Harvard-MIT Division of Health Sciences and Technology
HST.512: Genomic Medicine
Prof. Marco F. Ramoni

Machine Learning Methods for Microarray Data Analysis

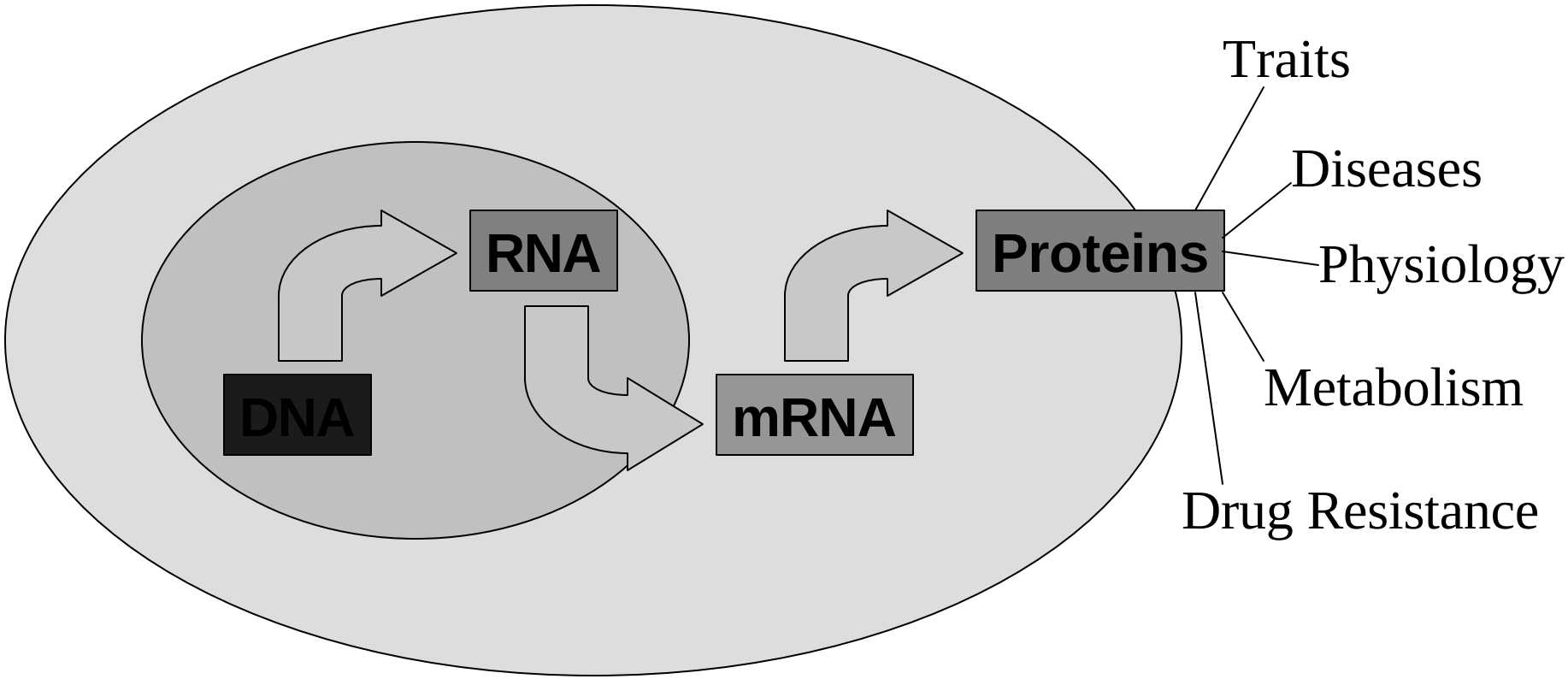
Marco F. Ramoni
Children's Hospital Informatics Program
Harvard Partners Center for Genetics and Genomics
Harvard Medical School

HST 512

Outline

- Supervised vs Unsupervised
- Supervised Classification
  - ✓ Definition
  - ✓ Supervision
  - ✓ Feature Selection
  - ✓ Differential Analysis
  - ✓ Normalization
- Prediction and Validation
  - ✓ Probabilistic
  - ✓ Voting Schemes
  - ✓ Independent Validation
  - ✓ Cross Validation
- Clustering
  - ✓ One-dimensional
  - ✓ Self Organizing Maps
  - ✓ Hierarchical
  - ✓ Bayesian
  - ✓ Temporal
- Bayesian networks
  - ✓ Definitions
  - ✓ Learning
  - ✓ Prediction
  - ✓ Validation

Central Dogma of Molecular Biology



Functional Genomics

- Goal: Elucidate functions and interactions of genes.
- Method: Gene expression is used to identify function.
- Tools: Characteristic tools of functional genomics:
  - ✓ High throughput platforms.
  - ✓ Computational and statistical data analysis.
- Style: The intellectual style is different:
  - ✓ Research is no longer hypothesis driven.
  - ✓ Research is based on exploratory analysis.
- Issue: Functional genomics is in search of a sound and accepted methodological paradigm.

Microarray Technology

Scope: Microarrays are reshaping molecular biology.

Task: Simultaneously measure the expression value of thousands of genes and, possibly, of entire genomes.

Definition: A microarray is a vector of probes measuring the expression values of an equal number of genes.

Measure: Microarrays measure gene expression values as abundance of mRNA.

Types: There are two main classes of microarrays:

cDNA: use entire transcripts;

Oligonucleotide: use representative gene segments.

Measuring Expression

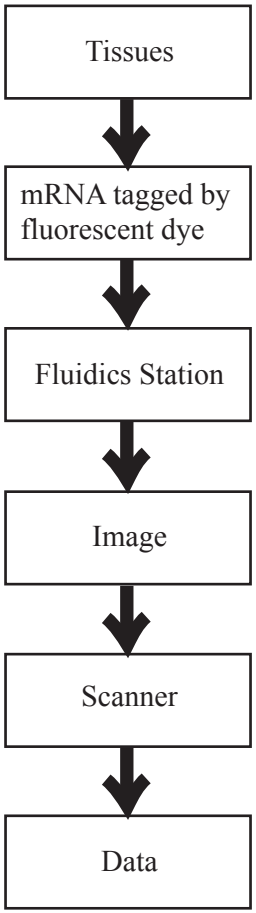
Rationale: Measurement of gene expression reverses the natural expression process.

Hybridization: Process of joining two complementary strands of DNA or one each of DNA and RNA to from a double-stranded molecule.

Artificial process: Backward the mRNA production.

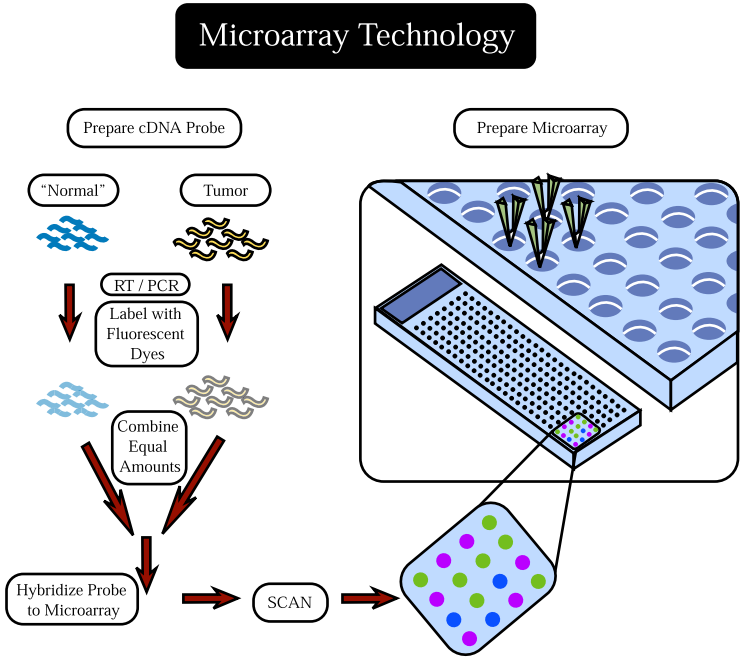
- ✓ DNA samples (probes) are on the microarray.
- ✓ Put cellular labeled mRNA on the microarray.
- ✓ Wait for the sample to hybridize (bind).
- ✓ Scan the image and, for each point, quantify the amount of hybridized mRNA.

From Tissues to Microarrays



cDNA microarrays

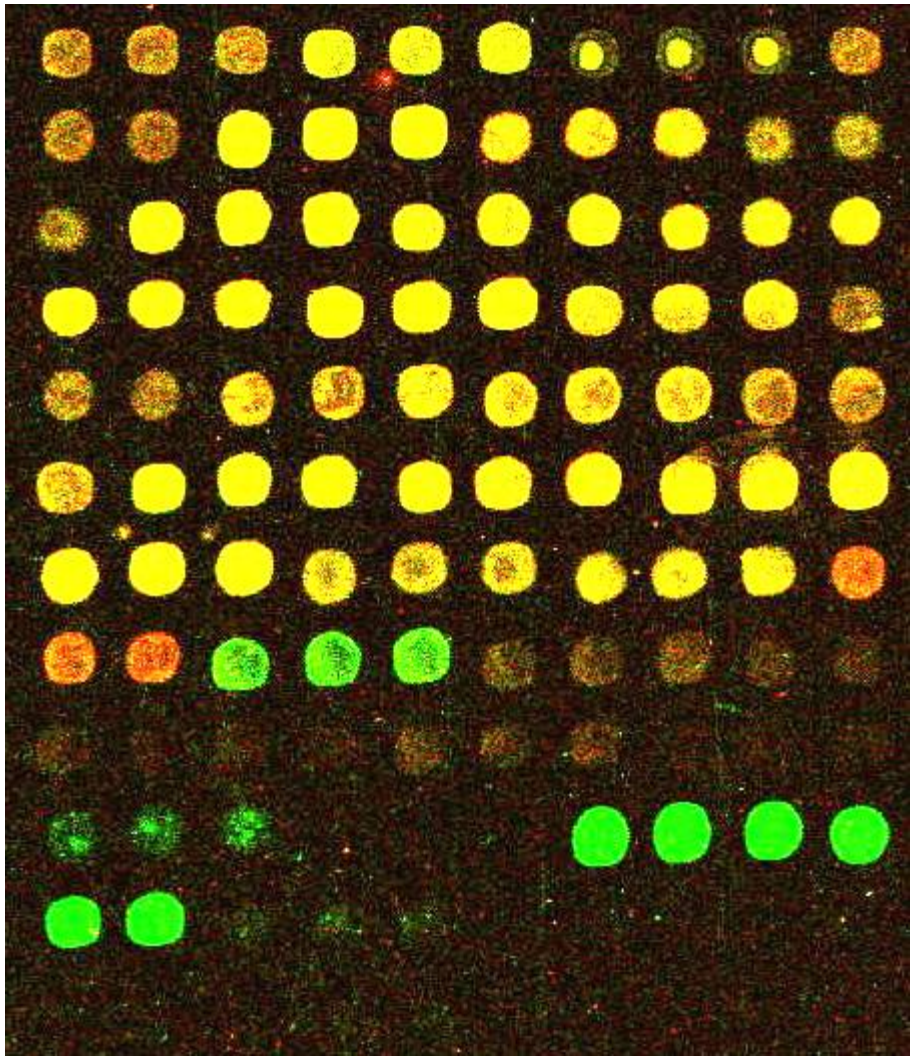
Fix, for each gene, many copies of two functional DNA on a glass.



The labeled probes are allowed to bind to complementary DNA strands on the microarray.

Fluorescent intensity in each probe measures which genes are present in which sample.

cDNA Microarray Data



Green: genetic material is present in the control but not in the treated sample.

Red: genetic material is only present in the treated sample but not in the control.

Yellow: genetic material is present in both samples.

Gray: genetic material is not contained in either samples.

Oligonucleotide Microarrays

- Oligonucleotide arrays : Affymetrix genechip.
- Represent a gene with a set of about 20 probe pairs:
  - ✓ Each probe (oligonucleotide) is a sequence of 25 pairs of bases, characteristic of one gene.
- Each probe pair is made by:

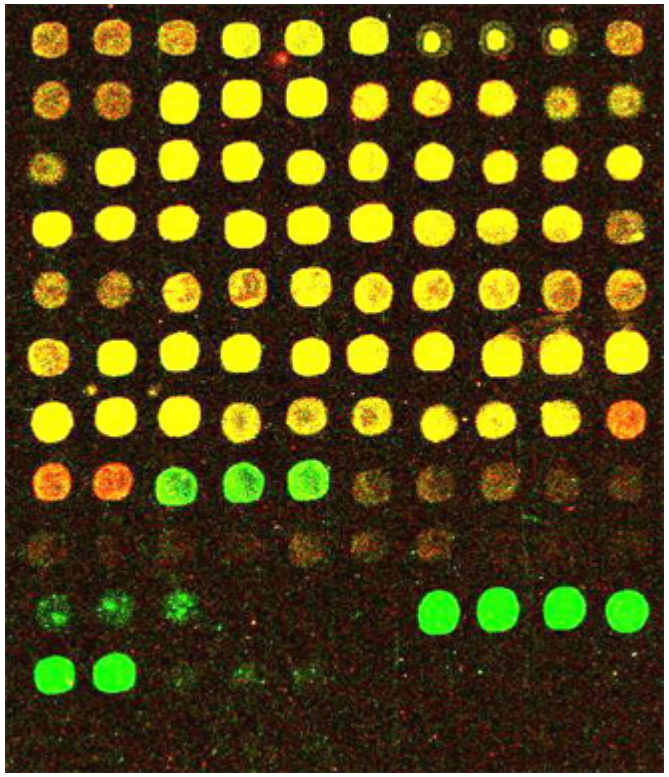
  Perfect match (PM): a probe that should hybridize.

  Mismatch (MM): a probe that should not hybridize, because the central base has been inverted.

PM	ATGAGCTGATGCCATGCCATGAGAG
MM	ATGAGCTGATGCGATGCCATGAGAG

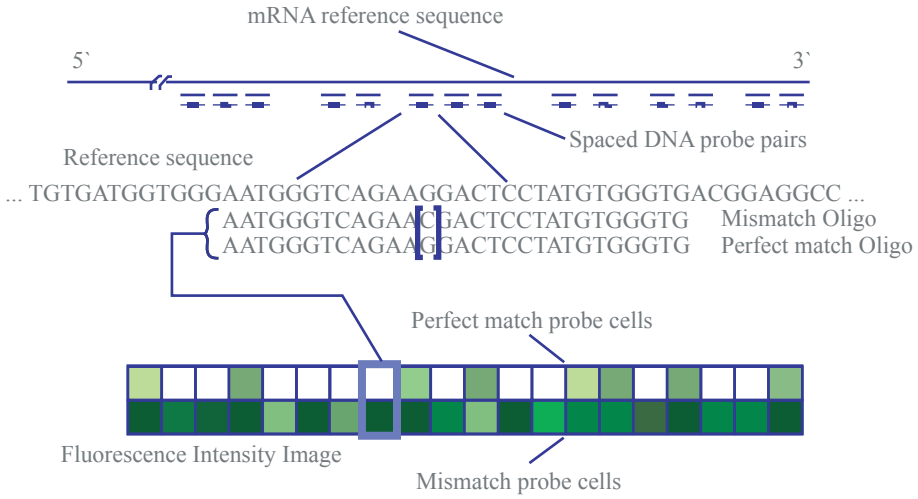
Oligonucleotide Microarray Data

Scanned microarray



Intensity: Gene expression level is quantified by the intensity of its cells in the scanned image.

Each cell measures the expression level of a probe.



Expression = avg(PM-MM)

Expression Measures

Definition: Expression is calculated by estimating the amount of hybridized mRNA for each probe as of quantity of its fluorescent emission.

Design: Different microarrays are designed differently:

- cDNA: Combine conditions in paired experiments.
- Oligonucleotide: Independent measures.

Experiments: Require different design per platform:

- cDNA: One array for an experimental unit.
- Oligonucleotide: 2 arrays for a experimental unit.

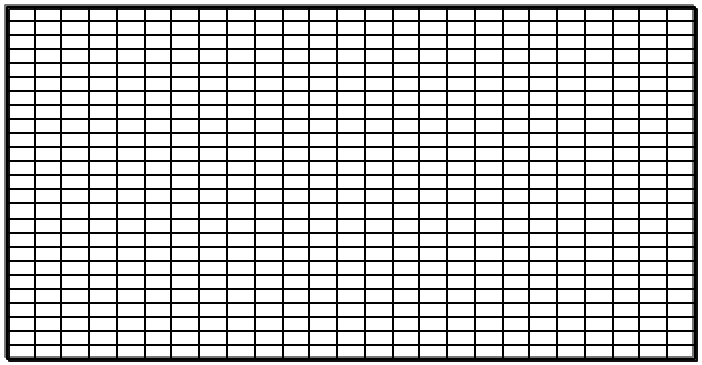
Statistical Challenges

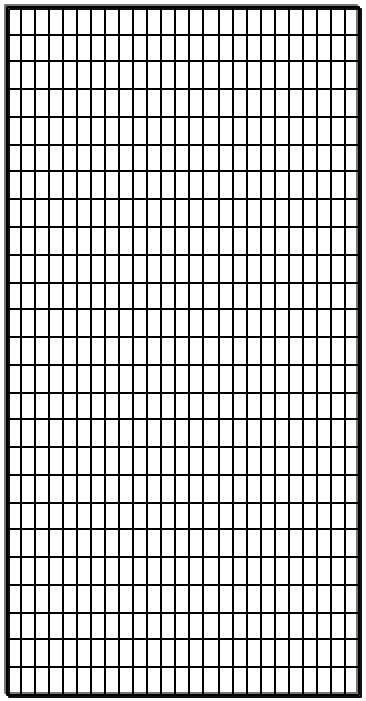
Small N large P: Many variables, few cases.

Noisy results: Measurements are vary variable.

Brittle conditions: Sensitive to small changes in factors.

Design: Platforms are designed without considering the analysis to be done.





Supervised vs Unsupervised

Elements: Features (genes) and a training signal (class).

Question: Which function best maps features to class?

Goal: Find a good predictive system of class (e.g. build a system able to take a patient and return a diagnosis).

Assumption: Different features are best predictor.

Task: Estimation (except for feature selection, the task of finding the best predictors).

Elements: Features (genes) but no training signal.

Question: Which features behave in a related (similar) way across experiments?

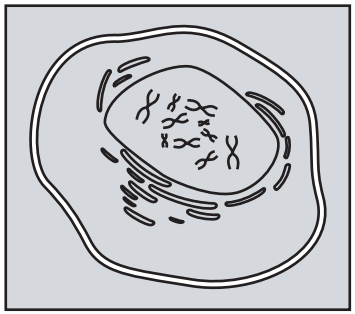
Goal: Understand interaction (e.g. how genes behave similarly under certain experimental conditions).

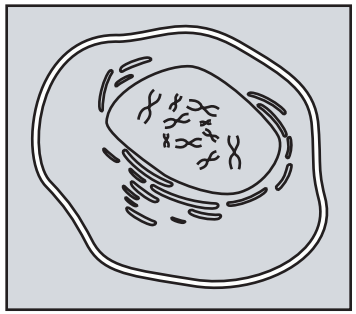
Assumption: Same behaviors mean same functional class.

Task: Model selection.

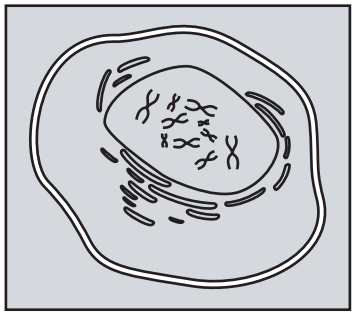
Comparative Experiments

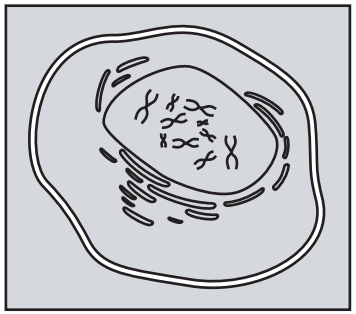
Healthy Cell



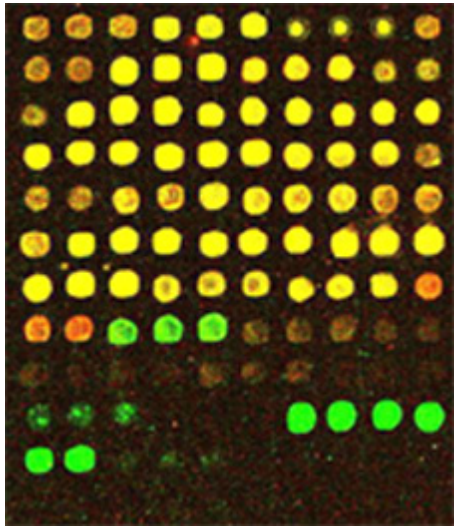


Tumor Cell





Sample 1 Sample 2 Sample 3 Sample 4



Samples k=1,…,n_i

genes g=1,…,G

Gene.Desc	S1	S2	S3	S4	S5
AFFX-BioC	88	283	309	12	168
hum_alu_a	15091	11038	16692	15763	18128
AFFX-Dap	311	134	378	268	118
AFFX-LysX	21	21	67	43	8
AFFX-HUM	215	116	476	155	122
AFFX-HUM	797	433	1474	415	483
AFFX-HUM	14538	615	5669	4850	1284
AFFX-HUM	9738	115	3272	2293	2731
AFFX-HUM	8529	1518	3668	2569	316
AFFX-HUM	15076	19448	27410	14920	14653
AFFX-HUM	11126	13568	16756	11439	15030
AFFX-HUM	17782	18112	23006	17633	17384
AFFX-HSA	16287	17926	22626	15770	16386

y_{gik}

Identify genes that are differentially expressed in two conditions i=A,B.

Comparative Experiments

Case Control: Asses how many times a gene is more (less) intense in one condition than in another.

Elements: Condition = training signal; genes = features.

Measure of differential expression:

$$\text{fold} = \frac{\bar{y}_{gA}}{\bar{y}_{gB}} \qquad \text{difference} = \frac{\bar{y}_{gA} - \bar{y}_{gB}}{s_g}$$

Threshold: decide a threshold, to select genes that are "significantly" differentially expressed.

Rationale: A particular experimental condition creates differences in expression for some genes.

Distribution Free Tests

Permutation tests to identify gene specific threshold:

SAM (Stanford) uses a statistic similar to the classical t-statistic. The parameter a is chosen to minimize the coefficient of variation.

GeneCluster (Whitehead) uses signal-to-noise ratio statistic.

Problem: p-values in multiple comparisons – corrections make impossible to identify any change.

$$t_g = \frac{\bar{y}_{gA} - \bar{y}_{gB}}{a + \sqrt{\frac{S_{gA}^{\ 2}}{n_A} + \frac{S_{gB}^{\ 2}}{n_B}}}$$

$$s2n_g = \frac{\bar{y}_{gA} - \bar{y}_{gB}}{\frac{S_{gA}}{\sqrt{n_A}} + \frac{S_{gB}}{\sqrt{n_B}}}$$

Supervised Classification

Goal: A predictive (diagnostic) model associating features to class.

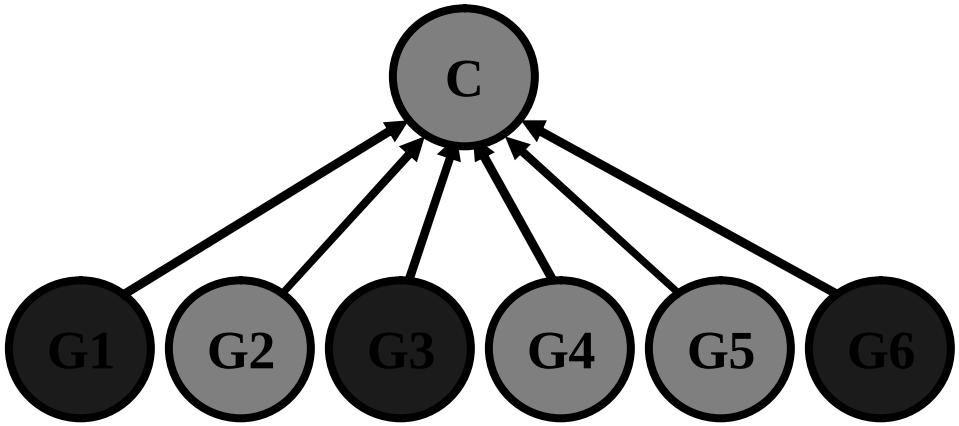
Rationale: Difference is an indicator of predictive power.

Components: Dataset of features and a training signal.

Features: Gene expression levels in different classes.

Training signal: The class label.

Feature selection: Find the best predictors to maximize accuracy.



Feature Selection

Task: Identify those genes that best predict the class.

Advantage I: Typically increases predictive accuracy.

Advantage II: More compact representation.

Advantage III: Provide insights into the process.

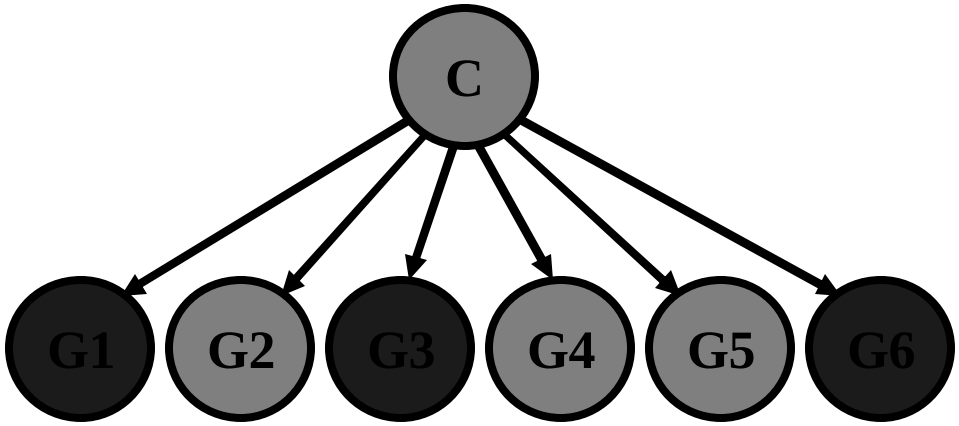
Type of Task: Model selection.

Differential analysis: A special case (binary) of feature (the most discriminating genes) selection.

Rationale: Since we cannot try all combinations, most different features should be the best at discriminating.

Parametric Methods

- A simple approach to prediction is to assume that the features (genes) are conditionally independent given the class.
- These models are called Naïve Bayes Classifiers.
- Estimate, for each gene, the probability density of the gene given each class: p(g|c).
- The challenge is to identify the right distribution.



Prediction

- Once a mapping function (or a model + function for feature selection) has been identified, we can use this function to classify new cases.
- Non parametric methods do not provide explicit functions to map features to class.
- Mixture of Experts is a weighted voting algorithm to make prediction from non parametric models.
- Intuitively, in a weighted voting algorithm:
  - ✓ Each gene casts a vote for one of the possible classes and.
  - ✓ This vote is weighted by a score assessing the reliability of the expert (in this case, the gene).
  - ✓ The class receiving the highest will be the predicted class.

Parametric Prediction

Analysis: Suppose the analysis leads to select a group of genes which are differentially expressed across the two conditions.

Prediction: we may want to classify new samples on the basis of their expression profile *z* (molecular diagnosis):

$$p(\text{class} = i \mid \text{sample molecular profile}) = \text{p}(\text{class} = i \mid z)$$

Bayes rule: choose the maximum probability classification.

$$p(\text{class} = i \mid z) \propto f(z \mid \text{class} = i)\, p(\text{class} = i)$$

$$f(z \mid \text{class} = i) = \prod_j \{ f(z_j \mid i, M_{gj})\, p(M_{gj}) + f(z_j \mid i, M_{lj})\, p(M_{lj}) \}$$

Assumptions: gene independent given class and parameters.

Predictive Validation

Prediction: To assess the validity of a classification system (either a function or a model + function), we can use an independent labeled data set and predict the class of each case with the generated system. Or split a sample in two sets:

Training set: a data set used to build the model/function;

Test set: a labeled data set to predict with the model/function.

Cross Validation: When an independent test set is not available, we can use cross validation:

1. Split the sample in k subsets;
2. Predict one subset using the other k-1 subsets to build the model/function;
3. Repeat the operation predicting the other sets.

Leave one out: for small samples, use single cases as k sets.

An Example

Example: Acute lymphoblastic leukemia (27) vs acute myeloid leukemia (11).

Method: Correlate gene profiles to an “extreme” dummy vector of 0s and 1s.

Results: 50 genes on each side.

Please see Figure 3b of Science. 1999 Oct 15; 286 (5439):531-7.

Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Calligiuri MA, Bloomfield CD, Lander ES.

Normalization

- An attempt to solve the problem of small sample size is to use "normalization" – a technique to reduce the variance.
- Normalization is an accepted procedure to balance the two channels of a cDNA microarray.
- When oligo microarrays were introduced, some tried to apply some form of variance reduction under the name of normalization to this new platform that has NO paired experiments.
- There are hundreds of different "normalization" methods.

Please see Figure 1 of Nat Rev Genet. 2001 Jun; 2(6):418-27.
Computational analysis of microarray data.
Quackenbush J.

Normalization?

Unsupervised Methods

- Differential experiments usually end up with:
  - ✓ A list of genes changed across the two conditions;
  - ✓ A “stochastic profile” of each condition.
- Useful to identify diagnostic profiles and prognostic models.
- They are not designed to tell us something about regulatory mechanisms, structures of cellular control.
- With supervised methods, we look only at relations between gene expression and experimental condition.
- Unsupervised methods answer different experimental questions.
- We use unsupervised methods when we are interested in finding the relationships between genes rather than the relationship between genes and a training signal (eg a disease).

One Dimensional Clustering

Strategy: Compute a table of pair-wise distances (eg, correlation, Euclidean distance, information measures) between genes.

Clustering: Use permutation tests to assess the cut point.

Relevance networks: Create a network of correlated genes and remove the links below the chosen threshold.

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Gene 1	1	0.2	0.8	-0.3	0.5	0.7	0.1
Gene 2		1	0.5	0.6	-0.2	-0.5	0.3
Gene 3			1	0.2	0.1	-0.2	0.1
Gene 4				1	0.9	0.4	0.3
Gene 5					1	0.1	-0.4
Gene 6						1	0.1
Gene 7							1

Please see Figure 2 of Proc Nati Acad Sci U S A. 2000 Oct 24;97(22):12182-6.

Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS.

Hierarchical Clustering

Components: Expression profiles, no training signal.

Method: Sort the expression profiles in a tree a using a pair-wise similarity measure (say, correlation) between all the profiles.

Model: Build a single tree merging all sequences. Use the mean of each set of merged sequences as representation of the joint to traverse the tree and proceed until all series are merged.

Abstraction: When two genes are merged, we need to create an abstract representation of their merging (average profile).

Recursion: The distance step is repeated at each merging until a single tree is created.

Clustering: Pick a threshold to break down the single tree into a set of clusters.

Eisen et al., *PNAS* (1998)

Dendrogram

Please refer to Curr Opin Mol Ther. 1999 Jun;1(3):344-58.

Modified oligonucleotides-synthesis, properties and applications.

Lyer RP, Roland A, Zhou W, Ghosh K.

Two Dimensional Clustering

- We want to discover an unknown set of patient classes based on an unknown set of gene functional classes.
- A two-dimensional optimization problem trying to simultaneously optimize distribution of genes and samples.
- Survival time (KL curves) were used as independent validation of patient clusters.

Please see Figures 1 and 5 of Nature. 2000 Feb 3;403(6769):503-11.

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.

Alizadeh AA, et al.

Bayesian Clustering

Problem: How do we decide that N genes are sufficiently similar become a cluster on their own?

Similarity: Profiles are “similar” when they are generated by the same stochastic process.

Example: EKGs are similar but not identical series generated by the a set of physiological process.

Clusters: Cluster profiles on the basis of their similarity is to group profiles generated by the same process.

Bayesian solution: The most probable set of generating processes responsible for the observed profiles.

Strategy: Compute posterior probability p(M|D) of each clustering model given the data and take the highest.

Ramoni et al., *PNAS* (2002)

Posterior Probability

- We want the most probable model given the data:

$$p(M_i \mid \Delta) = \frac{p(M_i, \Delta)}{p(\Delta)} = \frac{p(\Delta \mid M_i)\, p(M_i)}{p(\Delta)}$$

- But we use the same data for all models:

$$p(M_i|D) \propto p(D \mid M_i) p(M_i).$$

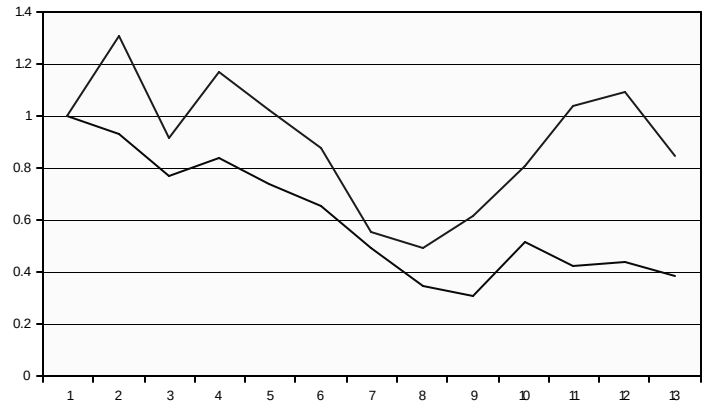
- We assume all models are a priori equally likely:

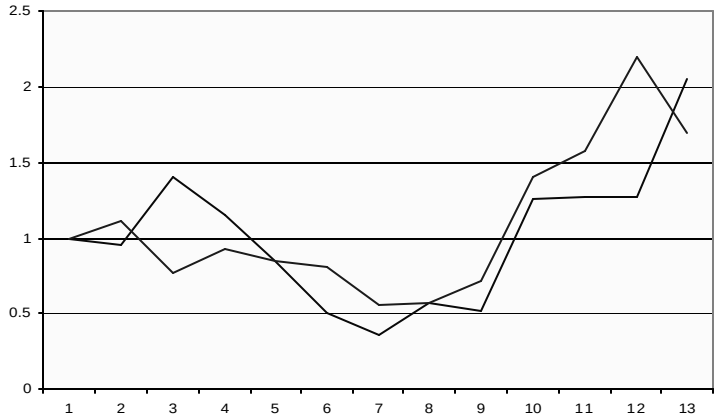
$$p(M_i|D) \propto p(D \mid M_i).$$

- This is the marginal likelihood, which gives the most probable model generating Δ.

Temporal Clustering

- A process developing along time (eg yeast cell cycle).
- Take microarray measurements along this process (2h for 24h).
- Cannot use standard similarity measures (eg correlation) because observations are not independent.
- Need a model able to take into account this dependence of observations
- Our perception of what is similar may be completely different under these new conditions.





Autoregressive Models

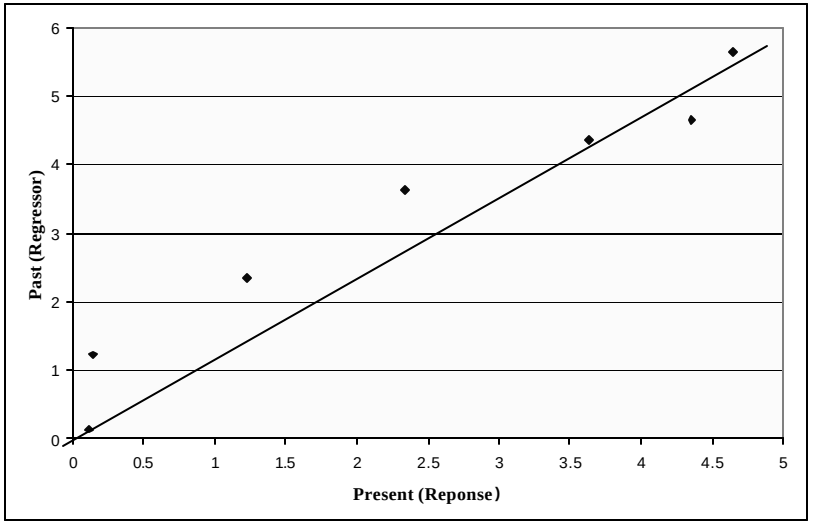
- Take a time series, of dependent observations:

$$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \ldots$$

- Under the assumption is that t_0 is independent of the remote past given the recent past:

$$P(x_t \mid x_0, x_1, \ldots, x_{t-1}) \quad \text{——} \quad P(x_t \mid x_{t-p}, \ldots, x_{t-1})$$

- The length of the recent past is the Markov Order *p*.



Networks

- Clustering rests on the assumption that genes behaving in similar ways belong to the same process.
- The result of a clustering model is to break down the set of all genes into boxes containing genes belonging to the same process.
- However, clustering tells us nothing about the internal mechanisms of this control structures: it provides boxes, not chains of command.
- To discover chains of command, we need to resort to a new approach: Bayesian networks.

Bayesian Networks

- Bayesian networks (also called Causal probabilistic networks) were originally developed to encode human experts' knowledge, to they are easily understandable by humans.
- Their two main features are:
  - ✓ The ability to represent causal knowledge to perform diagnosis, prediction, etc.
  - ✓ They are grounded in statistics and graph theory.
- Late '80s, people realize that the statistical foundations of Bayesian networks makes it possible to learn them from data rather than from experts.

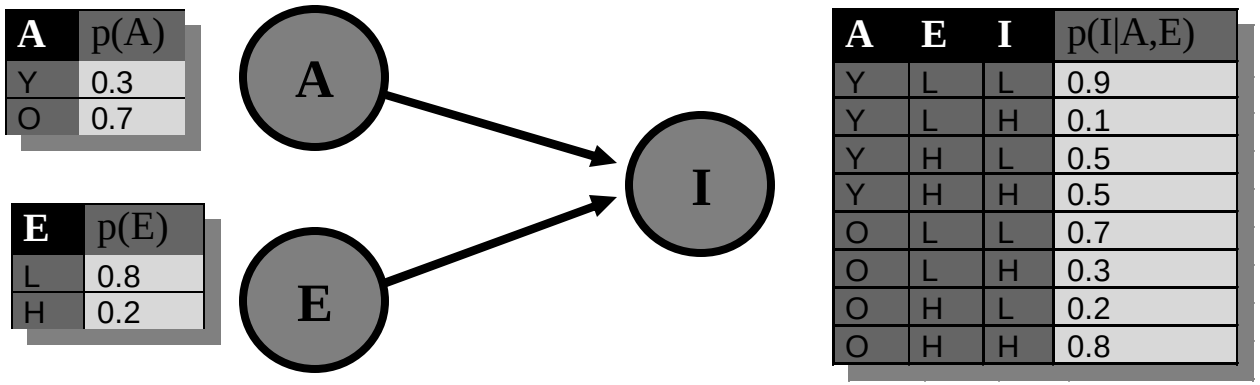
Components

Qualitative: A dependency graph made by:

Node: a variable X, with a set of states $\{x_1,\ldots,x_n\}$.

Arc: a dependency of a variable X on its parents Π.

Quantitative: The distributions of a variable X given each combination of states π_i of its parents Π.



A	p(A)
Y	0.3
O	0.7

E	p(E)
L	0.8
H	0.2

A	E	I	p(I|A,E)
Y	L	L	0.9
Y	L	H	0.1
Y	H	L	0.5
Y	H	H	0.5
O	L	L	0.7
O	L	H	0.3
O	H	L	0.2
O	H	H	0.8

A=Age; E=Education; I=Income

Learn the Structure

- In principle, the process of learning a Bayesian network structure involves:
  - ✓ Search strategy to explore the possible structures;
  - ✓ Scoring metric to select a structure.
- In practice, it also requires some smart heuristic to avoid the combinatorial explosion of all models:
  - ✓ Decomposability of the graph;
  - ✓ Finite horizon heuristic search strategies;
  - ✓ Methods to limit the risk of ending in local maxima.

An Application

Cases: 41 patients affect by leukemia.

- Genomic: expression measures on 72 genes;
- Clinical: 38 clinical phenotypes (3 used).

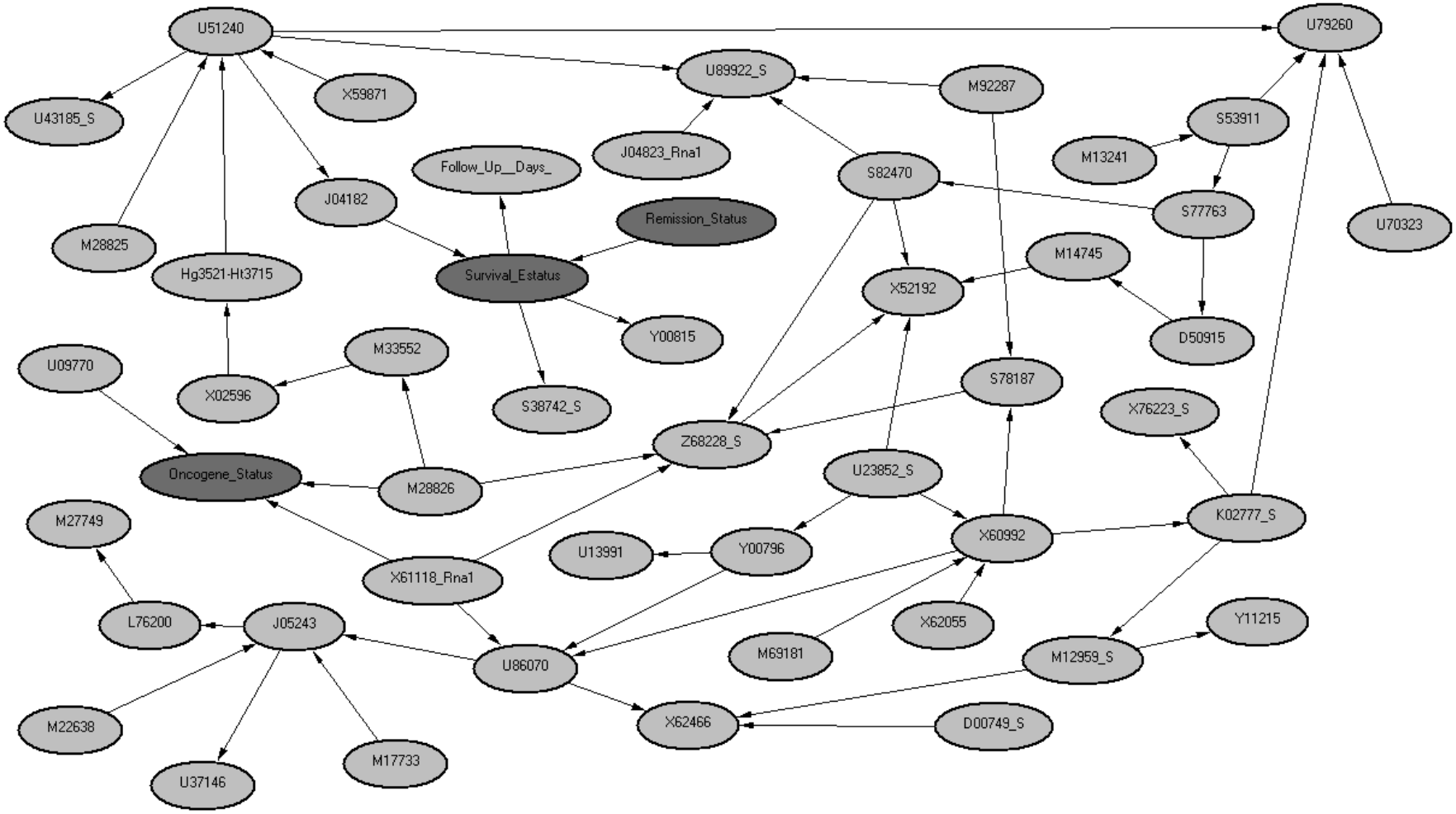
Representational Risks:

- Deterministic links: hide other links more interesting.
- Overfitting: Too many states for the available data.

Transformations:

- Definitional dependencies: if suspected, removed.
- Sparse phenotypes: consolidated (oncogene status).

The Network


U51240
U79260
U89922_S
X59871
M92287
U43185_S
S53911
M13241
J04823_Rna1
Follow_Up__Days_
S82470
J04182
S77763
U70323
Remission_Status
M28825
M14745
Hg3521-Ht3715
Survival_Estatus
X52192
Y00815
D50915
M33552
U09770
S78187
X02596
S38742_S
X76223_S
Z68228_S
Oncogene_Status
M28826
U23852_S
K02777_S
M27749
X60992
U13991
Y00796
X61118_Rna1
L76200
J05243
X62055
Y11215
M69181
M12959_S
U86070
M22638
X62466
D00749_S
M17733
U37146

Dependency Strength

Bayes factor: ratio between the probability of 2 models.

Threshold: To add a link, we need to gain at least 3 BF.

X61118_Rna1	M28826	U09770		1
U09770	X61118_Rna1	M28826	S53911	7
U09770	X61118_Rna1	M28826	M69181	56
X61118_Rna1	M28826	M17733		63
X61118_Rna1	M28826	S77763		315
U09770	X61118_Rna1	M28826	U43185_S	447
U09770	X61118_Rna1	M28826	J05243	447
X61118_Rna1	M28826			973
X61118_Rna1	M28826	S36742_S		1016
X61118_Rna1	M28826	J05243		1534
U09770	X61118_Rna1	M28826	M17733	1804
U09770	X61118_Rna1	M28826	Z68228_S	1807
U09770	X61118_Rna1	M28826	J04823_Rna1	3558
U09770	X61118_Rna1	M28826	U37146	3564
U09770	X61118_Rna1	M28826	X62055	3564
U09770	X61118_Rna1	M28826		3570
X61118_Rna1	M28826	Y11215		3933
X61118_Rna1	M28826			4254
U09770	X61118_Rna1	M28826	Survival_Estatus	7369
X61118_Rna1	M28826	Survival_Estatus		11093

Validation

Cross-validation: A form of predictive validation.

1. For each case, remove it from the database;
2. Use these data to learn the probability distributions of the network;
3. Use the quantified network to predict value on a variable of the removed case.

Validation parameters:

Correctness: Number of cases correctly predicted;

Coverage: Number of cases actually predicted;

Average Distance: How uncertain is a prediction.

Take Home Messages

- Machine learning methods are now an integral part of the new, genome-wide, biology.
- Genome-wide biology presents some new challenges to machine learning, such as the sample size of the experiments.
- Supervised and unsupervised methods answer different questions:
  - ✓ Supervised methods try to map a set of gene profiles to a predefined class.
  - ✓ Unsupervised methods try to dissect interactions of genes.
- Distance-based clustering rests on the assumption that genes with similar behavior also belong to the same process/function.
- There are methods to identify dependency structures from data.

Reading/Software List

Reviews:

P Sebastiani et al. Statistical Challenges in Functional Genomics. Statistical Science, 2003. http://genomethods.org/papers/statscience02.pdf.

IS Kohane et al. Microarrays for an Integrative Genomics. MIT Press, Cambridge, MA, 2002.

Software:

GeneCluster: http://www-genome.wi.mit.edu/cancer.

SAM: http://www-stat.stanford.edu/~tibs/SAM.

CAGED: http://genomethods.org/caged.

Assignment:

Supervised: Using GeneCluster or SAM;

Unsupervised: Using GeneCluster or CAGED.