Report on DIMACS Workshop on Machine Learning Techniques in Bioinformatics

Dates: July 11 - 12, 2006

Location: DIMACS Center, CoRE Building, Rutgers University

Organizers:
Dechang Chen, Uniformed Services University of the Health Services;
Xue-Wen Chen, University of Kansas;
Sorin Draghici, Wayne State University

Report Authors: Dechang Chen, Xue-Wen Chen, Sorin Draghici

Date of Report: November 14, 2006

(DIMACS is a collaborate project of Rutgers and Princeton Universities, AT&T Labs, Bell Labs, Telcordia Technologies and NEC Laboratories America, with affiliated members of Avaya Labs, Georgia Institute of Technology, HP Labs, IBM Research, Microsoft Research, Rensselaer Polytechnic Institute, Stevens Institute of Technology)

1. Introduction

Bioinformatics aims to solve biological problems by using techniques from mathematics, statistics, computer science, and machine learning. Recent years have observed the essential use of these techniques in this rapidly growing field. Examples of such applications include those to gene expression data analysis, gene-protein interactions, protein folding and structure prediction, genetic and molecular networks, sequence and structural motifs, genomics and proteomics, text mining in bioinformatics, and so on.. Bioinformatics provides opportunities for developing novel machine learning techniques; and machine learning plays a key role in advancing bioinformatics. The workshop was devoted to computational challenges of important biological problems. The goal of this workshop was to bring together researchers in both machine learning and bioinformatics to discuss state-of-the-art machine learning algorithms and their applications to various tasks in bioinformatics.

The Presentations are summarized below.

2. Presentations

Machine Learning as Applied to Structural Bioinformatics: Results and Challenges

Philip E. Bourne, University of California San Diego

Structures of proteins contribute greatly to researchers' understanding of living systems. However, current researchers are often locked into thinking about structure as a static entity. This talk introduced new methods to predict protein structure flexibility and the sites of protein-protein interactions. The method for **prediction of protein structure flexibility** uses a Gaussian Network Model to model short and long range movements and defines functional flexibility for a selected range of proteins. The results are then used as input to a support vector machine (SVM) to identify flexible regions important for protein function. The method for **prediction of protein-protein interaction sites** first derives structurally conserved residues from multiple structure alignments of the individual components of known complexes. The assigned conservation score is then weighted based on the crystallographic B factor to account for the structural flexibility that will result in a poor alignment. Sequence profile and accessible surface area information are then combined with the conservation score to predict protein–protein binding sites using SVM. The incorporation of the conservation score significantly improved the performance of the SVM. The results from the proposed method support the hypothesis that in many cases protein interfaces require some residues to provide rigidity to minimize the entropic cost upon complex formation. Additionally, Prof. Bourne listed some challenging bioinformatics problems for the machine learning community.

An Expectation Maximization Algorithm for Inferring the Evolution of Eukaryotic Gene Structure

Liran Carmel, NIH/NLM/NCBI

Recently, several evolutionary models, each incomplete in a different way, yielded very different predictions about evolution of introns. This talk formulated a more general and realistic model of intron gain and loss. The proposed model takes into account gene specific rates, branch specific rates, invariant sites and rate variability of both gain and loss. The parameters of the model are estimated using an Expectation Maximization (EM) algorithm. Using an extended intron-exon dataset including 18 complete genomes of eukaryotic species, the proposed method refutes both extreme views of intron early and intron late. A diverse kaleidoscope of events is proposed from this method, with different parts of the phylogenetic tree exhibiting dramatically different patterns of intron gain and loss.

A Machine Learning Approach for Predicting the EC Numbers of Proteins

James Howse, Los Alamos National Laboratory

The Enzyme Commission number (EC number) is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. This talk presented a principled machine learning

methodology for predicting the EC numbers of proteins. This proposed method first constructs a reference predictor whose results model those achieved by a human expert. Then it estimates the smallest possible prediction error (Bayes error) that can be achieved using the chosen feature space. Finally the proposed method constructs a predictor using a multi-class classifier whose generalization performance (performance on future data) and computational performance (training and testing time and memory) are well characterized. The results show that the method outperforms the reference predictor with high confidence. In addition, it shows that combining structure information with sequence information could produce smaller errors.

Machine Learning and Data Combination for Regulatory Pathway Prediction

Dustin T. Holloway, Mark Kon, and Charles DeLisi, Boston University

Using machine learning techniques to study biological regulatory pathways has attracted many interests. Support vector machines (SVM) based classifiers provide a robust framework for analysis of regulatory networks. Processing of classifier outputs can provide high quality predictions and biological insight into functions of particular transcription factors. This talk presented a supervised learning approach to site identification using SVM to combine 26 different data types. A comparison with the standard approach to site identification using position specific scoring matrices (PSSMs) for a set of 104 Saccharomyces cerevisiae regulators indicates that the proposed method is more sensitive (73% vs 20%) and has double the precision (positive predictive values of 90% and 45% respectively). Applying this method for each transcriptional regulator to all promoters in the yeast genome has led to thousands of new targets for further investigations. This method was also applied to predict other features such as functions of regulatory factors and promoter melting temperature curves.

How to Avoid Misinterpreting Microarray Data

Sungchul Ji, Rutgers University

DNA arrays have been successfully used to measure mRNA levels (TL, transcript levels), rates of gene expression (TR) and DNA levels (TD). In the literature, there have been interpretations of mRNA levels in terms of quantities other than mRNA levels themselves, such as rates of gene expression (TR). This interpretation assumes that TL and TR are linearly correlated so that TL and TR always change in parallel. This talk presented direct measurements of both TL and TR, indicating that this assumption is not always valid. An analysis of the TL and TR data on 5,184 genes of budding yeast measured at 6 time points after shifting glucose to galactose indicates that TL and TR can change independently. These observations can be readily accounted for if one assumes that TL is regulated not only by TR but also by TD. A mathematical equation was derived to model the relationship between TL and TR/TD. This equation can be used to calculate the TD values from TL and TR data on glycolytic and oxidative phosphorylation genes. The results demonstrate that transcript degradation plays a key regulatory role in RNA metabolism in budding yeast during glucose-galactose shift, thus establishing the concept of "degradational control" in analogy to the well-known "transcriptional control".

Learning the cis Regulatory Code by Predictive Modeling of Gene Regulation

Christina Leslie, Columbia University

A cell's gene regulatory network refers to the coordinated switching on and off of genes by regulatory proteins that bind to non-coding DNA. Studying the behavior of gene regulatory networks using high throughput genomic data has become one of the central problems in computational biology. Most work in this area has focused on learning structure from data such as finding clusters of potentially co-regulated genes. Instead of adopting the structure learning viewpoint, this talk focused on building predictive models of gene regulation. A prediction function for the regulatory response of genes is learnt by using a boosting algorithm to enable feature selection from a high dimensional search space while avoiding overfitting. In particular, by coupling the activity of a regulator in an experiment with the presence of motifs in the promoter region of a gene, motifs representing putative regulatory elements can be learnt from their differential expression pattern. This information is combined into a global predictive model for gene regulation. The proposed method is applied to both the environmental stress response and oxygen sensing and regulation systems in yeast. The predictions of which genes will be up or down regulated in held out (test) microarray data show a high accuracy. This method can also be used to predict true regulatory elements, and suggest interpretable biological hypotheses about regulatory mechanisms.

Genome Wide Tagging SNPs with Entropy Based Methods

Zhenqiu Liu, University of Maryland Medicine

It is widely hoped that the study of SNPs will provide a means of elucidating the genetic component of complex diseases and variable drug responses. Throughput technologies such as oligonucleotide array have produced enormous SNP data and put great challenges in genome wide disease association studies. This talk presented a new entropy based monte carlo method for optimally selecting minimum informative subsets of SNPs. The proposed method is based on a multilocus LD measure and the global optimization technique. One advantage of the method is that it can deal with a large number of SNPs without the block assumption. A comparison of the proposed method with other techniques was made, indicating that the method was much faster and performed well on the datasets used in the study.

Comparing the Performance of Several Popular Machine Learning Algorithms on Classifying TATA-box from Putative TATA Boxes

Raja Loganantharaj, University of Louisiana at Lafayette

To understand regulatory mechanisms, it is important to detect all the binding sites in promoter regions such as TATA boxes. A putative TATA-box is said to be detected in a promoter region of a DNA sequence, when the total weight of a subsequence exceeded a certain threshold when applied to TATA-box's position specific weighted matrix (PSWM). It has been estimated that a large number of putative TATA boxes occur in promoter sequences of many genomes. Identification of a true TATA box among putative TATA boxes will improve the accuracy of

determining a transcription start site (TSS) and hence the detection of a promoter. This talk presented an investigation of the effectiveness of several popular machine learning algorithms in discriminating TATA boxes from putative TATA boxes, namely Naïve Bayes, artificial neural networks, decision tree and support vector machine. Prediction accuracy, true and false positive, and Kappa values are used as metrics to compare the effectiveness. Previous work on discriminating TATA boxes from putative TATA boxes in plant genome has revealed that the neighborhood around a TATA box carries the information required to distinguish TATA boxes from putative TATA boxes. Empirical results on plant genome reveal that Naïve Bayes has outperformed other sophisticated machine learning algorithms in predicting TATA boxes from putative TATA boxes. The prediction accuracy and the true positive of Naïve Bayes increase with the length of flanking strings. The comparison of Kappa values, a measure of performance of different algorithms over a random prediction algorithm, indicates that Naïve Bayes performs better than other algorithms.

Modular Organization of Protein Interaction Network

Feng Luo, Clemson University

Accumulating evidence suggest that biological systems are composed of interacting functional modules. Identification of these modules is essential for the understanding of the structure, function and evolution of biological systems. These biological modules are reflected in different topological sub-networks in biological networks. This talk extended the degree concept from single vertices to sub-graphs, and proposed a formal definition of a module in a network. Two algorithms were presented to identify modules within biological networks. Applying the algorithms to the yeast core protein interaction network from the Database of Interaction Proteins (DIP) identifies statistical significant protein modules. Both algorithms allow the construction of an interconnecting web of modules to get insights into the high level relationship among modules.

Motif Refinement by Improving Information Content Scores Using Neighborhood Search

Chandan Reddy, Yao-Chung Weng and Hsiao-Dong Chiang, Cornell University

The goal of motif finding is to detect novel, overrepresented unknown signals in a set of sequences. Most widely used algorithms for finding motifs obtain a generative probabilistic representation of the overrepresented signals and try to discover profiles that maximize information content score. The major difficulty of these algorithms arises from the fact that the best motif corresponds to the global maximum of a non-convex continuous function. Algorithms like Expectation Maximization (EM) and Gibbs sampling are very sensitive to the initial guesses and only converge to the nearest local maximum. This talk described a novel optimization framework that searches the neighborhood regions of the initial alignments in a systematic manner to explore the neighborhood profiles. The results show that the popularly used EM algorithm can be significantly improved by the proposed neighborhood search. Based on experiments using both synthetic and real datasets, the proposed method demonstrates significant improvements in the information content scores of the probabilistic models. The proposed method also gives flexibility in using different local solvers and global methods that work well for some specific

datasets.

Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles

Aik Choon Tan, Daniel Q. Naiman, Lei Xu, Raimond L. Winslow and Donald Geman, Johns Hopkins University

Various studies have shown that cancer tissue samples can be successfully detected and classified by their gene expression patterns using machine learning approaches. One of the challenges in applying these techniques for classifying gene expression data is to extract accurate, readily interpretable rules providing biological insight as to how classification is performed. Current methods generate classifiers that are accurate but difficult to interpret. This talk introduced a new classifier, k-TSP (k Top Scoring Pairs), in order to address these problems. This method generates simple and accurate decision rules that only involve a small number of gene-to-gene expression comparisons. Comparisons have been made between this approach and other machine learning techniques for class prediction in 19 binary and multi-class gene expression datasets involving human cancers. Results from the comparisons show that the k-TSP classifier performs as efficiently as Prediction Analysis of Microarray and support vector machine, and outperforms other learning methods (decision trees, k-nearest neighbor and naïve Bayes). The proposed method can also be applied to cross-platform analysis. The classifier built from the marker gene pair, which simply compares relative expression values, achieves high accuracy, sensitivity and specificity on independent datasets generated using various array platforms.

Cancer Tissue Classification with Data-dependent Kernels

Anne Zhang, The University of Kansas

Microarray techniques enable the measurement of genes' expression profiles at genome scale and provide an unprecedented opportunity to characterize cells at molecular level. Classification of tissue samples according to their gene expression profiles can be a valuable diagnostic tool for diseases such as cancer. Currently, gene expression data sets are typically characterized by high dimensionality (a large number of genes) and small sample size, which makes the classification task quite challenging. This talk introduced a data-dependent kernel for cancer classification with microarray data. This kernel function is engineered so that the class separability of the training data is maximized. A data resampling scheme is introduced for kernel optimization to overcome the problem of overfitting. The effectiveness of this adaptive kernel for cancer classification is illustrated with a k-Nearest Neighbor (KNN) classifier. The data-dependent kernel leads to a significant increase in the accuracy of the KNN classifier based on experimental results. The kernel-based KNN scheme is shown to be competitive to, if not better than, more sophisticated classifiers such as support vector machines (SVMs) and the uncorrelated linear discriminant analysis (ULDA) in classifying the gene expression data.

3. Conclusions

The workshop was very well received. More than 50 people attended the workshop and participated in the discussion. Attendants from different backgrounds exchanged viewpoints and new ideas. Many talks aroused keen interests and discussions. The slides of the talks can be found at: http://dimacs.rutgers.edu/Workshops/MLTechniques/slides/slides.html

4. Acknowledgements

The organizers and the DIMACS Center acknowledge the support of the National Science Foundation under grant number CCF 05-14703 to Rutgers University. This special focus was jointly sponsored by the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), the Biological, Mathematical, and Physical Sciences Interfaces Institute for Quantitative Biology (BioMaPS), and the Rutgers Center for Molecular Biophysics and Biophysical Chemistry (MB Center).