# Machine Learning for Intelligent Bioinformatics – Part 1 Machine Learning Integration

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:* - The highly-interdisciplinary field of bioinformatics goal is to develop computing systems capable of analyzing molecular biology. We argue that bioinformatics has undergone a historical transition from the first phase to the second, now underway. The first phase was dominated by the use of traditional, intelligence-free computer programs such as database management systems, on the one hand, and by a small fraction of computational statistics, on the other hand. The second phase, now unfolding, heavily relies on artificial intelligence techniques such as probabilistic, nearest neighbor and genetic algorithm approaches, *inter alia*. In this first part of the present work, we describe both phases, emphasizing integration of alternative machine learning methods such as grammatical inference. This helps in constructing an overall framework including intelligent control described in the second part of the work, reported in an independent paper.

*Key-Words:* - Bioinformatics, Machine learning, Grammatical inference, Data mining.

## 1 Introduction

Our aim is to integrate machine learning and control theories in bioinformatics, under one unified perspective. The field of bioinformatics main objective is to develop computational systems for molecular biology. 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 [5]. 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 and by some computational statistics methods.

In the second phase, *ad hoc* artificial intelligence methods were used. In both phases, bioinformatics heavily relied on computers. Information retrieval and analysis require programs; some fairly straightforward and others extremely sophisticated [2]. Distribution of the information requires the facilities of computer networks and the World Wide Web (WWW). Computer science goal 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. Let us consider these 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 [14]. In our forthcoming discourse, we regard these two phases as two levels of understanding in increasing degree of complexity.

The paper is organized as follows. Section 2 deals with the problem formulation. This section answers the fundamental question: “How can we enhance the

bioinformatics level phase through grammatical inference?" 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. Section 5 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 [18], we suggest using the entry points available to computer and machine learning scientists. Specifically, the aim is to extend some works on grammar inference [11], [12] to bioinformatics and especially to contribute to the enhancements to the previously-described two levels of bioinformatics.

## 2.1 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, transcriptomic and proteomic 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.

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. The present work is divided into two parts. The first one, reported here, describes the integration of learning methods, with emphasis on grammatical inference. The second part of the work, described in an independent article, reports the integration of intelligent control, making possible the control of biological systems and discovery of novel drugs.

## 2.2 Two Levels of Bioinformatics Discipline

### 2.2.1 Intelligence-Free Programs

Intelligence-free programs characterize the first level in bioinformatics development. Standard DBMS are among these programs. The discipline of bioinformatics has reached the end of its first level.

### 2.2.2 Intelligence-Based Programs

Intelligence-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 bioinformatics In second level bioinformatics, the discipline, instead of being informed by just computer science and computational statistics, is also informed by artificial intelligence techniques and its heuristics. Indeed, for complexity reasons, 'brute force' use of heuristics-free algorithms is useless. 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 [14] that we enhance with grammatical inference.

# 3 Problem Solution

## 3.1 Computer Science Contributions

The proposed solution is to be considered under four different and complementary banners namely computer science as such, machine learning, data mining, formal grammars.

There are many ways in which computer science can help in molecular biology research and development [9]. We describe here the most important ones and show how computers can be useful in biology and therefore contribute to bioinformatics.

### 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 project 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 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..
(ii) *OMIM* [http://www.ncbi.nlm.nih.gov/omim/]. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes.
(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.
(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 scientific studies [21].

*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.
[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 Sequence Comparison**
*3.1.2.1 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.

*3.1.2.2 Sequence alignment*
For the retrieval of similar sequences, we need to measure the similarity of the probe sequence to every sequence in the database.
Some of the publicly-available programs (*e.g.,* BLAST [www.ncbi.nlm.nih.gov/blast/]) have taken years of development and have been finely tuned. Other available applications include various tools such as Align, ClustalW2, Kalign, MAFFT, MUSCLE, T-Coffee [http://www.ebi.ac.uk/Tools/].

**3.1.3 Protein Prediction**
*3.1.3.1 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 [20].

*3.1.3.2 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. 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 [1].

**3.1.4 Image Processing and Computer Graphics**
Many areas of biology rely on images for

communicating their results. 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.

# 4. Machine Learning Contributions

## 4.1 Machine Learning at Large

One of the major breakthroughs in bioinformatics was the applications of machine learning. As a long-developed field in artificial intelligence, machine learning 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 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 [2].

## 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 (DT), 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 [13].

## 4.3 Soft Computing and Bioinformatics

By soft computing, we usually refer to methods like neural networks (NNs), genetic algorithms (GAs), [3] and fuzzy systems along with hybrid methods including a combination of these methods. Soft computing methodologies, whether used in classification or clustering, occupy now a prominent position among the computer science approaches used in bioinformatics [2]. This is so, because there is an enormous amount of data available for processing and, fortunately, biologists have annotated some it.

## 4.4 Data Mining for Data Explosion Solution

### 4.4.1 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 [10]. 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/], *Weka* [www.cs.waikato.ac.nz/ml/weka/].

### 4.4.2 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 $A \rightarrow B$ 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 [13].

#### 4.4.3 Bioinformatics Issues and Data Mining

The main issues tackled by data miming in bioinformatics are protein structure prediction, gene finding, protein-protein interaction, and phylogenetics, amongst others [8]. To address the issue 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/]

### 4.5 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.

#### 4.5.1 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 [19]. 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), [6]. As an example, the language of RNA as a formal grammar that includes pseudoknots has been extensively studied [4], [16].

#### 4.5.2 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 observations, 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 [17] and on the study of hidden Markov models (HMMs), as special case of stochastic grammars to predict biological sequences functions [7], [15].

## 5 Impacts of Proposed Framework

We believe that the study and 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 considerable, not only on computer science as such but also on medicine, pharmacy and technology at large. We expect great impacts of our framework 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. *languages:* To extend the applicability of formal stochastic grammatical methods to bioinformatics.
iv. *Bioinformatics:* To further formalize bioinformatics problems and solutions.

## 6 Conclusion

It is highly expected that machine learning will uncover more useful structures hidden in biological sequences. The language incorporated in genes can be handled by advanced tools of grammatical inference. On top of actual query search methods now available, however intelligent these might be, future public bioinformatics databases have to include an array of "what-if" simulation scenarios capable of producing machine learning-oriented results. Further integration of diverse theories from data mining and grammatical inference into bioinformatics will remain indeed a challenging task for years to come.

**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] Akutsu T., "Algorithmic Aspects of Protein Threading", In Hsu, H.H. (Ed,) "*Advanced Data Mining Technologies in Bioinformatics*, Chap. 7, pp. 129-146, 2006.

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

[3] 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

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

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

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

[7] 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, and S. Kasif (Eds.), *Computational Methods in Molecular Biology*. Elsevier, Amsterdam, 45–63, 1998.

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

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

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

[11] 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.

[12] Hamdi-Cherif, C. (*alias* Kara-Mohammed), and A. Hamdi-Cherif, "*ILSGInf* : Inductive Learning System for Grammatical Inference". *WSEAS Trans. on Computers*, 6(6): 991-996, July 2007, ISSN 1109-2750, http://www.wseas.org

[13] 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.

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

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

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

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

[18] Searls, D. B. "Grand challenges in computational biology". In *Computational Methods in Molecular Biology*, S. L. Salzberg, D. B. Searls, and S. Kasif, (Eds.), Elsevier, Amsterdam, 1998.

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

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

[21] 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, Vol. 31, No. 1 345–347, 2003.*