



# **Cancer Microarray Analysis with Clustering Penalization**

---

**Shuangge Ma**

**Department of Epidemiology and Public Health  
Yale University**



# Microarray Technology

---

- High throughput technologies that allow for **simultaneous** monitoring of biochemical activity for thousands of genes.
- Examples
  - cDNA microarrays;
  - Affymetrix;
  - Protein microarrays;
  - ChIP microarrays.



# Microarray Technology

---

- **1991: Photolithographic printing (Affymetrix)**
- **1994: First cDNA collections developed at Stranford**
- **1995: Quantitative monitoring of gene expression patterns with a complementary DNA microarray**
- **1996: Commercialization of arrays (Affymetrix)**
- **1997: Genome- wide expression monitoring in yeast**
- **2000: Portraits/ Signatures of cancer**
- **2003: Introduction into clinical practices**
- **2004: Whole human genome on one microarray**



# Cancer Microarray Studies

---

- **Cancer is caused by a series of genetic mutations**
- **DNA microarray can survey the whole genome, and provides a powerful tool for identifying molecular signatures and accurately predicting the presence and progression of cancer**
- **“the molecular signatures so identified help reveal the biological spectrum of cancers, provide diagnostic tools as well as prognostic and predictive gene signatures, and may identify new therapeutic targets...”**



# Features of Cancer Microarray Studies

---

- **Studies usually have small sample sizes (<1000) and huge sets of genes (~10,000)**
- **Genes have the inherent clustering structure**
  - **Clusters are composed of genes with coordinated functions or correlated expressions**
  - **Statistical evidence: high correlations**
  - **Biological evidence: the concept of pathway**



# Related Studies: a **VERY** incomplete review

---

- **To tackle the “large p, small n” problem:**
  - **Marginal significance + FDR**
  - **Joint modeling: many dimension reduction and variable selection methods have been proposed.**



## Related Studies: a **VERY** incomplete review

---

- **To accounting for the clustering structure**
  - **Simple clustering based approaches (clustering + simple averaging)**
  - **Detection of marginally significant clusters (Geoman; Efron and Tibshirani; Subramanian)**
  - **Joint modeling: Hongzhe Li's group**



# **Our Understanding of Cancer Microarray**

---

- **Genes have the inherent clustering structure; Each cluster is composed of multiple genes;**
- **Cancer development and progression are caused by mutations or defects of multiple gene clusters;**
- **Within influential gene clusters, only a subset of genes are associated with the cancer clinical outcome of interest.**



# Our Plan

---

- **For cancer microarray data with cluster structure defined *a priori*:**
  - **Select influential gene clusters**
  - **Select influential genes within those influential clusters**
  - **Construct predictive models simultaneously**



# Clustering Penalized Estimation

---

- 1. Construct clusters of genes**
- 2. Penalized estimation**
  - Assume a statistical model linking gene expressions with cancer clinical outcomes**
  - Maximize the penalized objective functions**
  - Two-level selection (gene-cluster level and within-cluster-gene level) along with estimation**
- 3. Evaluation and validation**



# Construction of Gene Clusters

---

- 1. Retrieve pathological pathway information from databases such as KEGG, BioCarta and others.**
- 2. For genes without pathway information, construct clusters based on statistical correlations.**
  - K-means, Hierarchical, model based ...**
- 3. The final clusters are the union of 1 and 2.**



# Construction of Statistical Models

---

**Simple parametric or semiparametric models are suggested.**

- **Diagnostic type: logistic regression**
- **Prognostic type: Cox model**
- **Continuous markers: linear regression**



# Notations

---

- Let  $Z$  be the length  $d$  vector of gene expressions.
- Denote  $Z^1, \dots, Z^m$  as the gene expressions in cluster  $1, \dots, m$ .
- Assume that  $Y$  is associated with  $Z$  through a parametric or semiparametric model  $Y \sim \phi(\beta'Z)$ .
- Assume  $n$  iid observations  $(Y_1, Z_1), \dots, (Y_n, Z_n)$  are available.



# Binary Classification with Logistic Regression

$$\text{logit}(P(Y = 1 | Z)) = \alpha + \beta'Z$$

$$(\hat{\alpha}, \hat{\beta}) = \arg \max_{\alpha, \beta} R_n(\alpha, \beta)$$

$$R_n(\alpha, \beta) = \sum_{i=1}^n Y_i \log\left(\frac{\exp(\alpha + \beta'Z_i)}{1 + \exp(\alpha + \beta'Z_i)}\right) + (1 - Y_i) \log\left(\frac{1}{1 + \exp(\alpha + \beta'Z_i)}\right)$$

- **For simplicity, denote  $R_n(\alpha, \beta)$  as  $R_n(\beta)$**



# Penalized Estimation

---

- **Define**  $\|\beta^j\|_1 = |\beta^{j1}| + \dots + |\beta^{jp_j}|$

$$\hat{\beta} = \arg \max \left\{ R_n(\beta) - \lambda_n \sum_{j=1}^m \|\beta^j\|_1^\gamma \right\}$$



# Penalized Estimation

---

- **The proposed penalty is a composite penalty.**
- **It is a “combination” of bridge and Lasso penalty: bridge at the cluster level and Lasso at the gene level.**
- **It is closely related to the group Lasso penalty.**

# Computation via group Lasso

$$S_n(\beta, \theta_1, \dots, \theta_m) = R_n(\beta) - \left\{ \sum_{j=1}^m \theta_j^{1-\frac{1}{\gamma}} \|\beta^j\|_1 + \tau \sum_{j=1}^m \theta_j \right\}$$

where  $\tau$  is a penalty parameter.

$$\lambda_n = \tau^{1-\gamma} \gamma^{-\gamma} (1-\gamma)^{\gamma-1}$$

$\hat{\beta}$  maximizes the group bridge objective function  
if and only if

$$(\hat{\beta}, \hat{\theta}_1, \dots, \hat{\theta}_m) = \arg \max S_n(\beta, \theta_1, \dots, \theta_m)$$

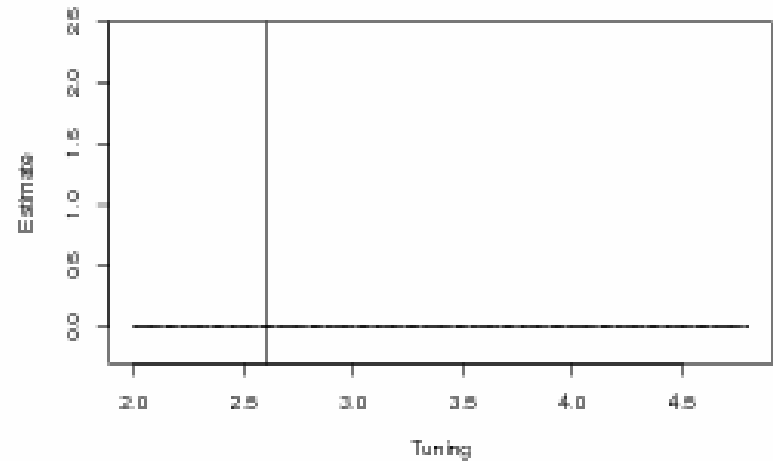
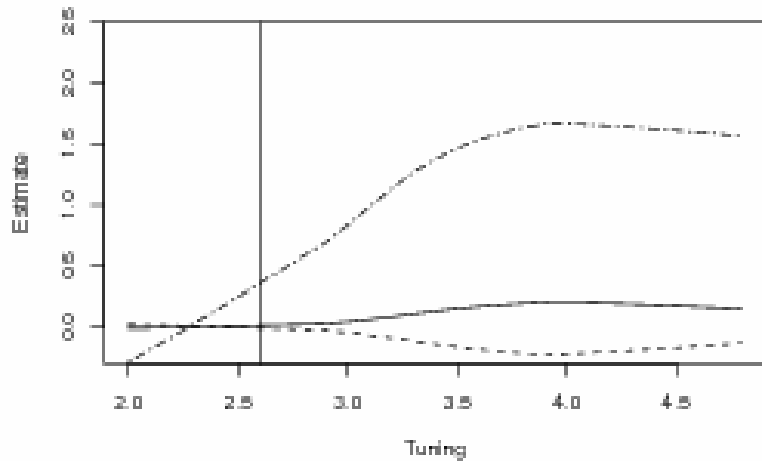
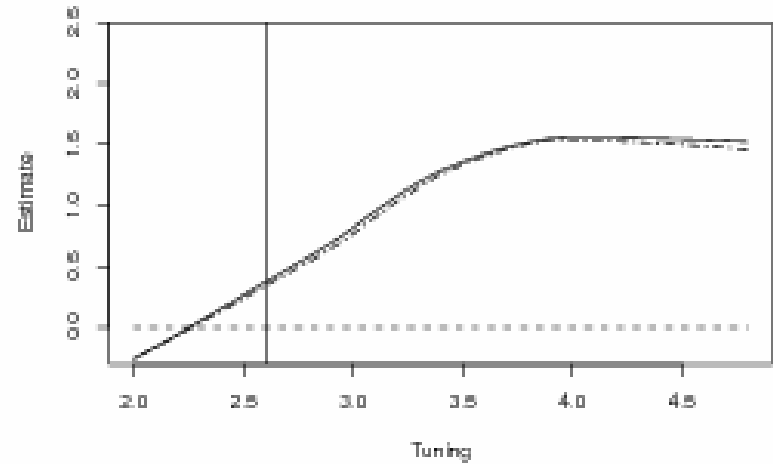
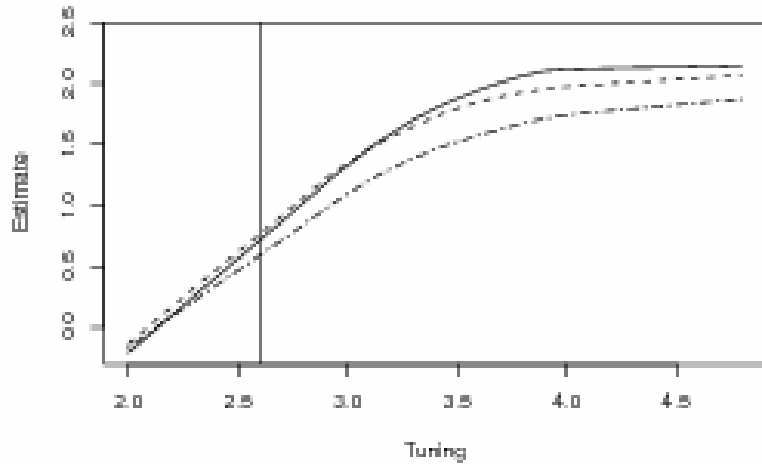


# Tuning Parameter Selection

---

- **Cross validation!**

# A Graphic Representation





# Follicular Lymphoma Study

---

- **Follicular lymphoma is the second common form of non-Hodgkin's lymphoma, accounting for 22 percent of all cases.**
- **Tumor-biopsy specimens from 191 untreated patients were obtained.**
- **Median age at diagnosis: 51 years.**
- **Median follow up time: 6.6 years.**
- **Gene expression measured with Affymetrix U133A and U133B genechips.**



# Follicular Lymphoma Study

---

- **With the hybrid clustering, a total of 130 clusters are constructed.**
- **Among them, 12 are constructed based on statistical correlations.**
- **16 genes are selected, representing 4 different gene clusters.**

# Follicular Lymphoma Study

UNIQID	Symbol	Cluster	Est.
1100790		Cancer Gene II	0.009
1104365	EBF	Cancer Gene II	-0.322
1106389		Cancer Gene II	-0.032
1109193	ANKRD13	Cancer Gene II	-0.055
1112339		Cancer Gene II	-0.014
1137071	TRA2A	Cancer Gene II	0.225
1119299	ENO2	Glycolysis/Gluconeogenesis	-0.198
1119350	ALDH2	Glycolysis/Gluconeogenesis	-0.161
1112764	IFNGR1	Jak-STAT signaling pathway	-0.035
1098405	IL7R	Jak-STAT signaling pathway	-0.295
1100582	CREB3L2	Melanogenesis	0.163
1097846	CREB1	Melanogenesis	-0.138
1132548	CREB1	Melanogenesis	-0.339
1128804	FZD3	Melanogenesis	0.151
1101010	GNAS	Melanogenesis	0.189
1116700	CAMK2D	Melanogenesis	0.239



## Follicular Lymphoma Study

---

- **Early B cell factor (EBF) is a transcription factor suggested to be involved in the transcriptional control of several B cell restricted genes. EBF is also essential for B lymphocyte development.**
- **ENO2 is also known as NSE. The frequency of a high NSE serum value in acute and lymphoma type adult T-cell leukemia (ATL) suggests that ATL cells preferentially produce NSE compared with other NHL cells. NSE may have a role in development of pyothorax-associated lymphoma (PAL).**
- **Gene IFNGR1 has been carefully investigated as one of the lymphoma signature genes in Lan et al. (2006).**



# Follicular Lymphoma Study

---

- **Aldehyde dehydrogenase (ALDH; gene *aldh3a2*) plays a significant role in the metabolism of many biological substances. It has been shown to be related to lymphoma in animal models.**
- **In situ hybridization has shown that lymphoma cells express IL7R. The protein encoded by this gene is a receptor for interleukine 7 (IL7). This protein has been shown to play a critical role in the V(D) J recombination during lymphocyte development.**
- **Mutations in *GNAS1*, the human *GNA* gene, result in Alright hereditary osteodystrophy (AHO), which may suggest its more general role in cancer.**



# Follicular Lymphoma Study

---

- **CREB has also been implicated in the pathogenesis of lymphomas. CREB binds the CRE site in the promoter of translocated bcl-2 in follicular lymphoma with the t(14;18) translocation but not normal alleles in both follicular and transformed lymphomas.**
- **Gene GNAS codes recombinant lymphoma associated protein (LAP). GNAS also plays a role in diseases other than leukemias and lymphomas.**



# Summary

---

- **We propose using clustering penalization in cancer microarray studies.**
- **Limited** data analyses show that the proposed approach can identify a small number of gene clusters and genes, with biologically meaningful implications and satisfactory prediction.



# Conclusion

---