Multiclass Analysis of Microarray Data
Robert Stengel
October 24, 2007

- Background
- Small, round, blue-cell tumor example
- Application to colon cancer, metastases, and normal tissue

Lecture for MOL 457, Computational Aspects of Molecular Biology, Princeton University

Production of a Single Protein

- Up-regulation of mRNA transcripts in tumor cells
  - Causal input (tumor enhancing gene)
  - Defensive response (tumor suppressor gene)
  - Bystander effect
  - Tissue effect
  - Artifact

- Down-regulation of mRNA transcripts in tumor cells
  - Present, but mutated (and, therefore, not detected)
  - Eliminated or suppressed by tumor growth
  - Tissue effect
  - Artifact

Critical Dynamic Processes in the Pathogenesis of Cancer

Putative Paradigm for Microarray Analysis: mRNA Transcript Expression Level Infers Cell Function & Carcinogenesis
Limitations of Microarray Analysis

- Human genome microarray probe sets describe wild-type nucleotide sequences
  - They do not see mutations
- Microarray data are “noisy”
  - Biological variation
  - Measurement error
- mRNA expression level alone is not sufficient to determine function
  - Inhibition and amplification of biological processes are multi-gene/ - protein effects
  - Ancillary data are required to interpret significance of transcript levels
- Realistic goal is to classify samples by statistical analysis of transcript expression levels

Typical Pairs of Colon Cancer Microarray Expression Levels
(Alon, Notterman et al, 1999)

- Samples not well differentiated in individual transcript clusters (overlapping)

Classification of Data

- Data set characterized by two features
Clustering of Data

- How many clusters?

Discriminants of Data

- Where are the boundaries between sets?

The Data Set Revealed

The discriminant is the Delaware River

Towns and Crossroads of Pennsylvania and New Jersey
Another Classification Example

- Party for returning graduate alumni/ae, Reunions, 1999
- From this picture, who are the:
  - Grad alums
  - Current students
  - Spouses
  - Children
  - Visitors from abroad
  - Hosts
  - Oldest alums
  - Youngest alums
  - Party crashers?
- What features are used to classify?

Supervised and Unsupervised Learning of Discriminants

- Learning depends on “closeness” of related features
- Previously unknown correlations or features are detected
- Classification occurs after learning via exogenous knowledge
- Same answer given for all questions

... and What if the Data are Distorted?

- Best line or curve may classify with significant error
- Best plane or surface classifies with equal or less error

Effect of Feature Set Dimension

- Learning depends on prior definition and knowledge of class
- Complex correlation between features revealed
- Classification is inherent in learning
- Different answers given for different questions
Characteristics of Classification Features

- Strong feature
  - Individual feature provides good classification
  - Minimal overlap of feature values in each class
  - Significant difference in class mean values
  - Low variance in class is desirable

- Additional features
  - Orthogonal feature (low correlation) may add new information to the set
  - Co-expressed feature (high correlation) is redundant; averaging may reduce error

But All Data Sets Are Not Gaussian!

- Association between \( t \) and \( p \)-value is different for Gaussian and non-Gaussian distributions
- All finite data sets have means and standard deviations that describe the average and the spread
- Classification goal is to find transcripts whose expression levels are significantly different, not to compute \( p \)-value
- Large value of \( t \) virtually assures small \( p \)-value for comparison of any unimodal distributions
  - Actual \( p \)-value is of little concern if it is “small enough”
- Statistical significance can be confirmed by standard techniques (e.g., Bonferroni correction, Benjamini-Hochberg procedure)

Two-Sample Student \( t \) Test for Transcript Selection

- \( t \) compares mean values of two data sets
  - \( |t| \) is reduced by uncertainty in the data sets (\( \sigma \))
  - \( |t| \) is increased by number of points in the data sets (\( n \))

\[
t = \frac{(m_A - m_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}
\]

- \( m \) = mean value of data set
- \( \sigma \) = standard deviation of data set
- \( n \) = number of points in data set

- \( t \) test applied to mean tumor/normal expression levels of individual transcripts
  - \( |t| > 3, m_A \neq m_B \) with \( \geq 99.7\% \) confidence
    - Error \( p \leq 0.003 \) (Gaussian) \( [n > 25] \)

Example of Transcript-by-Transcript Tumor/Normal Classification by \( t \) Test

(data from Alon, Notterman et al, 1999)

- 1,151 probe sets* are over/under-expressed in tumor/normal comparison, \( p \leq 0.003 \)
- Genetically dissimilar samples are apparent
  - “Cancer-positive probe sets”
  - “Cancer-negative probe sets”

* Each probe set represents one mRNA transcript
Correlation Matrices

- Probe set correlation
- Sample correlation

Class Prediction and Evaluation Using Ensemble Mean Values
(data from Alon, Notterman, 1999)

- Single-feature prediction
  - Cancer-positive genes: 2 errors
  - Cancer-negative genes: 1 error

- Two-feature prediction (cross-plot): 1 error (mislabeling?)

Ensemble Mean Values

- Treat each probe set (row) as a redundant, corrupted measurement of the same tumor/normal indicator
  \[ z_{ij} = k_i y + \varepsilon_{ij}, \quad i = 1, m, \quad j = 1, n \]

- Compute column averages for each sample sub-group (i.e., sum each column and divide by \( n \))
  \[ \bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij} \]

- Random variable sums approach Gaussian distribution by central limit theorem as \( n \to \infty \)

Clustering of Sample Averages for Primary Colon Cancer vs. Normal Mucosa
[NCI Program Project Grant (PPG), 144-sample set]

- 144 samples, 3,437 probe sets analyzed
- 47 primary colon cancer
- 22 normal mucosa
- Affymetrix HG-U133A GeneChip
- All transcripts “Present” in all samples

# Probe sets
Up: 1067
Down: 290
Constant: 19
Clustering of Sample Averages for Primary Polyp vs. Normal Mucosa (PPG, 144-sample set)

![Graph showing clustering of sample averages for primary polyp vs. normal mucosa.]

Clustering of Sample Averages for Primary Polyp vs. Primary Colon Cancer (PPG, 144-sample set)

![Graph showing clustering of sample averages for primary polyp vs. primary colon cancer.]

Identified Cancer Links in Selected Transcript Training Sets (PPG, 144 samples)

- **Primary Colon Cancer** vs. Normal Mucosa
  - Over-Expressed (10) 7 (70%)
  - Under-Expressed (10) 5 (50%)
  - Both (20) 12 (60%)

- **Primary Polyp** vs. Norma Mucosa
  - Over-Expressed (10) 7 (70%)
  - Under-Expressed (10) 5 (50%)
  - Both (20) 12 (60%)

- **Primary Colon Cancer** vs. Primary Polyp
  - Over-Expressed (10) 7 (70%)
  - Under-Expressed (10) 6 (60%)
  - Both (20) 13 (65%)

**Total**: 21 (70%) Over-Expressed, 16 (53%) Under-Expressed, 37 (62%) Both

---

Small, Round Blue-Cell Tumor Classification Example

- Childhood cancers, including:
  - Ewing’s sarcoma (EWS)
  - Burkitt’s Lymphoma (BL)
  - Neuroblastoma (NB)
  - Rhabdomyosarcoma (RMS)

- cDNA microarray analysis presented by J. Khan, et al., 2001
  - 96 transcripts chosen from 2,308 probes for training
  - 63 EWS, BL, NB, and RMS training samples
  - Classification with principal component analysis and a linear neural network
  - Source of data for my analysis

Desmoplastic small, round blue-cell tumors
Overview of Present SRBCT Analysis

- Transcript selection by t test
  - 96 transcripts, 12 highest and lowest t values for each class
  - Overlap with Khan set: 32 transcripts
- Ensemble averaging of highest and lowest t values for each class
- Cross-plot of ensemble averages
- Classification by sigmoidal neural network
- Validation of neural network
  - Novel set simulation
  - Leave-one-out simulation

Comparison of Present SRBCT Set with Khan Top 10

- EWS Student t Value
- BL Student t Value
- NB Student t Value
- RMS Student t Value
- Most Significant t Value
- Khan Gene Class

Khan selection distinguished by strong, consistent relation to t values

Ranking by EWS t Values
(Top and Bottom 12)

- 24 transcripts selected from 12 highest and lowest t values for EWS vs. remainder

<table>
<thead>
<tr>
<th>Image ID</th>
<th>Transcription Description</th>
<th>EWS t Value</th>
<th>BL t Value</th>
<th>NB t Value</th>
<th>RMS t Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>770394</td>
<td>5' flanking of IgH, receptor, transporter, alpha</td>
<td>-9.41</td>
<td>-7.41</td>
<td>-6.17</td>
<td>-5.38</td>
</tr>
<tr>
<td>74846</td>
<td>30 kDa polypeptide chain of 5,496 kDa protein</td>
<td>-7.41</td>
<td>-6.17</td>
<td>-5.38</td>
<td>-4.79</td>
</tr>
<tr>
<td>295985</td>
<td>30 kDa polypeptide chain of 5,496 kDa protein</td>
<td>-7.41</td>
<td>-6.17</td>
<td>-5.38</td>
<td>-4.79</td>
</tr>
</tbody>
</table>

Strong t-value distinctions between classes

Cluster of SRBCT Ensemble Averages

- Repeated for BL vs. remainder, NB vs. remainder, and RMS vs. remainder
Artificial Neural Networks

Neural Network Training Set

- Each input row is an ensemble average for a transcript set, normalized in (−1, +1)

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>...</th>
<th>Sample 62</th>
<th>Sample 63</th>
</tr>
</thead>
<tbody>
<tr>
<td>EWS</td>
<td>EWS+</td>
<td>EWS+</td>
<td>...</td>
<td>EWS+</td>
<td>EWS+</td>
</tr>
<tr>
<td>EWS−</td>
<td>EWS−</td>
<td>EWS−</td>
<td>...</td>
<td>EWS−</td>
<td>EWS−</td>
</tr>
<tr>
<td>BL±</td>
<td>BL±</td>
<td>BL±</td>
<td>...</td>
<td>BL±</td>
<td>BL±</td>
</tr>
<tr>
<td>BL−</td>
<td>BL−</td>
<td>BL−</td>
<td>...</td>
<td>BL−</td>
<td>BL−</td>
</tr>
<tr>
<td>NB±</td>
<td>NB±</td>
<td>NB±</td>
<td>...</td>
<td>NB±</td>
<td>NB±</td>
</tr>
<tr>
<td>NB−</td>
<td>NB−</td>
<td>NB−</td>
<td>...</td>
<td>NB−</td>
<td>NB−</td>
</tr>
<tr>
<td>RMS±</td>
<td>RMS±</td>
<td>RMS±</td>
<td>...</td>
<td>RMS±</td>
<td>RMS±</td>
</tr>
<tr>
<td>RMS−</td>
<td>RMS−</td>
<td>RMS−</td>
<td>...</td>
<td>RMS−</td>
<td>RMS−</td>
</tr>
</tbody>
</table>

SRBCT Neural Network Training and Application

- Neural network
  - 8 ensemble-average inputs
  - various # of sigmoidal neurons
  - 4 linear neurons
  - 4 outputs
- Training accuracy
  - Train on all 63 samples
  - Test on all 63 samples
  - 100% accuracy

Leave-One-Out Validation of SRBCT Neural Network

- Remove a single sample
- Train on remaining samples (125 times)
- Evaluate class of the removed sample
- Repeat for each of 63 samples
- 6 sigmoids: 99.96% accuracy (3 errors in 7,875 trials)
- 12 sigmoids: 99.99% accuracy (1 error in 7,875 trials)
**Novel-Set Validation of SRBCT Neural Network**

- Network always chooses one of four classes (i.e., “unknown” is not an option)
- Test on 25 novel samples (400 times each)
  - 5 EWS
  - 5 BL
  - 5 NB
  - 5 RMS
  - 5 samples of unknown class
- 99.96% accuracy on first 20 novel samples (3 errors in 8,000 trials)
- 0% accuracy on unknown classes

**Observations of SRBCT Classification using Ensemble Averages**

- t test identified strong features for classification in this data set
- Neural networks easily classified the four data types
- Caveat: Small, round blue-cell tumors occur in different tissue types
  - Ewing’s sarcoma: Bone tissue
  - Burkitt’s Lymphoma: Lymph nodes
  - Neuroblastoma: Nerve tissue
  - Rhabdomyosarcoma: Soft tissue
- Genetic variation may be linked to tissue differences as well as tumor genetics

**260-Sample PPG Colon Cancer Data Set**

- Primary colon cancer (130)
- Primary polyp (31)
- Normal mucosa (24)
- Liver metastasis (33)
- Normal liver (12)
- Lung metastasis (20)
- Normal lung (5)
- Microadenoma (4)
- High grade dysplasia (1)

**9-Class Colon Cancer Neural Network**

- Neural network
  - 18 ensemble-average inputs
  - various # of sigmoidal neurons
  - 9 linear neurons
  - 9 outputs
  - 12-48 transcripts in each ensemble average

- Affymetrix HGU-133A GeneChip
- 22,283 probe sets
- 17,455 probe sets verified
- Transcripts are “Present” in 80% of samples
- 7,343 probe sets analyzed
Training and Validation Accuracy of 9-Class Colon Cancer Neural Network

9-class PPG set is more heterogeneous than the 4-class SRBCT set
Generalization is better with fewer neurons

Primary Tumor vs. Polyp
Positive vs. negative meta-transcripts
Sample correlation
Significant correlation between tumors and polyps

Classification of Primary Tumors, Polyps, and Normal Mucosa
Primary colon cancer (130), primary polyp (31), normal mucosa (24)

Training Set
Leave-One-Out

<table>
<thead>
<tr>
<th>Transcripts in</th>
<th>Neurons</th>
<th>Trials</th>
<th>Training Set Correct</th>
<th>Leave-One-Out Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td></td>
<td></td>
<td>% Correct</td>
<td>% Correct</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>260</td>
<td>176</td>
<td>151</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>260</td>
<td>188</td>
<td>156</td>
</tr>
<tr>
<td>48</td>
<td>4</td>
<td>260</td>
<td>179</td>
<td>173</td>
</tr>
<tr>
<td>48</td>
<td>9</td>
<td>260</td>
<td>194</td>
<td>170</td>
</tr>
<tr>
<td>48</td>
<td>18</td>
<td>260</td>
<td>234</td>
<td>172</td>
</tr>
</tbody>
</table>

% Correct

<table>
<thead>
<tr>
<th>Transcripts in</th>
<th>Neurons</th>
<th>Trials</th>
<th>Training Set Correct</th>
<th>Leave-One-Out Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td></td>
<td></td>
<td>% Correct</td>
<td>% Correct</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>370</td>
<td>323</td>
<td>310</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>370</td>
<td>328</td>
<td>314</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>370</td>
<td>296</td>
<td>298</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>370</td>
<td>336</td>
<td>320</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>370</td>
<td>336</td>
<td>319</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>370</td>
<td>333</td>
<td>319</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>370</td>
<td>334</td>
<td>323</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>370</td>
<td>311</td>
<td>314</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>370</td>
<td>330</td>
<td>320</td>
</tr>
</tbody>
</table>

Classification of Primary Tumors, Metastases, and Normal Tissue

Primary colon cancer (130), normal mucosa (24), liver metastasis (33), normal liver (12), lung metastasis (20), normal lung (5)
Elimination of organ-specific effects on tumor comparisons

<table>
<thead>
<tr>
<th>Transcripts in</th>
<th>Neurons</th>
<th>Trials</th>
<th>Training Set Correct</th>
<th>Leave-One-Out Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td></td>
<td></td>
<td>% Correct</td>
<td>% Correct</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>448</td>
<td>260</td>
<td>275</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>448</td>
<td>364</td>
<td>343</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>448</td>
<td>427</td>
<td>367</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>448</td>
<td>427</td>
<td>379</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>448</td>
<td>433</td>
<td>377</td>
</tr>
</tbody>
</table>

% Correct
Classification of Primary Colon Cancer According to Gender

- Male (61), female (44)

<table>
<thead>
<tr>
<th>Transcripts in Training Set</th>
<th>Leave-One-Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble Neurons Trials</td>
<td>Correct</td>
</tr>
<tr>
<td>6 4 210</td>
<td>168</td>
</tr>
<tr>
<td>12 2 210</td>
<td>178</td>
</tr>
<tr>
<td>12 4 210</td>
<td>178</td>
</tr>
<tr>
<td>12 8 210</td>
<td>178</td>
</tr>
<tr>
<td>12 9 210</td>
<td>179</td>
</tr>
<tr>
<td>24 2 210</td>
<td>184</td>
</tr>
<tr>
<td>24 4 210</td>
<td>184</td>
</tr>
</tbody>
</table>

50 annotated samples

- Implication that genetic distinction occurs in the 60s

Future Work on PPG Colon Cancer Data

- Classification by outcome
- Analysis of complete data set
- Orthogonal statistics of the data set
- Alternative targets and data sets for classification of data, e.g.,
  - Clinical data
  - Pathology data
  - Mutated DNA drawn from same sample set
- Identification of functional pathways

Conclusions

- Simple tests for transcript selection
  - Mean (t test)
  - Covariance (correlation matrices)
- Multiclass classification using neural networks
- Heterogeneity of colon cancer data
- Confounding effect of tissue-specific transcripts
- Supervised classification according to
  - Primary tumor, polyp, normal mucosa
  - Primary tumor, metastasis, normal tissue
  - Patient gender and age
Acknowledgments

- Gunter Schemmann, Princeton University
- Daniel Notterman, Princeton University
- Francis Barany, Cornell Weill Medical School
- Phillip Paty, Memorial Sloan Kettering Cancer Center
- Eytan Domany, Weizmann Institute of Science
- Jürg Ott, Rockefeller University
- Arnold Levine, Institute for Advanced Study
- Hao Liu, Bristol-Meyers-Squibb (formerly with UMDNJ)
- Douglas Welsh, Princeton University
- ... and members of the Simons Center for Systems Biology, Institute for Advanced Study