

TUTORIAL PROPOSAL FOR ISMB02

Title

DNA MICROARRAYS AND GENE REGULATION

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. He is the author of over 100 scientific papers and several books including:

- Bioinformatics: the Machine Learning Approach (MIT Press, Second Edition 2001)
- The Shattered Self—The End of Natural Evolution (MIT Press, 2001)
- DNA Microarrays and Gene Regulation (Cambridge University Press, 2002, in press)

Length

Half-Day

Reference/Notes

Tutorial notes will consist of the book:

P. Baldi and G. Wesley Hatfield
"DNA Microarrays and Gene Regulation"
Cambridge University Press (2002).

The book will be sold by Cambridge University Press to ISMB for this purpose with a 30% discount price. A similar approach was successfully used at the ISMB98 and ISMB01 conferences with the book Bioinformatics: the Machine Learning Approach.

Target Audience

Broad audience including students and researchers, biologists interested in an overview of algorithms and statistical modeling techniques for the analysis of DNA microarray data (differential expression, clustering, visualization, modeling of regulatory networks), computational scientists interested in modern data mining/machine learning techniques and the opportunities and applications arising from the application of high-throughput technologies to biology. No particular background expected although some preliminary basic knowledge of bioinformatics and DNA microarrays would be helpful.

Motivation and Goals

DNA microarrays are an important modern high-throughput technique which, together with sequencing and proteomics techniques, is in the process of generating an exponentially increasing amount of data and profoundly transforming biomedical sciences and technologies. Like the invention of the microscope a few centuries ago, DNA microarrays provide entirely new vistas on the workings of cells and gene regulatory networks by providing snapshots of the level of expression of all the genes in a cell at a given time. As genome sequencing and DNA microarray projects continue to advance unabated, the emphasis progressively switches from the accumulation of data to its interpretation, from the study of single genes/proteins in isolation to genomics, proteomics, and “systems biology.”

Biologists are not well prepared to analyze the data produced by DNA microarray projects. Not only are the terabytes of data numerical in nature and therefore hard to interpret with the “naked eye”, but the technology is also at a relatively early stage of development hence noisy and with a lack of standards that results in heterogeneous formats. To analyze, visualize, store, and navigate microarray data modern computer/statistics/bioinformatics methods have become absolutely essential in this area, from the understanding of life and evolution, to the discovery of new drugs and therapies.

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 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. Data mining and statistical machine-learning approaches (e.g. Bayesian probabilistic approaches, graphical models, clustering, neural networks, etc), 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.

Once basic issues of noise, calibration, and normalization have been dealt with, the analysis of DNA microarray data can proceed with at least four different levels. To begin with, there is the level of differential analysis where genes are analyzed one by one independently of each other to detect changes in expression across different conditions. This problem is already challenging due to the noise and low repetition characteristic of many DNA microarray experiments. The next level of analysis involves visualizing correlations in the data, using statistical techniques such as PCA (principal component analysis) and more generally, unsupervised methods for feature discovery such as clustering (including k-means and hierarchical clustering) and supervised discrimination methods, such as artificial neural networks and SVMs (support vector machines). The third level of analysis combines DNA microarray data with sequence data by searching, for instance, for common regulatory motifs located in the upstream regions of co-regulated genes detected by clustering methods. Finally, at the “systems” level, one tries to model and understand gene regulatory networks. A number of different computational techniques are useful in this analysis

ranging from Boolean networks, to Bayesian networks, to ordinary and partial differential equations, to qualitative analysis, to rule-based methods.

The tutorial will provide a broad overview of the problems, methods available, and results across all four levels of analysis together with demonstration of software and Internet resources.

Detailed Outline

1. A Brief History of Genomics

2. DNA Array Formats

In situ Synthesized Oligonucleotide Arrays

Pre-synthesized DNA arrays

Filter-based DNA Arrays

Non-conventional Gene Expression Profiling Technologies

3. DNA Array Readout Methods

Reading data from a Fluorescent signal

Reading data from a Radioactive signal

4. Gene Expression Profiling Experiments: Problems, Pitfalls and Solutions

Primary Sources of Experimental and Biological Variation

Differences among samples

RNA isolation procedures

Special Considerations for Gene Expression Profiling in Bacteria

A comparison of the use of targets prepared with message-specific primers and targets prepared with random hexamer primers from total RNA preparation for use with pre-synthesized DNA arrays

Rapid Turnover of mRNA in Bacterial Cells

Preparation of bacterial targets for in situ synthesized DNA arrays

Target preparation with Non-polyadenylated mRNA from bacterial cells

Problems Associated with Target preparation with Polyadenylated mRNA from Eukaryotic Cells

A Total RNA Solution for Target preparation from Eukaryotic Cells

Target cDNA synthesis and radioactive-labeling for pre-synthesized DNA arrays

Data acquisition for Nylon Filter experiments

Data acquisition for Affymetrix GeneChip™ glass slide experiments

Normalization methods (Normalization to Total or Ribosomal RNA, Normalization to

Housekeeping Genes, Normalization to a Reference RNA, Normalization by Global Scaling)

5. Statistical Analysis of Array Data: Inferring Changes

Problems and Common Approaches

Probabilistic Modeling of Array Data

The Bayesian Probabilistic Framework

Gaussian Model for Array Data

The Conjugate Prior

Full-Bayesian Treatment Versus Hypothesis Testing

Parameter Point Estimates
Hyperparameter Point Estimates and Implementation
Simulations
Extensions

6. Statistical Analysis of Array Data: Dimensionality Reduction, Clustering, and Regulatory Regions

Problems and Approaches
Visualization, Dimensionality Reduction, and Principal Component Analysis
Clustering Overview
Data Types
Supervised/Unsupervised
Similarity/Distance
Number of Clusters
Cost Function and Probabilistic Interpretation
Hierarchical Clustering
Hierarchical Clustering Algorithm
Tree Visualization
K-Means, Mixture Models, and EM Algorithm
K-Means Algorithm
Mixtures Models and EM
DNA Arrays and Regulatory Regions

7. Systems Biology

Systems Biology and Gene Regulation
The Molecular World: Representation and Simulation
Metabolic Networks
Protein Networks (Signaling)
Regulatory Networks
Computational Models of Regulatory Networks
Discrete Models: Boolean Networks
Continuous Models: Ordinary Differential Equations
Continuous Models: Partial Differential Equations (Reaction-Diffusion Models)
Qualitative Modeling and Piece-Wise Linear Models
Limitations and Learning in Continuous Models
Probabilistic Models: Bayesian Networks
The Search for General Principles. Scaling Laws.

8. Software and Internet Resources

Academic and Commercial Softwares
Data Bases
Internet Resources

