TUTORIAL PROPOSAL FOR ISMB01

Title
BIOINFORMATICS: THE MACHINE LEARNING APPROACH

Tutor
Pierre Baldi

Pierre Baldi is the Director of the Institute for Genomics and Bioinformatics at the University of California, Irvine. He is a Professor in the Department of Information and Computer Science and the Department of Biological Chemistry. He has taught several courses in bioinformatics at UCI and elsewhere and presented ISMB tutorials in previous years.

Length
Half-Day

Abstract
Machine learning approaches play a significant role in bioinformatics due to the abundance of highly variable data and the lack of comprehensive theories. This tutorial provides a broad overview of machine learning approaches in bioinformatics. Specifically, the main topics are:

1. The Bayesian probabilistic framework for modeling and induction as the common foundation for all machine learning algorithms.
2. A presentation of a number of important classes of models routinely used in bioinformatics such as neural networks, hidden Markov models, stochastic grammars, and belief networks, as well as the corresponding learning algorithms.
3. Specific applications: (a) neural networks for the prediction of protein structural and functional features; (b) hidden Markov models for data base searches, multiple alignments, pattern discovery, and gene finding; (c) stochastic grammars for RNA modeling; (d) Bayesian t-test and clustering methods for DNA microarrays.

Reference/Notes
Tutorial notes will consist of the book:

P. Baldi and S. Brunak
"Bioinformatics: the Machine Learning Approach"
MIT Press, Second Edition, (2001).

The book will be sold by MIT Press to ISMB for this purpose at a 30% discount price. This scheme was used at the ISMB98 conference in Montreal.

Target Audience
Broad audience including students and researchers, biologists interested in an overview of machine learning methods and their scope of application to bioinformatics problems, computer scientists and physical scientists interested in machine learning methods and their applications to biology. No

particular background expected although some preliminary basic knowledge of bioinformatics problems and/or machine learning would be helpful.

Motivation and Goals

Computational analysis of biological sequences---linear descriptions of protein, DNA and RNA molecules---has completely changed its character since the late 1980s. The main driving force behind the changes has been the advent of new, efficient experimental techniques, primarily DNA sequencing (but also DNA microarrays and other high throughput technologies), that have led to an exponential growth of data. As genome and other sequencing projects continue to advance unabated, the emphasis progressively switches from the accumulation of data to its interpretation, fromr the study of single genes/proteins in isolation to genomics, proteomics, and ``system biology.''

Computational tools for classifying sequences, detecting weak similarities, separating protein coding regions from noncoding regions in DNA sequences, predicting molecular structure and function, and reconstructing the underlying evolutionary history have become an essential component of the research process. This is essential to our understanding of life and evolution, as well as to the discovery of new drugs and therapies. Bioinformatics is emerging as a strategic discipline at the frontier between biology and computer science, impacting medicine, biotechnology, and society in many ways.

Large databases of biological information create both challenging data-mining problems and opportunities, each requiring new ideas. In this regard, conventional computer science algorithms have been useful, but are increasingly unable to address many of the most interesting sequence analysis problems. This is due to the inherent complexity of biological systems, brought about by evolutionary tinkering, and to our lack of a comprehensive theory of life's organization at the molecular level. Machine-learning approaches (e.g. neural networks, hidden Markov models, belief networks), on the other hand, are ideally suited for domains characterized by the presence of large amounts of data, ``noisy" patterns, and the absence of general theories. The fundamental idea behind these approaches is to learn the theory automatically from the data, through a process of inference, model fitting, or learning from examples. Thus they form a viable complementary approach to conventional methods. The aim of this book is to present a broad overview of bioinformatics from a machine-learning perspective.

Machine-learning methods are computationally intensive and benefit greatly from progress in computer speed. It is remarkable that both computer speed and sequence volume have been growing at roughly the same rate since the late 1980s, and continue to do so, at the moment doubling about every 15 to 18 months. To the novice, machine-learning methods may appear as a bag of unrelated techniques. 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 framework---that is fueling the machine-learning expansion, in bioinformatics and elsewhere.

An often-met criticism of machine-learning techniques is that they are ``black box'' approaches: one cannot always pin down exactly how a complex neural network, or hidden Markov model, reaches a particular answer. We will address such legitimate concerns both within the general probabilistic framework and from a practical standpoint. It is important to realize, however, that many other techniques in contemporary molecular biology are used on a purely empirical basis. The polymerase chain reaction, for example, for all its usefulness and sensitivity, is still somewhat of a black box technique. Many of its adjustable parameters are chosen on a trial-and-error basis. The movement and mobility of sequences through matrices in gels is another area where the pragmatic success and usefulness are attracting more attention than the lack of detailed understanding of the underlying physical phenomena. Also, the molecular basis for the pharmacological effect of most drugs remains largely unknown. Ultimately the proof is in the pudding. In this tutorial, we will show how machine-learning methods yield good puddings and are elegant at the same time.

Detailed Outline

1 INTRODUCTION: BIOLOGICAL PROBLEMS AND DATA
1.1 Biological Data in Digital Symbol Sequences
  Database Annotation Quality
  Database Redundancy
1.2 Genomes---Diversity, Size, and Structure
1.3 Proteins and Proteomes
  From Genome to Proteome
  Protein Length Distributions
  Protein Function
1.4 On the Information Content of Biological Sequences
  Information and Information Reduction
  Alignment Versus Prediction: When are Alignments Reliable?
  Prediction of Functional Features
  Global and Local Alignments and Substitution Matrix Entropies
  Consensus Sequences and Sequence Logos
1.5 Prediction of Molecular Function and Structure
  Sequence-based Analysis

2 MACHINE LEARNING FOUNDATIONS: THE PROBABILISTIC FRAMEWORK
2.1 Introduction: Bayesian Modeling
2.2 The Cox--Jaynes Axioms
2.3 Bayesian Inference and Induction
  Priors (Maximum Entropy, Group Theoretic Considerations, Useful
  Practical Priors: Gaussian, Gamm a, and Dirichlet)
  Data Likelihood
  Parameter Estimation and Model Selection
  Prediction, Marginalization of Nuisance Parameters, and Class Comparison
  Ockham's Razor
  Minimum Description Length
2.4 Model Structures: Graphical Models and Other Tricks
  Graphical Models and Independence

Hidden Variables
Hierarchical Modeling
Hybrid Modeling/Parameterization
Exponential Family of Distributions

3 PROBABILISTIC MODELING AND INFERENCE: EXAMPLES
The Simplest Sequence Models
The Single Die Model with Sequence Data
The Single Die Model with Counts Data
The Multiple Dice Model with Sequence Data

4 MACHINE LEARNING ALGORITHMS
4.1 Introduction
4.2 Dynamic Programming
4.3 Gradient Descent
Random Direction Descent
4.4 EM/GEM Algorithms
4.5 Markov Chain Monte Carlo Methods
Markov Chains
Gibbs Sampling
Metropolis Algorithm
4.6 Simulated Annealing
4.7 Evolutionary and Genetic Algorithms
4.8 Learning Algorithms: Miscellaneous Aspects
Control of Model Complexity
On-Line/Batch Learning
Training/Test/Validation
Early Stopping
Ensembles
Balancing and Weighting Schemes

5 NEURAL NETWORKS: THE THEORY
5.1 Introduction
5.2 Universal Approximation Properties
5.3 Priors and Likelihoods
Priors
Gaussian Regression
Binomial Classification
Multinomial Classification
The General Exponential Family Case
5.4 Learning Algorithms: Backpropagation

6 NEURAL NETWORKS: APPLICATIONS
6.1 Sequence Encoding and Output Interpretation
6.2 Prediction of Protein Secondary Structure
Secondary Structure Prediction Using MLPs

Prediction Based on Evolutionary Information and Amino Acid Composition
More Recent Work on Protein Secondary Structure Prediction
6.3 Prediction of Signal Peptides and Their Cleavage Sites
SignalP
6.4 Applications for DNA and RNA Nucleotide Sequences
The Structure and Origin of the Genetic Code
Eukaryotic Gene Finding and Intron Splice Site Prediction
Gene Structure Prediction by Sensor Integration
Prediction of Intron Splice Sites by Combining Local and Global Sequence Information
Doing Sequence Analysis by Inspecting the Order in Which Neural Networks Learn

7 OTHER MODELS: THEORY AND APPLICATIONS
7.1 Hidden Markov Models: The Theory
Introduction
HMM Definition
HMMs for Biological Sequences
Prior Information and Initialization
Likelihood and Basic Learning Algorithms
Multiple Alignments
Database Mining and Classification
Structural Analysis and Pattern Discovery
7.2 Hidden Markov Models: Applications
Protein Applications (G-Protein-Coupled Receptors)
DNA (Periodic Patterns in Exons and Introns, Promoter Regions)
7.3 Hybrid Systems: Hidden Markov Models and Neural Networks
7.4 Stochastic Grammars
7.5 Bayesian Networks

8 DNA MICROARRAYS
8.1 Microarrays
Introduction
Gene Expression and Regulation
Technological Problems
Data Analysis Problems
Inferring Gene Changes
Gene Clustering
Inferring Regulatory Elements and Regulatory Networks
8.2 Bayesian Inference of Gene Changes
Probabilistic Framework
Cyber-T Software
8.3 Gene Clustering
Hierarchical Clustering
K-Means
Other Approaches
8.4 Gene Regulation
Regulatory regions

Modeling and Inferring Regulatory Networks

9 INTERNET RESOURCES AND PUBLIC DATA BASES
A Rapidly Changing Set of Resources
Databases over Databases and Tools
Databases over Databases
Databases
Selected Machine Learning Prediction Servers
Prediction of Protein Structure from Sequence
Gene Finding and Intron Splice Site Prediction
Other Prediction Servers
Internet Courses