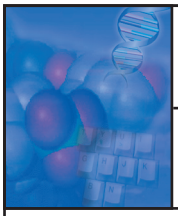

# Biology Applications for Computational & Mathematical Methods

## Eytan Domany, Chair

### Controlling the False Discovery Rate in Behavioral Genetics and Microarrays Analysis

**Yoav Benjamini**, Dept. of Statistics and Operations Research, Tel Aviv University

Any study based on statistical evidence is prone to produce false discoveries. Testing statistical hypotheses at the traditional 0.05 level is a way to limit the probability of producing such a false discovery when it is not real. Alas when many such tests are conducted in a study, the probability of producing a false discovery increases dramatically. On the other hand, limiting this probability in the traditional way incurs deterioration in the probability of detecting true discoveries.

The control of the false discovery rate (FDR) has been suggested as an intermediate approach to the problem. Procedures that control the FDR at a desired level are gaining popularity in areas of science where the problems encountered are large. Two such areas will be discussed in this talk – behavioral genetics and gene expression microarray data. The two will serve as a vehicle to discuss the FDR approach, the simple procedures that may be used to control it, and their properties.

### Large-Scale Clustering of Protein Sequences: Some Theory and Some Practice

**Nati Linial**, Dept. of Computer Science, Hebrew University of Jerusalem

In this presentation, I will review some of our recent work on the large-scale classification and analysis of proteins. I will explain some of the algorithmic concepts that underlie our work. Among the algorithmic tools that we employ are: Spectral data analysis, combinatorial algorithms to detect significant domains and small distortion metric embeddings.

### Cluster Analysis of Gene Expression Data

**Gad Getz**, Dept. of Physics and Complex Systems, Weizmann Institute of Science

A single microarray experiment allows simultaneous measurement of the expression level of thousands of genes. A typical experiment uses a few tens of such microarrays, each focusing on one sample - such as material extracted from a particular tumor. Hence the results of such an experiment contain several hundred thousand numbers, that come in the form of a table, of several thousand rows (one for each gene) and 50 - 100 columns (one for each sample).

We developed and applied a “data mining” method, called Coupled Two-Way Clustering} (CTWC), to extract biologically relevant data from such matrices. By an iterative clustering procedure the method reveals correlations that involve small subsets of genes and samples. The method can identify sets of correlated genes which usually belong to the same biological process or pathway, and uncover types and sub-types of diseases. The algorithm is designed to link cellular conditions to their relevant genes by finding conditional correlation among them. I will present results obtained from analyses of several types of cancer.

The CTWC method was applied by others and by us to numerous data types including gene expression data, antigen reactivity data, analysis of sugar compounds, document categorization and low-temperature phases of short range spin glasses.