

# Chapter 1

## Biomedical Databases and Data Mining

Claudia Plant<sup>1</sup>, Christian Böhm<sup>1,2</sup>

<sup>1</sup>*Florida State University, Tallahassee, FL, USA*

<sup>2</sup>*University of Munich, Germany*

Medicine, biology, and life sciences are very data intensive disciplines. All kinds of data are produced in tremendous amounts, e.g. text, semi-structured data, images, time-series, data streams, and often very high dimensional feature vectors. Modern devices for data acquisition allow to record more and more information. High resolution mass spectrometry, for example, allows us to measure and quantify hundreds of metabolites or peptides from one biosample. At the level of raw mass spectra even hundreds of thousands of features are measured per sample. A further application scenario is patient monitoring: Modern wearable sensors record various parameters of vital functions and are suitable for long-term patient monitoring.

The data explosion in biomedicine challenges the database community to provide solutions for effective data storage, processing, and exchange. A major additional challenge is to support the so called *knowledge discovery process*, i.e. to extract as much as possible useful knowledge from the data. Data from proteomic spectra for example has shown the potential of superior results for early stage cancer detection than traditional biomarkers [Petricoin *et al.* (2002b,a)]. However, only very few among several thousands of features are relevant for diagnosis. One challenge in the emerging field of proteomics is to identify biomarkers for various diseases, characterizing e.g. different types and stages of cancer. It is still a long way towards to the clinical use of diagnostic tests but one important goal is to identify

significant features (biomarkers) from very high dimensional biological data sets. In the monitoring scenario, it is essential to efficiently identify unusual patterns in streaming time series of sensor measurements. These suspicious observations can be shown to an expert for further analysis.

Going beyond the original intention of the data acquisition, biomedical data may contain valuable, previously unknown information which even can be out of the scope of the original study. The inspiration behind the research area called *data mining* is to reveal such information. High-dimensional data, for example in proteomics and metabolomics, may exhibit various groups of instances, representing unknown sub-stages of a complex disease. In time series of sensor measurements there may be undiscovered correlations between patterns which are characteristic for an abnormal physiological process.

## 1.1 Databases and Knowledge Discovery in Biomedicine

It is a long way from the masses of raw data to information useful to biomedical research and patient management. As foundation for all further steps, the data needs to be suitably organized to guarantee accessibility and long term conservation. Since the earliest days of electronic computing, the need for software systems structuring and organizing data has promoted vital research activities in the area of databases. In biomedical applications, huge amounts of data need to be organized preserving privacy and allowing for efficient search and retrieval. The data are commonly of various sources, type, quality and format. For biology, the community of bioinformatics has identified the design, development and long-term management of biological databases as a core research area [Birney and Clamp (2004)]. Biological databases typically contain raw and aggregated data from scientific experiments and literature from the areas of genomics, proteomics, metabolomics, and phylogenetics. Medical databases, as well as clinical and health information systems typically contain phenotypic data from patients and are managed by local hospitals or on a national basis. Medical databases are often focusing on a certain disease and tend to be less standardized than biological databases. Information exchange across health care providers or even across countries is still often difficult. Recently, much research effort is spent towards standardization and exchangeability of medical data, as reflected by numerous projects focussing on the electronic patient record.

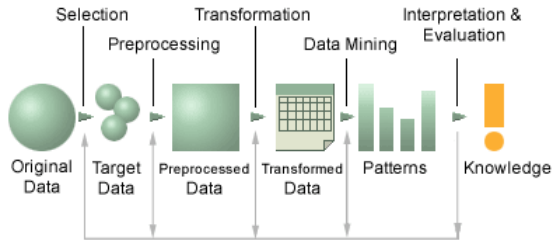


Fig. 1.1: The KDD process.

To support biomedical research on complex diseases, typically data from several biological and/or medical databases need to be integrated.

To meet the requirements of huge amounts of data, the research areas of *knowledge discovery in databases (KDD)* and *data mining* have emerged in the recent years, with multiple books, e.g. [Ester and Sander (2000); Witten and Frank (2005); Han and Kamber (2001)] and numerous papers, e.g. [Fayyad *et al.* (1996); Ng and Han (1994); Ester *et al.* (1996); Papadimitriou *et al.* (2003)], surveys and theses, e.g. [Murtagh (1983); Pan (2006)] to mention a few. Often the terms data mining and knowledge discovery are used interchangeably, however, in the strict sense data mining is one step in the KDD process, which is defined as follows in [Fayyad *et al.* (1996)]: ***Knowledge discovery in databases is the non-trivial process of identifying novel, potentially useful, and ultimately understandable patterns in data.***

Figure 1.1 gives an overview on the KDD process which comprises five major steps:

- (1) **Data Selection.** As a first step the data needs to be carefully selected. Selection criteria include e.g. data availability, quality, type, format, semantics. The selection of high quality data semantically corresponding to the goal of the discovery process is essential for the following steps.
- (2) **Preprocessing.** The target data often requires preprocessing. Suitable strategies on scaling and normalization of the features and strategies to handle missing attribute values have to be selected and applied.
- (3) **Transformation.** To reduce the dimensionality of the data, dimensionality reduction techniques, which derive transformed representa-

tions of the original features, e.g. Principal Component Analysis [Jolliffe (1986)], Independent Component Analysis [Hyvärinen and Oja (2000)], can be applied. Alternatively, feature selection techniques can be used. Feature selection techniques reduce the dimensionality by identifying features which are useful for the goal of the discovery process.

- (4) **Data Mining.** Depending on the goal of the discovery process, a suitable data mining method is selected. The decision on the data mining algorithm is not easy, since for most tasks there is a huge variety of possibilities. The selected algorithm is applied on the preprocessed and transformed data. For most data mining algorithms it is also a non-trivial task to find appropriate parameter settings.
- (5) **Interpretation and Evaluation.** The results of the data mining algorithm are analyzed and interpreted. If the results are not satisfactory, there may be the need to go back one or more steps. In fact, the KDD process is an iterative process.

As a crucial step of the KDD process, data mining requires the selection of a suitable data mining algorithm w.r.t. the goal of the discovery process. Following a common characterization [Han and Kamber (2001)], the diverse data mining methods can be categorized as follows:

- **Clustering:** Find a partitioning of the objects of the data sets into groups (clusters) while maximizing the similarity of the objects in a common cluster and minimizing the similarity of the objects in different clusters.
- **Outlier Detection:** Find objects in the data set which are exceptional, i. e. which do not correspond to the general characteristics or model of the data.
- **Classification:** Learn a function, model or other method from a subset of the data objects to assign a data object to one of several predefined classes.
- **Association Analysis:** Find subsets of the attributes or subsets of attribute ranges which occur frequently together in the data set (called frequent item sets). Derive so called association rules from the frequent item sets. The association rules describe common properties of the data.
- **Evolution Analysis:** Discover and describe regularities or trends for objects with properties that change over time.
- **Characterization and Discrimination:** Summarize general proper-

ties of the data set or of a set of features (characterization). Compare different subsets of the data to other subsets (discrimination).

Another common characterization is to distinguish between *supervised* and *unsupervised* data mining methods. Supervised data mining requires that an attribute is selected by the user which has to be learned by the system for future prediction. The attribute to be learned is often called the class attribute. Specifying suitable values for the class attribute requires domain knowledge and is often done by human experts. The most prominent example for supervised data mining is the task of classification. A classifier is trained on a data set of labeled instances with known values of the class attribute, the so-called training data set. From the training data set, the classifier extracts information to learn a method to predict the value of class attribute of a novel unlabeled instance.

For unsupervised data mining, no class information is required. The most familiar example is clustering. Clustering algorithms aim at finding unknown classes of the data set without any a priori knowledge. This book provides contributions to supervised and unsupervised data mining.

An important issue, especially in the context of biomedical applications, is the performance of data mining methods. In the definition of data mining given in [Fayyad *et al.* (1996)] the performance aspect is explicitly highlighted:

*Data Mining is a step in the KDD process consisting of applying data analysis algorithms, that, under acceptable efficiency limitations, produce a particular enumeration of patterns over the data.*

## 1.2 Outline of this Book

This book features research contributions from internationally leading experts addressing various aspects along the way from data to knowledge in biomedicine. Chapters 2 to 8 are dedicated to biomedical databases, data warehouses, and information systems. Central aspects discussed include modeling of such systems, privacy, data exchange data integration, and data quality.

Chapter 2 by Darshan S. Dillon *et al.* presents a methodology for modeling multi-agent systems. Multi-agent systems are distributed software sys-

tems consisting of autonomous intelligent components called agents which allow supporting complex high-level tasks including data collection and information retrieval. The DYNASTAT methodology purely relies on the standardized modeling language UML 2.2. The authors demonstrate the potential of DYNASTAT in three scenarios: information retrieval on diseases, data collection and data mining, and a system for the general health care practitioner.

In Chapter 3 Fusheng Wang et al. propose SciPort, an extensible data management platform for biomedical research. The web-based platform allows researchers to collect, manage, browse and exchange scientific data. Based on native XML, SciPort supports comprehensive user-friendly queries implemented with XQuery. By using ontologies, SciPort provides semantic enabled data management. To share data, SciPort provides a central server-based light-weight approach to integrate data sources across distributed research institutions. SciPort has been successfully used for translational biomedical research consortia and large scale research collaboration.

Grigorios Loukides et al. focus in Chapter 4 on the challenges of anonymizing and integrating clinical and genomic data which is essential for knowledge discovery. The authors outline DIANOVA (Data Integration ANonymization and View Auditing), a framework for privacy preserving data integration and exchange. DIANOVA is intended to support anonymized integrated views of the database which are useful for genome wide association studies.

In Chapter 5 Mahesh Visvanathan and Gerald H. Lushington provide insights into the challenges emerging from combining mathematical modeling and biological knowledge. For the long-term goal of drug target discovery, the authors integrated biological knowledge from online databases and mathematical knowledge about pathways into a common database. This database allows to systematically study the basic nature of pathways, taking the TNF $\alpha$ -pathway as an example.

Chapter 6 by Sandra Geisler et al. presents a case study in ontology-based data integration. For data management in clinical trials, OnTrIS is presented, an ontology-based data integration system. Based on a core ontology, the user can derive an adapted ontology which triggers the creation of an integrated database. The system has been successfully tested on a pilot trial within the project 'MyHeart' on prevention and monitoring cardiovascular diseases.

In Chapter 7 Francesca Cordero et al. introduce the BIOBITS data warehouse for comparative genomic studies. In particular, a computational

genomic comparison of the *Ca. Glomeribacter Gigasporarum* Bacterium and the Arbuscular Mycorrhizal Fungi genome is supported by BIOBITS. The genomic and proteomic components of the biological problem are represented by a double star schema. The data warehouse also contains two data mining modules: a case-based reasoning and a clustering component. Case-based reasoning is essential for retrieving information at various levels of abstraction and clustering provides the possibility to annotate genetic sequences.

Matteo Bertoni et al. present in Chapter 8 a real life case study on data quality in medicine: A large scale medical database containing several ten million records of clinical and administrative data from hospitals in the Bologna area (Italy) is subjected to an extensive analysis of data quality. Clinical data are privacy-sensitive, heterogeneous, produced by different information systems and mainly intended for patient care. Major quality problems have been found, e.g. wrong genders, missing data, inconsistencies, and useless database columns which are mostly due to low motivation and training of the personnel and missing database constraints. Data quality deserves more attention when designing hospital databases.

Chapters 9 and 10 present solutions for efficient similarity search in large biomedical image databases. Efficient similarity search is a building block for knowledge discovery and data mining. Chapter 9 by Marc Wichterich et al. proposes an approach for efficient computation of the Earth Movers Distance (EMD) which is an established measure for effective image retrieval but very time consuming to compute. The EMD between two images is defined as the minimal amount of changes required to transform the feature representation of one image to the other and the evaluation required solving a linear optimization problem. To speed up computation, the authors propose a dimensionality reduction technique for EMD incorporating domain-specific aspects. The evaluation demonstrates that the approach successfully supports evaluating the EMD on two biomedical image data sets with very different characteristics.

In Chapter 10 Andreas Wichert and Pedro Santos present a system supporting similarity search on clinical records. The system comprises the Subspace Tree, an indexing structure to support content-based image retrieval. In contrast to comparison methods, the Subspace Tree does not partition the data space but the distances among subspaces. Experiments with X-ray images on the prototype system demonstrate that the indexing technique successfully copes with the very high dimensionality of the image data.

Chapters 11 to 15 present contributions implementing various stages of the KDD process in a large variety of exciting biomedical applications. As an approach belonging to the data transformation step, Michael Netzer and Christian Baumgartner present in Chapter 11 a framework for ensemble feature selection in biomedical applications. Stacked feature ranking, a recently developed ensemble feature selection technique is applied for distinguishing alcoholic from non-alcoholic fatty liver disease from breath gas analysis. Both diseases are generally difficult to distinguish but require different treatment. Breath gas samples have been analyzed by mass-spectrometry. The feature selection approach yielded some breath gas maker candidates well corresponding to domain knowledge as well as several novel findings.

The following two chapters are dedicated to unsupervised data mining for investigating the very complex mechanisms in certain diseases: breast cancer and somatoform pain disorder. Chapter 12 by F. Javier Lopez et al. is dedicated to fuzzy association rule mining from breast cancer genomic data. For breast cancer some well studied major prognostic factors exist, including e.g. the size and the histological grade of the tumor. The authors integrate the prognostic factors with whole genome microarray data to study the associations between these two types of data. An analysis on 2,751 patients revealed several interesting rules partly confirming and partly extending previous knowledge.

Bianca Wackersreuther et al. present in Chapter 13 a graph-mining approach to detect co-activated networks in the human brain. Different anatomical regions of the human brain form distinct functional networks fulfilling certain tasks. Functional magnetic resonance images from a study on somatoform pain disorder have been studied. After modeling the brain of each subject as a graph, the proposed algorithm searches for so called motifs, which correspond to frequently occurring sub-graphs. The evaluation demonstrates that certain motifs are much more prevalent in patients than in healthy controls which supports the hypothesis that somatoform pain disorder manifests itself in altered brain function.

Chapters 14 and 15 focus on supervised mining medical data in two scenarios for highest importance to the health care system especially considering the demographic shift. Chapter 14 by Marie Persson et al. studies the interesting question whether it is possible to predict the need of surgery of a patient send to hospital from the referral documents issued by the general practitioner. Sufficiently accurate predictions would allow estimating the surgery demand for different departments of the hospital and thus more

efficient resource planning for the hospital and less waiting time for the patients. On a sample data set from Blekinge hospital in Sweden promising results have been obtained with different classifiers.

Philipp Kranen et al. present in Chapter 15 an adaptive multi-step approach for scalable emergency detection from remote health monitoring data. To minimize communication costs and energy consumption of the mobile devices, a technique combining several index-based Bayes-tree classifiers is presented: Potentially critical events are identified by an efficient pre-classifier. Only for critical events, detailed sensor data are sent to a more powerful classifier at the next layer. The experimental evaluation demonstrates that this approach is suitable for multi-step anytime classification of huge amounts of streaming data.