

Tutor: Zoltan Szallasi, M.D. is an assistant professor at the Department of Pharmacology, Uniformed Services University. He has been active in both the theoretical and experimental aspects of genetic network analysis of cancer. His lab is conducting cDNA microarray based large-scale gene expression measurements in breast cancer and developing experimental and computational analytical tools in order to achieve a more complex understanding of this disease.

Teaching experience: Zoltan Szallasi has been giving his genetic network tutorial for two years at several meetings including ISMB'99 (Heidelberg), PSB2000 and PSB2001 (Hawaii), and the First International Conference on Systems Biology (ICSB2000) in Tokyo. The tutorial is updated on a regular basis in order to incorporate the most recent relevant research achievements.

Motivation and goals of the tutorial: Technological developments such as cDNA microarray based RNA quantitation and proteomics have opened the opportunity for massively parallel biological data acquisition. This has shifted our attention towards a more complex understanding of molecular biology. In addition to determining the roles of individual genes, genetic network analysis enables us to study cells as a complex network of biochemical factors. The aim of this tutorial is to provide an introduction to this fast developing complex research field and provide an overview of its experimental, theoretical and computational foundations.

Audience: The tutorial is designed to benefit both molecular biologists who seek a good intuitive understanding of how theory and computer science can advance their work AND computer scientists who are looking for a better understanding of molecular biological data and ways to help biologists analyze genetic networks. By the end of the tutorial attendants can expect the following: to be able to make informed decisions about the appropriate experimental technologies they need to apply to solve a particular problem.; to be able to estimate the amount and nature of information they can obtain in large scale molecular biology; to be able to locate literature, softwares or collaborators for computational analysis of massively parallel experimental data sets. Participants will also understand the potential practical use of modeling and theory in experimental biology.

In addition to the tutorial the tutor is offering a well-written , up-to-date version of the material. (See attached a recent version of the tutorial handout.)

Length : Half day (approximately 3 hours)

Detailed outline of the presentation:

**1. Introduction: combinatorial biology, information content in biology, reductionism**

**2. Massively parallel data acquisition in biology, type of data, information content:**

- A brief review of massively parallel data acquisition methods
- The limitations of useful information content of these methods.

**3. The conceptual framework of genetic network analysis:**

- Estimating the size and complexity of real genetic networks.

**4. Genetic network modeling:**

- The relevance of genetic network modeling for experimental biology.
- small scale (1-3 genes) genetic network modeling to study the dynamics and stability of gene expression
- medium scale (10-100 genes) genetic network modeling in order to perform predictive forward modeling
- large scale (100-10000 genes) or ensemble approaches in order to study the overall behavior of genetic networks.
- Boolean genetic networks modeling
- Continuous differential equation based modeling
- Stochastic genetic network modeling.

**5. Reverse engineering of genetic regulatory networks:**

- Principles of reverse engineering
- Reverse engineering on continuous data - additive models solved by genetic algorithms, weight matrices etc.
- Reverse engineering on discrete data - REVEAL algorithm
- The information requirement of successful reverse engineering.
- Reverse engineering of chemical pathways using correlation matrices.

**6. Classification and cluster analysis in gene expression matrices:**

- the curse of dimensionality, internal data structure of biological measurements
- data preprocessing: multidimensional scaling, principal component analysis.
- The potential uses of data preprocessing

- Detailed discussion of unsupervised clustering methods highlighting their advantages and limitations, including:
  - Bottom up hierarchical clustering
  - Top down hierarchical clustering
  - Self organizing maps
  - K-means
  - Cluster identification via connectivity kernels. (CLICK)
  - Gene shaving
  - Percolation clustering
  
- Supervised learning methods - support vector machines.
  
- The use of cluster analysis in biology:
  - predicting novel upstream gene regulatory elements (Tavazie et al 1999)
  - predicting novel tumor subclasses (Golub et al, 1999)

### **7. Generative models in the analysis of gene expression matrices:**

- Statistical analysis of gene expression matrices is further complicated by gene co-regulation. Introducing generative models that will estimate the likelihood of chance separators and chance clusters in highly co-regulated gene expression matrices.

### **8. Systems approach to genetic networks and biology:**

- robustness as a key feature of genetic networks
- the possible effect of robustness on reverse engineering and internal data structure of gene expression matrices.