Automated Data Analysis

Richard E. Higgs, Jr.

Molecular Informatics
Lilly Research Laboratories

higgs@lilly.com
Objective

Convey candid observations from someone on a road from data to decisions...
Agenda

• Session 1
  – Accessing Data (*first things first*)
  – Analyzing Data (*statistics, statistics, statistics*)

• Session 2
  – Implementation (*Linux changing rules of game*)
  – Case Studies
    • HPLC-MS of complex mixtures
    • FTMS
    • Proteomics
Accessing Data

• Raw files
  – Contains maximum amount of information
  – Proprietary issues
  – Vulnerable to format changes by vendors
  – Multiple vendors
Accessing Data

- netCDF converted files
  - Self describing binary format
  - Some loss of information (vendor issue)
  - Limited to MS and LC-MS
  - Standard data format
  - Libraries available for many analysis tools
    - ANDI library for C/C++
    - R/Splus, MATLAB, etc. compatible
Accessing Data

• ASCII text files
  – Report files
    • Lots of information lost
    • Easy to read (e.g. Perl)
  – New technologies (standards) may help
    • XML
    • File size may still be an issue for your application
Data Storage Options

• Storing raw MS data in relational databases
  – Overhead problems
  – Terabyte databases still a challenge?

• Storing raw MS data on disk
  – Less overhead
  – File organization issues

• Compromise solution
  – Some of both
Instructing Instrument

• Automatic sample list generation
  – Sample identification (file naming)
  – Instrument method
  – Integrated with relational experiment database
Accessing Data

Instrument

Autosampler

Control PC

Relational Database
- experiment conditions
- filenames

File Server
- spectra

Compute Server
- analyses

network
Analyzing Data

• Noise and variability in all measurements

• Statisticians
  – Uniquely trained to analyze data with noise
  – Statistics = Science of decision making

• Seek the help of data analysis specialists
  – Data analysis is a complex, rapidly changing field
  – Much of what you may need has been invented
Analyzing Data

• Classes of problems
  – Data transformation
    • Filtering, cleaning, summarizing
  – Inference
    • Making decisions
  – Multivariate Curve Resolution
    • Identifying components in complex mixture
  – Prediction
    • Accuracy vs. interpretation
  – Data mining: dredging for patterns
    • Hypothesis generation
Data Transformation

• For MS data:
  – Digital filtering
    • Reduce noise
    • Estimate baseline
  – Wavelet smoothing
    • Filter noise for “spikey” signals
    • Compression
Wavelet Smoothing

NMR Spectrum

Waveshrink Estimate of NMR Spectrum
Inference – Making Decisions

• Good experimental design
  – Screening designs (fractional factorial)
  – Response surface

• Underutilized in MS method development

• Watch for randomization restrictions in execution of experiment
  – Split plot / split-split plot
Multivariate Curve Resolution

- Mathematical model of two-way data

\[
\begin{bmatrix}
\text{D} \\
\end{bmatrix} =
\begin{bmatrix}
\text{C} \\
\end{bmatrix} 
\begin{bmatrix}
\text{S} \\
\end{bmatrix}
\]

- Initial estimate of C matrix using factor analysis

- Solve for S using matrix algebra: \( S = (C'C)^{-1}D \)

- Solve for C using matrix algebra: \( C = D(SS')^{-1} \)

- Iteration

Single ion chromatograms
Relative concentration profiles for compounds
Spectra for each compound
Multivariate Curve Resolution

• Many proposed methods
  – Evolving factor analysis
  – Windowed factor analysis
  – Alternating least squares

• MATLAB examples (Roma Tauler)
Multivariate Curve Resolution

- Mass spectrometry poses unique challenges
  - Dynamic range
• “All models are wrong but some are useful.”
  – G.E.P. Box

• Goal: make accurate predictions using model

• Interpretation of model not of interest

• Accuracy*Interpretability < Breiman’s Constant
  – Leo Breiman
# Prediction

## Short List of Tools

<table>
<thead>
<tr>
<th>Classification</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discriminant analysis</td>
<td>x</td>
</tr>
<tr>
<td>(linear, quadratic, flexible)</td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>x</td>
</tr>
<tr>
<td>Multiple linear regression</td>
<td>x</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>x</td>
</tr>
<tr>
<td>Projection pursuit regression</td>
<td>x</td>
</tr>
<tr>
<td>MARS</td>
<td></td>
</tr>
<tr>
<td>Partial Least Squares</td>
<td>x</td>
</tr>
<tr>
<td>Neural networks</td>
<td>x</td>
</tr>
<tr>
<td>k-Nearest neighbor</td>
<td>x</td>
</tr>
<tr>
<td>Kernel methods (support vector machine)</td>
<td>x</td>
</tr>
<tr>
<td>Decision trees</td>
<td>x</td>
</tr>
</tbody>
</table>

**Note:** The table indicates whether a tool is suitable for classification (X) or regression (x).
Prediction

- Decision Trees
  - Fast, capable of handling large data sets
  - Recursively partition data into homogeneous sets
  - Can capture nonlinear behavior
  - Can capture interactions

```
If clogP<4
  If MW<400
    active
  If MW<400
    inactive
  If # rings > 1
    inactive
  If clogP<5.5
    active
  inactive
```
Prediction

• Support Vector Machines
  – Promising results thus far in several domains
  – Advantages
    • Principled way to build highly nonlinear models
    • Incorporation of domain specific information
Prediction

• Ensemble Methods
  – Average predictions from several models to improve overall prediction accuracy
  – Bagging
  – Boosting
  – Revolutionized prediction accuracy
Prediction

**Overfitting**
Analogous to fitting high degree polynomials to a calibration curve
Prediction

• Overfitting is a huge problem 🚸
  – Must be vigilant about using independent test data to assess model performance
  – Leave-one-out cross-validation is too optimistic!
  – Leave $\frac{1}{2}$ out cross-validation preferred

• If test data used to modify model, prediction performance will be biased
Prediction

- **Prediction Error**
  - **Model Complexity**: Low - High
  - **Training Sample**: Low Bias - Low Variance
  - **Test Sample**: High Bias - High Variance

- **Prediction Error** vs. **Model Complexity** graph
  - **Low**: Low Bias, Low Variance
  - **High**: Low Bias, High Variance
Data Mining

• Finding "useful" patterns or relationships in observational databases

• Spurious correlations abound

• Hypothesis generation
  – Confirmed by designed experiments
Data Mining

• "Unsupervised Methods"
  – Clustering
  – Association rules
  – No clear measure of success
  – Proliferation of methods since effectiveness is a matter of opinion
Analyzing Data - Take Home

• Beware of “fuzzy neural genetic …”

• Have multiple tools at your disposal
  – “To a little boy with a hammer, all the world’s a nail.”
  – Seek collaborators with big tool boxes

• Stay focused and principled
  – What’s the essence of the problem
  – What method best solves the problem
  – Not "...how can I apply tool X here..."
Implementing

• Hardware
  – Symmetric Multiprocessor (SMP)
    • Multiple processors all sharing same memory space
    • Tens of processors, tens of Gbytes of memory, Terabytes of disk
    • Good for big memory jobs, file serving
    • Operating system handles load balancing
    • Expensive, infrequent upgrades
Example System (Lilly Natural Products):
Sun E6000
Ten 333 MHz UltraSparc processors
4 GBytes RAM
1.5 TeraBytes disk
Implementing

• Hardware
  – Linux clusters
    • Unbeatable single CPU price/performance
    • Scalable and less expensive than SMPs
    • Requires queuing software
    • Small memory jobs
    • Unbeatable overall performance for many applications
Linux Cluster (version 1.0)

40 Desktop PCs on HPLC carts (733 MHz PIII)

Simple and inexpensive

No centralized admin.

"Total Cost of Ownership" lower than some will claim
Linux Cluster (version 1.1)

20 CPUs
• 866 MHz PIII
• 512 MB RAM

500 Gbytes Raid-5 Disk
Implementing

• **Software**
  
  – General tools (we do GNU, www.gnu.org)
    
    • gcc, ghostscript, bash, gzip, emacs
    
    • Consistent development environment

  – Queuing software for clusters
    
    • LSF – full featured, $$ (www.platform.com)
    
    • Grid Engine - less features, no $$ (www.sun.com)
Implementing

• Software
  – Experimental design & interactive analysis
    • JMP (www.sas.com)

  – Rapid prototyping
    • MATLAB (www.mathworks.com)
    • R / Splus (www.r-project.org, www.insightful.com)
    • Perl
Implementing

• Software
  – Decision trees
    • R / Splus
    • SAS Enterprise Miner
    • Salford Systems (www.salford-systems.com)
  – Production deployment
    • SAS (www.sas.com)
    • Perl
    • C/C++
Case Study #1 - Complex Mixtures

• Business problem
  – Construct purified natural product library
  – Strategy
    • Use analytical measurements to direct purification

• Data analysis issues
  – Tracking and storage of ~$10^6$ MS injections
  – Diversity analysis of complex mixtures
  – Data mining (searching) of database
Case Study #1 - Complex Mixtures

• Path we took
  – SMP Sun Server
  – 1.5 Terabytes of netCDF files
  – Database associating experiments with filenames

• Strategy
  • Automated analysis required
  • Visualization needed to validate algorithms
Case Study #1 - Complex Mixtures

Sample Tracking Database

Navigator LC-MS System

network

Sun E6000 SMP System
10 CPUs
4 Gbytes RAM
1.5 Tbytes

SAS, Splus, Perl, C, PV-Wave
GNU tools
Case Study #1 - Complex Mixtures

Two Similar Red Algae (by +MS)

Unknown Red Alga (San Simeon, CA Jun '95)

Unknown Red Alga (Montano De Oro, CA Jun '95)
Case Study #1 - Complex Mixtures

Two Dissimilar Sponges (by +MS)

Verongula rigida

Unknown
Encrusting Sponge
Case Study #2 - FTMS

• Business problem
  – Screen traditionally difficult targets
  – Strategy
    • Use FTMS to directly detect bound complexes

• Data analysis issues
  – Discrete wavelet transform to compress spectra
  – Detecting target:ligand complexes
    • Matched filtering
Case Study #2 - FTMS

• Path we took
  – Bruker binary files
    • Cooperative vendor
  – Linux cluster
    • Saved > $400K relative to SMP solution
  – C++
    • Custom code
    • FFTW library
    • Wave++ library
  – No user interface required
    • Populate existing HTS databases
Case Study #2 - FTMS

File Server (Sun Workstation)

Local network Switch

Master Compute Node
20 Node Compute Cluster
866 MHz Pentium III's
512 MB RAM

500 Gbyte Disk

FTMS Control PC

FTMS

HTS Results Database

Lilly Network

= 100 Mbit/sec ethernet
= 100 Mbyte/sec Fibre channel
Case Study #2 - FTMS

Matched Filtering (5' 16S A-site)

Noiseless Spectrum

Noisy Spectrum (SNR=0.473)

(-5) Matched Filter Output
Case Study #3 - Proteomics

• Business problem
  – Estimate quality of "triple-play" ion trap MS data
    • Zoom scan
    • MS/MS data
  – Data sets with high MS quality and poor sequence database matches are of interest

• Data analysis issues
  – Robust estimation of noise levels
  – Integration into existing web environment
Case Study #3 - Proteomics

• Path we took
  – Report files (*.zta, *.dta) files
  – Windows NT required
  – Tight integration required
    • Use Perl as the glue
  – R statistical computing environment
    • Rapid development
Case Study #3 - Proteomics

Windows NT System

- R Statistical Computing Package
- Perl
- Sequest
- Web Server
- Summary MS files
  - *.dta, *.zta
- LC/Q Mass Spec
Case Study #3 - Proteomics

rat_serum_Urea_30kDa_C4_fr4.0590.0590.2
901.2 [1354.5] (2+) ZTA_Score=0.77

Y_box_protein 1 [Rattus norvegicus]
901.18 (2+) 36 [0.9633 292.4]
Resources

**General Programming**

**Experimental Design**

**Multivariate Curve Resolution**
R. Tauler, Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*, 30, 133-146 (1995).
Resources

Statistics / Machine Learning


www.jmlr.org
Resources

Support Vector Machines
www.kernel-machines.org
www.support-vector.net

Wavelets