Integrating Machine Learning in Intelligent Bioinformatics

ABOUBEKEUR HAMDI-CHERIF [(1)]
Computer College
Computer Science Department
Qassim University
PO Box 6688 – 51452 Buraydah
SAUDI ARABIA

[(1)] Permanent Address : Université Ferhat Abbas Setif (UFAS)
Faculty of Engineering
Computer Science Department
19000 Setif
ALGERIA
[(1),]email: shrief@qu.edu.sa , elhamdi62@gmail.com

Abstract: - Machine learning is the adaptive process that makes computers improve from experience, by example, and by analogy. Learning capabilities are essential for automatically enhancing the performance of a computational system over time on the basis of previous history. Bioinformatics is the interdisciplinary science of interpreting biological data using information technology and computer science. The field of bioinformatics main objective is to develop relevant computational systems for biological purposes. In this paper, we study how machine learning can help in developing better bioinformatics methods and tools in a coherent manner. We attempt to integrate the multitude of existing methods and tools in a unifying framework as a prelude to showing how machine learning can uncover even more useful structures hidden in biological sequences.

Key-Words: - Intelligent Bioinformatics, Machine learning, Soft computing, Data mining, Grammatical inference.

1 Introduction

Our aim is to contribute to the integration of machine learning and intelligent bioinformatics, under one unified perspective. Although, the interplay between machine learning and bioinformatics has been reported earlier [23] [25], there is no integration between the different approaches. Some efforts have concentrated on specific topics of soft computing methods such as neural networks, genetic algorithms and fuzzy systems [29]. Other works concentrated on modeling methods, such as supervised classification, clustering and probabilistic graphical models for knowledge discovery, as well as deterministic and stochastic heuristics for optimization. Applications in genomics, proteomics, systems biology, evolution and text mining have also been studied extensively [23].

These efforts remain too specific and do not provide a coherent framework for further integration of novel machine learning and/or bioinformatics methods as they arise. They also lack reference to existing on-line repositories, to grammar inference as a machine learning method and to data mining issues, of paramount importance to researchers. Thus, the present framework that meets these deficiencies.

Machine learning is an adaptive process whereby computers can improve from experience, by example, and by analogy. Learning capabilities are therefore essential for automatically improving the performance of a computational system over time on the basis of previous history. A basic learning model typically consists of the following four components:

- learning element, responsible for improving its performance,
- performance element, or decision support system (DSS) responsible for the choice of actions to be taken,
- critical element, which tells the learning element whether the criteria are met within some critical boundaries,

And

- problem generator, responsible for suggesting actions that could lead to new or informative experiences [4].

Bioinformatics is the interdisciplinary science of computational interpretation of biological data, heavily relying on computer science. The field of bioinformatics main objective is to develop relevant computational systems. The importance of this new field of inquiry is constantly growing since researchers continue to generate and integrate large quantities of genomic, proteomic, and other data. We show how machine learning can help in developing better bioinformatics methods and tools in a coherent manner.

The paper is organized as follows. Section 2 deals with problem formulation. This section gives the basic outlines for answering the fundamental question: "How can we enhance the bioinformatics field through machine learning methods?" Section 3 describes the contributions of computer science to bioinformatics and how it can help molecular biology in research and development. Section 4 describes machine learning issues in bioinformatics. Data mining and grammatical inference are described in Section 5. Section 6 is devoted to the possible impacts the proposed framework is thought to induce on future bioinformatics. The paper ends with a conclusion summing up the main results and pointing towards some potential future developments.

2 Problem Formulation

Not pretending to delve into all the intricacies of the highly complex and interdisciplinary field described earlier, we suggest using the entry points available to computer and machine learning scientists. Specifically, the aim is to extend some works on grammar inference [17], [18] to bioinformatics. We first begin by explaining how the field of bioinformatics can be stratified into at least two levels. This stratification motivates for the introduction of intelligent programs.

2.1 Stratifying Bioinformatics

The stratification of bioinformatics can be done through the study of its evolving history. What is the historical unfolding of bioinformatics? One of the central factors promoting the importance of biology is its relationship with medicine. Fundamental progress in medicine depends on elucidating some of the mysteries that occur in the biological sciences. Biology depended on chemistry to make major contributions. This led to the development of biochemistry. Similarly, biophysics came out of the need to explain biological phenomena at the atomic level and their fundamental forces. The huge amount of data gathered by biologists, with the need to interpret it, required tools that have been developed by computer scientists. That is how the interdisciplinary field of bioinformatics came into being [8]. It is believed that this expanding field has undergone a historical transition from the first phase to the second, now underway. In the first phase, the field was dominated by intelligence-free computer programs such as database management systems (DBMSs) and by some relatively computational statistics methods.

In the second phase, *ad hoc* artificial intelligence (AI) methods were used. In both phases, bioinformatics heavily relied on computers, entailing algorithms and systems development. Computer science goal, on the other hand, is making most effective use of information technology hardware through the design and implementation of efficient algorithms. Some areas of theoretical computer science relate most directly to bioinformatics. Information retrieval and analysis require programs; some fairly straightforward and others extremely sophisticated [4]. Finally, distribution of the information requires the facilities of computer networks and the World Wide Web (WWW). Let us consider these different components with reference to a specific biological / bioinformatics problem, namely: 'Retrieve from a database all sequences similar to a probe sequence'. This query spans the first and second phases of bioinformatics [21]. In our forthcoming discourse, we regard these two phases as the mapping of two nested levels of understanding in increasing degree of complexity.

2.2 Two Levels of Bioinformatics Discipline

2.2.1 Intelligence-Free Programs

Intelligence-free programs characterize the first level in bioinformatics development. Standard DBMSs are among these programs. Because of the obvious limitations of traditional DBMS technology, the discipline of bioinformatics has reached saturation of its first level.

2.2.2 Intelligence-Based Programs

Intelligence-based or AI-based programs are the characteristics of the second level in bioinformatics development. The motivation behind this paper is to describe the principles that enhance the existing second level of bioinformatics in which the discipline, instead of being informed by just computer science and computational statistics, is also informed by Artificial Intelligence (AI) techniques and its heuristics. Indeed, for complexity reasons, 'brute force' use of heuristics-free algorithms is pointless. Clearly, a more 'intelligent', *i.e.* heuristics-based approach is required to solve these increasingly difficult bioinformatics problems, such as gene expression analysis and protein structure prediction. Hence the second level bioinformatics [21], further enhanced by grammatical inference, expanded in the sequel.

2.3 Issue of Data Explosion

It is widely recognized that the field of biology, like most sciences, is literally in the midst of a deluge of data. A series of technical advances in recent years has increased the amount of data that biologists can record about different aspects of an organism at the genomic, proteomic and transcriptomic levels. This data is, of course, vital for advancing our knowledge. In recent years, bioinformatics has allowed biologists to make full use of the breakthroughs in computer science and computational statistics in analyzing this data [14].

Fortunately, as the volume of data grows, the techniques used have become more sophisticated to cater for large-scale data and noise. Also, given the growth in biological data, there is a need to extract information that was not previously known from these databases to supplement current knowledge. Large databases may contain interesting patterns that, if identified and authenticated by further laboratory and clinical work, can lead to novel theories about the causes of various diseases and also possibly to the design of new drugs for their treatment. All these issues dictate the urge for novel integrated frameworks [15] whose solution components are reported in the following sections.

3 Problem Solution

The proposed solution is to be considered under four different and complementary banners namely computer science as such, machine learning, data mining, and formal grammars. The main aim is the integration, within bioinformatics, of learning methods, with emphasis on grammatical inference.

3.1 Computer Science Contributions

There are many ways in which computer science can help in molecular biology research and development. We describe here the most important ones and show how computers can be useful in biology and therefore contribute to bioinformatics [15].

3.1.1 Database Technology

3.1.1.1 Data Order of Magnitudes

The discovery of the structure of deoxyribonucleic acid (DNA), as a building bloc of living species, was a turning point in the history of science, culture and society. The use of computer technology for storing DNA sequence information and constructing the correct DNA sequences from fragments identified by restriction enzymes (enzymes which break up the DNA at certain points) was one of the first applications, arising from the different bioinformatics sequencing projects. One of the major projects is perhaps the Human Genome Project whose goals were to make the sequencing of human DNA.
[http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml]

3.1.1.2 Main Bioinformatics Databases

Many databases are now available on the Web and provide relevant information from which it is now possible to extract appropriate bioinformatics characteristics. Among these databases are the following:

(i) *MedLine* [http://www.ncbi.nlm.nih.gov/pubmed]: (Medical Literature Analysis and Retrieval System Online) is a bibliographic database of life sciences and biomedical information. Compiled by the United States National Library of Medicine (NLM), *MedLine* is freely available on the Internet and searchable via *PubMed* and NLM's National Center for Biotechnology Information's *Entrez* system.

(ii) *OMIM* [http://www.ncbi.nlm.nih.gov/omim/]. OMIM (Online Mendelian Inheritance in Man) is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in *OMIM* contain information on all known mendelian disorders and over 12,000 genes. It is updated daily, and the entries contain various links to other genetics resources. *OMIM*® and *Online Mendelian Inheritance in Man*® are registered trademarks of the Johns Hopkins University.

(iii) *PDB* [http://www.rcsb.org/pdb/home/home.do]. The Protein Data Bank (PDB) archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the *wwPDB*, the RCSB PDB curates and annotates PDB data according to agreed-upon standards. The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

(iv) *PIR* [http://pir.georgetown.edu/pirwww/about/] The Protein Information Resource (PIR) is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and other close scientific studies. PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information. Prior to that, the NBRF compiled the first comprehensive collection of macromolecular sequences in the *Atlas of Protein Sequence and Structure*, published from 1965-1978. For over four decades, PIR has provided protein databases and analysis tools freely accessible to the scientific community including the Protein Sequence Database (PSD) [37].

3.1.1.3 Database Structuring – The GO Project

Large databases need to be structured and organized using a common 'ontology', or set of terms which are related structurally to each other, so that researchers can access data from different databases using the same 'query language'. The *Gene Ontology (GO)* project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.

The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from *GO* Consortium members, as well as tools to access and process this data. It is hoped that *GO* will be used by the bioinformatics community so that a common way of referring to genes and their products emerges. [http://www.geneontology.org/].

3.1.1.4 Database Maintenance

Once databases of genomes are created, there is a need for maintaining these databases and for checking that their contents are error-free and valid as researchers add new information. Anomalies and inconsistencies must be identified and actions taken to ensure that the databases are as consistent as possible.

3.1.2 Images and Graphics

3.1.2.1 Image Processing

Many areas of biology rely on images for communicating their results. Image processing tools and techniques are required for describing, analyzing, manipulating and searching for features within these images. Computer graphics are used to visualize 2D or 3D images such as proteins.

3.1.2.2 Genomic Signal Processing (GSP)

The conversion of symbolic sequences into complex genomic signals allows using signal processing methods for the handling and analysis of nucleotide sequences. This methodology reveals surprising regularities, both locally and at a global scale, allowing the prediction of nucleotides in a sequence, when knowing the preceding ones. Such experiments have a major biological significance, as they explore the possibility and the efficiency of error correction in processes like replication, transcription and translation.

To account for these genomic regularities and to make sound processing of genomic signals, an engineering discipline is now emerging, namely genomic signal processing (GSP). Since regulatory decisions within the cell utilize numerous inputs, analytical and numerical tools are necessary to model the multivariate influences on decision-making produced by complex genetic networks. Signal processing approaches such as detection, prediction and classification have been used in the recent past to construct genetic regulatory networks capable of modeling and simulating genetic behavior. One of the objectives of network modeling is to use the network to design different intervention approaches for affecting the time evolution of the gene activity profile of the network. More specifically, one is interested in intervening to help the network avoid undesirable disease-related states [9].

3.2 Bioinformatics Contributions

In the last decade or so, rapid developments in genomics and proteomics have generated a huge amount of data. Often, drawing conclusions from these data requires sophisticated computational methods and corresponding *ad hoc* tools. We briefly review the most important ones with their respective tools.

3.2.1 Omics

In cellular and molecular biology, forming nouns ending with –omics has the sense of "all constituents considered collectively", such as genomics and proteomics [23].

3.2.1.1 Genomics

Genomics, or the study of the genomes of organisms, represents one of the most important domains in bioinformatics. As the number of sequences available is increasing exponentially, these data need to be processed in order to obtain useful information. As a first step, from genome sequences, we can extract the location and structure of the genes. More recently, the identification of regulatory elements and non-coding RNA genes is also being tackled from a computational point of view. Sequence information is also used for gene function and RNA secondary structure prediction.

3.2.1.2 Proteomics

Proteomics is study of proteins, particularly their structures and functions. Proteins are very complex macromolecules with thousands of atoms and bounds. Hence, the number of possible structures is huge. If the genes contain the information, proteins are the operators that transform this information into life. Proteins play a very important role in the life process, and their three-dimensional (3D) structure is a key feature in their functionality. In the proteomic domain, the main application of computational methods is protein structure prediction, a very complicated combinatorial problem where optimization techniques are required. In proteomics, as in the case of genomics, machine learning techniques are applied for protein function prediction.

3.2.2 Sequence Comparison

3.2.2.1 Comparative Genome Analysis

Once genome sequences are stored and accessed, there is a need for comparative genome analysis across databases so that the organization of genomes can be studied. Such analyses may uncover relationships between model organisms, crops, domestic animals and humans. Visualization tools and techniques are required to conduct these analyses. Computational approaches to genome comparison have recently become a common research topic in computer science. A public collection of case studies and demonstrations is growing, ranging from whole genome comparisons to gene expression analysis [10]. This has increased the introduction of different ideas, including concepts from systems and control, information theory, strings analysis and data mining. It is anticipated that computational approaches will foster a standard topic for research and teaching. The main available tools are:

a) The NCBI's Comparative Genomics Development Site offers various resources such as :
   - Whole genomic alignment.
   - Multiple and pairwise alignment.
   - Regulation of co-expressed genes.
   - Identification of conserved transcription factor binding sites (cTFBS).
   - Various other resources.
   [http://www.dcode.org/]
b) The Integrated Microbial Genomes (IMG) system [27] serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life in an integrated context. Plasmids that are not part of a specific microbial genome sequencing project and phage genomes are also included into IMG in order to increase its genomic context for comparative analysis.
   [http://img.jgi.doe.gov/cgi-bin/pub/main.cgi].

3.2.2.2 Sequence alignment

For the retrieval of similar sequences, we need to measure the similarity of the probe sequence to every sequence in the database. It is possible to do much better than the naïve approach of checking every pair of positions in every possible juxtaposition, a method that even without allowing gaps would require a time proportional to the product of the number of characters in the probe sequence times the number of characters in the database [12]. A specialty in computer science known colloquially as 'stringology' focuses on developing efficient methods for this type of problem, and analyzing their effective performance.

Some of the publicly-available programs for sequence alignment have taken years of development and have been finely tuned. Among these programs, we find:

- BLAST [www.ncbi.nlm.nih.gov/blast/])
- FASTA (http://www.ebi.ac.uk/fasta33/).
- GAPSA pairwise sequence alignment based on genetic algorithms [5].

Other multiple sequence alignment programs include:

- ClustalW2 (http://www.ebi.ac.uk/clustalw/),
- DiAlign (http://www.genomatix.de/cgibin/dialign/dialign.pl)
- MultAlin (http://prodes.toulouse.inra.fr/multalin/multalin.html)

Various other tools include Align, Kalign, MAFFT, MUSCLE, T-Coffee [http://www.ebi.ac.uk/Tools/].

3.2.3 Protein Prediction

3.2.3.1 The basic protein problem

As protein sequences are incrementally added to protein databases, and while these are not growing as quickly as genomic databases, there is a need to store protein sequences and their structure as well as their function. Even if a common vocabulary for describing proteins is accepted, there is a major need to link protein sequences with their DNA source sequences, given the problems of introns and non-coding DNA. There is also a need for tools that can predict the structure of a protein from its sequence of amino acids [33]. The main available tools are:

- *Gene finding*

a) Genscan [http://genes.mit.edu/GENSCAN.html]
b) GenomeScan [http://genes.mit.edu/genomescan.html]
c) GeneMark [http://exon.biology.gatech.edu/]

- *Protein domain analysis and identification*

a) Pfam [http://pfam.sanger.ac.uk/]
b) BLOCKS [http://blocks.fhcrc.org/]
c) ProDom [http://prodom.prabi.fr/prodom/current/html/home.php]

- *Pattern identification*

a) GibbsSampler [http://bayesweb.wadsworth.org/gibbs/gibbs.html]
b) AlignACE [http://atlas.med.harvard.edu/cgi-bin/alignace.pl]
c) MEME or GLAM2 [http://meme.sdsc.edu/meme4_3_0/intro.html]

3.2.3.2 Protein threading

Protein threading is one of the most powerful approaches to protein structure prediction, *i.e.* to infer three-dimensional (3-D) protein structure for a given protein sequence. Like many issues in proteomics, protein threading boils down to an optimization problem. Optimal solutions can be obtained in polynomial time using simple dynamic programming algorithms if profile type score functions are employed. However, this problem is computationally hard (NP-hard) if score functions include pairwise interaction preferences between amino acid residues. Therefore, various algorithms have been developed for finding optimal or near-optimal solutions. Algorithms are now available including those involving protein threading with constraints, comparison of RNA secondary structures and protein structure alignment [3]. The main available tools are described below. For more details see : [http://en.wikipedia.org/wiki/Protein_structure_prediction_software]

- *Homology modeling*

a) LOMETS (LocalMeta-Threading-Server) [http://zhanglab.ccmb.med.umich.edu/LOMETS]
b) 3D-JIGSAW [http://bmm.cancerresearchuk.org/~3djigsaw/]
c) Biskit, A Python platform for structural bioinformatics [http://biskit.pasteur.fr/]
d) CPHModel [http://www.cbs.dtu.dk/services/CPHmodels/]
e) ESyPred3D [24] [http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/]
f) Hhpred [http://toolkit.tuebingen.mpg.de/hhpred]
g) MODELLER and its GUI-based version EasyModeller, [http://salilab.org/modeller/]
h) ROBETTA [http://robetta.bakerlab.org/]
i) Swiss-Model [http://swissmodel.expasy.org/]

- *Threading/fold recognition and prediction*

a) RAPTOR RApid Protein Threading by Operation Research technique [38] [http://www.cs.uwaterloo.ca/~j3xu/RAPTOR_form.htm]
b) SUPERFAMILY, a hidden Markov modeling library and genome assignments server, [http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/]
c) LOOP: Learning, Observing and Outputting Protein Patterns [37] [http://cbsu.tc.cornell.edu/software/loopp/]
d) 3D-PSSM – Phyre [http://www.sbg.bio.ic.ac.uk/~3dpssm/]
e) MUSTER: http://zhanglab.ccmb.med.umich.edu/MUSTER
f) PredictProtein [http://www.predictprotein.org/]
g) SwissModel [http://swissmodel.expasy.org/]

- *Ab initio structure prediction*

a) I-TASSER [http://zhanglab.ccmb.med.umich.edu/I-TASSER/]
b) ROBETTA [http://robetta.bakerlab.org/] also used in homology modeling as indicated above.
c) BHAGEERATH:
A computational protocol for modeling and predicting protein structures at the atomic level. An Energy Based Protein Structure Prediction Server

[http://www.scfbio-iitd.res.in/bhageerath/index.jsp]

- *Secondary structure prediction*

a) PREDATOR:
Knowledge-based database comparison [http://mobyle.pasteur.fr/cgi-bin/portal.py?form=predator]

b) GOR:
Information theory/Bayesian inference, [http://gor.bb.iastate.edu/]

c) PSIPRED
Two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST
[http://bioinf4.cs.ucl.ac.uk:3000/psipred/]

d) YASSPP:
Cascaded SVM-based predictor using PSI BLAST profiles
[http://glaros.dtc.umn.edu/yasspp/]

e) NetSurfP
Profile-based neural network
[http://www.cbs.dtu.dk/services/NetSurfP/]

4. Machine Learning Contributions

4.1 Machine Learning Methodology

One of the major breakthroughs in bioinformatics is the application of machine learning. As a long-developed field in artificial intelligence (AI), machine learning (ML) focuses on automatic learning from data set(s). A suitable *a priori* model with many parameters is built first for a certain domain problem and an error measure is defined. A learning (training) procedure is then used to adjust the parameters according to the predefined error measure. This is a classical data-fitting issue as the purpose here is to fit the data into the model.

There are different theories for the learning procedure, including gradient decent, expectation maximization (EM) algorithms, simulated annealing, and evolutionary algorithms. The learning procedure is repeated until the error measure ideally reaches zero or at is least minimized. After the learning procedure is completed with the training data, the parameters are set and kept unchanged and the model can be used to predict or classify new data samples, during the test phase. Further adjustment of the parameters might be undertaken if the test phase gives poor performance.

Important issues in machine learning include the learning speed, the guarantee of convergence, and how the data can be learned incrementally. Regarding bioinformatics applications, many machine learning methods have been implemented to address the various issues [4].

4.2 Supervised *vs.* Unsupervised Learning

As far as relevant machine learning is concerned, there are basically two categories of learning schemes, namely supervised learning and unsupervised learning. Supervised learning learns the data with a known answer at hand. Meaning, the parameters are modified according to the difference of the real (actual) output and the desired known output, or expected answer. The classification problem falls into this category. On the other hand, unsupervised learning learns without any knowledge of the outcome. Clustering belongs to this category. It finds data with similar attributes and aggregates them in the same cluster.

The main familiar machine learning methods such as decision trees learning (DTL), and support vector machines (SVM) have proved very useful in addressing both classification and clustering problems. But machine learning techniques usually handles relatively small data sets because the learning procedure is normally very time-consuming. To apply the techniques to data mining tasks, the problem with handling large data sets must be overcome [19]. One of the ways to reduce data is to use Support Vector Machines (SVM) method explained in the next subsection.

4.3 Data Reduction

Some of the bioinformatics problems such as gene expression usually require high dimensional data. This issue constitutes a serious problem in several machine learning methods. Dimensionality reduction can be used, although it often leads to information loss and performance degradation. Usually, there is a tradeoff between dimensionality reduction and information loss.

4.3.1 Use of SVM

4.3.1.1 Basic methodology and applications

Basically SVM separates data points into two groups of linear separable data sets. It attempts to find out the optimal hyperplane by minimizing an upper bound of the errors and maximizing the margin between the separated hyperplane and data. It was first invented for binary classification problems based on statistical theory. A maximal separating hyperplane is built by SVM to map input vectors to a higher dimensional space. Two parallel hyperplanes are built and the data are separated on each side of the hyperplane. SVMs deliver state-of-the-art performance in real-world applications such as text categorization, hand-written character recognition, image classification, biosequences analysis, among others. Their first introduction in the early 1990s lead to a recent explosion of applications and deepening theoretical analysis, that has now established SVMs, along with neural networks as one of the standard tools for machine learning and data mining [11].

4.3.1.2 SVM use in bioinformatics

Since the support vector machine (SVM) is well known as a training algorithm for learning classification from data, it is widely used for the applications of classification and pattern recognition problems in bioinformatics [11].

When the misclassification rates of SVMs are compared with those of other machine learning approaches, SVMs are found to be the best performing methods. In addition to their use for evaluating microarray expression data, SVMs have been shown to perform well in multiple areas of biological analysis, including detecting remote protein homologies. In this respect, variants of SVMs using a new kernel function have also been developed. The kernel function is derived from a generative statistical model for a protein family, in this case a hidden Markov model (HMM). This general approach of combining generative models like HMMs with discriminative methods such as SVMs may have applications in other areas of biosequence analysis as well [20]. Fortunately, SVMs are shown to be able to overcome the data reduction problem as they can generalize high dimensional data well [36].

4.3.2 Use of radial basis functions (RBF)

It’s well known that the computing time to train multilayer perceptrons (MLPs) is very long because of weight space of the neural networks and small amount of adjustment of the weights for convergence. The matter becomes worse when the size of training data set is large, which is common in data mining tasks. Moreover, depending on samples, the performance of neural networks changes. So, in order to determine appropriate sample sizes for multilayer perceptrons a standard method suggests the use of radial basis function networks (RBFNs) that work as a guide for reducing the data [34].

4.4 ML Interactions with Bioinformatics

Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. These are typical processes handled by machine learning. A particularly active area of research in bioinformatics is the application and development of machine learning techniques to biological problems. As stressed in Section 3.2 above, examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, and statistical modeling of protein-protein interaction, among others. Each of these tasks can be expressed within the framework of machine learning. Accounting for these developments and interactions, a special issue was devoted to the subject. It is shown that machine learning can provide powerful tools for analyzing, predicting, and understanding data from emerging genomic and proteomic technologies. The papers submitted to this special issue provide strong evidence that this is the case [25]. However, integration between these works is lacking.

4.5 Soft Computing and Bioinformatics

By soft computing, we usually refer to methods like neural networks (NNs), genetic algorithms (GAs), [5] and fuzzy systems along with hybrid methods including a combination of some of these methods [1]. Soft computing methodologies, whether used in classification or clustering, occupy now a prominent position among the computer science approaches used in bioinformatics [4]. This is so, because there is an enormous amount of data available for processing. Fortunately, biologists have annotated some of these data. Machine learning and computer science researchers have developed general methods to deal with its reduction [34], and specific methods for bioinformatics [36]. Further improvements of SVM have also been proposed [7] [28].

5. Data Mining and Grammatical Inference Issues

5.1 Data Mining for Data Explosion Solution

5.1.1 Defining Data Mining Field

Data without the pertinent knowledge that conveys it might be useless. Knowledge can be seen as the patterns or characteristics of the data. Therefore, it is much more valuable than data on its own. Indeed, pure or raw data are sometimes meaningless because what we want is the knowledge hidden in the data and not the data *per se.* That is why a new technology field has emerged in the mid 1990’s to deal with the discovery of knowledge from data. It is called knowledge discovery in databases (KDD) or simply data mining (DM). Uncovering hidden information is the fundamental goal of data mining.

A distinctive aspect of bioinformatics is its extensive use of the Web and the manipulation of huge data. Indeed, the immense databases containing DNA sequences and 3D protein structures are available to almost any researcher. That is why data mining tools are unavoidable in bioinformatics [16]. Now some freely-available platforms implementing most data mining algorithms are available on the Web;
e.g. Tanagra [http://eric.univ-lyon2.fr/~ricco/tanagra/],
and *Weka* [www.cs.waikato.ac.nz/ml/weka/].

5.1.2 Data Mining Process and Tasks

(i) Process

Data mining process is based on the following steps: data collection, data preprocessing, data mining proper, information interpretation, and visualization.

(ii) Tasks

The data mining tasks are categorized as follows.

- *Classification* decides the class/group for each data sample. For example, iris species can be classified based on their measurement.
- *Clustering* is to group similar data into a finite set of separate clusters/categories. This task is also referred to as *segmentation*. In machine learning, it requires *unsupervised learning i.e.* that the clusters are not known in advance.
- *Association*, *link analysis* or *affinity analysis* is to tell whether a set of data is dependent on the rest of the set. An association rule can be written as *A* → *B* where both A and B are a data set.
- *Summarization* or *characterization* is a simple description of a large data set. It is desirable that representative information of a data set can be obtained, so that we can easily have a general view of the data set.
- *Text mining* is used when the data to be mined are text instead of numerical data. It originated from information retrieval (IR) of the library science. Keywords are used to find related documents from a document database. For more advanced applications of text mining, classification, clustering, and association techniques are utilized. Text mining can facilitate the re-examination of the biology literature to link the facts that were already known [19].

5.1.3 Bioinformatics Issues and Data Mining

One of the most challenging problems in bioinformatics is to extract knowledge from the huge volume of data, provided by newly developed technologies. For this reason, data mining has become an important tool to carry out this task**.**

The main issues tackled by data mining in bioinformatics are protein structure prediction, gene finding, protein-protein interaction, and phylogenetics, amongst others [14]. A Special issue on data mining for bioinformatics has been published in 2005 by *IEEE Intelligent Systems* [26]. To solve the problem of scattered data, the bioinformatics community has developed a myriad of application programs accessible through the Internet. The National Center for Biotechnology Information (*NCBI*) represents one out of many places where it is possible to find a full array of public tools regarding biomedical and genomic information.
[http://www.ncbi.nlm.nih.gov/]. Some of these tools have been described in Section 3.2 above.

5.1.4 UIMA Approach and Expert Systems

Another approach to handle data explosion consists in using the so-called Unstructured Information Management Architecture (UIMA). It is a component software architecture for the development, discovery, composition, and deployment of multi-modal analytics for the analysis of unstructured information and its integration with search technologies developed by IBM [www.research.ibm.com/UIMA]. A link between UIMA and knowledge-based expert systems (KBES) has been tested. A UIMA-KBES has been developed, which provides facilities for both data and process management. A case study involving a multiple alignment expert system prototype called AlexSys has also been implemented [2].

5.1.4.1 Advantages of UIMA-KBES

Traditionally, bioinformatics analyses rely on independent task-specific services and applications, using different input and output formats frequently not designed to inter-operate. In general, such analyses were performed by experts who manually verified the results obtained at each step in the process. Handling the various applications used to study this information presents a major data management and analysis challenge to researchers. Because of its inherent declarative approach, UIMA-KBES approach is capable of processing the large-scale heterogeneous data in order to extract the pertinent information.

5.1.4.2 Disadvantages of UIMA-KBES

Like any KBES system, the knowledge elicitation process represents a “bottle-neck”, usually difficult, subjective and prone to noisy data.

5.2 Formal Grammars Contribution

Methods based on formal language, statistical, and machine learning theories have been developed for modeling and simulating biological sequences such as DNA, RNA, and proteins. We give here a summary of relevant contributions.

5.2.1 From DNA/RNA and Proteins to Grammars

The biological sequences representing DNA, RNA, or proteins can indeed be seen as sentences derived from a formal grammar [32]. When we view DNA, RNA, or protein sequences as strings or formal languages on alphabets of four nucleotides A,C,G,T or A,C,G,U or 20 amino acids, respectively, a grammatical representation and an inference method can be applied to various problems for biological sequence analyses. Especially, the increasing numbers of yielded DNA and RNA need the development of a grammatical system, especially stochastic grammars such as hidden Markov models (HMMs), [13]. As an example, the language of RNA as a formal grammar that includes pseudoknots has been extensively studied [6], [30].

5.2.2 Grammatical Inference for Bioinformatics

Grammatical inference, also known as grammatical induction or syntactic pattern recognition, refers to the process of automatically learning a formal grammar, usually in the form of re-write rules or productions, from a set of examples, thus constructing a model which accounts for the characteristics of the observed objects, *e.g.* positive examples and eventually negative examples.

Grammatical inference is to be distinguished from traditional decision rules and other such methods principally by the nature of the resulting model, which in the case of grammatical inference relies heavily on hierarchical substitutions. Whereas a traditional decision rule set is directed toward assessing object classification, a grammatical rule set is oriented toward the generation of examples. In this sense, the grammatical inference problem can be said to seek a generative model, while the decision rule problem seeks a descriptive model.

A prominent line of research is to focus on stochastic grammars in biological sequences [31] and on the study of hidden Markov models (HMMs), as special case of stochastic grammars to predict biological sequences functions [13], [22].

6 Impacts of Proposed Framework

We believe that the integration of previously-described theories will advance our knowledge of biological processes based on the most powerful theoretical and technological tools available to computer and machine learning scientists, entailing a better understanding of molecular biology. The impacts on many fields of research are expected to be important, not only on computer science as such but also on medicine, pharmacy and technology at large. We expect impacts of our study on the following fields of research and technology.

i. *DBMS:* More structured organization of data for efficient response to queries. For instance, to develop ways to index or otherwise preprocess the data to make sequence-similarity searches more efficient.

ii. *Human Computer Interaction (HCI)*: To provide interfaces that will assist the user in framing and executing queries.

iii. *Formal Languages:* To extend the applicability of formal stochastic grammatical methods to bioinformatics.

iv. *Bioinformatics:* To further formalize bioinformatics problems and solutions.

7 Conclusion

We have attempted an integration between machine learning and bioinformatics. On top of the multitude of existing methods and tools, it remains highly expected that machine learning will uncover even more useful structures hidden in biological sequences. In addition to actual query search methods now available, however intelligent these might be, future public bioinformatics databases ought to include an array of "what-if" simulation scenarios capable of producing machine learning-oriented results. On the other hand, the language incorporated in genes can be handled by advanced tools of grammatical inference. Further integration of diverse theories from data mining and grammatical inference into bioinformatics will remain indeed a challenging task for years to come. The present work can serve as a pointer to some of the most representative contributions in the field and can be considered as an original categorization and classification of the machine learning methods within bioinformatics. As a logical continuation of the present work, described in an independent article, we report the integration of intelligent control, making possible the control of biological systems and discovery of novel drugs.

Acknowledgments

This work has been gratefully supported by the Deanship of Scientific Research, Qassim University, Saudi Arabia, under Contract # SR-D-007-30.

References:

[1] Adeli, H. (1995) *"Machine learning : neural networks, genetic algorithms, and fuzzy systems"*, Wiley, 1995.

[2] Aniba M.R., S. Siguenza, A. Friedrich, F. Plewniak, O, Poch, A. Marchler-Bauer, J.D. Thompson "Knowledge-based expert systems and a proof-of-concept case study for multiple sequence alignment construction and analysis", *Brief. In Bionf.*, 10(1):11-23, Oct. 2008.

[3] Akutsu T. "Algorithmic aspects of protein threading", In Hsu, H.H. (Ed,) "*Advanced Data Mining Technologies in Bioinformatics*, Chap. 7, pp. 129-146, 2006.

[4] Baldi P., S. Brunak. "*Bioinformatics: The Machine Learning Approach*", MIT Press, 2002.

[5] Ben Othman, M., A. Hamdi-Cherif, Gamil A. Azim. "Genetic algorithms and scalar product for pairwise sequence alignment"*, Int. J. of Computers,* 2(1):134-147, 2008, http://www.naun.org

[6] Cai I., R.L. Malmberg, Y. Wu. "Stochastic modeling of RNA pseudoknotted structures: A grammatical approach", *Bioinformatics*, 19:i66-i73, 2003.

[7] Chen Y., J. Z. Wang, "Support vector learning for fuzzy rule-based classification systems", *IEEE Transactions on Fuzzy Systems,* 11(6):716-728, December 2003.

[8] Cohen J. "Bioinformatics—An introduction for computer scientists". *ACM Computing Surveys*, 36(2):122-158, June 2004.

[9] Cristea P., V. Mladenov, R. Tuduce, G. Tsenov, S. Petrakieva "Neural Networks for Prediction of Nucleotide Sequences by using Genomic Signals", *WSEAS Trans. On Systems,* 7(7):637-644, July 2008.

[10] Cristianini N., M. Hahn. "*Introduction to Computational Genomics",* Cambridge Univ. Press, 2006. [www.computational-genomics.net]

[11] Cristianini N., J. Shawe-Taylor, *"An Introduction to Support Vector Machines"*, Cambridge Univ. Press, 2000. [www.support-vector.net].

[12] Durbin R., S. Eddy, A. Krogh, G. Mitchison. *"Biological Sequence Analysis"*, Cambridge Univ. Press, 1998.

[13] Delcher A., S. Kasif, R. D. Fleischmann, A. Peterson, A. Krogh. "An introduction to hidden Markov models for biological sequences", In S. L. Salzberg, D. B. Searls, S. Kasif (Eds.), *Computational Methods in Molecular Biology*, Elsevier, Amsterdam, 45–63, 1998.

[14] Goodman N. "Biological data becomes computer literate: new advances in bioinformatics", *Curr. Op. Biotech*, 13:66–71, 2002.

[15] Gusfield D. *"Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology"*, Cambridge Univ. Press, 1997.

[16] Hand D. J., Mannila, H., Smyth, P. "*Principles of Data Mining*", MIT Press, 2000.

[17] Hamdi-Cherif C. (*alias* Kara-Mohammed), A. Hamdi-Cherif. "Apprentissage Inductif de Grammaires: Le système GASRIA. (Inductive Learning for Grammars: The GASRIA System)", *Revue d'Intelligence Artificielle*, Hermes-Lavoisier Edition, Paris, France, 21(2): 223-253, March-April 2007,http//ria.revuesonline.com, http://www.revuesonline.com.

[18] Hamdi-Cherif C. (*alias* Kara-Mohammed), A. Hamdi-Cherif. "*ILSGInf* : Inductive Learning System for Grammatical Inference". *WSEAS Trans. on Computers*, 6(6): 991-996, July 2007, http://www.wseas.org

[19] Hsu H.H. "Introduction to data mining in bioinformatics", In Hsu, H.H. (Ed,) "*Advanced Data Mining Technologies in Bioinformatics*, Chap. 1, pp. 1-12, 2006.

[20] Jaakkola T., M. Diekhans , D. Haussler "A discriminative framework for detecting remote protein homologies", http://www.cse.ucsc.edu/research/compbio/research.html, 1999.

[21] Keedwell E., A. Narayanan *"Intelligent Bioinformatics - The Application of AI Techniques to Bioinformatics Problems"*, John Wiley & Sons Ltd, 2005.

[22] Krogh A., M. Brown, I.S. Mian, K. Sjolander, D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling". *J. Molec. Biol.*, 235:1501-1531, 1994.

[23] Larranaga P., B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, Jos, A. Lozano, R. Armananzas, G. Santafe, A. Perez, V. Robles, "Machine Learning in Bioinformatics", *Briefings in Bioinf.* 7(1):86-112, 2005.

[24] Lambert C, N. Leonard, X. De Bolle, E. Depiereux. "ESyPred3D: Prediction of proteins 3D structures", *Bioinformatics,* 18(9):1250-1256, 2002.

[25] Ling C. X., W. S. Noble, Q. Yang "Guest editors' introduction to the special issue: machine learning for bioinformatics—Part 1", *IEEE/ACM Trans. On Comput. Biol. And Bioinf.,* 2(2), 2005.

[26] Li J., L. Wong, Q. Yang. *Special Issue on Data Mining for Bioinf., IEEE Intel. Syst.*, 20(6), 2005.

[27] Markowitz V. M., I-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, I.

Anderson, A. Lykidis, K. Mavromatis, N.N. Ivanova, N.C. Kyrpides "The integrated microbial genomes system: an expanding comparative analysis resource", *Nucleic Acids Res.*, 1–9, 2009.

[28] *Papadimitriou S., K. Terzidis.* "Classification process analysis of bioinformatics data with a support vector fuzzy inference system" *Proc. of the 8th WSEAS Int. Conf. on Neural Networks,* Vancouver, British Columbia, Canada, June 19-21, p. 90-95, 2007.

[29] Prompramote S., Y. Chen, Y.-P. Ph. Chen "Machine learning in bioinformatics", In *Bioinformatics Technologies,* Springer, p. 117-153, 2005.

[30] Rivas E, S. Eddy, "The language of RNA : A formal grammar that includes pseudoknots," *Bioinformatics*,16:334-340, 2000.

[31] Sakakibara Y. "Grammatical inference in bioinformatics". *IEEE Trans. on Patt. Anal. and Mach. Intell.* 27(7):1051-1062, 2005.

[32] Searls D. B. "The language of genes". *Nature,* 420:211–217, November 2002.

[33] Seeman N.C. "DNA in a material world". *Nature*, 421:427-430, January 2003

[34] Sug, H. "Empirical determination of sample sizes for multi-layer perceptrons by simple RBF networks", *WSEAS Trans. on Computers*, 8(9): 1504-1513, Sept. 2009, http://www.wseas.org

[35] Teodorescu O., T. Galor, J. Pillard., R. Elber, "Enriching the sequence substitution matrix by structural information", *Proteins: Structure, Function and Bioinformatics,* 54:41-48, 2004.

[36] G. Valentini, "Supervised gene expression data analysis using support vector machines and multi-layer perceptrons", *Sixth Int.l Conf. on Knowledge-Based Intel. Inf. & Eng. Syst. KES'2002* , *Special Session Machine Learning in Bioinformatics*, 2002.

[37] Wu C. H., L.-S.L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R.S. Ledley, B.E. Suzek, C.R. Vinayaka, J. Zhang and W. C. Barker. "The protein information resource", *Nucleic Acids Research,* 31(1):345–347, 2003.

[38] Xu J., M. Li, D. Kim, Y. J. Xu "RAPTOR: optimal protein threading by linear programming". *Bioinf. Comput. Biol.*, 1(1):95-117, 2003.