# Novel Analysis Methods for Biosignature Data

Eric Mjolsness, Tobias Mann, and Barbara Wold*

JPL Machine Learning Systems Group
*Caltech Division of Biology

Biomolecular signatures of the future are likely include multiple modalities with the data characteristics of recent microarray gene expression instrumentation: sensitivity to thousands of molecular species, reusability enabling one to take data under hundreds of conditions including many time courses, and the problem of inferring aspects of potentially large networks of biochemical reactions far from equilibrium. We will discuss our current statistical, algorithmic and software approaches to these generalizable aspects of the microarray data, emphasising the need for objective comparisons between different methods.

As an example, we have considered statistical data models consisting of a simple set of clusters each modeled by a Gaussian distribution of diagonal covariance, fit to data by the Expectation-Maximization algorithm using cross-validation to choose the optimal number of clusters. This provides a statistically well-founded methodology for cluster analysis. We find that the log-likelihood scores of the flat clustering experiments have a broad plateau rather than a single peak as a function of cluster number, and that the plateau indicates that many classes are present in the data. Another model has two-level hierarchical Gaussian clusters: superclusters and subclusters. We find that the flat-clustering plateau is close to that expected if the data were actually distributed according to such a two-level hierarchical Gaussian model. We also show how cluster-mean time courses from gene expression data can be fit to continuous-time gene regulation network models.