Integrated Analysis of Gene Expression Data

(Tutorial Proposal for ISMB 2001)

Yike Guo

Reader in Computing Science
Dept. of Computing
Imperial College
180 Queen's Gate
London SW7 2BZ
UK

John Sgouros

Head of Computational Genome Analysis Lab.
Imperial Cancer Research Fund.
44 Lincoln's Inn Fields
London WC2A 3PX
UK

Tutors :

Dr.Yike Guo is the Reader in Computing Science, Dept. of Computing, Imperial College. He is also the technical director of the parallel computing center of Imperial College, and the head of the data mining group of the university. He has been leading a large research group to develop new technology and software systems for genomic data analysis. In particular, the Kensington Discovery Edition system developed by the group and commercialized by InforSense Limited, an Imperial College spin off company founded by Dr. Yike Guo, has been adopted by various research institutions and companies as a platform for large scale integrated genomic data analysis. Dr. Yike Guo is a leading computer scientist in the field of distributed data mining. He won international prizes for his pioneer work in the internet based distributed data mining. He has been invited to many conferences and organizations to present his work. He also has been teaching data mining, bioinformatics in Imperial College and University of Tsukuba of Japan since 1997.

Dr. John Sgouros is the Head of Computational Genome Analysis Lab of Imperial Cancer Research Fund. He is also an Senior Scientist and Head of Bioinformatics at Turku Centre for Biotechnology, Finland. He is also the bioinformatics architect of joint Sanger Centre - ICRF - Ludwig Institute for Cancer Research microarray facility consortium (Hinxton, UK). John Sgouros has been organizing large scale genomic data analysis for various organization. He has been invited to speak in many bioinformatics and genomics conferences. He also teaches intensively in University of London on computational genomics. He has been collaborating with Dr. Yike Guo on large scale integrated gene expression data analysis and its applications since 1999.

Tutorial presentation:

Due to the advance of high throughput DNA microarray technologies, gene expression information has increased exponentially and will continue to do so for the foreseeable future. Conventional means of storing, analysing and comparing related data are already overburdened. Moreover, the rich information in gene expression and its wide biological implication requires new technologies of analysing data that employ sophisticated statistical and machine learning algorithms, powerful computers and intensive interaction with other data source such as sequence data and metabolic pathway information to discover complex gene expression patterns and to correlate gene expression patterns with other biological process to gain a comprehensive understanding of cell physiology. The goal of this tutorial is to provide a comprehensive survey on the state-of-the art analysis technologies for large scale gene expression analysis. In particular, the tutorial will emphasis on the integration of gene expression analysis with other information for advanced biology knowledge discovery.

The content of the tutorial will include

- Basic gene expression analysis technology such as clustering, northern analysis and classification;
- Advanced gene expression analysis technology such as gene regulatory network discovery;
- Integrated gene expression analysis including combining expression analysis with motif analysis and promoter predication, combining expression analysis with metabolic pathway;
- Applications of integrated analysis including function annotation of gene expression patterns and drug discovery

Intended audience:

Data analysts , biologists and computer scientists working in gene expression related analysis. The tutorial aims to be introductory and requires only basic knowledge and experience on statistics and molecular biology. The tutorial will be taught with illustration through live demos of all the analysis.

Length :

Prefer to be a one day tutorial but also can do a half day tutorial by focusing mainly on integrated analysis.

Detailed outline of the presentation:

Part I : Gene expression and gene expression analysis. This part will cover the concept of gene expression profiling, data warehousing, data quality and normalization, basic query models, basic analysis technology.

Part II : Co-expression analysis by clustering . In this part, we will introduce various clustering algorithms including distance-base clustering (K-means, hieratical clustering), self organize map

(SOM), density base clustering (EM and Autoclass) and principle component analysis (PCA). We will give examples of gene expression analysis base on these algorithms. Also we will discuss some special designed gene expression analysis methods based on these methods.

Part III : Advanced gene expression analysis : Basic technologies of advanced gene expression analysis such as gene regulatory network discovery will be introduced. Also we will introduce classification technology for identifying gene expression signature for biological facts (e.g. disease or drug effects).

PartIV: Integrated Gene Expression Analysis : This is the most important part of the tutorial. Analyzing gene expression data requires more discovery driven exploratory approach rather than conventional hypothesis-driven statistical study. A useful approach is to integrate expression analysis with "external" information such as sequences to understand the structure of co-expressed genes and to discover interesting motif in promoter regions or investigate important mutation. Also, by mapping the changes we observed in the mRNAs encoding enzymes onto metabolic pathways, we can infer the flow of metabolites through the systems.
The technology of automatically pathway discovery by generating statistical information from gene expression data to score possible pathway structures will also be introduced.

Part V : Applications of integrated analysis. This part we will mainly introduce the technology of pattern scoring which use "external" biology or literature information to assess and to verify the patterns generated from gene expression analysis. Such an approach can also be used in providing biology semantics to each discovered patterns (function annotation for patterns).