ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Machine Learning in Bioinformatics

JAVAED MOHAMMED
Department of Computer Science New York Institute of Technology Old Westbury, NY

Manuscript Info

Manuscript History:

Received: 25 October 2014
Final Accepted: 20 November 2014
Published Online: December 2014

Key words:
Algorithm, Machine Learning, Optimization, Integration and Interaction, Bioinformatics, Genomics, transcriptomi-cs.

***Corresponding Author**

Javed Mohammed

Abstract

Bioinformatics is the application of computational techniques to analyze the information associated with biomolecules on a large-scale, has now firmly established itself as a discipline in molecular biology, and encompasses a wide range of subject areas from structural biology, genomics to gene expression studies. The availability of new, highly effective tools such as particular genomics, transcriptomics, proteomics and metabolomics for biological exploration is dramatically changing the way one performs research in bioinformatics. Taking advantage of this wealth of "genomic" information has become a norm for whoever ambitions to remain competitive in sciences in general. Machine learning naturally appears as one of the main drivers of progress in this context. In this review we provide an introduction and overview of the current state of the field. We discuss the main principles that underpin bioinformatics analyses using machine language.

Introduction

Machine learning methods are computationally intensive and benefit greatly from progress in computer speed[8]. It is remarkable that both computer speed and sequence volume have been growing at roughly the same rate since the late 1980s, doubling every 16 months or so. More recently, with the completion of the first draft of the Human Genome Project and the advent of high throughput technologies such as DNA microarrays, biological data has been growing even faster, doubling about every 6 to 8 months, and further increasing the pressure towards bioinformatics[9]. Two main paradigms exist in the field of machine learning: supervised and unsupervised learning. Both have potential applications in biology. In supervised learning, objects in a given collection are classified using a set of attributes, or features. The result of the classification process is a set of rules that prescribe assignments of objects to classes based solely on values of features. In a biological context, examples of object-to-class mappings are tissue gene expression profiles to disease group, and protein sequences to their secondary structures.

To the novice, machine-learning methods may appear as a bag of unrelated techniques but they are not. On the theoretical side, a unifying framework for all machine-learning methods also has emerged since the late 1980s. This is the Bayesian probabilistic framework for modeling and inference. In our minds, in fact, there is little difference between machine learning and Bayesian modeling and inference, except for the emphasis on computers and number crunching implicit in the first term. It is the confluence of all three factors data, computers, and theoretical probabilistic framework that is fueling the machine-learning expansion, in bioinformatics and elsewhere. And it is fair to say that bioinformatics and machine learning methods have started to have a significant impact in biology and medicine. As genome and other sequencing projects continue to advance unabated, the emphasis progressively switches from the accumulation of data to its interpretation. Our ability in the future to make new biological discoveries will depend strongly on our ability to combine and correlate diverse data sets along multiple dimensions and scales, rather than a continued effort focused in traditional areas. Sequence data will have to be integrated with structure and function data, with gene expression data, with pathways data, with phenotypic and clinical data, and so forth. Basic research within bioinformatics will have to deal with these issues of system and integrative biology, in the situation where the amount of data is growing exponentially.

The large amounts of data create a critical need for theoretical, algorithmic, and software advances in storing, retrieving, networking, processing, analyzing, navigating, and visualizing biological information. In turn, bioinformatics have inspired computer science advances with new concepts, including genetic algorithms, artificial neural networks, computer viruses and synthetic immune systems, DNA computing, artificial life, and hybrid VLSI-DNA gene chips. This cross-fertilization has enriched both fields and will continue to do so in the coming decades. In fact, all the boundaries between carbon-based and silicon-based information processing systems, whether conceptual or material, have begun to shrink.

2. Background Studies

The history of relations between biology and the field of machine learning is long and complex. An early technique [4] for machine learning called the perceptron constituted an attempt to model actual neuronal behavior, and the field of artificial neural network (ANN) design emerged from this attempt. Early work on the analysis of translation initiation sequences [6] employed the perceptron to define criteria for start sites in Escherichia coli. Further artificial neural network architectures such as the adaptive resonance theory (ART) [5] and neocognitron [7] were 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 for measuring their reliability, and it is natural to hope that machine learning methods will improve the efficiency of discovery and understanding in the mounting volume and complexity of biological data.

Some representative applications of machine learning in computational and systems biology include: Identifying the protein-coding genes from genomic DNA sequences; Predicting the functions of a protein from its primary sequence Identifying functionally important sites from the protein’s amino acid sequence and, when available, from the protein’s structure; Classifying protein sequences into structural classes; Identifying functional modules and genetic networks from gene expression data.

These applications collectively span the entire spectrum of machine learning problems including supervised learning, unsupervised learning and system identification. For example, protein function prediction can be formulated as a supervised learning problem: given a dataset of protein sequences with experimentally determined function labels, induce a classifier that correctly labels a novel protein sequence. The problem of identifying functional modules from gene expression data can be formulated as an unsupervised learning problem: given expression measurements of a set of genes under different conditions, and a distance metric for measuring the similarity or distance between expression profiles of a pair of genes, identify clusters of genes that are co-expressed. The problem of constructing gene networks from gene expression data can be formulated as a system identification problem: given expression measurements of a set of genes under different conditions, and available background knowledge or assumptions, construct a model that explains the observed gene expression measurements and predicts the effects of experimental perturbations.

3. Applications for Treatment by Machine-Learning Approaches

3.1 Sequence-based Analysis

In most cases, single-stranded sequences are used, no matter whether the object in the cellular environment is DNA or RNA. One exception is the of structural elements of DNA, such as bend ability or intrinsic bending potential, which must be based on a true double-stranded interpretation of the double helix.

3.1.1 Intron Splice Sites and Branch Points in Eukaryotic pre-mRNA

Intervening sequences that interrupt the genes of RNA and proteins are characterized, but not unambiguously defined, by local features at the splice junctions. Introns in protein-encoding genes present the most significant computational challenge.

3.1.2 Gene Finding in Prokaryotes and Eukaryotes

Machine-learning techniques have been applied to almost all steps in computational gene finding, including the assignment of translation start and stop, quantification of reading frame potential, frame interruption of splice sites, exon assignment, gene modeling, and assembly. Usually, highly diverse combinations of machine-learning approaches have been incorporated in individual methods

3.1.3 Recognition of Promoters Transcription Initiation and Termination

Initiation of transcription is the first step in gene expression and constitutes an important point of control in the organism. The initiation event takes place when RNA polymerase the enzyme that catalyzes production of RNA from the DNA template recognizes and binds to certain DNA sequences called promoters. This prediction problem is hard due to both the large variable distance between various DNA signals that are the substrate for the recognition by the polymerase, and the many other factors involved in regulation of the expression level.

3.1.4 Gene Expression Levels

This problem may be addressed by predicting the strength of known promoter signals if the expression levels associated with their genes have been determined experimentally. Alternatively, the expression level of genes may be predicted from the sequence of the coding sequence, where the codon usage and/or in some cases, the corresponding codon adaption indices, have been used to encode the sequence statistics

3.1.5 Prediction of DNA Bending and Bend Ability

Many transactions are influenced and determined by the flexibility of the double helix. Transcription initiation is one of them, and prediction of transcription initiation or curvature/ bendability from the sequence would therefore be valuable in the context of understanding a large range of DNA-related phenomena.

3.1.6 Sequence Clustering and Cluster Topology

Because sequence data are notoriously redundant, it is important to have clustering techniques that will put sequences into groups, and also to estimate the intergroup distances at the same time. Both neural networks, in the form of self-organizing maps, and hidden Markov models have been very useful for doing this. One advantage over other clustering techniques has been the unproblematic treatment of large data sets comprising thousands of sequences.

3.1.7 Prediction of RNA Secondary Structure

The most powerful methods for computing and ranking potential secondary structures of mRNA, tRNA, and rRNA are based on the minimization of the free energy in the interaction between base pairs and between pairs of base pairs and their stacking energies [1][3].

3.1.8 Other Functional Sites and Classes of DNA and RNA

Many different types of sites have been considered for separate prediction, including branch points in introns, ribosome binding sites, motifs in protein–DNA interactions, other regulatory signals, DNA helix categories, restriction sites, DNA melting points, reading frame-interrupting deletions in EST sequences, classification of ribosomal RNAs according to phylogenetic classes, and classification of tRNA sequences according to species.

3.1.9 Protein Structure Prediction

This area has boosted the application of machine-learning techniques within sequence analysis, most notably through the work on prediction of protein secondary structure of Qian and Sejnowski[2]. Virtually all aspects of protein structure have been tackled by machine learning. Among the specific elements that have been predicted are categories of secondary structure, distance constraints between residues, fold class, secondary structure class or content, disulfide bridges between cysteineresidues, family membership, helical trans -membrane regions and topology of the membrane crossing, membrane protein class, MHC motifs, and solvent accessibility.

3.1.10 Protein Function Prediction

Functionally related features that have been considered for prediction are intracellular localization, signal peptide cleavage sites, de novo design of signal peptide cleavage sites, signal anchors, glycosylation signals for attachment of carbohydrates, phosphorylation and other modifications related to posttranslational, various binding sites and active sites in proteins.

3.1.11 Protein Degradation

In all organisms proteins are degraded and recycled. In organisms with an immune system the specificity of the degradation is essential for its function and the successful discrimination between self and nonself. Different degradation pathways are active; in several of them proteins are unfolded prior to proteolytic cleavage, and therefore the specificity is presumably strongly related to the pattern in the sequence and not to its 3Dstructure. This general problem has therefore quite naturally been attacked by machine-learning techniques, the main problem being the limited amount of experimentally characterized data.

4. Current and Future Needs

Particular concern of sophisticated integration of extremely diverse sets of data. These novel types of data originate from a variety of experimental techniques of which many are capable of data production at the levels of entire cells, organs, organisms, or even populations.

5. Conclusion

Modern biology can benefit from the advancements made in the area of machine learning. With the current deluge of data, computational methods have become indispensable to biological investigations. Caution should be taken when judging the superiority of some machine learning approaches over other categories of methods. It is argued that the success or failure of machine learning approaches on a given problem is sometimes a matter of the quality indices used to evaluate the results, and these may vary strongly with the expertise of the user. Of special concern with supervised applications is that all steps involved in the classifier design should be cross validated to obtain an unbiased estimate for classifier accuracy. For instance, selecting the features using all available data and subsequently cross-validating the classifier training will produce an optimistically biased error estimate. Bioinformatics has not only provided greater depth to biological investigations, but added the dimension of breadth as well. In this way, we are able to examine individual systems in detail and also compare them with those that are related in order to uncover common principles that apply across many systems and highlight unusual features that are unique to some. In this review we provide an introduction and overview of the current state of the field. We discuss the main principles that underpin bioinformatics analyses using machine language.

REFERENCES

[1] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequencesusing thermodynamic and auxiliary information. Nucl. Acids Res., 9:133–148,1981.

[2] N. Qian and T. J. Sejnowski.Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol., 202:865–884, 1988.

[3] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonherffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures.Monatsheftef.Chemie, 125:167–188, 1994.

[4] Vapnik VN (1998) Statistical learning theory. New York: Wiley. 736 p.

[5] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999)Molecular classification of

cancer: Class discovery and class predication bygene expression monitoring. Science 286: 531–537.

[6] Jirapech-Umpai T, Aitken S (2005) Feature selection and classification formicroarray data analysis: Evolutionary methods for identifying predictivegenes. BMC Bioinformatics 6: 148.

[7] Seiffert U, Jain LC, Schweizer P, editors. Bioinformatics usingcomputational intelligence paradigms. Berlin: Springer. pp. 119–141.

[8] Hoffman, Andreas (2001) Learning Bipedal Locomotion by Demonstration. Bio-Machines, A.I.Lab,MIT

[9] John D Van Horn* and Arthur W. Toga Human Neuroimaging as a “Big Data” Science. Brain Imaging Behav. Jun 2014; 8(2): 323–331.