Analysis of MALDI-TOF Data: from Data Preprocessing to Model Validation for Survival Outcome

Heidi Chen, Ph.D.
Cancer Biostatistics Center
Vanderbilt University School of Medicine
March 20, 2009
Outline

• MALDI-TOF
• Data preprocessing for raw spectra
• Build a prediction model from training set
• Model validation
MALDI-TOF

Step 1: Sample preparation

Sample Preparation for MALDI-TOF MS
MALDI-TOF mass spectrometry

Matrix
Assisted
Laser
Desorption
Ionization

Matrix-biomolecule co-crystal

Time
Of
Flight

laser
mirror
target
acceleration electrode
detector

acceleration distance
field-free drift distance
Principle Idea of MALDI-TOF MS

• Upon laser irradiation all molecules obtain similar energy
• Convert electric energy to kinetic energy
• Time Of Flight (TOF) separates ions based on size (mass/charge, m/z)

\[
\text{TOF} : \ (m/z)^{1/2}
\]
Raw Spectra
• MALDI-TOF

• Data preprocessing for raw spectra

• Build a prediction model from training set

• Model validation
Baseline correction

Denoise Normalization

Feature Detection

Calibration
(1) m/z values around some known proteins
(2) show clear bell-shape
Convolution Based Calibration

- Calibrate each spectrum with the known peaks. Max $h(t)$ happens when $f$ and $g$ overlap the most.

- The optimum shift is obtained by maximizing the sum of convolution values on the multiple peak locations.

Note: all process are on the time domain.

$$f(t) : \text{observed peak}$$
$$g(t) : \text{ideal peak (normal distribution)}$$

$$h(t) = (f * g)(t) \equiv \int_{-\infty}^{t} f(\tau)g(t-\tau)d\tau = \int_{-\infty}^{t} f(t-\tau)g(\tau)d\tau$$
Calibration

Before Calibration

After Calibration
Wavelet Denoising

\[ S = A + D \]
Wavelet Decomposition Tree

\[ S = A_1 + D_1 \]
\[ = A_2 + D_2 + D_1 \]
\[ = A_3 + D_3 + D_2 + D_1 \]
Wavelet Denoise

Remove noise by thresholding
Baseline Correction

Spline curve to fit the local minima
Peak Detection

(1) local maxima
(2) pass signal/noise cutoff to filter out small peaks
Peak Distribution
Common Peaks Finding

Peak location: local maximum
Boundaries: adjacent local minima
Filter: > 5% of spectra show peaks

Kernel Density of Peaks Distribution
Feature Quantification

- Normalization: standardize the AUC for all spectra to the median AUC

- Peak intensity: AUC within peak boundaries
Threshold Selection

- Denoise: wavelet threshold
- Peak detection: signal to noise ratio
- Common peak finding: bandwidth of KDS
Training Set

• 35 pretreatment (EGFR+VEGF) serum samples from stage IIIB/IV NSCLC; 14 male, 21 female; age range 36-72

• Experimental Design

<table>
<thead>
<tr>
<th></th>
<th>Day1</th>
<th>Day2</th>
<th>Day3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication</td>
<td>3 replications each pt</td>
<td>3 replications each pt</td>
<td>3 replications each pt</td>
</tr>
</tbody>
</table>

105 samples were randomly spotted in two 64-well plates each day.
Training Set

290 good spectra; 25 bad spectra

174 features (3000-20,000 mz) after data preprocessing
Variance Components
• MALDI-TOF
• Data preprocessing for raw spectra
• Build a survival prediction model from training set
• Model validation
Procedure of Constructing Survival Prediction Model from Training Set

(1) Feature selection
- CPH model
- FDR cutoff 0.05 : 11 features associated with survival

(2) Create a compound score as a prediction index

\[ s_i = \sum_{j=1}^{k} (\text{sign of } \beta_j) w_j x_{j,i} \quad w_j : \text{Wald statistics} \]

(3) Prediction model : predicted hazard rate

\[ h_i(t \mid s_i) = h_0(t) \exp(0.0235 s_i) \]
- MALDI-TOF
- Data preprocessing for raw spectra
- Build a prediction model from training set
- Model validation
Overfitting

Training set

Independent test set
Model Validation

Goal: evaluate how the model performs in the future dataset

• External validation: independent test set
• Internal validation: training set
External Validation

Data Preprocessing for Independent Test Set

- Calibration
- Baseline correction
- Wavelet denoise
- Normalization : median AUC from training set
- Peak location and boundaries from training set
Prediction for Independent Test Set

Freeze $w_1, w_2, \ldots, w_k$ and $\varphi$ from training set

Compound score as a predictor

$$s'_i = \sum_{j=1}^{k} \text{(sign of } \beta_j \text{)} \, w_j x'_{j,i}$$

Association with survival outcome.

CPH : predicted hazard rate

$$h'_i(t) = h_0(t) \exp(0.0235 \, s'_i)$$
Validation of Predictive Model

• What to validate?
  - Predictive ability

• C-index (Harrel et al. 1982): measure the agreement between predicted and observed survival time for two subjects. C-index ranges from 0 to 1.
  1: perfect prediction; 0.5: random prediction; 0: opposite prediction
Independent Test Sets

• Eastern Cooperative Oncology Group (ECOG) (n=82)
• Italian (n=66)
KM Survival

Training set: (n=35)

Italian B: (n=66)

ECOG: (n=82)

Cox model: p < 0.001
# C-index

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>Test set</strong></td>
<td>0.62 (Italian)</td>
</tr>
</tbody>
</table>

**Training C-index** : the C-index on the training set

**Generalized C-index** : the C-index on the independent test set
Internal Validation

Goal: estimate the generalized C-index through the internal validation process

- Data splitting
- K-fold cross validation
- Bootstrap

Focus on the procedure after data preprocessing
(1) Build a prediction model based on training set
(2) Compound score for test set: winners and Wald statistics from training set
(3) Generalized C-index: calculate C-index from test set
K-fold Cross Validation

Sample size $n$ divided into $K = 5$ parts,

C-index for each test set

Combine C-index from all test sets to get the estimate the generalized C-index
Bootstrap

Original training set (n)

Sampling with replacement

Bootstrap training
Samples (n)

Bootstrap test sample: obs not in bootstrap training
Bootstrap

• Generalized C-index

\[ C^{(0.632+)} = (1 - W) C^{\text{training}} + W \frac{1}{m} \sum_{i=1}^{m} C^{T_i(\text{test})} \]

\( \gamma \) (non-informative C-index) = 0.5

\[ R \text{ (relative overfitting rate)} : \quad \frac{\frac{1}{m} \sum_{i=1}^{m} C^{T_i(\text{test})} - C^{\text{training}}}{\gamma - C^{\text{training}}} \quad R \in [0, 1] \]

\[ W(\text{weight}) = \frac{0.632}{1 - 0.368R} \quad w \in [0.632, 1] \]

C-index(0.632+) ranges from C-index(0.632) if there is minimum overfitting (R=0) to \( \frac{1}{m} \sum_{i=1}^{m} C^{T_i(\text{test})} \) if there is maximum overfitting (R=1)
## C-index

<table>
<thead>
<tr>
<th>Method</th>
<th>NSCLC (n=35)</th>
<th>Overfit Example (n=77)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>0.71</td>
<td>0.53</td>
</tr>
<tr>
<td>Indep test set</td>
<td>0.62 (Italian)</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Reproducibility of MALDI-TOF MS

<table>
<thead>
<tr>
<th>Winners of Case Study (11) EGFR+VEGF</th>
<th>Winners of JNCI 2007 (8) EGFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>4121</td>
<td>5843</td>
</tr>
<tr>
<td>4596</td>
<td>11445</td>
</tr>
<tr>
<td>4720</td>
<td>11529</td>
</tr>
<tr>
<td>4821</td>
<td>11685</td>
</tr>
<tr>
<td>5720</td>
<td>11759</td>
</tr>
<tr>
<td>5841</td>
<td>11903</td>
</tr>
<tr>
<td>11441</td>
<td>12452</td>
</tr>
<tr>
<td>11528</td>
<td>12579</td>
</tr>
<tr>
<td>11684</td>
<td></td>
</tr>
<tr>
<td>11731</td>
<td></td>
</tr>
<tr>
<td>11902</td>
<td></td>
</tr>
</tbody>
</table>
Take Home Message

- Good experimental design
- Precisely follow the protocol of MALDI-TOF
- MALDI-TOF can detect the true signal
Acknowledgements

Stuart Salmon              Dean Billheimer
Shuo Chen                  Yu Shyr
Roy Herbst                 Ju-Whei Lee
Anne Tsao                   Pierre Massion
Hai Tran                   Julie Brahmer
Alan Sandler                Joan Schiller
David Carbone              Thao P. Dang