

**Functional Genomics in 4 Hours:
A Practical Guide to Creating Your Own High-Throughput Pipeline**

**Atul Butte, MD
Isaac Kohane, MD, PhD
Children's Hospital Boston and Harvard Medical School**

Half-day Tutorial Proposal for ISMB 2002

Abstract

The massively parallel acquisition of RNA expression data is rapidly becoming streamlined and dropping in price. In the near future we can expect that biologists and clinicians in many institutions will be routinely measuring such data. Therefore analysis of these data sets to characterize biological systems, identify high-yield candidate genes/ESTs for further biological investigation, or quantify a patient's health risks, to just name a few tasks, will become a standard part of the investigational armamentarium. Many algorithms have been developed to take RNA expression data sets and generate clusters that are putatively reflective of functional dependencies. These algorithms range in complexity from simple fold-difference calculations to comprehensive pair-wise comparisons and model construction. This tutorial is designed to teach the basics of the various bioinformatics methodologies available to analyze RNA expression data sets, yet will approach the subject from a practical standpoint, so that attendees can immediately put these algorithms to use.

Content Description (4 parts):

1. Review

The first part of the tutorial will be the most didactic. It will include a review of:

- Brief descriptions of the commonly used types of microarrays, and how use of the different types impacts downstream analysis.
- Typical flow of an investigation of the functional genomics of a biological domain going from hypothesis generation to hypothesis validation.
- Nature and format of the expression data files generated by the two most common technologies will be described. Particular emphasis will be placed on the different characteristics of these measurement systems and how normalization of the data sets can be approached (and common mistakes).
- Quantifying and compensating for noise and irreproducibility in microarray measurements.
- Description of the most frequently used clustering techniques. Their strengths and weakness will be summarized. The questions for which each might be better suited will be addressed as well as reasonable approaches to the interpretation of results generated by these techniques.
- Review of several instances of the clustering techniques applied to various biological systems.
- What to do next after a list of genes is obtained; how to link genes to other national databases.
- A review of the current relevant biological ontologies, taxonomies, and data models and their application to microarray analysis and storage.

2. Example Analysis

Several publicly available data sets will be introduced. The instructors will lead the participants step by step through several analyses of these data sets. They will provide a very concrete sense of what is involved in performing the analyses introduced in the Review part of the tutorial.

3. Human Genome Project

An update on the Human Genome Project will be presented, including coverage of the draft genome release. Implications of the findings from the Human Genome Project will be put in the context of functional genomics and microarray analysis.

4. Questions and Answers

During this segment of the tutorial, participants will be encouraged to explore how they might use these techniques in domains that are of interest to them. Also, the instructors will moderate a more detailed discussion of the problems associated with each of the techniques reviewed and where the current research challenges lie.

Goals of the tutorial:

By the end of the session, attendees will be able to:

1. Be able to explain the different types of genomic clustering available, including intervention fold differences, self-organizing maps, phylogenetic-type trees, and know the advantages and disadvantages of each.
2. Know how to calculate correlation coefficient, mutual information, entropy, and other measures of information and dissimilarity.
3. Be able to interpret the results of each clustering method, and know what possible next steps are available in analyzing the results.
4. Understand all the types of experiments done with microarrays to date, and the potential variety of experiments possible.
5. Know how to get more information about a resulting list of genes.

Who should attend:

- Bioinformaticians
- Medical informaticians and clinicians looking to get involved in bioinformatics
- Students
- Information Technology professionals

Coverage:

- Basic 30%
- Intermediate 40%
- Advanced 30%

Instructors:

Both instructors have previously presented and taught at PSB, AMIA, and ISMB tutorials. This tutorial was presented by the same instructors at AMIA 2001 and PSB 2002 (as well as ISMB 1999 in an earlier version) and received excellent reviews. Isaac Kohane and Atul Butte are coauthors on an upcoming book titled "Microarrays for an Integrative Genomics", which will be the first book published on microarray analysis (to appear in July 2002 by MIT Press). This tutorial will be taught based on this book.

The instructors for this course are involved in investigations of gene expression with collaborators in multiple academic centers in Boston and elsewhere. These collaborations involve the study of the functional genomics of organ transplantation and rejection, diabetes, cardiac disease, tumorigenesis, neurodevelopment, and neuromuscular disease, to just mention a few of the established application domains. The Children's Hospital Informatics Program is funded by five institutes of NIH (NIDDK, NLM, NIAID, NHLBI, and NINDS) to perform and analyze microarrays in functional genomics, as well as many other funding agencies.

Isaac S. Kohane, MD, PhD
Associate Professor of Pediatrics, Harvard Medical School
Director, Children's Hospital Informatics Program

Isaac (Zak) Kohane is the director of the Children's Hospital Informatics Program and Associate Professor of Pediatrics at Harvard Medical School. Dr. Kohane is leading multiple collaborations at Harvard Medical School and its hospital affiliates in the elucidation of regulatory networks of genes and the interaction between genotype and phenotype using a variety of bioinformatics techniques. Application domains he is currently involved in include tumorigenesis, neurodevelopment, neuro-endocrinology and transplantation biology. Dr. Kohane's research builds on his doctoral work in computer science on decision support and subsequent research in machine learning applied to biomedicine. Dr. Kohane has also led the development of cryptographic health identification systems and automated personal health records. He has published over 50 papers in biomedical informatics. Dr. Kohane has chaired several national meetings including the two most recent Spring Symposia on Artificial Intelligence in Medicine at Stanford University and the session on Linking Phenotype to Genotype at the Pacific Symposium on Biocomputing. He is also a founder of the Center for Outcomes and Policy Research at the Dana Farber Cancer Institute, founder and Associate Director for the Center for Genetic Epidemiology at Harvard Medical School. He is a Fellow of the American College of Medical Informatics and a Fellow of the Society for Pediatric Research. He is Associate Editor for Bioinformatics for the Journal of Biomedical Informatics and on the editorial board of the Journal of the American Medical Informatics Associations. Dr. Kohane is also a practicing pediatric endocrinologist at Children's Hospital in Boston.

Atul Butte, MD
Instructor in Pediatrics, Harvard Medical School
Assistant in Pediatric Endocrinology, Children's Hospital, Boston

Atul Butte is currently on staff in the Children's Hospital Informatics Program, is a practicing pediatric endocrinologist at Children's Hospital, Boston, and is an Instructor at Harvard Medical School. Dr. Butte received his undergraduate degree in Computer Science from Brown University in 1991, and worked in several stints as a software engineer at Apple Computer (on the System 7 team) and Microsoft Corporation (on the Excel team). He graduated from the Brown University School of Medicine in 1995, during which he worked as a research fellow at NIDDK through the Howard Hughes/NIH Research Scholars Program. He completed his residency in Pediatrics and Fellowship in Pediatric Endocrinology in 2001, both at Children's Hospital, Boston. During his research work under Dr. Isaac Kohane, he developed a novel methodology for analyzing large data sets of RNA expression, called Relevance Networks. This technique was published in the Proceedings of the National Academy of Science (2000, 97:12182). Dr. Butte is an inventor on five pending US patents, including one for Relevance Networks. Dr. Butte's recent awards include the 2001 American Association for Cancer Research / Pharmacia Scholar-In-Training Award and the 2001 Lawson Wilkins Pediatric Endocrine Society NovoNordisk Clinical Scholar Award. Dr. Butte's research is supported by grants from NIDDK, NHLBI, NINDS, NLM, the Endocrine Fellows Foundation, the Genentech Center for Clinical Research and Education, the Lawson Wilkins Pediatric Endocrinology Society, and Merck.