MICROARRAY GENE EXPRESSION DATA WITH LINKED PHENOTYPES:  *Prediction*
Outline

- Phenotypes
- Prediction
- **Continuous**: Ro1 / Kappa Opioid
- **Survival**: Diffuse large-B-cell lymphoma
- Analytic methods:
  - gene harvesting
  - penalized approaches
- Conclusions
Phenotypes

• None -- clustering, unsupervised analyses
• Categoric -- classification / discrimination
• Continuous -- regression
• Survival -- time-to-event outcome subject to (right) censoring
Prediction

- Objectives: accuracy, interpretability
- Develop model on training/learning data
- Assess model’s predictive performance on independent test/validation data
Transgenic mice -- overexpress Ro1 receptor
This hyperactivity induces cardiomyopathy
Reversible -- with use of receptor antagonist
Interest in determining gene expression profiles that are predictive of Ro1 levels
Affymetrix Mu6500 array / 30 mice
Redfern et al., PNAS, 2000
Gene Harvesting

• Hastie et al., *Genome Biology*, 2001

• Handles continuous, survival and categorical phenotypes

• Seeks to avoid “lists of genes” results

• GoMiner, MAPPFinder, EASE, ...

• Can be prone to artifactual solutions: Segal et al., *J Computational Biology* 2003
1. Cluster genes using hierarchical clustering

2. Get average expression profiles from all $p - 1$ clusters: these serve as covariates in addition to the $p$ individual genes

3. Use of clusters as covariates biases toward correlated sets of genes; reduces overfitting

4. Forward stepwise regression algorithm up to a prescribed number of terms

5. Provision for interactions with included terms

6. Model choice by cross-validation (???)
Clustering genes

E.g. p=5

Cluster 6=(1,2)
Cluster 7=(1,2,3)
Cluster 8=(4,5)
Cluster 9=(1,2,3,4,5)

Let p = number of genes.
Perform hierarchical clustering which will produce (2p-1) clusters of genes.
Simple Linear Regression

- **Data**: \( \{(x_i, y_i)\}_{i=1}^{n} \)
  
  \( x_i \): explanatory / feature variable; covariate
  
  \( y_i \): response variable; outcome; output
  
  sample of \( n \) covariate, outcome pairs.

- **Linear Model**: \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \)

- **Errors**: \( \epsilon \) – mean zero, constant variance \( \sigma^2 \)

- **Coefficients**: \( \beta_0 \) – intercept; \( \beta_1 \) – slope

- **Estimation**: least squares – select coefficients to minimize the *Residual Sum of Squares* (RSS)
  
  \[
  RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2
  \]

- **Solution**: \( \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \)
Ro1 / Harvesting / Average Linkage

<table>
<thead>
<tr>
<th>Step</th>
<th>Node</th>
<th>Parent</th>
<th>Score</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6295</td>
<td>0</td>
<td>22.40</td>
<td>687</td>
</tr>
<tr>
<td>2</td>
<td>1380</td>
<td>6295</td>
<td>19.67</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>663</td>
<td>0</td>
<td>15.62</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3374</td>
<td>663</td>
<td>10.69</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1702</td>
<td>0</td>
<td>12.92</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>6268</td>
<td>663</td>
<td>11.27</td>
<td>83</td>
</tr>
</tbody>
</table>

\[ y = \beta_0 + \beta_1 \bar{x}_{Node6295} + \beta_2 (\bar{x}_{Node1380} \cdot \bar{x}_{Node6295}) + \ldots \]
Gene Harvesting: Ro1

Correlations: Node 6295
Gene Harvesting: Ro1

Node score = 22.4!
Ro1: 10-fold CV Error Variance

Constrained Harvesting
Training Error
suppressor of K+ transport defect 3
eukaryotic translation elongation factor 1 alpha 2
myoglobin
EST homologous to ATP synthase coupling factor 6
translationally–controlled tumor protein 1
EST homologous to ATP synthase gamma chain
delta–aminolevulinate dehydratase
lipoprotein lipase
EST homologous to ATP synthase coupling factor 6
myogoblin
eukaryotic translation elongation factor 1 alpha 2
suppressor of K+ transport defect 3
<table>
<thead>
<tr>
<th>Step</th>
<th>Node</th>
<th>Parent</th>
<th>Score</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>g3655</td>
<td>0</td>
<td>21.97</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2050</td>
<td>g3655</td>
<td>20.62</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>g900</td>
<td>g3655</td>
<td>16.91</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>g1324</td>
<td>g3655</td>
<td>16.01</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>g1105</td>
<td>g3655</td>
<td>24.34</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>g230</td>
<td>g3655</td>
<td>12.44</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ y = \beta_0 + \beta_1 x_{Gene3655} + \beta_2 (\bar{x}_{Node2050} \cdot x_{Gene3655}) + \cdots \]
Harvesting Summary

- Can produce artifactual solutions
- Remedied by restricting eligible clusters
- Sensitive to choice of clustering algorithm
- Underlying forward selection scheme arguably too greedy for microarray setting
Diffuse Large B Cell Lymphoma

- Aggressive malignancy of mature B lymphocytes.
- Most common type of lymphoma in adults.
  - Annual US incidence exceeds 25,000 cases.
  - 40% of non-Hodgkins lymphoma cases.
- Despite treatment advances less than 50% of patients achieve lasting remission.
- Clinical features modest predictors of response.
Diffuse Large B Cell Lymphoma

- The combination of lack of good clinical predictors of survival outcomes and underlying disease heterogeneity spawned several microarrays studies of DLBCL.

- Objective: devise a molecular profile that predicts survival and can be used for therapy evaluation / selection.
Diffuse Large B Cell Lymphoma

- Alizadeh et al., *Nature*, 2000
  - lymphochip / unsupervised / subtypes.
- Shipp et al., *Nature Medicine*, 2002
  - Affy HU6800 / supervised - binary outcome.
- Wright et al., *PNAS*, 2002
  - reconcile the above / compound covariates.
- Lossos et al., *NEJM*, 2004
  - RT-PCR / Cox PH / 6 gene predictor.
**Diffuse Large B Cell Lymphoma**

- **Rosenwald et al., NEJM, 2002**

- ‘Lymphochip’ cDNA microarray composed of genes expressed in lymphoid cells and/or implicated in cancer, immune function

- Here 7399 features representing 4128 genes

- Survival (from chemotherapy) obtained for 240 patients partitioned into training (160) and validation (80) datasets
• Hierarchical clustering of expression data used to identify ‘signatures’ -- by inspection -- ‘fluid’

• uses training and test datasets

• tight clustering: Tseng Wong; Biometrics, 2004

• Univariate Cox proportional hazards was used to identify significant genes

• Where possible these genes were assigned to the previously identified signatures

• Signature averages used in multivariate PH model

• At best, poor-man’s gene harvesting
Subgroup of Diffuse Large-B-Cell Lymphoma

- Germinal-center B-cell-like
- Type 3
- Activated B-cell-like

Proliferation signature
- MHC class II signature
- Germinal-center B-cell signature
- Lymph-node signature

Relative Level of Expression (× median value)
thyroxine−binding globulin precursor expression
Survival (years)

<table>
<thead>
<tr>
<th>coef</th>
<th>exp(coef)</th>
<th>se(coef)</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>thyroxine-binding</td>
<td>-0.202</td>
<td>0.817</td>
<td>0.0435</td>
<td>-4.65</td>
</tr>
</tbody>
</table>

Rsquare= 0.093  (max possible= 0.997 )

Score (logrank) test = 22.4  on 1 df,  p=2.27e-06
>harvest(clust.aves, expr.mat, y = survtime, ic = status, method = "survival", alpha = 0.1, maxterms = 6)

Adding term= 1  (Node) = 3498  Size= 14  Score= 3.827
Adding term= 2  (Node) = 3383  Size= 9  Score= 3.062
Adding term= 3  (Node) = 2357  Size= 2  Score= 3.610
Adding term= 4  (Node) = -5730  Size= 1  Score= 3.342
Adding term= 5  (Node) = 3450  Size= 2  Score= 2.758
Adding term= 6  (Node) = 6223  Size= 4  Score= 2.470

**alpha** biases toward selecting clusters

**Score** moderated z statistic

**3498** contains exclusively MHC class II genes:

*all*  DP alpha, DR alpha, DR beta, DP beta

*none*  DM alpha, DQ alpha, DM beta, DQ beta

MHC class II foremost amongst the four chosen signatures
Artifactual solutions arise when clusters are not tight and cluster average associates with outcome.

While not an issue we have biased toward selecting (larger) clusters.
• When we turn this off ($\alpha = 0$) only singletons (original genes) are selected

• At this juncture it becomes apt to consider refinements to forward stepwise approaches

• The $p \gg n$ microarray setting warrants some form of regularization: cf DLDA
Lasso Regression

- Tibshirani *JRSSB*, 1996
- $L_1$ penalty on coefficients captures good features of ridge (prediction accuracy) and subset (interpretation) regression:

$$
\min_{\beta} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|
$$

coefficients often exactly zero  cf selection
• Has been extended to Cox PH.

• Estimation/computational problems in microarray settings.

• These can be solved by (i) use of least angle regression (LARS), and (ii) use of residual-based approximations.
Top genes selected by direct or approximate methods coincide -- first 4 genes (CV) selected:

- X00452: MHC class II, DQ alpha 1
  *MHC signature*

- AA805575: thyroxine-binding globulin precursor
  *Germinal-center B-cell signature*

- X59812: cytochrome P450, subfamily
  *Lymph-node signature*

- LC_29222: *Lymph-node signature* [not on Hu6800]
Predictive Performance

- Risk scores based on fitted model computed for patients in (withheld) test dataset.
- Low/High risk strata created by thresholding.
- Statistical testing of between strata differences.
- Many limitations: arbitrariness of stratification, p-values don’t capture explained variation.
Lymphochip --- Test Data

Years Post Therapy

L1 Penalized Cox PH

Lossos --- 6 Gene Model

Rosenwald --- Signatures

Harvesting --- MHC Class II
<table>
<thead>
<tr>
<th>Method</th>
<th>#{high risk}</th>
<th>log-rank p</th>
<th>Rsquared</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Penalized Cox PH</td>
<td>36</td>
<td>0.0004</td>
<td>0.138</td>
</tr>
<tr>
<td>Harvesting -- alpha = 0</td>
<td>33</td>
<td>0.20</td>
<td>0.025</td>
</tr>
<tr>
<td>Lossos -- 6 Gene Model</td>
<td>42</td>
<td>0.0136</td>
<td>0.074</td>
</tr>
<tr>
<td>Rosenwald -- Signatures</td>
<td>32</td>
<td>0.0005</td>
<td>0.128</td>
</tr>
</tbody>
</table>
Lymphochip --- Test Data

- L1 Penalized Cox PH Lossos
- 6 Gene Model
- Rosenwald Signatures
- Harvesting MHC Class II

Years Post Therapy

Time-Dep ROC Area
**Affymetrix -- Lossos Processing**

![Graph](image)

- **Time-Dep ROC Area**
- **Years Post Therapy**

- **LARS / 3 LC Genes**
- **LARS / 6 Affy Genes**
- **Lossos -- 6 Genes**
Conclusions

• Gene harvesting formalizes *ad hoc* approaches based on clustering

• Can be prone to artifact; too greedy

• $L_1$ penalized Cox proportional hazards avoids greediness and, for DLBCL, exhibits relatively good predictive performance

• LARS / devices make computationally viable
Conclusions

• More refined methods for evaluating predictive performance warranted
• Time-dependent ROC curves have merit
• Molecular profiling of DLBCL provides only modest prediction of post therapy survival
Software

- harvest(...) 
- lars(...) 
- coxph(...) 
- roc.KM.calc(...)