An introduction into Data Mining in Bioinformatics.

**Introduction**  
Over recent years the studies in proteomic, genomics and various other biological researches has generated an increasingly large amount of biological data. Drawing conclusions from this data requires sophisticated computational analysis in order to interpret the data. One of the most active areas of inferring structure and principles of biological datasets is the use of data mining to solve biological problems. Some typical examples of biological analysis performed by data mining involve protein structure prediction, gene classification, analysis of mutations in cancer and gene expressions. As biological data and research become ever more vast, it is important that the application of data mining progresses in order to continue the development of an active area of research within bioinformatics. This essay aims to draw information from varied academic sources in order to discuss an overview of data mining, bioinformatics, the application of data mining in bioinformatics and a conclusive summary.

**Data mining**

Data mining is the method extracting information for the use of learning patterns and models from large extensive datasets. Data mining itself involves the uses of machine learning, statistics, artificial intelligence, database sets, pattern recognition and visualisation (Li, 2011). Often referred to as Knowledge Discovery in Databases (KDD) or Intelligent Data Analysis (IDA) (Raza, n.d.), the data mining process is not just limited to bioinformatics and is used in many differing industries to provide data intelligence. The application of data mining and machine learning models can involve varied systems, Kononenko and Kukar (2013) identify

*“Machine learning systems may be rules, functions, relations, equation systems, probability distributions and other knowledge representations.”*

This intelligence or knowledge discovery gained from data mining has a vast amount of aims, including the likes of forecasting, validation, diagnosis and simulations (Guillet, 2007). Typically the process for knowledge discovery (see Figure 1) through databases includes the storing and processing of data, application of algorithms, visualisation/interpretation of results (Kononenko and Kukar, 2013)



Figure 1: Process of Knowledge Discovery through Data Mining

It’s important to state that the process of data mining or KDD encompasses a multitude of techniques, such as machine learning. As a result the process of data mining includes many steps needed to be repeated and refined in order to provide accuracy and solutions within data analysis, meaning there is currently no standard framework of carrying out data mining.

Figure 2: Phases of CRISP-DM Process Model for Data Mining

However, CRISP-DM (Cross Industry Standard Process for Data Mining), defines one standard framework for the process of data mining across multiple industries containing phases, generic tasks, specialised tasks, and process instances (Chalaris et al., 2014) (see figure 2). Jain (2012) discusses that the main tasks for data mining are:  
1. Classification: Classifies a data item to a predefined class  
2. Estimation: Determining a value for unknown continuous variables  
3. Prediction: Records classified according to estimated future behaviour  
4. Association: Defining items that are together  
5. Clustering: Defining a population into subgroups or clusters  
6. Description & Visualisation: Representing data  
Typically speaking, this process and the definition of Data Mining defines the extraction of knowledge. Where we define machine learning within data mining is the automatic data mining methods used, Kononenko and Kukar (2013) state that

*“Machine Learning cannot be seen as a true subset of data mining, as it also compasses the other fields, not utilised for data mining”*

Following this, knowledge is gained through the use of differing machine learning methods used include: classification, regression, clustering, learning of associations, logical relations and equations (Kononenko and Kukar, 2013) (see figure 3)

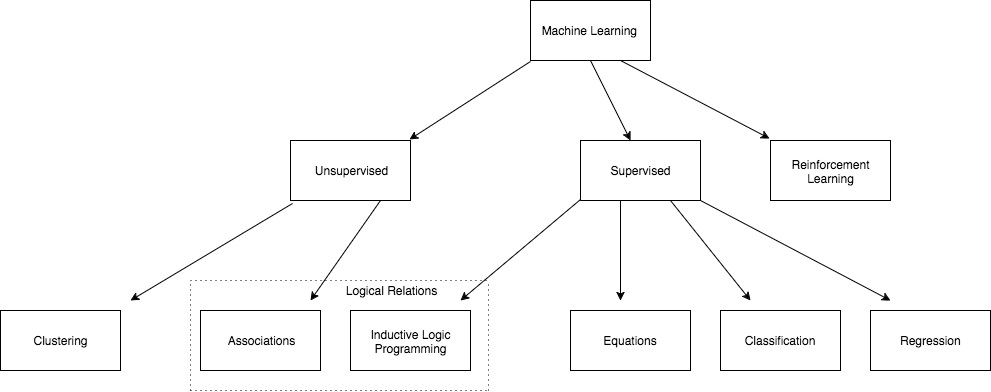


Figure 3: Methods of Machine Learning

As seen in Figure 3, Machine learning can be catergorised into unsupervised or supervised learning models. Unsupervised learning models involve data mining algorithms identifying patterns and structures within the variables of a data set, i.e clustering (Larose and Larose, 2014). Supervised learning defines where the variable is specified or provided in order for the algorithms to predict based off of these, i.e regression (Larose and Larose, 2014).

**Bioinformatics**

Bioinformatics deals with the storage, gathering, simulation and analysis of biological data for the use of informatic tools such as data mining. Zaki, Karypis and Yang (p. 1, 2007) discuss informatics as being the handling science of biological data involving the likes of sequences, molecules, gene expressions and pathways. Additionally Fogel, Corne and Pan (2008), define bioinformatics as:

*“Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store , organise, archive analyse, or visualise such data.”*

It’s also important to state that bioinformatics is also broadly speaking, the research of life itself. As Tramontano (2007), defines

*“…we could define bioinformatics as the science that analyzes biological data with computer tools in order to formulate hypotheses on the processes underlying life”*

Over resent years the development of technology both computationally, medically and within biology has allowed for data to be developed and accumulated at an extrodonary rate, and thus the interpritation of this information has rapidly grown (Ramsden, 2015). This highly interdisiplinary field, encompasses many differenciating subfields of study; Ramsden, (2015) specifies that DNA squencies is one of the most widely researched areas of analysis in bioinformatics. As a general rule, bioinformatic data is often divided into three main categories, these being: sequence data, structural data and functional data (Tramontano, 2007). Moreover, this data contains differing biological entities, genes or proteins, which means that whilst knowledge discorvery is a large part of bioinformatics, data management is also a primary concern (Chen, 2014)

**Application of Data Mining in Bioinformatics**

As discussed bioinformatics is an increasingly data rich industry and thus using data mining techniques helps to propose proactive research within specific fields of the biomedical industry. Additionally this allows for researchers to develop a better understanding of biological mechanisms in order to discover new treatments within healthcare and knowledge of life. In recent years the computational process of discovering predictions, patterns and defining hypothesis from bioinformatics research has vastly grown (Fogel, Corne and Pan, 2008). Raza (2010), explains that data mining within bioinformatics has an abundance of applications including that of “*gene finding, protein function domain detection, function motif detection and protein function inference”.*Data banks such as the Protein Data Bank (PDB) have millions of records of varied bioinformatics, for example PDB has 12823 positions of each atom in a known protein (RCSB Protein Data Bank, 2017). The methods of clustering, classification, association rules and the likes discussed previously are applied to this data in order to predict sequence outputs and create a hypothesis based on the results. Though these results may not be exact, as that would require a physical model, the application of data mining allows for a faster result.

**Conclusion**

The extensively vast science of data mining within the domain of bioinformatics is a seemly ideal fit due to the ever growing and developing scope of biological data. As this area of research is so extensive it is apparent that attributes of biological databases propose a large amount of challenges. Improving the quality and the accuracy of conclusions drawn from data mining is ever more key due to these challenges. As a result it is important for the future directions of research to adapt for the integration of new bioinformatics databases in order to provide more methods of effective research.