# Artificial Intelligent and Machine Learning Methods in Bioinformatics and Medical Informatics

Noor A. Jebril and Qasem Abu Al-Haija

**Abstract**

Recently, the machine learning techniques have been widely adopted in the field of bioinformatics and medical informatics. Generally, the main purpose of machine learning is to develop algorithms that can learn and improve over time and can be utilized for predictions in hindcast and forecast applications. Computational intelligence has been significantly employed to develop optimization and prediction solutions for several bioinformatics and medical informatics techniques in which it utilized various computational methodologies to address complex real-world problems and promises to enable computers to help humans in analyzing large complex data sets. Its approaches have been widely applied in biomedical fields, and there are many applications that use the machine learning, such as genomics, proteomics, systems biology, evolution and text mining, which are also discussed. In this chapter, we provide a comprehensive study of the use of artificial intelligent and machine learning methods in bioinformatics and medical informatics, including AI and its learning processes, machine learning and its applications for health informatics, text mining methods, and many other related topics.

**Glossary**

| | |
|---|---|
| AI | Artificial intelligence. |
| IT | Information technology. |
| ML | Machine learning. |
| NN | Neural networks. |
| EHR | Electronic health records. |
| ANI | Artificial narrow intelligence. |
| AGI | Artificial general intelligence. |
| NLP | Natural language processing. |
| SR | Speech recognition. |
| ES | Expert systems. |
| AI-R | Artificial intelligence for robotics. |
| PML | Probabilistic machine learning. |
| MLD | Machine learning data. |
| aML | automated Machine Learning. |
| iML | interactive Machine learning. |
| HCI | Human–computer interaction. |
| KDD | Knowledge discovery/data. |
| RNA | Ribonucleic acid. |
| DNA | Microarray deoxyribonucleic acid. |
| cDNA | complementary Microarray deoxyribonucleic acid. |
| mRNA | messenger ribonucleic acid. |
| MIAME | Minimum information about a microarray experiment. |
| FGED | Functional Genomics Data Society. |
| FRG/BKG | Foreground/background. |
| MS | Mass spectrometry. |
| SVM | Support vector machine. |
| RBF | Radial basis function. |
| CART | Classification and regression tree. |
| OOB | Out of the bag. |
| TM | Text mining. |
| IR | Information retrieval. |
| DC | Document classification. |
| NER/NEN | Named entity recognition/normalization. |
| ART | Adaptive resonance theory. |
| DNN | Deep neural network. |

N. A. Jebril
Department of Computer Sciences, King Faisal University, Hufof, Al-Ahsa, Saudi Arabia

Q. Abu Al-Haija (✉)
Department of Computer Information and Systems Engineering, Tennessee State University, Nashville, TN, USA
e-mail: qasem.abualhaija@uop.edu.jo



J. Aljaâ am et al. (eds.), *Emerging Technologies in Biomedical Engineering and Sustainable TeleMedicine*, Advances in Science, Technology & Innovation, https://doi.org/10.1007/978-3-030-14647-4_2

## 1 Introduction

Due to the rapid advances in the digital computing technology where computers become core parts of any modern industry and automated applications, it becomes a necessity to involve the machine with latest trend in the current industry. This in turn led to moving ahead toward the smart industry by adopting intelligent machines and computational systems that can replace or reduce the human intervention at several points of execution or production. Thus, the *artificial intelligence (AI)* can be interpreted naturally as the work of a machine that a human could have done using his intelligence [1, 2]. Currently, the artificial intelligence is gaining an appreciated amount of interest as almost all major IT companies are spending millions on the development and implementation of the artificial intelligence considering its criticality to their future situation. Providing personal relationships with machines is the latest trend in product-based industries and is believed to be booming more recently [3].

Artificial intelligence (AI) occupies significant part of computer science applications that use algorithms, inference, pattern matching, grammar, deep learning and cognitive computing for approximate results without direct human input. Even though the artificial intelligence can determine meaningful relationships in raw data, however, the use of artificial intelligence creates a non-easy challenge for researchers which can face complex troubles that are hard to fix or almost impossible to fix. Nevertheless, it can be used to support the diagnosis, treatment and prediction of outcomes in many medical situations which increased the possibility for artificial intelligence to be applied in almost every scoop of medical fields, including drug expansion, patient observation and personalized patient processing plans.

Artificial intelligence can be designed in a way that mimics the task of *neural networks (NN)* in the human brain. This was made possible by utilizing multiple layers of nonlinear processing units to "teach" itself how to comprehend data/record classification or make predictions. Thus, artificial intelligence can synthesize electronic health records (EHR) data and unstructured data to build predictions about patient health. For instance, artificial intelligence software can speedily read a retinal image or flag cases for follow-up when multiple manual reviews would be too heavy. Doctors benefit from having more time and concise data to make better patient decisions. Artificial intelligence solutions for healthcare currently run toward enhancing outcomes of patients and decreasing healthcare budgets. Figure 1 illustrates the amount of market spending on AI application in the healthcare sector during the period from 2013 to the mid of 2018 [4]. According to the figure, it can be clearly seen that the healthcare AI is increasingly gaining a significant funding as it accumulated around 4300 million USD from 2013 to 2018 through 576 deals which pushed it on top of among AI-based industries.

The adoption of artificial intelligence in medicine started early since 1972, where the researchers at Stanford University developed the MYCIN program, which is an early expert system, or artificial intelligence (AI) program, for treating blood infections. This project has been revisited again in the 1980s, where people at Stanford University resumed the work with artificial intelligence for medical care via developing newer version of their medical expert system via the Stanford University Medical Experimental Computer —Artificial Intelligence in Medicine (SUMEX-AIM) project. Indeed, AI has acquired its importance as "*the next big thing*" for decades but its widespread practical uses have only begun to take off in the 2000s; for instance, AI has attracted more than $17 billion in investments since 2009 and is likely to reach $36.8 billion by 2025 [5].

AI neural networks have been used efficiently to handle raw data blocks and learn how to organize those data using the most important variables in predicting health outcomes. Today, artificial intelligence techniques such as IBM Watson are used at Memorial Sloan-Kettering Cancer Center to support diagnosis and create management plans for oncological patients [5]. Watson is achieving these plans by collecting millions of medical reports, patient records, clinical trials and medical journals. Watson's findings routinely refer to "medical patients" in certain cases [5]. IBM also joined CVS Health in the treatment of chronic diseases using artificial intelligence technology. Johnson & Johnson and IBM AI is used to analyze scientific research to find new links to drug development [5].

Other examples of AI currently used in medicine include patient care in radiology [5]. AI can quickly search and interpret billions of data points—text or images—inside the patient's electronic medical record. It can be done using other patient-like conditions and through the latest medical research. In genomics [5], AI can extract unstructured data from peer-reviewed literature to continuously improve its knowledge base. It gives various information and modern clinical content—based on the latest approved handling options, including targeted options, immune cells, occupational guidance and clinical trial options based on biological indicators, genome databases and related publications.

In a more realistic way, the best way of achieving AI in medical care applications is by adoption of the *machine learning,* which is a method of data analysis that automates analytical model building. Machine learning (ML) is an artificial intelligence field in which the system depends on learning from the data, identifying patterns, taking the decisions with minimal involvement of human and then determining or predicting a set of events. Recently, machine learning (ML) research focuses more on the selection or


**Fig. 1** Funding of artificial intelligence of healthcare hit a historic high in second quarter of 2018 (Q2'18), disclosed equity funding, Q1'13–Q2'18

improvements of algorithms that learn how to make data-based predictions and implement experiments on the basis of such algorithms [6] which resulted in several emerging applications in the field of bioinformatics.

*Bioinformatics* deals with computational and mathematical methods for understanding and manipulating biological data [7]. Prior to the development of machine learning algorithms, bioinformatics algorithms had to be programmed by hand, which, for problems such as protein structure prediction, proved to be very difficult [8]. The recent machine learning techniques such as deep learning algorithms employ learning automatic features, which means that based on the data set alone, the ability of the algorithm is to learn how to integrate multiple features of input data into a more abstract set of features that are further learned [9]. This multilayered approach to learning patterns in input data allows these systems to make fully complex predictions when trained on large data sets. In recent years, the size and number of available biological data sets has increased, enabling biotechnological researchers to benefit from these automated learning systems [9].

## 2 Artificial Intelligence and Machine Learning

*Artificial intelligence* (AI), the term first formulated by John McCarthy, is the field of science that concerned with the development of computer systems to accomplish missions that would need human intelligence. It includes things like planning, understanding language, recognizing objects and sounds, learning and problem-solving [10]. The technology of artificial intelligence passes through three levels of integration: artificial narrow intelligence (ANI), artificial general intelligence (AGI) and artificial super intelligence (ASI). The emergence of these levels is illustrated in Fig. 2. ANI is the first level that can decide only in one field; AGI is the second level and it can reach the intelligence level of a human, i.e., it can imagine, plan, solve problems, think abstractly, understand the complicated ideas, learn quickly and learn from experience [11]; and the last level is ASI. It is much smarter than the best human brain in practically every area, including scientific creativity, general wisdom and social skills [12].

The artificial intelligence has been used efficiently to develop solutions for wide range of applications, especially those based on human thinking to provide a proper decision. AI is contributing to several fields with many significant relation and commonalities between them. The most common fields AI being used nowadays are shown in Fig. 3, including the machine learning (ML) approaches, natural language processing (NLP) applications, speech recognition (SR) techniques, expert systems (ES), AI for robotics (AI-R), which provides solutions for optimization and planning as well as scheduling problems, and also provide artificial algorithms for machine vision applications.

In addition, the artificial intelligence has been employed for the last couple of years to develop significant solutions in even more sensitive situations such as the medical care applications where it can develop the course of care for patients with chronic diseases and suggests accurate treatments for complex diseases and improves adherence to substances in clinical trials [5]. Artificial intelligence can be used in a variety of ways in medicine, and the following are four examples:

- Clinical data annotated: Almost 80% of healthcare data is unorganized, and AI can read and understand unorganized data. AI's ability to address natural language allows for clinical text to be read from any source, medical, social concepts, classification and coding.

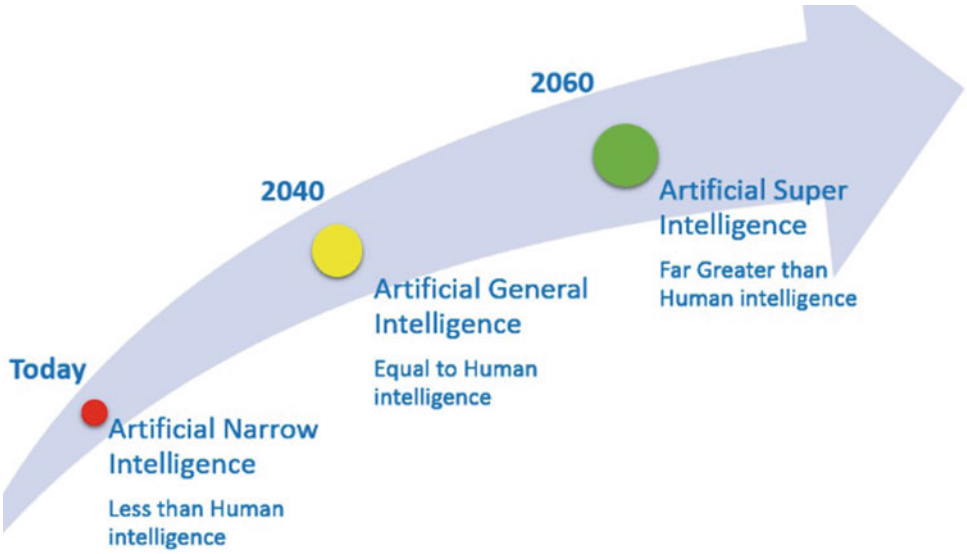



**Fig. 2** Future evolution of artificial intelligence

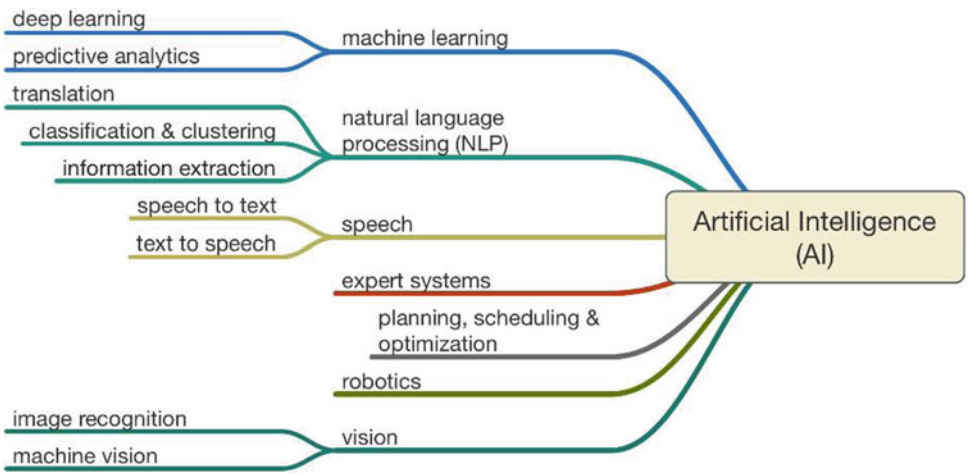



**Fig. 3** Fields of artificial intelligence

- Visions for patient data: Artificial intelligence can identify problems in historical medical records of patients—whether in structured or unstructured text. It summarizes the history of their care about these problems and can provide a concise summary of patient records.
- Patient similarity: AI can determine the clinical similarity measure among patients. This allows researchers to create dynamic patient groups rather than fixed patient groups. It also allows understanding what the path of care works best for a group of patients.
- Medical visions: Using AI techniques, researchers can find information in uneducated medical literature to support hypotheses—assist to find new insights. AI can access a full area of medical literature, such as Medellin, and identify documents that are medically linked to any combination of medical concepts.

In more realistic situations, the AI for machine learning field provided more precise methods for medical applications that are heavily based on the data analysis that automates analytical model building. The applicable machine learning algorithms fall into four main categories that are briefly illustrated in Fig. 4. The techniques used in these categories differ mainly in the learning process rule which defines the mathematical models of updating the weights and bias levels of the neural network when a network is simulated in a specific data environment.

## 2.1 Supervised Learning

In this technique, the learning process depends on the comparison between the calculated outputs and the desired (i.e., expected) outputs since the desired output is already

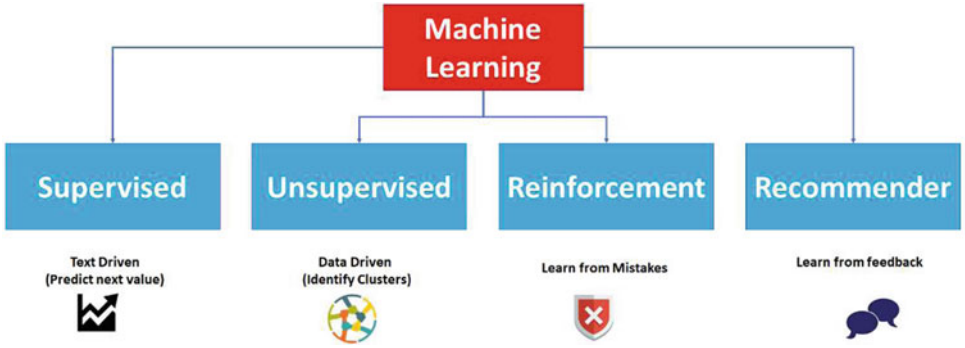



**Fig. 4** Types of machine learning

known prior to the initialization of learning process, i.e., learning refers to the calculation of the error and then adjusts the error to achieve the expected output. Supervised learning is the main methodology in machine learning and it also has central importance in the processing of multimedia data [13].

### 2.2 Unsupervised Learning

In this technique, the desired output is not known before the starting of the learning process, i.e., unsupervised learning, and thus the network learns by its own and that is done by discovering and adopting based on the input pattern. If the data are split into various clusters and then the learning is called a clustering algorithm such as Google News (*news.google.com*), which is a well-known example where clustering is used and where in Google News, the network groups new stories on the web and puts them into collective news stories [6].

### 2.3 Reinforcement Learning

In this technique, the reinforcement learning process depends on the output with how an agent will take behavior in an environment to maximize some notion of long-term reward. A reward is given for the correct output and a penalty is issued for the incorrect output. Reinforcement learning is different from the supervised learning problem in that correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected [6].

### 2.4 Recommender Systems

In this technique, the recommender systems learning process depends on the online users who can customize their sites to check customer desires. There are mainly two approaches: content-based recommendation and collaborative recommendation, which assist the user for getting and mining data, producing intelligent and novel recommendations, ethics. Most e-commerce site uses this system [14].

## 3 Machine Learning for Health Informatics

The area of *machine learning* (ML) has emerged very fast as a development technical domain at the interchange of informatics and statistics [15, 16] closely associated with data science and knowledge discovery, especially the *probabilistic machine learning* (PML) that is highly helpful for health informatics where most problems include treatment with uncertainty. The theoretical fundamental of probabilistic machine learning [17, 18] was first initiated by the mathematician, Thomas Bayes (1701–1761), and since then, probabilistic inference has widely affected the artificial intelligence and statistical learning methods while the inverse probability allowed to conclude unknowns, learn from data and make predictions [19, 20].

The recent advances in machine learning have been driven by the expansion of new learning algorithms and theory that can accommodate the newer technologies raised from the constant collapse of data, meanwhile, reduced computation. The adoption of data-intensive *machine learning data algorithms (MLD)* can be found in all areas of health informatics employment and is specifically useful for brain information from the basic research of intelligence understanding [21] to a more complex domain of certain brain informatics research [22]. Implementation of machine learning methods in biomedical and health has enabled to drive to more evidence-based decisions making and assist in going to personal medicine [23]. Indeed, the realization of any scientific area can be discussed by the questions it studies, for instance, the machine learning area trying to answer the question: "*How can we construct algorithms that automatically get better through experience? What are the primary laws that control all learning processes?*" [24].

The challenge in almost all machine learning techniques is to find the related structural patterns/temporal patterns (i.e., knowledge) in such data, which are often hidden and

unavailable to the human expert. However, the problem is that majority of biomedical data sets are poorly structured and unmanaged [25], with most data are in dimensions well above three, and although human experts are excellent at pattern recognition for dimensions <= 3, like data build manual analysis is often impossible. Therefore, most fellows from the machine learning community are focusing on *automated machine learning* (aML) with the great goal of getting humans out of the loop and a realistic example of best practice can exist in autonomous vehicles [26]. Nevertheless, biomedical data sets are full of uncertainty and incompleteness [27] and may include missing data, noisy data, dirty data, undesirable data, and most of all, some medical problems are difficult, which makes it difficult to implement the complete automated approaches.

Another noticeable issue is the complexity of advanced machine learning algorithms which has been arrested by non-experts from the application of these solutions. Thus, the integration of knowledge from the area expert and the interaction of the area expert with the data can be greatly enhanced by the strengthening of pipeline knowledge discovery process. Hence, the *interactive machine learning* (iML) places the "human-in-the-loop" to authorize what a human nor a computer could not do on their own. This concept is supported by a synergistic combination of two field methodologies that provide ideal conditions for solving such problems: Human–computer interaction (HCI) and knowledge discovery/data mining (KDD), in order to support human intelligence with machine intelligence to find new insights unknown in data HCI-KDD approach [28].

The enormous growth of the amount of available biological data raises two problems: the first is the efficient storage and management of information, and the second problem is the extraction of useful information from such data. The second trouble is one of the major challenges in computer biology, which demand the development of tools and methods that can transform all these heterogeneous data into biological information about the underlying mechanism. These tools and methods must enable us to go further than describing data and providing knowledge in the form of testable models to be able to gain system predictions.

There are many biological fields where machine learning techniques can be implemented to extract knowledge from data. Figure 5 displays a diagram for the major biological problems in which arithmetic methods are applied. The problems are categorized into six different areas: genomics, proteomics, microarrays, systems biology, evolution and text mining. The category named "Other application groups" combines the remaining issues. These categories should be understood in a very general way, particularly, genomics and proteomics, which in this review are considered a study of nucleotide chains and proteins, respectively.

## 3.1 Genomics

*Genomics* is one of the most substantial fields in bioinformatics. According to Fig. 6, the number of available sequences is steadily growing where these data must be processed in order to gain useful information. Using the genome sequences, we can extract the location and structure of the genes [29] and recently, regulatory elements [30–32] and irregular ribosomal genes [33] have been specified from an accounting point of view. Sequencing information is also used in gene function and prediction of secondary ribonucleic acid (RNA) structure. If genes include information, proteins are the factor that converts this information into life. Proteins play a very important part in the life process, and their three-dimensional (3D) structure is a major feature of their functions.

In the area of proteomic, the major application of arithmetic methods is to predict protein structure. *Proteins* are massive complex molecules containing thousands of atoms and boundaries. Thus, the number of potential structures is huge. This makes predicting of protein structure a very complicated fusion problem where optimization techniques are required. In proteomics, as in genomics, machine learning techniques are implemented to predict protein function.

## 3.2 Microarray

Another interesting application of computational methods in biology is complicated, experimental data management. *Microarray* articles are the best known (but not the only) domain where this type of data is collected. Microarrays can be utilized to define gene expression patterns in a specific cell or tissue. *Microarray deoxyribonucleic acid* (DNA) is a set of microscopic DNA sequences (oligos) connected to a solid surface. These sequences appear as part of a large library of genes in the cell [34].

The states of the gene as shown in Fig. 7 can be classified according to these points [34]:

- If the gene is active within the cell, then the complementary DNA (cDNA) (resulting from the transcript of the messenger RNA (mRNA)) will bind to its complementary oligo.
- If the cDNA has been characterized by fluoridation, it can determine the complementary oligo (along with the gene it represents).
- If cDNA is classified from healthy and diseased cells with different fluorophores, comparisons of gene expression can be made.
- Only active genes in a diseased or natural state will be of particular interest to scientists.

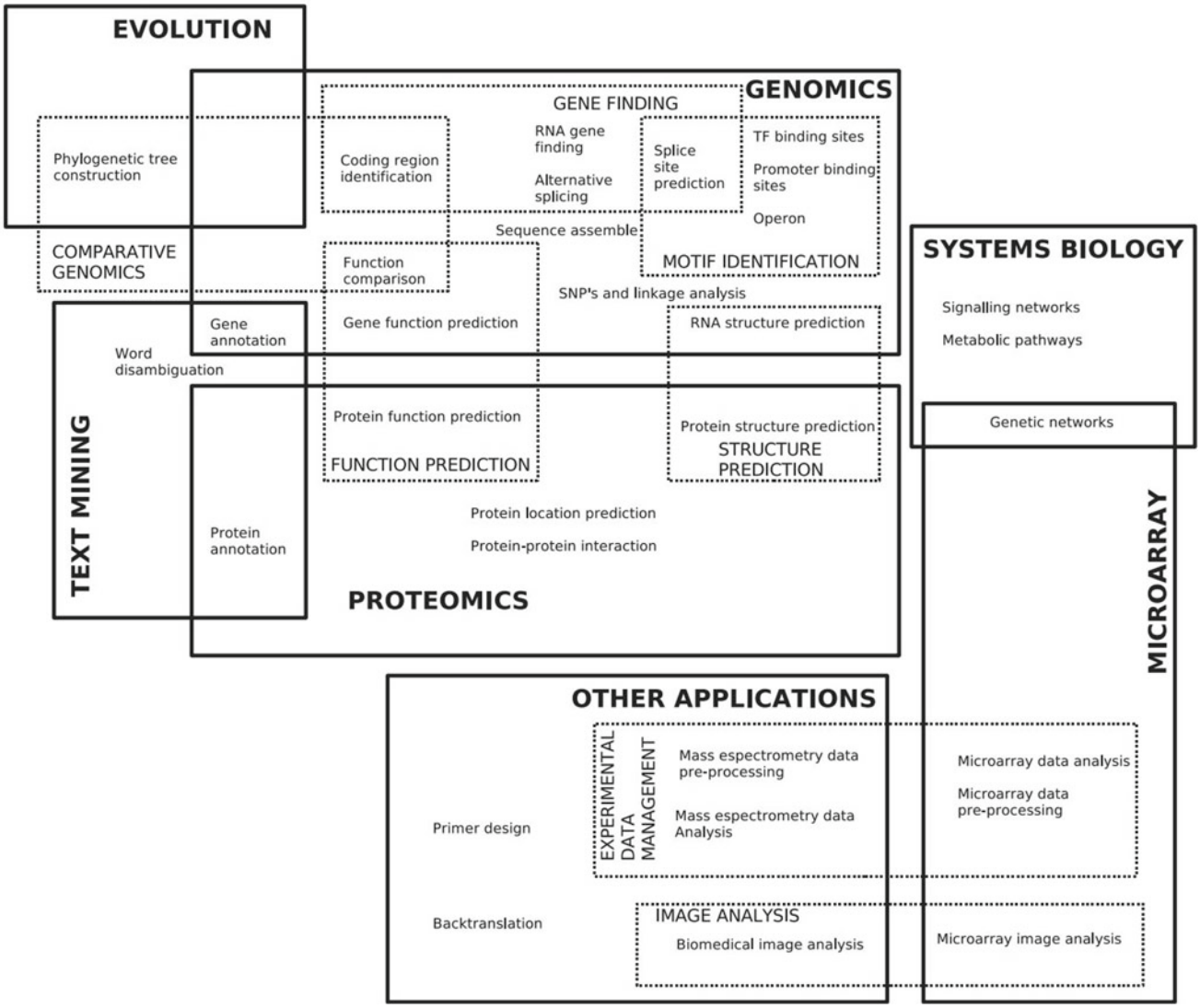



**Fig. 5** Classification of the topics where machine learning methods are applied

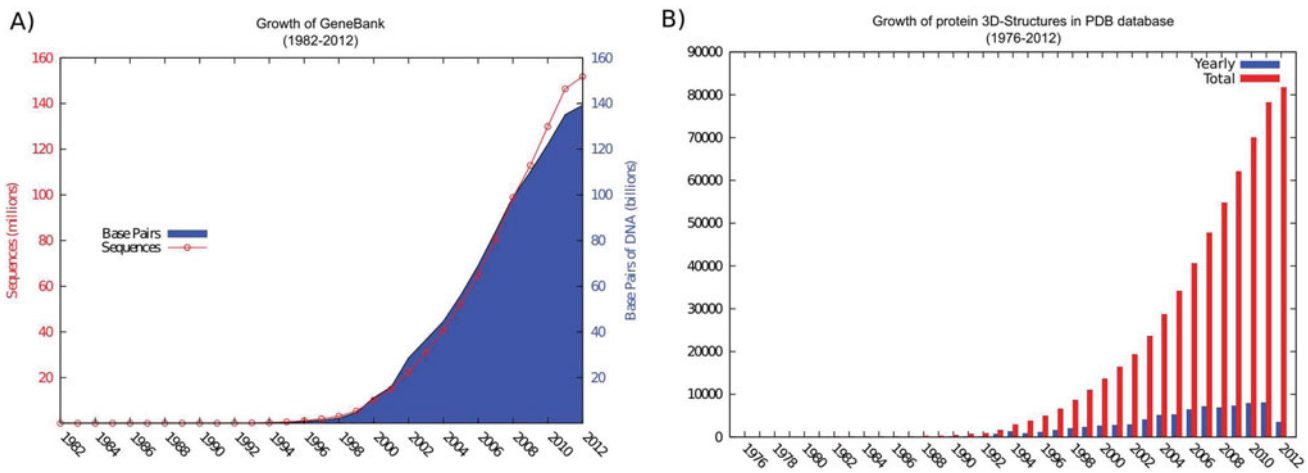



**Fig. 6** Growth of public databases: **A** Evolution of the GenBank database size and **B** 3D structures of proteins

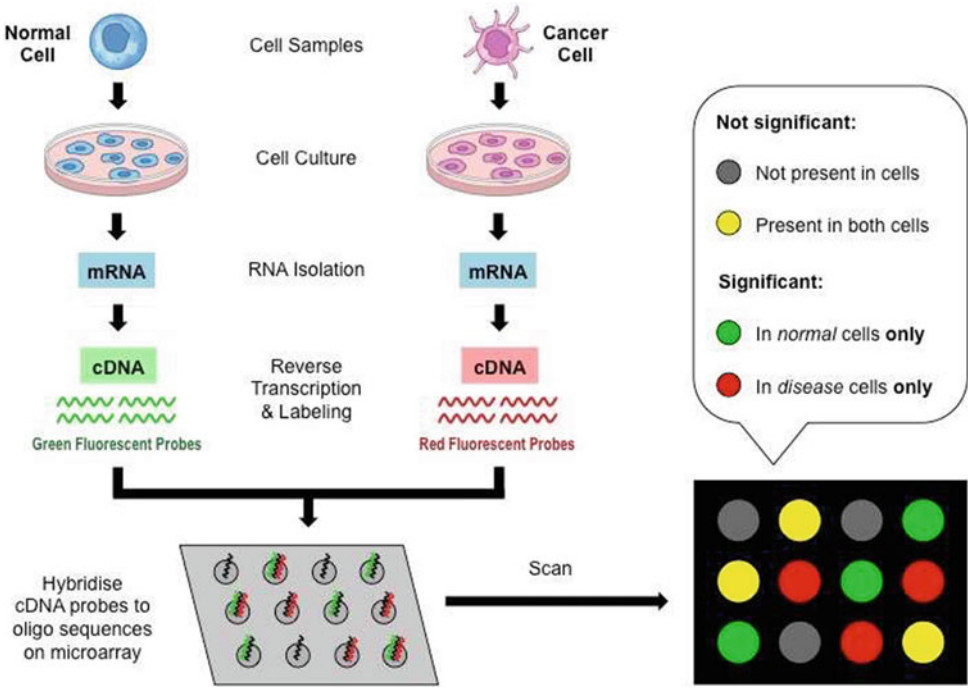



**Fig. 7** DNA microarray overview [34]

Complicated empirical data raise two different troubles. First, data must be processed in advance, i.e., modified data to be used appropriately by machine learning algorithms. Second is data analysis, which relies on what we are looking for. In the case of microarray data, the most popular application is to determine the pattern of expression, classification and induction of the genetic network. Microarray framework for data management is shown in Fig. 8. Generally, the objective of microarray image analysis is to extract density descriptors from each point representing the levels of gene expression and input properties for the upcoming analysis and then the biological results are plotted using the data mining and statistical results of all extracted features. The DNA microarray image analysis components include grid alignment problem, foreground separation, quality assurance, quantification and normalization [35].

In addition, management of data should satisfy to a *minimum information about a microarray experiment* (MIAME), which is a standard created by the Functional Genomics Data Society (FGED) for reporting microarray experiments [36]. It is prepared to assign all the information needed to unambiguously understand the results of the experiment and to potentially reproduce the experiment. While the standard determines the content required for compliant reports, it does not specify the format in which this data should be presented; there are several file formats for representing this data, both public and subscription-based repositories for such experiments [37].

In short, the process in Fig. 8 can be summarized as follows: it starts by the laser scanning of the images (data) which generates accurate 2D microarray-DNA images. Then, the outcome is automatically derived from the data (a machine learning perspective) to subsequent model fitting, and the alignment of the microarray grid produces a set of cut-off rows at each point that realize a correct introduction as well as the separation of the foreground, which are the main processing steps of DNA microarray images that impact the quality of the gene expression information, and therefore affect the confidence in any biological conclusions derived from the data. Thus, understanding the microarray data processing steps becomes crucial for optimal data analysis in the microarray. Then the unreliable microarray cells are eliminated. Finally, an image of the average sample values in each grid cell is extracted using a special mask and colored in a red, green and blue space with the color assigned to each cluster/pixel. The statistics for each cluster can be found in the text area. Finally, the output of all steps is the statistical behavior of measurements and thus the test of the hypotheses or knowledge must be performed automatically from the data (machine learning perspective) for subsequent model fitting and draw biologically meaningful conclusions [35].

The microarray data processing framework contains issues related to data management (MIAME-compliant database) [38]; however, the main concern is of mechanism by which all stains should be reliably determined without any human intervention dependent on human preparation once. One-time setup is to incorporate any prior knowledge of microarray image layout into network alignment algorithms in order to reduce their parameter search

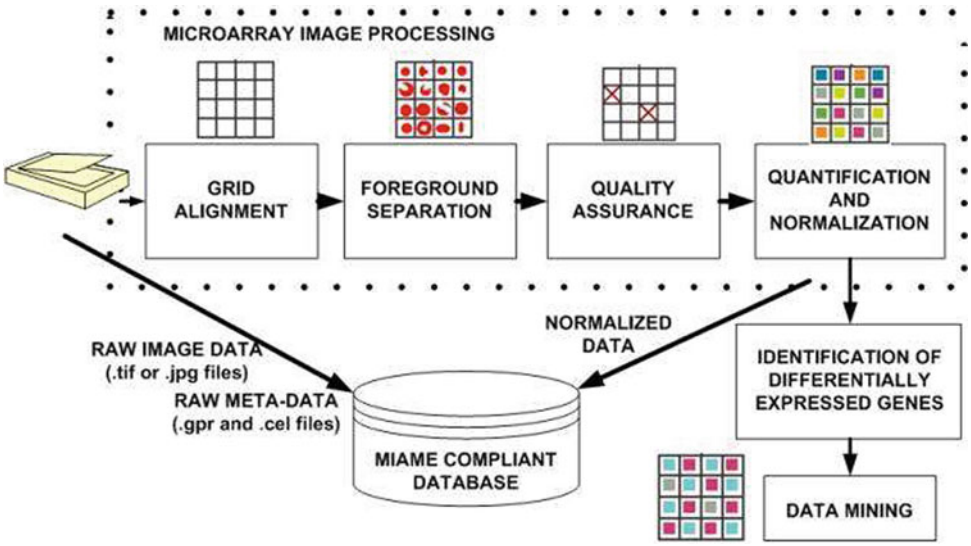



**Fig. 8** Microarray data processing workflow [35]

space. This method is usually data-driven and has multiple internal arithmetic coefficients improved in the parameter search area to compensate for differences in microarray image [35].

## Grid Line Algorithm

The grid line algorithm [39] is the first step of microarray data processing workflow. It is based on data and it aims to find a set of mutually orthogonal lines intersecting at the center of each grid cell obtained from DNA microarray scanners. The points generally have two states: the varying radii and their position deviates from a perfect grid placement. The grid cell is defined as a one-point area that is part of a set of two-dimensional arrays of points.

## Foreground Separation Algorithm

The foreground separation algorithm [35] is the second step of microarray data processing workflow. It can be implemented by several methods, including foreground separation using spatial templates [40], foreground separation using intensity-based clustering [41], foreground using intensity-based segmentation [42], foreground separation using spatial and intensity information (hybrid methods) [43], and foreground separation from multichannel microarray images [35].

## Quality Assurance Process

The quality assurance process is the third step of microarray data processing workflow where the spot quality assessment is implemented to identify grid cells that include valid spots and to eliminate invalid spots from further analysis. It is also used to determine the invalid or defective spots, for example, as deviations from the "ideal" microarray image and the deviation threshold values separating valid and invalid spot categories. Indeed, there are two types of background variation criteria: local and global background variability. The local metrics can detect the existence of noise in a grid cell while global standards provide indicators of differences across a complete microarray slide. The quality metric $q$ for background according to the designed formulas at [44]:

$$q_{BKG}^{LOC\&GLOB1} = \frac{\mu_{BKG}^{GLOBAL}}{\mu_{BKG}^{LOCAL} + \mu_{BKG}^{GLOBAL}}; q_{BKG}^{LOC\&GLOB2} = \frac{m_{BKG}^{GLOBAL}}{m_{BKG}^{LOCAL} + m_{BKG}^{GLOBAL}} \quad (1)$$

where $q$ the quality is metric, $m$ is a median and $\mu$ is a mean. The notations, $FRG$ refers to foreground and $BKG$ to background.

## Quantification and Normalization

After a set of valid spots and two sets of image pixels labeled as foreground and background each spot is calculated, and the descriptors are extracted to evaluate the gene regulation in a process called *data quantification* [45] (also called spot feature extraction), where the feature should be directly proportional to the mRNA quantity in the solution that was deposited in a spot and should perform as the deposited gene expression level. During the step of data preparation, the fluorescent intensity measurements are assessed or distorted differently according to some linear or nonlinear functions. Thus, a normalization process for the descriptors of extracted points is desirable.

In general, spot descriptors are split into two classes: the absolute-relative descriptors and the statistical-deterministic descriptors. It is important to understand the experimental structure of the microarray in expression of the outcome of the gene expression. The intensity of the raw microarray

cannot be explained as an absolute measurement due to the random and systematic fluctuations in microarray image data preparation. Thus, in the experiments of cDNA, one is concerned with the statistical difference in the levels of gene expression between the probe and the target (also referred to as the test and reference) which is a hybrid mRNA mixture to the array and library in the array. Based on these considerations, the relative statistics descriptions will focus on the forms of microarray spot descriptors by using ratio or logarithmic ratio provided as follows [46]:

$$des^{X}_{RATIO} = \frac{X^{CHANNEL\,0}_{FRG}}{X^{CHANNEL\,1}_{FRG}} \tag{2}$$

$$des^{X\ WRT\ BKG}_{LOG\ RATIO} = \log_2\left(\frac{X^{CHANNEL\,0}_{FRG} - X^{CHANNEL\,0}_{BKG}}{X^{CHANNEL\,1}_{FRG} - X^{CHANNEL\,1}_{BKG}}\right) \tag{3}$$

where $X$ is the symbol for sample mean or median or mode, the subscripts *FRG* and *BKG* refer to foreground and background, respectively, and the superscript *CHANNEL* refers to red or green microarray laser scans. While Eq. (1) represents a direct ratio of absolute values, Eq. (2) is a logarithmic ratio of relative differences (*XWRTBKG* stands for $X$ with respect to background). Also, the normalization via statistical descriptors is common technique that can be applied by either division or subtraction of statistical descriptors in which Z-transformation would normalize intensities but would not compensate for labeling nonlinearity. This method is modeled in Eq. (4) where $\mu$ is the mean and $\sigma$ is the standard deviation of an entire image [46]:

$$I^{NORM\ STAT}_{Z-TRANSFORM}(row, col) = \frac{I(row, col) - \mu}{\sigma} \tag{4}$$

### 3.3 Proteomics

One of the major aims of late biology is to know the relationships between the context of the structure and function of genomic information. Different *mass spectrometry* (MS) techniques attempt to provide qualitative results that describe relationships. The isotope labeling and fluorescent labeling techniques have been used in the quantitative analysis of proteins. However, researchers are turning to non-discriminatory methods because they are faster and simpler [47–49].

*Peptides* are produced by enzymatic digestion of the protein mixture and then the development of these peptides for training [50]. The label-free method makes use of several peptides to characterize MS tryptic observations to estimate the relative amount of protein [51]. However, the spectral count of the possibility of a peptide can be confused to be observed [52, 53]. A series of studies have found that we can determine the protein based on one note or a few peptides that have been detected preferentially [47]. Some research has also found that different types of peptide likelihood detection may differ from others. Peptide physicochemical properties can affect the discovery of final MS due to many factors such as peptide length, mass, average flexibility indices, net charge and other properties that can affect peptide compliance [54]. This variation must be considered to estimate the quantity; otherwise errors may occur in the assessment of abundance of protein.

To estimate the quantity of protein, we can use the number of peptides detected to indicate abundance, and two classifiers can be used to classify peptides into two types called *proteotypic* and *unposerved*. The peptides are produced through different platforms and must be classified separately. The likelihood of each peptide can be deduced for its original protein and then if we are trying to identify a new protein and quantify it, the likelihood can be very useful for accurate prediction. Some peptides are more readily identified than others that can be observed from the experiments given in Table 1. Thus, the preferentially observed peptides are called as *proteotypic peptides* and therefore, the main two question regarded to proteotypic peptides are: (1) "*What characteristics distinguishing the frequently observed peptides from peptides in the same protein sample but remain unknown?"* and (2) *"What peptide characteristic can be applied to all living organisms?"*. Moreover, if we find the characteristics of the outside, is it possible to predict whether the peptide is *proteotypic* using the protein's sequence. To implement these targets, the *proteotypic* peptides are taken from four different platforms and distinguished with physicochemical properties [55].

To determine the characteristic that controls a peptide's proteotypic inclination, 544 different parameters of the physicochemical properties of amino acids is evaluated, including the water-resistant index, containing hydrophobicity index, residue volume and transfer-free energy to surface [54]. Given that both the total and the average value can contribute to the inclination at the same time, we used a value to describe the peptides, resulting in 1088-dimension property vector in the peptide. There are two methods used in this approach: Support vector machine and random forests. *Support vector machine* (SVM) sets a hyper-plane to classify the given pattern; its basic idea is given in Fig. 9. In addition, this approach used transformation method for getting better performance and kernel functions. The input features vector space can be converted by nonlinear function applications to a high-dimensional space where the best hyper-plane level can be learned which can resolve more complex classification among set that have not been identified at all in the original space [56, 57].

Figure 9 displays two state of vectors that are separated by hyper-plane: the white circle which presents the positive

**Table 1** The genuine and imposter distance distribution [54]

| Peptide example | | K | L | I | G | D | Total | Average |
|---|---|---|---|---|---|---|---|---|
| 1 | Amino acid composition | 0.68 | 0.98 | 1.02 | 0 | 0.76 | 3.44 | 0.688 |
| 2 | Relative mutability | 6.6 | 7.4 | 4.5 | 8.4 | 5.5 | 32.4 | 6.48 |
| 3 | Melting point | 56 | 40 | 96 | 49 | 106 | 347 | 69.4 |
| 4 | Optical rotation | 224 | 337 | 284 | 290 | 270 | 1405 | 281 |
| 5 | Steric parameter | 14.6 | –11 | 12.4 | 0 | 5.05 | 21.05 | 4.21 |

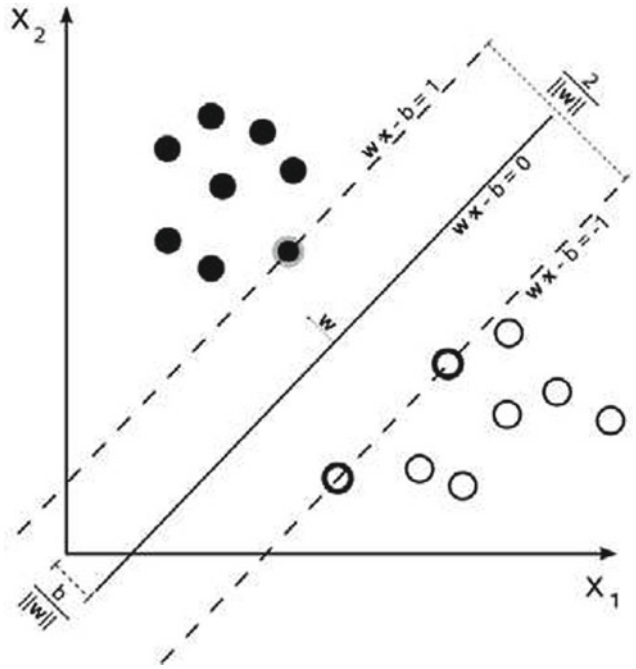



**Fig. 9** Support vector machine

vectors and the black presents the negative vectors. The kernel function is used to avoid the overlap circles that called overfitting, which convert the input space to a higher dimensional space. The kernel function plays an important role in assigning the input space implicitly to a larger dimension space for features, and in this situation, the separation can be better in that the original model will lead to overfitting learning. There are many types of kernel functions [56, 57], such as the polynomial kernel function that is widely used. The function is:

$$K(x_i^T + x_j + 1)^p, \text{ where } p \text{ is a positive constant} \quad (5)$$

Or the Gaussian radial basis function (RBF) kernel given by

$$K(x_i + x_j) = \exp\left(-\gamma ||x_i + x_j||^2\right) \quad (6)$$

Sometimes parameterized using:

$$\gamma = 1/2\sigma^2 \quad (7)$$

where $\sigma > 0$ is a constant that defines the kernel width. Furthermore, there are some other kernel functions that can be used into various states such as the hyperbolic tangent function. Under the use of kernel function, the discriminant function in an SVM classifier is:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \quad (8)$$

where $K(xi, x)$ is the kernel function, are the support vectors specified from the training data, is the class indicator (e.g., +1 and –1 for a two class problem) related with each , $N$ is the number of supporting vectors specified through training process and $b$ is a scalar representing the perpendicular distance of the hyper-plane from origin. The data of the training set are support vectors that are used by SVM classifier to make the decision and predict the label (positive or negative) of samples. In fact, they consist of those training examples that are most difficult to classify.

The features in Table 1 are used to extract different vector data and then generate an input space. These vectors can be described as positive or negative, representing traditional and non-observed peptides. There are two parts of vectors separated by SVM: the first part is applied for training and the second part is applied to predict the performance of the training model. For each peptide in the training set, the vectors have 1088 features as a representation, for example, considering the sample of the amino acid composition of peptides varies from 0 to 4 while the melting point can vary up to several hundred centigrade. As a result of these different ranges, the corresponding features of the rating may dominate the classification and invalidate the effects of other features.

To avoid this problem, each feature vector is normalized to [–1, 1] and after calculating the features, matrix is generated. Suppose: $X = (x_1, x_2, \ldots, x_k)$, where $k$ is the number of features, $X_k = (x_{1k}, x_{2k}, \ldots, x_{nk})$, where $n$ is the number of peptides in training set and $\max(X_k)$ is the maximum value in $x_k$ and $min(x_k)$ is the minimum value in $x_k$. The normalization method is shown as below:

$$x_{(i,k)} = -1 + 2\frac{x_{(i,k)} - \min(x_k)}{\max(x_k) - \min(x_k)} \quad (9)$$

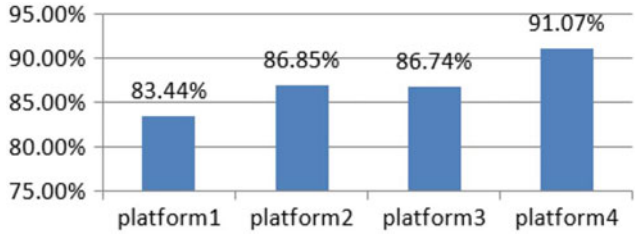



**Fig. 10** The accuracy under the SVM classification [55]

After normalization, we categorized the classification, and prediction accuracy is used to define the performance of the trainer model. The four-accuracy platforms are as given in Fig. 10 which shows that the accuracy of training model has an impressive performance in peptide identification and some additions can improve it to better results: parameter selection. Parameter set is evaluated to get the best classification under the grid search of cost and gamma and thus the cross-validation accuracy for four platforms varies from 85 to 90%.

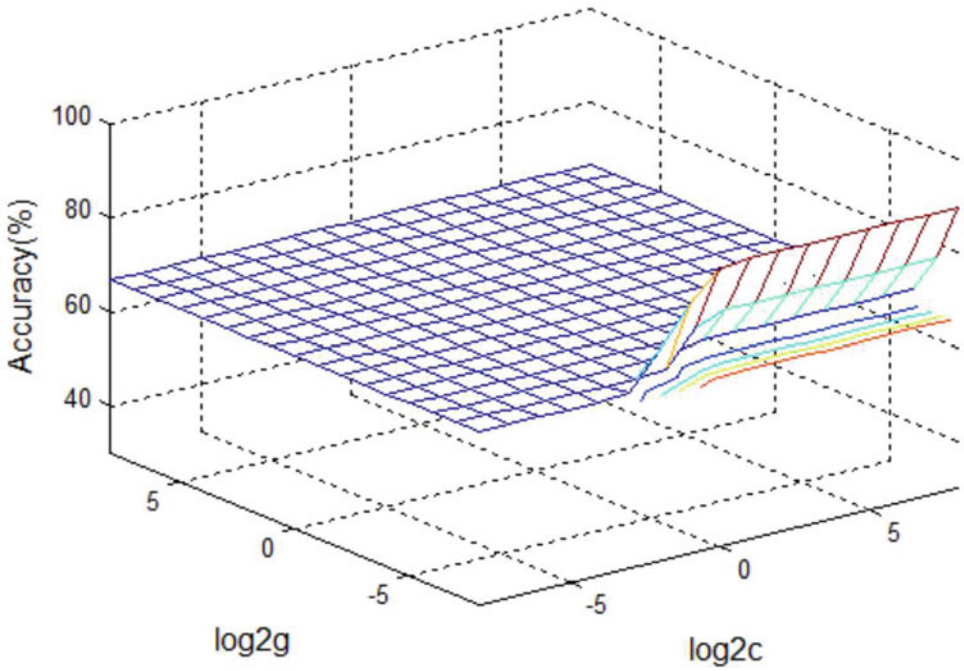
Figure 11 illustrates the parameter chosen for platform 4 with $g$ representing gamma for radial basis function (RBF) kernel function and $c$ represents cost function of SVM classification. In this optimum parameter's selection, the best $c = 4$ and $g = 0.0039$, and the best accuracy is 89%. In this case, the best model has the best robustness. In conclusion, the classification of SVM can lead to high accuracy in prediction of approximately 90%, and the process of parameterization can improve its performance.

Random forests is the other method used in Proteomics where it consists of many individual classification trees [58]. Each tree gives a single classification and the final result relies on how the results of these trees are classified. *Breiman procedure* [59] must conform to the following algorithms: assume that the training set number is $N$, and the experiment with $n$ cases in the bootstrap method, which means a sample with the substitution. The number of sample training set is approximately $2/3$ to the original training set number. Construct classification and regression tree (CART) for each bootstrap training set. These trees are unknown. Assuming there are $M$ features for each input vector, the best features of $M$ is chosen for split use. $M$ is the only parameter that can be modified.

Since each classification gives a result, then aggregate it and the mode of classification results is the final classification. One benefit of the bootstrap sampling method is that $1/3$ of the original training set is not selected and that these samples can be used in prediction. Random forests give discretionary error in training, which is called out of the bag (OOB) error. Berman [59] stated that the experience indicates an OOB error in an unbiased estimate as a cross-validation error. The random forest method is employed to train a classifier, and there are 100 trees trained to use in prediction and the $m$ try was set to be 32 ($= \sqrt{1088}$Features. The results are listed in Fig. 12.

The accuracy of random forest is cross-checking accuracy [59]. Random forest accuracy is better than SVM prediction accuracy, which means that random forest is the most appropriate classifier for this condition. In addition, random forest classifier can give functionality to choose features, which is very useful for some classifications. However, in this case, the less features may reduce the accuracy and all

**Fig. 11** The parameter selection for platform 4 [55]



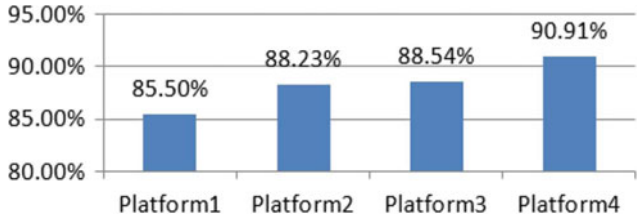



**Fig. 12** The accuracy under the random forest classification [55]

1088 dimensions must be used. It is still very convenient because the computer is fully capable of solving the algorithm in a few hours. As the result for all the number of observed peptides is usually influenced by their physico-chemical property, it makes the correction necessary for accurate prediction.

## 3.4 Text Mining Methods

*Text mining* (TM) is the process of exploring and analyzing large amounts of non-structural text data with the help of software that can identify concepts, patterns, subjects, keywords and other features in the data. One of the main objectives of text mining is to find the related information in the text by converting it into data that can be analyzed. However, this definition does not cover the real relevance, efficiency and role that text mining plays in bioinformatics. In the past decade, the research articles collected in public repositories, such as PubMed [60], are growing exponentially. Figure 13 shows that the number of publications per year on bioinformatics doubles in 2018 compared with 2001.

The applications of text mining require to achieve the linguistic analysis by using *natural language processing* (NLP) algorithms to resolve ambiguity in the human language. The NLP algorithms contain part-of-speech tagging, disambiguation and other fundamental methods that have been modified or better customized to classify text mining in bioinformatics and biomedical literature. Examples of the most text mining applications in bioinformatics include [60]:

- *Information retrieval* (IR) dedicated to obtaining related information from a set of information resources and user query.
- *Document classification* (DC) which defined one or more categories to a document.
- *Named entity recognition/normalization* (NER/NEN) dedicated to extracting the so-called machine-read or semi-structured entity.
- *Summarize* (SUM) which compiles the input text that covers all the contents of the analyzed documents.

In practice, there are several methods to be used to assess the text extraction applications that exist. For instance, ROUGE metric is used to evaluate text summarization systems to estimate the similarity of the resulting summary with the so-called gold summary at syntactic level by matching the n-grams, or at semantic level [61] by evaluating notion covered by the generated summary. There are many variables for ROUGE, where the next level comes with the scale that estimates the number of n-gram in both the gold and the summary that was created [60].

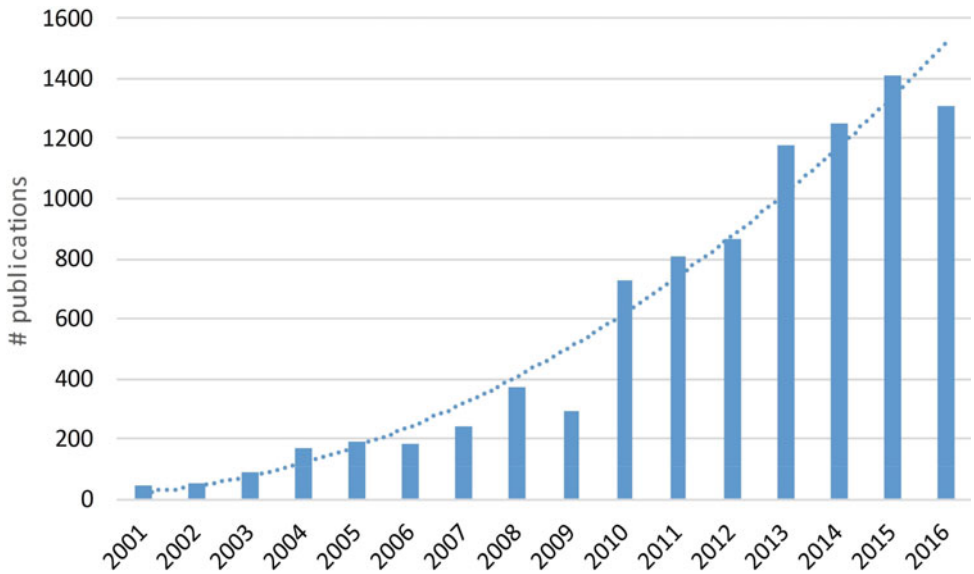



**Fig. 13** The number of publications in PubMed on text mining

$$\text{ROUGE} = \frac{\#(\text{relevent n - gram retrieved in Generate Summaries})}{\#(\text{relevent n - gram in Gold Summaries})} \tag{10}$$

However, the most commonly used measures to evaluate information retrieval, document classification, NER and other applications can be summarized as follows [62]:

- *Precision (P)* which is a part of all recovered documents that have been labeled as relevant

$$\begin{aligned}\text{Precision} &= \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} \\ &= P(\text{relevant}|\text{retrieved})\end{aligned} \tag{11}$$

- *Recall (R)* which is a part of all relevant documents that have been retrieved effectively:

$$\begin{aligned}\text{Recall} &= \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} \\ &= R(\text{retrieved}|\text{relevant})\end{aligned} \tag{12}$$

For those two measures (i.e., precision and recall), their relevancy can be judged according to the confusion or error matrix provided in Table 2, then:

$$P = \frac{tp}{(tp + fp)} R = \frac{tp}{(tp + fn)} \tag{13}$$

- *Balanced F-measure (F)* which attempts to reduce precision and recall to a single measurement.

$$F = 2\frac{(PR)}{(P + R)} \tag{14}$$

- *Accuracy (A)* which is the number of correct answers divided by the total number of answers. In terms of matrix confusion above

$$\text{Accuracy} = \frac{(tp + tn)}{(tp + fp + fn + tn)} \tag{15}$$

- *Error (E)* which is the proportion of incorrectly identified instances:

$$Error = 1 - Accuracy \; or \; E = 1 - A \; (\%) \tag{16}$$

### 3.5 Systems Biology

It has been a while when the relationship between biology and the field of machine learning started; indeed it's a historical and long-complex relationship. Early technique for machine learning called perceptron [63] was an attempt to plot the actual neurological behavior and the field of artificial neural network (ANN) design emerged from this attempt. More artificial neural network architecture has been inspired, such as adaptive resonance theory (ART) [64] and neocognitron [65] inspired from the organization of the visual nervous system. In the intervening years, the flexibility of machine learning techniques has grown along with mathematical frameworks to measure their reliability, and hopefully that machine learning methods will improve the efficiency of discovery and understanding on increasing the size and complexity of biological data.

Supervised and unsupervised learning methods of machine learning are used in applications of biology. In *supervised learning*, objects are grouped into a particular group using a set of attributes or features. The result of a classification process is a set of rules that specify object assignments for existing categories only to attribute values. In a biological context, examples of object-to-class assignments are images of the embodiment of genetic tissues into a group of diseases and the sequence of the protein into its secondary structures. Figure 14 shows the main difference between supervised and unsupervised learning by illustrating data samples for both cases. According to the figure, the features in these samples are levels of term of individual genes measured in tissue samples and the presence/absence

**Table 2** The confusion matrix

| | Relevant | Non-relevant |
|---|---|---|
| Retrieved | *True Positive (tp)* | *False Positive (fp)* |
| Non-retrieved | *False Negative (fn)* | *True Negative (tn)* |

of a specific amino acid code at a given position in the protein sequence, respectively.

The aim of supervised learning is to create a system that can accurately predict the membership of the new object category based on available features. In addition to predicting class properties such as class label (like classical discriminant analysis), supervised techniques can also be applied to predict continuous object properties (like regression analysis). In any supervised learning application, it would be helpful for the classification algorithm to return a "doubt" value (indicating that it is not obvious which one of several possible categories the object should be assigned to) or "externally" (indicating that this object is different from anything previously observed such that the appropriateness of any decision on class membership is questionable). In contrast to the supervised framework, in *unsupervised learning*, there are no predefined labels for the objects and in this case the unsupervised learning aims to search data and find similarities between objects. Similarities are utilized to identify groups of objects, referred to as *clusters*. In other words, unsupervised learning aims to reveal natural groups in the data. Thus, the two models can vary informally as follows: In supervised learning, the data comes with class labels and how to associate labeled data with classes; but in unsupervised learning, all data is unlabeled, and the learning procedure consists of both identify labels and link things to them.

In some applications, such as protein structure classification, only a few labeled samples (protein sequences with a known structural class) are available, while many other samples (sequences) with an unknown class are also available. In such status, *semi-supervised techniques* can be implemented to get a best classifier than can be gained if the labeled samples are used only [66]. This is potential, for instance, by making the cluster assumption, that is, class labels can be reliably divert from labeled to unlabeled objects that are "nearby" in feature space. Life science applications of unsupervised and/or supervised machine learning techniques abound in literature. For instance, gene expression data were used successfully to classify patients in different clinical groups and distinguish new disease groups [67–70] while allowing the genetic code to predict the structure of the secondary protein [71]. Continuous variable prediction was used with machine learning algorithms to estimate bias in microarray data [72].

Finally, Fig. 15 provides a machine learning workflow [73] that shows the difference between supervised and unsupervised learning in biological cases in which the whole process can be summarized in four steps labeled in the figure from **(A) to (D)**.

A. Data preparation stage includes four steps: data preprocessing, feature extraction, model learning and evaluation. It is common to refer to a single data sample including all common variables and features such as $x$ input (vector of numbers) and distinguish it from the output response variable $y$ (one number) when available.
B. The learning method: supervised machine learning methods relate input features $x$ to an output label $y$, whereas unsupervised method learns factors about $x$ without observed labels. The goals of supervised machine learning model are to learn a function $f(x) = y$ from a list of training pairs $(x_1, y_1), (x_2, y_2)$ for which data are listed. One model application in biology is to predict the sensitivity of a cancer cell line when exposed to a selection drug [73].
C. Raw input data is of high dimensions and associated with the corresponding label in a complex manner which challenges many classical machine learning algorithms (left plot). Alternatively, higher-level features extracted using a deep model may be better able to distinguish between layers (right plot). As the $x$ inputs, calculated from raw data, appear what the "model sees around the world", their choice is highly problem-specific.

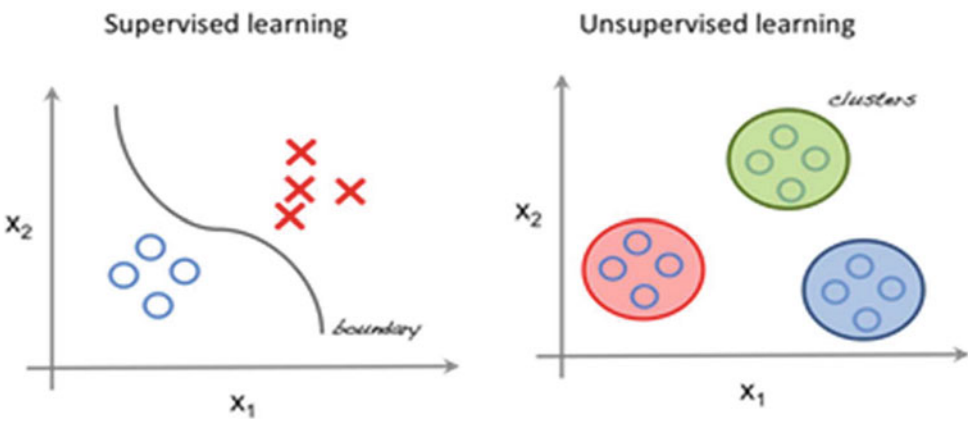



**Fig. 14** Supervised versus unsupervised learning in a biological context

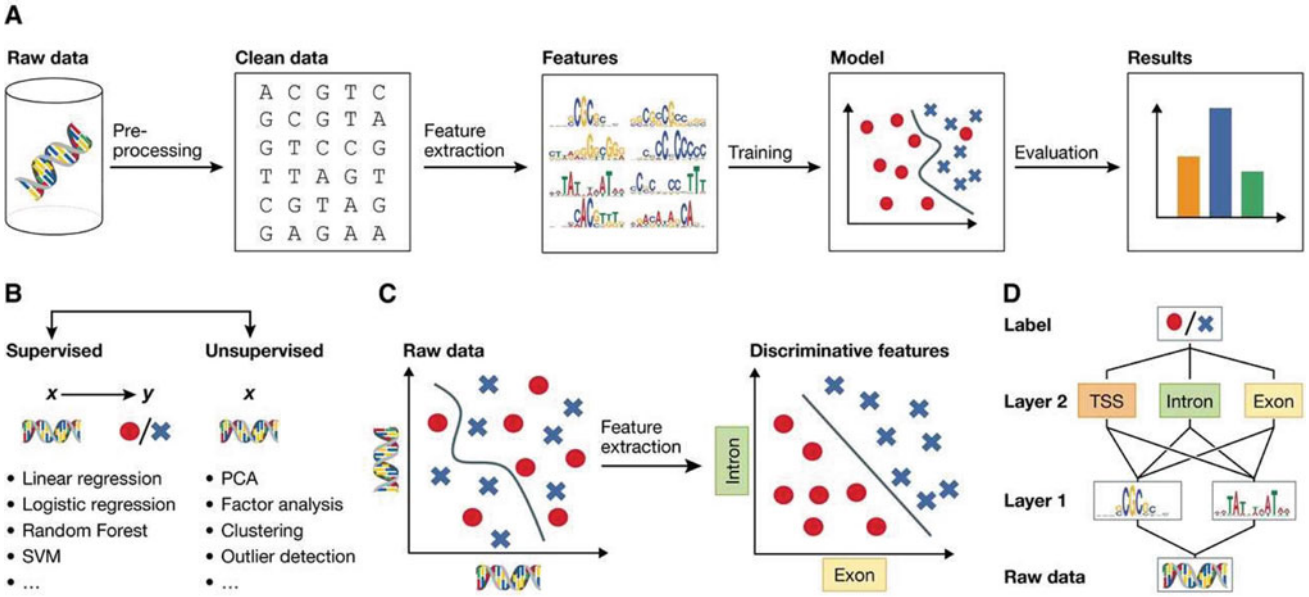



**Fig. 15** The complete framework of machine learning workflow for biological applications

D. Deep networks utilize a hierarchical structure to learn abstract representations of raw data increasingly. The discovery of most informative features is fundamental to the performance, but the process can be labor-intensive and need knowledge of the field. This bottleneck is particularly limited by high-dimensional data; even methods of selecting arithmetic properties are not scaled to assess the advantage of the large number of possible input combinations. A major advance in machine learning is the automation of this critical phase by learning an appropriate representation of data with deep artificial neural networks [74]. In short, the *deep neural network* transfers the raw data in the lower layer (input) and converts it into increasingly simulated representations by sequentially combining the output from the previous layer in a data-based way and encapsulating the very complex tasks in the process.

## 4 Review Questions

By the end of this chapter, you are encouraged to answer the following review questions:

1. Briefly, describe each of the following concepts: Expert system, Bioinformatics, Recommender systems, Reinforcement learning, Probabilistic machine learning, Genomics, MIAME, Support vector machine.
2. What are the three levels of integration for artificial intelligence?
3. What is the distinction between bioinformatics and biomedical engineering fields?
4. What is the distinction between artificial intelligence and machine learning fields?
5. Discuss the importance of Stanford applications in the field bioinformatics?
6. Provide one practical example for each of the following: natural language processing, AI for robotics, AI for speech processing, AI for clinical data annotated and AI for patient similarity
7. List two differences between supervised and unsupervised learning. Give two examples for each.
8. How can we construct algorithms that automatically get better through experience?
9. What are the primary laws that control all learning processes?
10. What is the distinction between automated machine learning (aML) and interactive machine learning (iML)?
11. What are the major biological problems in which arithmetic methods are applied as an AI solution?
12. Why predicting the protein structure is a very complicated problem?
13. Explain in steps how the DNA microarray works?

14. Develop a mathematical numerical example for the quality metric q for background quality assurance process of microarray process?

15. Why do we need a normalization process for the quantified data of microarray process? Justify your answer with numerical examples.

16. What characteristics distinguishing the frequently observed peptides from peptides in the same protein sample but remain unknown?

17. What peptide characteristic can be applied to all living organisms?

18. What is meant by text mining process? Explain by example its role in bioinformatics?

19. List six examples of text mining applications in bioinformatics?

20. Given the following values: Precision = 90% and Recall = 75%. Calculate the information retrieval error percent and the balanced F-measure.

21. Discuss the semi-supervised techniques data classification for bioinformatics applications?

22. What is the main purpose of deep neural network (DNN) in the machine learning workflow for biological applications?

## References


1. Brunette E.S, Flemmer RC, Flemmer CL, (2009). *A review of artificial intelligence*. Proceedings of 4th International Conference on Autonomous Robots and Agents (ICARA 2009), pp: 385–392.
2. Boden M.A (1998). *Creativity and artificial intelligence*. Artificial Intelligence 103: 347–356.
3. Müller V.C, Bostrom N. (2014). *Future progress in artificial intelligence*. AI Matters 1: 9–11.
4. Research Report. (2018). *The AI Industry Series: Top Healthcare AI Trends to Watch*. Retrieved online: https://www.cbinsights.com/research/report/.
5. IBM Watson Health (2018). Artificial Intelligence in medicine. Technical Report. Retrieved online: https://www.ibm.com/watson-health/learn/.
6. Sumit D. et. al. (2015): *Applications of Artificial Intelligence in Machine Learning: Review and Prospect*. International Journal of Computer Applications, Vol. 115, No. 9, April 2015.
7. Rahul C. Deo. (2018). *Machine Learning in Medicine*. Circulation. 2015 Nov 17; 132(20): 1920–1930. https://doi.org/10.1161/circulationaha.115.001593.
8. Yuedong Y. (2016). *Sixty-five years of the long march in protein secondary structure prediction: the final stretch.* Briefings in Bioinformatics, Volume 19, Issue 3, 1 May 2018, Pages 482–494, https://doi.org/10.1093/bib/bbw129.
9. Pedro. L. (2006). Machine learning in bioinformatics. Briefings in Bioinformatics, Volume 7, Issue 1, 1 March 2006, Pages 86–112, https://doi.org/10.1093/bib/bbk007.
10. Stuart R., Peter N. (2009). *Artificial Intelligence: A Modern Approach*. Pearson; 3 edition (December 11, 2009).
11. Linda S. G. (1997). *Mainstream Science on Intelligence: An Editorial With 52 Signatories*. Ablex Publishing Corporation.
12. Nick B. (2006). *How long before superintelligence?* Linguistic and Philosophical Investigations, 2006. - pp. 11–30.
13. Padraig C, Matthieu C., Sarah J.D Delany. (2008). *Supervised Learning: Machine Learning techniques for Multimedia*. Springer, Case studies on Organizational and retrieval.
14. Ricci F., Rokach L., Shapira B. (2010). *Recommender Systems Handbook*. Boston, MA: Springer. ISBN 9780387858197.
15. Jordan MI, Mitchell TM (2015). *Machine learning: trends, perspectives, and prospects*. Science, Vol. 349 No. 6245, P. p. 255–260.
16. LeCun Y, Bengio Y, Hinton G (2015). *Deep learning*. Nature, Vol. 521, No. 7553, P.p.:436–444.
17. Bayes T. (1763). *An essay towards solving a problem in the doctrine of chances (posthumous communicated by Richard Price*). Philosophical Transactions of the Royal Society of London 53 (1763), 370–418.].
18. Barnard GA, Bayes T (1958). *Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances*. Biometrika, Volume 45, Issue 3–4, 1 December 1958, Pages 293–295, https://doi.org/10.1093/biomet/45.3-4.293.
19. Hastie T, Tibshirani R, Friedman J (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd edition. Springer, New York.
20. Murphy KP (2012). *Machine learning: a probabilistic perspective*. MIT press, Cambridge.
21. Silver D, et. al. (2016). *Mastering the game of go with deep neural networks and tree search*. Nature, Vol. 529, No. 7587, P.p. 484–489.
22. Zhong N, et. al. (2007). *Web intelligence meets brain informatics*. Zhong N, Liu JM, Yao YY, Wu JL, Lu SF, Li KC (eds) Web intelligence meets brain informatics., Lecture Notes in Artificial Intelligence 4845, Springer, Berlin, pp 1–31.
23. Holzinger A (2014). *Trends in interactive knowledge discovery for personalized medicine: cognitive science meets machine learning*. IEEE Intelligence Inform Bull 15(1):6–14.
24. Mitchell TM (1997). *Machine learning*. McGraw Hill, New York.
25. Holzinger A, Dehmer M, Jurisica I. (2014). *Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions*. BMC Bio inform 15 (S6):I1.
26. Spinrad N. (2014). *Google car takes the test.* Nature, Vol. 514, No.7523, P.p. 528–528.
27. Holzinger A (2014). *Biomedical informatics: discovering knowledge in big data*. Springer, New York.
28. Holzinger A (2013). *Human–computer interaction and knowledge discovery (HCI-KDD): what is the benefit of bringing those two fields to work together?* Multidisciplinary research and practice for information systems., Springer Lecture Notes in Computer Science LNCS 8127Springer, Heidelberg, pp 319–328.
29. Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.
30. Ohler, W., Liao, C., Niemann, H. & Rubin, G. M. *Computational analysis of core promoters in the Drosophila genome*. Genome Biol. 3, RESEARCH0087 (2002).

31. Degroeve, S., Baets, B. D., de Peer, Y. V. & Rouzé, P. *Feature subset selection for splice site prediction*. Bioinformatics 18, S75–S83 (2002).
32. Bucher, P. (1990). *Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences*. J. Mol. Biol. 4, 563–578.
33. Heintzman, N. et al. (2007). *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome*. Nature Genet. 39, 311–318.
34. BioNinja. *Microarrays*. Retrieved online: http://ib.bioninja.com.au/higher-level/topic-7-nucleic-acids/72-transcription-and-gene/.
35. Peter B, Lei L. and Mark B. (2007*). DNA Microarray Image Processing*. DNA Array Image Anal. Nuts Bolts (Nuts Bolts Ser), Pages 1–77.
36. Adams R. M, B. Stancampiano, M. McKenna and D. Small. (2002). *Case Study: A Virtual Environment for Genomic Data Visualization*. IEEE Transactions on Visualization, Boston, MA, USA (published as CD).
37. Affymetrix Inc., *Gene Chip Arrays*. Retrieved online: http://www.affymetrix.com/index.affx.
38. Brazma A., et. al, (2001). *Minimum Information About a Microarray Experiment (MIAME)–toward standards for microarray data*, Nat. Genet. 29, 365–371, December 2001.
39. Whitfield CW, Cziko AM, Robinson GE. (2003). *Gene expression profiles in the brain predict behavior in individual honey bees*. Science. Vol. 302, pages 296–9.
40. Jain A. N., et. al. (2002). *Fully Automated Quantification of Microarray Image Data*. Genome Research, Vol. 12, No. 2, Feb 2002, pp. 325–332.
41. Steinfath M., et. al. (2001). *Automated image analysis for array hybridization experiments*. Bioinformatics, Vol. 17, pages 634–641.
42. Russ J. (1999). *The Image Processing Handbook: Third Edition*. CRC Press with IEEE Press. Published by CRC Press LLC. 1999.
43. Liew A W-C., H. Yan, and M. Yang. (2003). *Robust Adaptive Spot Segmentation of DNA Microarray Images.* Pattern Recognition 36, pages 1251–1254.
44. Axon Instruments Inc., *GenePix Pro*. Retrieved Online at: http://www.axon.com/GN_Genomics.html.
45. Dodd L. E., et. al., (2004). *Correcting Log Ratios for Signal Saturation in cDNA Microarrays*. Bioinformatics, Vol. 20, No. 16, pp. 2685–2693.
46. Kamberova G., S. Shah (editors) (2002). *DNA Array Image Analysis - Nuts and Bolts*. Data Analysis Tools for DNA Microarrays, DNA Press LLC, MA, 2002.
47. P. Mallick, et. al. (2007*). Computational prediction of proteotypic peptides for quantitative proteomics*. Nat Biotech, 25(I): I 25-I 31.
48. W. H. Zhu, J. W Smith and C. M. Huang (2010). *Mass Spectrometry-Based LabelFree Quantitative Proteomics*. Journal of Biomedicine and Biotechnology, vol. 2010, article ID: 840518.
49. S. E. Ong and M. Mann (2005). *Mass spectrometry-based proteomics turns quantitative.* Nature Chemical Biology, vol. 1, pp. 252–262.
50. R. Aebersold and M. Mann. (2003*). Mass spectrometry-based proteomics.* Nature, 422, pp. 198–207.
51. L. N. Mueller, M. Y. Brusniak, D. R. Mani and R. Aebersold (2008). *An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data*. J. Proteome Res., 7 (01), pp. 51–61.
52. P. Lu, C. Vogel, R. Wang, X. Yao and E. M. Marcotte. (2007). *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation*. Nature Biotechnology, 25, pp. 117–124.
53. J. C. Braisted, et. al., (2008) The Apex Quantitative Proteomics Tool: Generating protein quantitation estimates from LC-MS/MS proteomics results, BMC Bioinformatics 9:529.
54. S. Kawashima and M. Kanehisa. (2000). *AA_index: amino acid index database*. Nucleic Acids Res.vol. 28, no. 374.
55. Biao H., Baochang Z., Yan. F., (2013). *Discovery of Proteomics based on Machine Learning*. Quantitative Biology > Quantitative Methods.
56. B. M. Webb-Robertson, et. al., (2008). *A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics*, Bioinformatics Corrigendum, vol.26, no.13, pp. 1677–1683.
57. C. Cortes and V. Vladimir. (1995). *Support-Vector Networks*, Machine Learning, Vol. 20, Issue. 3, pp. 273–297.
58. E. Alpaydin (2004). *Introduction to Machine Learning*, MIT Press.
59. L. Brieman (2001). *Random Forest*, Machine Learning, vol.45 issue.1, pp. 5–32.
60. Fie. Z., et. al, (2013). Biomedical text mining and its applications in cancer research. Journal of Biomedical Informatics 46 (2013) 200–211.
61. Frawley W.J, Piatetsky S.G, Matheus C. J. (1992). *Knowledge discovery in databases: an overview*. AI Mag 1992;13:57–70.
62. Agarwal S, Liu F, Yu H. (2011). *Simple and efficient machine learning frameworks for identifying protein–protein interaction relevant articles and experimental methods used to study the interactions*. BMC Bioinformatics 2011;12 (Suppl.8):S10.
63. Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev 65: 386–408.
64. Carpenter GA, Grossberg S (1988). *The art of adaptive pattern recognition by a self-organizing neural network*. Computer 21: 77–88.
65. Fukushima K (1980). N*eocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*. Biol Cybern 36: 193–202.
66. Weston J, et al. (2005) *Semi-supervised protein classification using cluster kernels*. Bioinformatics 21: 3241–3247.
67. Alizadeh A.A, et al. (2000). *Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling*. Nature, Vol. 403, pages 503–510.
68. Perou CM, et. al. (1999). *Distinctive gene expression patterns in human mammary epithelial cells and breast cancers*. Proc Natl Acad Sci U S A 96: 9212–9217.
69. Alon U, et al. (1999). *Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by nucleotide arrays*. Proc Natl Acad Sci U S A 96: 6745–6750.
70. Ross DT, et al. (2000). *Systematic variation in gene expression patterns in human cancer cell lines*. Nat Genet 24: 227–235.
71. Rost B, Sander C (1994) *Combining evolutionary information and neural networks to predict protein secondary structure*. Proteins 19: 55–72.
72. Tarca AL, Cooke JE, Mackay J (2005) *A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data*. Bioinformatics 21: 2674–2683.
73. Christof A., et. al., (2016). Deep learning for computational biology. Published online 29.07.2016 Molecular Systems Biology (2016) 12, 878, https://doi.org/10.15252/msb.20156651.
74. Tin-Chih T.C, Cheng L. L., Hong D. L., (2018). *Advanced Artificial Neural Networks*. MDPI, Algorithms 2018, 11, 102; https://doi.org/10.3390/a11070102.