AM6: ISMB2006 Tutorial

Title: Integration and analysis of diverse genomic data

Topic Area:

- Database and Data Integration 20%
- Systems Biology (including Pathways and Networks) 50%
- Machine Learning and Artificial Intelligence 30%

Main Presenter:

- Title Dr
- Full name Olga G Troyanskaya
- Affiliation Princeton University
- Mailing Address Dept. of Computer Science, 35 Olden Street, Princeton, NJ 08544 USA
- Email address ogt@princeton.edu
- Telephone number – work and cell if available, with country and city codes 1-609-258 1749 (work)
- Fax number 1-609-258 1771
- Home page URL http://function.princeton.edu/

- Teaching experience:
  In the past three years, I have developed and taught two bioinformatics courses at Princeton University: Analysis and Visualization of Large-Scale Biological Data and Computational Modeling of Biological Networks. Both of these courses include advanced undergraduates, graduate students, and postdoctoral fellows at Princeton. In addition, I teach the microarray analysis lectures and workshops for the Advanced Bioinformatics course at the Cold Spring Harbor Laboratory and developed and taught a four-week bioinformatics course focused on microarray analysis at the California State University at Hayward.
- Earlier tutorial presentations and feedback:
  I have taught this tutorial "Integration and analysis of diverse genomic data" at ISMB 2005 in Detroit. The tutorial was extremely popular (it sold out very early on) and it got great feedback. This inspired me to update it and propose to offer the updated tutorial at ISMB 2006.

50-word abstract:

In the recent years, multiple types of high-throughput functional genomic data have become available that facilitate rapid functional annotation and pathway modeling in the sequenced genomes. However, genomic data sacrifice specificity for scale compared to traditional experimental methods, yielding large quantities of relatively lower quality measurements. This problem has generated much interest in bioinformatics in the past two years, as sophisticated computational methods are necessary for accurate functional interpretation of these large-scale datasets. This tutorial will present an overview of recently developed methods for integrated analysis of functional genomic data and outline current challenges in the field. The focus will be on the development and use of such methods for gene function prediction, understanding of protein regulation, and modeling of biological networks.

Tutorial level: Introductory to intermediate. This tutorial will serve as a thorough introduction to data integration in functional genomics, but some advanced issues will also be discussed (starting from appropriate fundamentals).

Prior knowledge required:
This tutorial will be self-contained and assume no prior background in the field of data integration or biological data analysis. No specific computational or biological background will be assumed, and the audience may include computer scientists, statisticians, bioinformaticians, and computationally savvy biologists. The audience should be familiar with basic biological concepts (e.g. regulation, transcription, etc) and basic computation (probability).

All concepts will be introduced on an intuitive level, so a biologist or a computer scientist will be comfortable with the material. Building on this introductory material, state-of-the-art methods for data integration will be introduced with special emphasis on assumptions, limitations, and strengths of each method. Finally, open problems in the field will be discussed.

Suitability of this tutorial for ISMB:
In the past five years, we have witnessed an explosion of multiple types of high-throughput functional genomic data. These data have the potential to facilitate rapid functional annotation and pathway modelling in the sequenced genomes. Gene expression microarrays are the most commonly available source of such data, but increasing amounts of other data, including protein-protein interactions, sequence, and localization data, are being generated. However, genomic data often sacrifice specificity for scale compared to traditional experimental methods, yielding very large quantities of relatively lower quality measurements.

This problem has generated much interest in bioinformatics in the past three years, as sophisticated computational methods are necessary for accurate functional interpretation of these large-scale datasets. This problem spans a wide range of areas within bioinformatics: from databases to machine learning to data analysis and modelling and its effective solution could have enormous impact on our understanding of how cells function. This tutorial will present an overview of recently developed methods that integrate the analysis of microarray, sequence, interaction, localization, and literature data and outline current challenges in the field. The focus will be on the development and use of such methods for gene function prediction, understanding of protein regulation, and modelling of biological networks. This tutorial will be of interest to computational researchers and students interested in contributing to the field of data integration and analysis of heterogeneous data and to biologists with some computational background who are interested in using the methods on their experimental data and understanding their properties and limitations.

Profile of Presenter
Olga Troyanskaya is an Assistant Professor in the Department of Computer Science and the Lewis-Sigler Institute for Integrative Genomics at Princeton University. Her research focus is on the analysis and modeling of heterogeneous biological data and on microarray data analysis. Her Ph.D. is in biomedical informatics from Stanford University, where her work included developing a Bayesian system for heterogeneous data integration for gene function prediction and methods development for robust

analysis of microarray data. Her laboratory at Princeton University is developing novel machine learning methods for data integration, gene function prediction, and biological pathway modeling. Her presentations on these topics include publications and multiple presentations at workshops, symposia, and seminars. She has also recently written an invited review paper on the topic of data integration for *Briefings in Bioinformatics*.

Dr. Troyanskaya has developed and taught two bioinformatics courses at Princeton University: Analysis and Visualization of Large-Scale Biological Data (a course she developed and teaches every year) and Computational Modelling of Biological Networks (a course she co-developed and co-teaches with another faculty member). Both of these courses include advanced undergraduates, graduate students, and postdoctoral fellows at Princeton. In addition, Dr. Troyanskaya teaches microarray analysis lectures and workshops for the Bioinformatics course at the Cold Spring Harbor Laboratory and developed and taught a four-week bioinformatics course focused on microarray analysis at the California State University at Hayward. She has also given a popular tutorial on data integration at ISMB 2005.

Tutorial Outline:

I. Introduction and overview of functional genomics data

   a. Goals and challenges for data integration
      i. Noise levels in genomic data, and early studies indicating that integrated analysis increases specificity
      ii. Data integration on different levels: databases, analysis, modelling
      iii. Challenges for data integration
         1. noise in datasets
         2. representing datasets in a coherent way
         3. integration of databases
         4. public data availability
         5. challenges for multi-cellular organisms

   b. Available experimental data
      i. Microarray data
         1. expression arrays
         2. ChIP-chip experiments
         3. protein arrays
      ii. Protein-protein interactions data (yeast two hybrid, affinity precipitation)
      iii. Genetic interactions (synthetic interactions, synthetic lethality)
      iv. Localization data
      v. Sequence data
      vi. Structure data
      vii. Biomedical literature
      viii. Public databases that provide the data

II. Introduction to methodology (no prior background assumed)

   a. Decision trees

b. Bayesian networks
c. Support Vector Machines (SVM)
d. Graph algorithms

III. Data integration for gene function prediction

a. Function prediction from “guilt by association” principle
b. Proof-of-principle: Intersection-based methods (e.g. Marcotte et al. 1999)
c. Methods based on decision trees (e.g. Clare et al. 2003, Zhang et al. 2004)
d. SVM-based methods (e.g. Lanckriet et al. 2004)
e. Bayesian methods (e.g. Troyanskaya et al. 2003)
f. Graph-based methods (e.g. Karaoz et al. 2004)
g. Methods based on biomedical literature
    i. Using curated data – Gene Ontology
    ii. Methods based on natural language processing of biomedical literature (a brief discussion) (e.g. Raychaudhuri et al. 2003)

IV. Data integration to study gene regulation

a. Integrating expression with sequence data for motif discovery (e.g. Roth et al. 1998, Bussemaker et al. 2001, Liu et al. 2001)
    i. Biological basis for these algorithms
    ii. Introduction to expectation-maximization and gibbs-sampling based algorithms
    iii. State-of-the-art algorithms
b. Identifying regulatory modules (e.g. Bar-Joseph et al. 2003, Kato et al. 2004, Segal et al. 2003)

V. Analysis and modelling of biological networks based on heterogeneous data

a. Definition of biological networks
b. Current progress in integrated analysis and modelling of biological networks (e.g. Hartemink et al. 2002, Nariai et al. 2004, Tanay et al. 2004, Myers et al. 2005)

VI. Open problems in data integration

Extensive bibliography will be provided to the participants of this tutorial.