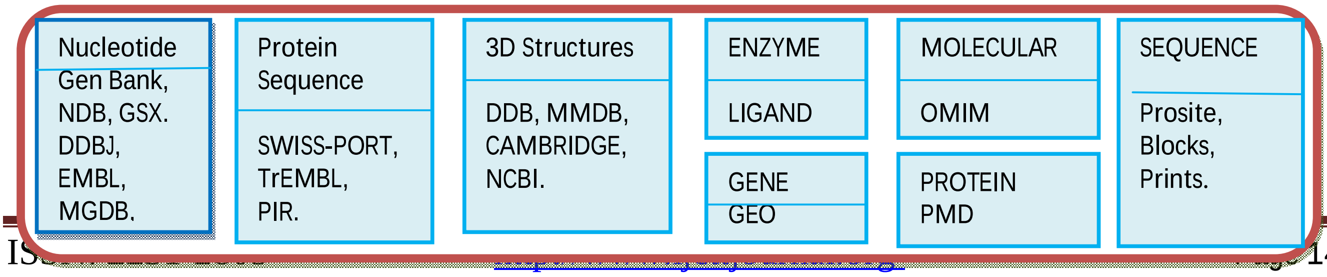
[11] Karasavvas, K.A., Baldock, R., Burger, A. (2004) "Bioinformatics integration and agent technology". *Journal of Biomedical Informatics* 37:205±219.

[12] Saltz, J. et al. caGrid: "Design and implementation of the core architecture of the cancer biomedical informatics grid". *Bioinformatics*

[13] Brown D. "Introduction to Data Mining for Medical Informatics. Clinics in Laboratory Medicine", Volume 28, Issue 1, March 2008, Pages 1-7, Clinical Data Mining and Warehousing

[14] Cios, J.K (2001) *"Medical Data Mining and Knowledge Discovery"* NY: Physica- Verlag Heidelberg.

[15] Harrison J.H. "Introduction to the Mining of Clinical Data. Clinics in Laboratory Medicine" Volume 28, Issue 1, March 2008, Pages 1-7, Clinical Data Mining and Warehousing



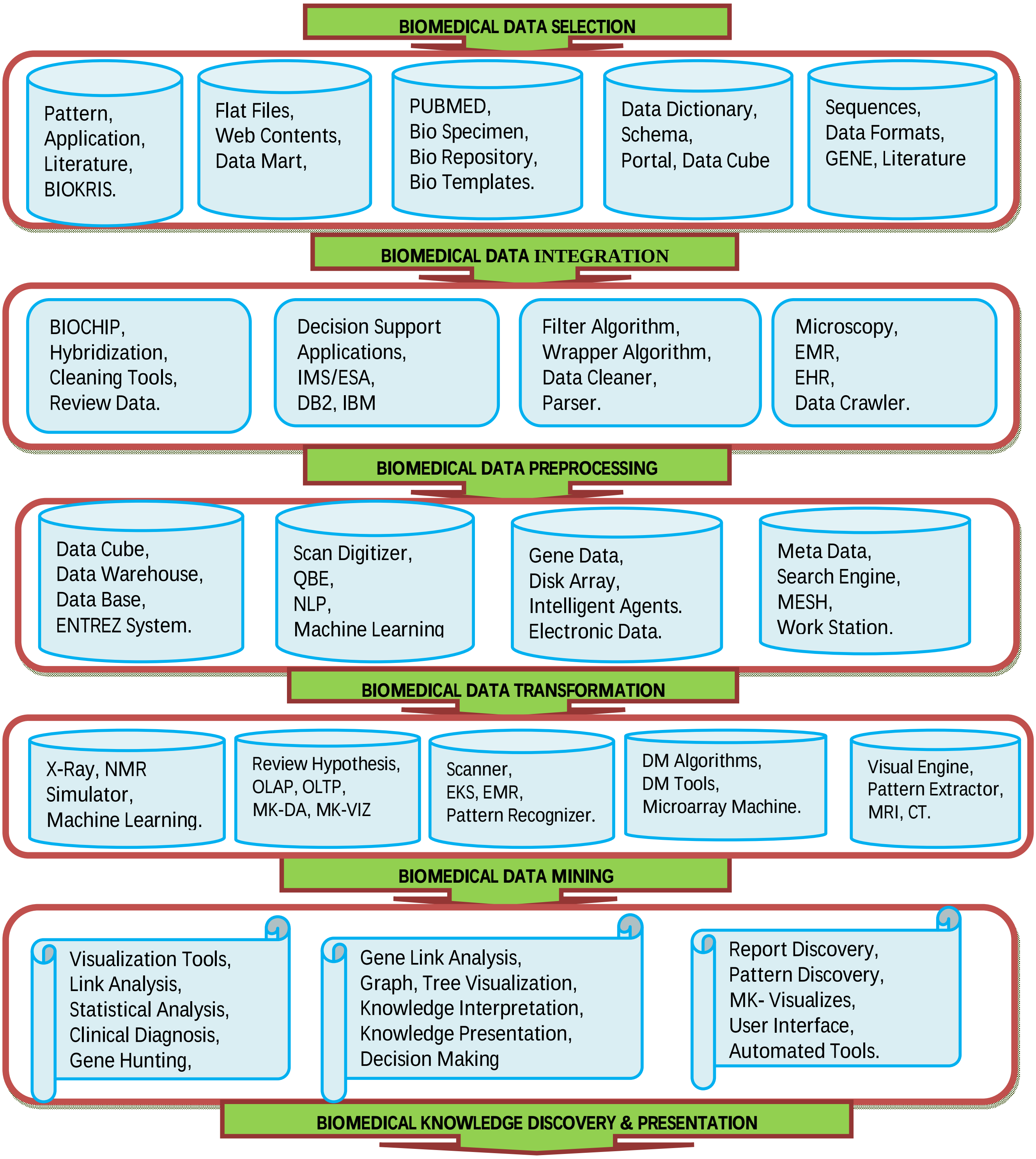

BIOMEDICAL DATA SELECTION
Pattern, Application, Literature, BIOKRIS.
Flat Files, Web Contents, Data Mart,
PUBMED, Bio Specimen, Bio Repository, Bio Templates.
Data Dictionary, Schema, Portal, Data Cube
Sequences, Data Formats, GENE, Literature
BIOMEDICAL DATA INTEGRATION
BIOCHIP, Hybridization, Cleaning Tools, Review Data.
Decision Support Applications, IMS/ESA, DB2, IBM
Filter Algorithm, Wrapper Algorithm, Data Cleaner, Parser.
Microscopy, EMR, EHR, Data Crawler.
BIOMEDICAL DATA PREPROCESSING
Data Cube, Data Warehouse, Data Base, ENTREZ System.
Scan Digitizer, QBE, NLP, Machine Learning
Gene Data, Disk Array, Intelligent Agents. Electronic Data.
Meta Data, Search Engine, MESH, Work Station.
BIOMEDICAL DATA TRANSFORMATION
X-Ray, NMR Simulator, Machine Learning.
Review Hypothesis, OLAP, OLTP, MK-DA, MK-VIZ
Scanner, EKS, EMR, Pattern Recognizer.
DM Algorithms, DM Tools, Microarray Machine.
Visual Engine, Pattern Extractor, MRI, CT.
BIOMEDICAL DATA MINING
Visualization Tools, Link Analysis, Statistical Analysis, Clinical Diagnosis, Gene Hunting,
Gene Link Analysis, Graph, Tree Visualization, Knowledge Interpretation, Knowledge Presentation, Decision Making
Report Discovery, Pattern Discovery, MK- Visualizes, User Interface, Automated Tools.
BIOMEDICAL KNOWLEDGE DISCOVERY & PRESENTATION

Fig. 1 General Biomedical Knowledge with Data Mining Framework