## Predictive Genomics, Biology, Medicine

### Learning theory: SLT – what is it?

Parametric statistics – small number of parameters – appropriate to small amounts of data

**Ex.** Find mean m and standard deviation s for a normal distribution from sample data.

Nonparametric statistics – large number of parameters – appropriate to large amounts of data

**Ex.** Neural Network, RBF network, support vector machine

### Genomics: Current interests:

**New algorithms for classification of and prediction from microarray gene expression data.**

Genome: about 50,000 genes

Gene expression in cell reflects physiological factors and processes.

Discovery of patterns in gene expression data: major computational challenge.

Includes genome and genetic regulation and expression information.

Information important in diagnosing physiological factors, e.g.:

- nature of disease, e.g. tumor
- state and prognosis for a genetically inherited disease

**Technology:** new, error-prone - statistical analysis must tease apart errors as well as many physiological factors present. Current methods of classification may not be as effective or accurate as they can be.

Understanding physiological correlates of gene expression (hence protein expression) promises to provide insight into conditions and diseases whose etiologies have been difficult to understand, e.g.:

- autism
- multiple sclerosis
- muscular dystrophy
- propensities for cancers and arteriosclerosis,
- Alzheimer's disease

Preliminary results have been obtained in these areas.

Purpose of project: work on aspects of such an approach.

Our work involves modeling, simulation, and algorithm based approaches to classification and prediction of cell physiology from microarray information.

Major aspect: deal with numerical simulations and their complexity.

- Emphasize accuracy of statistical models
- Computed algorithm discovery methods will search for algorithms appropriate to models.
- Subarray cocluster patterns (patterns occurring for subsets of genes and of the population).
- Computational demands require the high performance resources of Center for Computational Science at BU
- Statistical models of microarray experiments: Gene Expression Data Simulator (GEDS) at University of Pittsburgh

- Error of classification, prediction algorithms calculated with Monte Carlo simulations on GEDS

- Algorithms for discovery of subarray coclusters, testing sparse data for underlying distribution families, extending regression-attraction algorithm.

- Will also develop local numerical algorithms for "customized" predictions for individuals from microarrays.

Collaboration:

- Boston University (Mathematics and Statistics, Microarray Resource at the Medical School, and Center for Computational Science; Bioinformatics Program

- University of Massachusetts Lowell (Mathematics and Statistics)

- University of Pittsburgh Medical School (Medical School microarray core laboratory; UPCI Cancer Biomarkers Laboratory, PittArray core laboratory)

- Ben Gurion University in Israel (BGU Human Molecular Genetics Lab, Computer Science Department's Bioinformatics Program)

Goals: answer questions -

1. **Can computer implementations of microarray models be used to improve them?**

2. **Can model and parameter determination be accomplished computationally?**
3. **Can statistical algorithms to solve the models be tested, developed, and improved on such models?**

4. **Can statistical methods improve the yield of microarray information for small numbers of subjects?**

5. **Can sub-patterns (patterns in subsets of the genome *and* population) in microarray data be verified, discovered, and used?**

6. **What are maximal levels of information which can be obtained from gene expression information? Can we obtain probabilities that a queried patient belongs to a given trained group together with confidence bounds?**

7. **What can simulation of the genetic expression profile of cancer cells reveal about potential responses to therapies?**

*Statistical methods* work better when they incorporate *biological models* as a priori information.

Strategy: divide translation of physiological models into algorithms into three parts:

- biological modeling
- statistical modeling
- algorithm development

From biology to statistical modeling: two way process

biology → statistical model → simulated biological data (with scalable microarray simulation).

After statistical model is decided on: find algorithms which solve model – given complexity of good algorithms, we will use Monte Carlo to gauge efficiency via microarray simulator.

**Further study:** automated algorithm development via search methods within algorithm classes

## Co-regulation of genes:

**Many new methods** (e.g. graph-theoretic methods of cataloguing coregulation from published literature)

Will study *automated methods* of incrementing statistical models with information, and incorporating models into simulator. Simulator will allow testing models, via comparisons of simulator and biological data. Such objective tests of models do not presently exist.

More specific aims:

**1:** Develop tests of statistical models of microarrays through comparisons with computational simulations, and develop new models and methodologies on this basis, and discover and modify algorithms for these models.

**2:** Develop optimal robust classification algorithms for microarrays based on models, with probability estimates of classification membership and confidence bounds, and develop statistical methods for reducing patient sample sizes necessary.

**3:** Test classification algorithms and improve statistical properties through Monte Carlo simulation of accuracies, and use (low and eventually high dimensional) search techniques to find better algorithms.

**4:** Study search algorithms for discovering subarrays containing patterns not visible in full arrays.

**5:** Test and apply these methodologies to existing cancer databases for differentiating cancer gene expression information.

**6:** Apply these methods to develop software for practitioners using microarrays.

Outcomes: new tools for differentiating microarray clusters will be available

1. Information based on clustering of interacting genes and sub-populations affected by them will be obtainable from microarray analysis.
2. More accurate statistical models of microarrays will be implementable and testable using microarray simulator under development at the University of Pittsburgh.
3. Classification algorithms with tunable parameters (appropriate to different biological models) will be available, with class probability estmates and confidence bounds.
4. Applications of the above techniques to the development of diagnostic tools for differentiating cancer gene expression information will be developed
5. Open source software implementing this work will be available.

Emphasize implementable algorithms for diagnosis, classification, and prediction.

Differentiation of cancer gene expression profiles has the potential to greatly improve the use of cancer therapies.

Extend also to psychiatric drugs in which responsiveness to therapies seems often to be individual parameter.

Approach: separation of model from algorithm.

Modeling: biological problem

Once model is found, finding algorithm which decides which class (e.g., metastatic or non-metastatic tumors) microarray comes from becomes a

purely statistical and computational; notions of complexity and optimality then become appropriate and well-defined.

**Correspondingly, errors from classification algorithms can be broken down into two parts: model error and algorithmic error.**

Model error: biology not correctly modeled

Algorithmic error: correct statistical model exists, but classification algorithms developed for model have associated error making them worse than optimal algorithms.

Jim Lyons-Weiler and team (Pittsburgh) have developed microarray simulation tool (located at http://bioinformatics.upmc.edu/GE2/index.html), in which model can be adjusted, and algorithms can be simulated.

More complicated algorithms Claudio Rebbi, director of BU's Center for Scientific Computation.

**Differential Expression.** Jittering occurs between $t_1$ and $t_2$ in Fig. 1b. Differential expression occurs as an independent process after jittering. For any differentially expressed gene *i*, expression values in sample group **A** are individually and stochastically shifted (all up or all down) using replacement values $x_i + \Delta x_i$. The maximum value of $\Delta x_i$ is determined by parameter $\Delta X_{AB,}$ which is expressed in units of standard deviation (sd). We typically simulate over a range of ($\Delta X_{AB}$ 0.25 to 3.0).
Each leaf in Fig. 1 represents a sample in group A or B in our simulated data set.

**Fig. 1.** A Case vs. Control Pattern of Inheritance Model in Detail. Between group correlations specified by $\Delta r_{AB}$ and within group correlations are specified by $\Delta r_A$ for group A and $\Delta r_B$ for group B.

Fig. 1A.

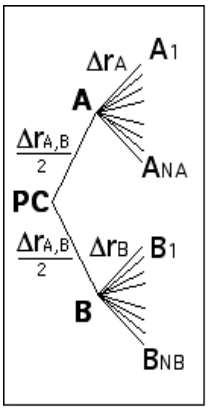



Fig. 1B.

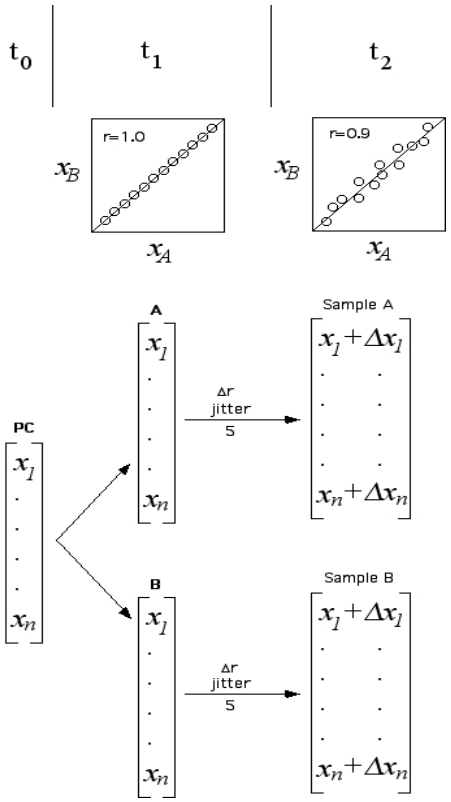
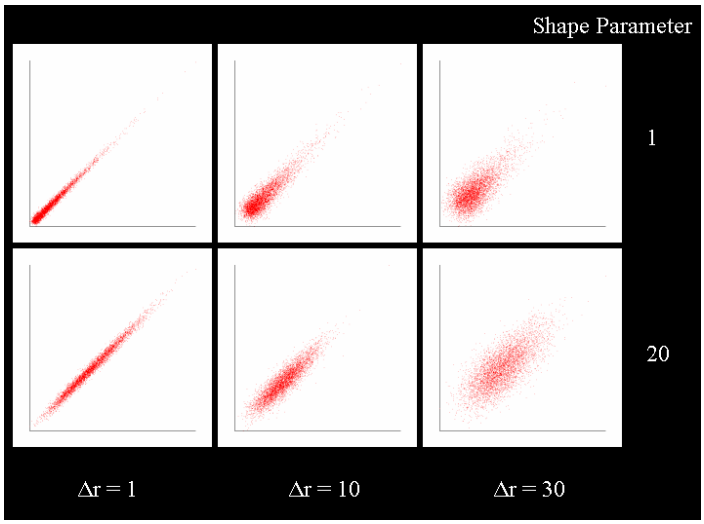



**FIG. 2.** Outcome of a jittering process to produce correlations between two arbitrary samples *i* and *j* (1,000 genes). Each biplot represents expression levels for *i* and *j* drawn from two gamma distribution shape parameter values (1= skewed; 20 = normal) over the range of expected correlation between *i* and *j* (determined by $\Delta r_{ij}$) to demonstrate the type of data that can be generated by the Gene Expression Data Simulator. In jittering, random genes are selected to be changed stochastically by a maximum amount v1. The between-sample correlation is measured, and if the target *r* is achieved, jittering stops. If not, then another gene is selected to be changed. The process continues until the target correlation is achieved.




**Example of the bivariate output.** Modeled expression intensities were generated for two samples for three levels of Δr under two gamma distribution shapes (Fig. 2). This result demonstrates that the simulator can approximate very well biologically realistic data sets with stochastic error and the desired correlation.

## PREDICTIVE MEDICINE:

Cancer markers: Size of tumor, past historical information, patient biomarkers, genomic information

Microarray markup language → biomarker markup language (need for NIH-approved standardized language)

Goal: database into which all kinds of information can be integrated.

Inference engine: dichotomy – mini-engine and meta-engine (boosting and bagging algorithms)

Medical applications: patient state is time dependent;

x = uncontrolled variables (e.g., cancer etiology, individual biomarkers and genetic markers)

y = controlled variables (patient treatment, drugs administered, etc.)

$$z = (x,y)$$

$$z(t+1) = f(z(t))$$

Learning: Discover the function f(t) from databases of examples

Control theory: how to adjust y(t) (controlled variables) so that disease history z(t) progresses as well as possible?

Financial mathematics – algorithms there also apply to control theory aspects here

Stochastic differential equations

$$d\mathbf{x}/dt = \mathbf{B}'(t) + \mathbf{b}(\mathbf{x})$$

C. Rebbi, J. Luciano: simulations (Matlab nlinfit program suffices) – psychiatric data, simulated cancer data