



HPLC and Principal Components Analysis for Identification of Mycobacterial Strain

Scott Ramos

Infometrix, Inc.
Bothell, WA

Marie Scandone

Bio-Rad Laboratories, Inc.
Philadelphia, PA

58th Pittsburgh Conference
Chicago, IL
Presentation #1310-8



Agenda of Presentation

- Background on Mycobacteria
- Chromatography and alignment
- PCA and stages of analysis
- Classification results
- Discussion
- Conclusions



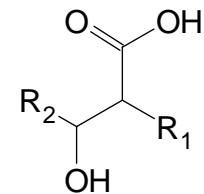
Background

- Tuberculosis has re-emerged as a global public health issue
- As much as 1/3 of world population may soon be infected
- Infection by drug-resistant strains is increasing
- HIV prevalence increases incidence of TB
- Infection by multiple strains of TB and/or by multiple species
- Concerns with respect to MOTT (Mycobacteria other than TB)

Mycobacteria Analysis

- Traditional
 - Biochemical tests
 - Laboratory options
 - Subjective interpretation
 - Time and cost
- DNA Probes
 - 16s RNA
 - Very selective and accurate
 - Only selected strains
- Chromatography
 - Mycolic acids - specificity
 - Variation in chain length and constituents
 - Some species indistinguishable

Mycolic Acids:
 α -branched, β -hydroxy long chain fatty acids



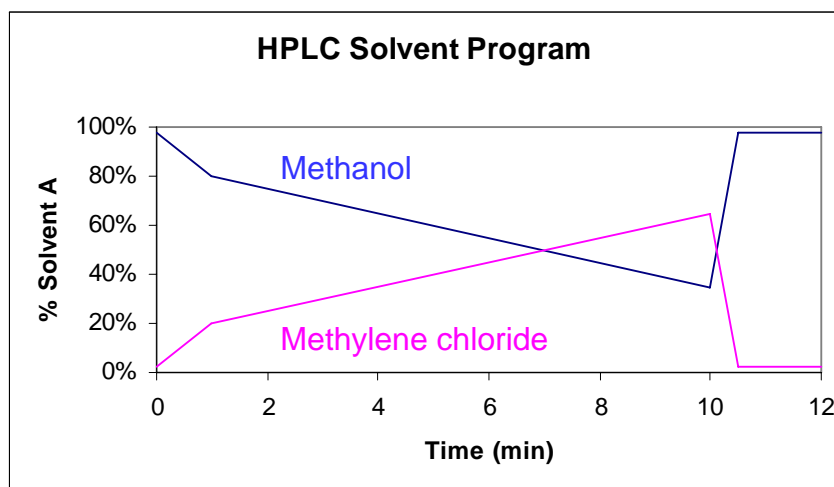


Chromatography for Mycobacteria Analysis

- TLC
 - First
 - Lack of automation
- GC
 - Rapid, reproducible, good resolution
 - Ignores Mycolic acids (60-90 carbons)
- HPLC
 - Species specificity
 - Adequate resolution

HPLC

- Rapid analysis
- UV or Fluorescence detection
- Peak heights/areas
 - Needs internal standards
 - Peak ID table transfer
- Whole chromatograms
 - Data rate
 - RT reproducibility (needs alignment)
 - File format

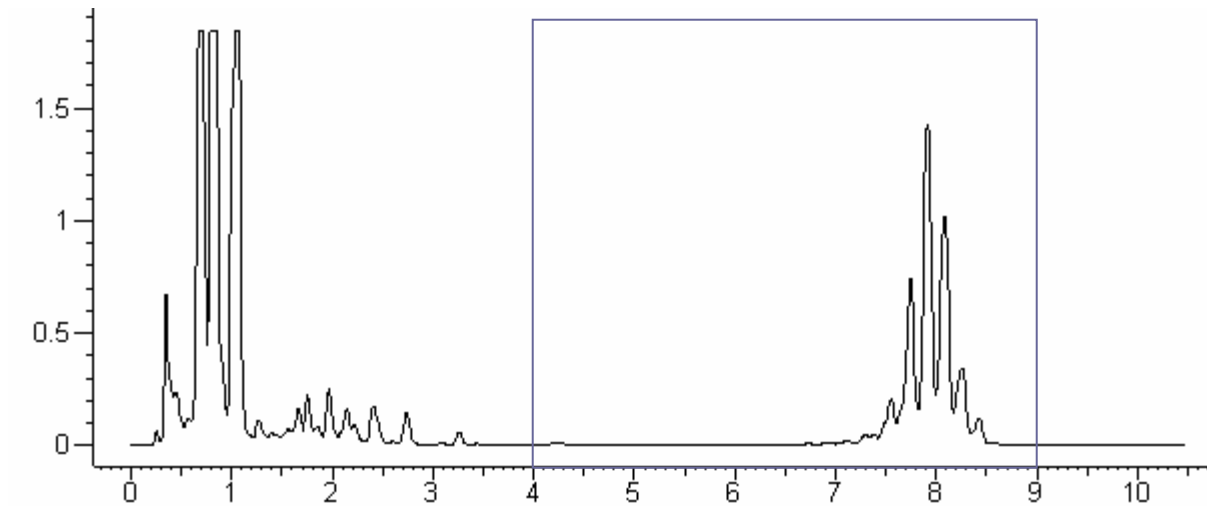




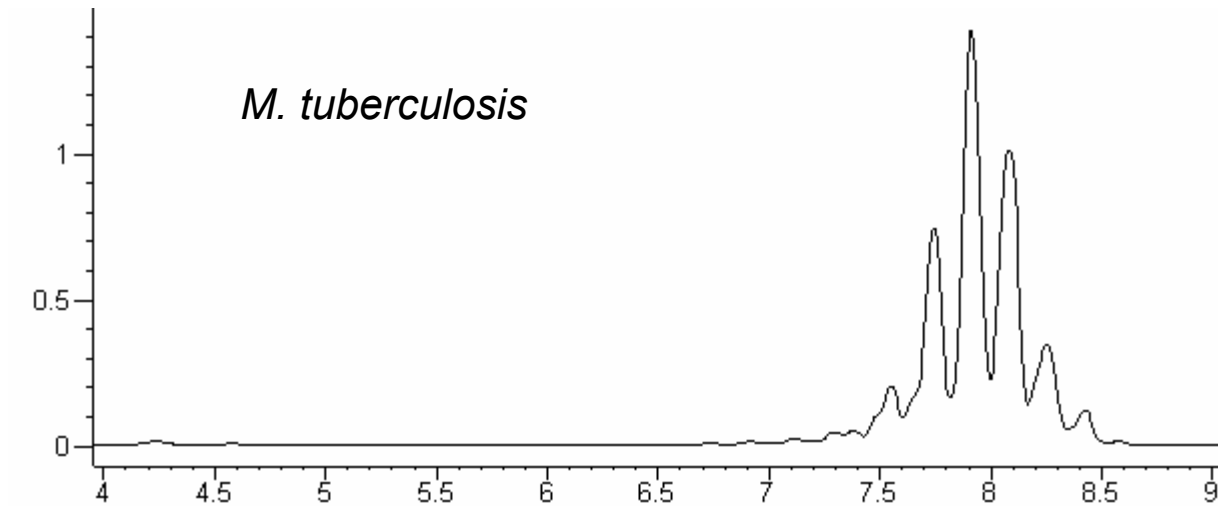
Chromatographic Data

- Data Source - HPLC Mycobacteria Users Group round robin study (5 laboratories)
 - Data from one lab, 322 samples in 23 strains
 - Data exported into CDF format from Agilent LC
 - Whole chromatograms imported into KnowItAll[®], stored as database
- Processing in KnowItAll 7.5 (Bio-Rad) and in Pirouette[®] 4.0 (Infometrix)

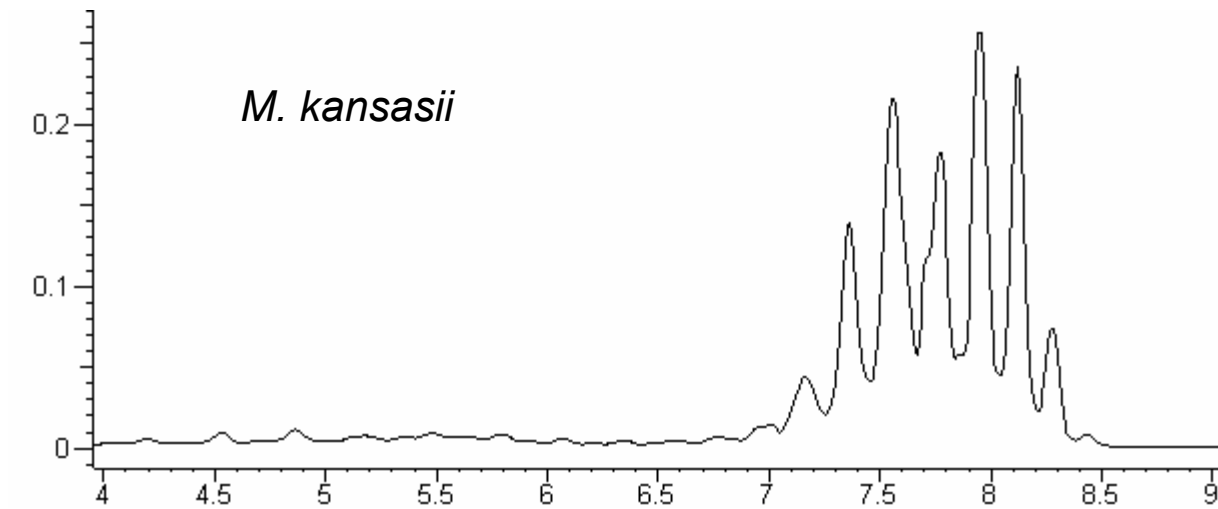
Mycobacteria Chromatogram



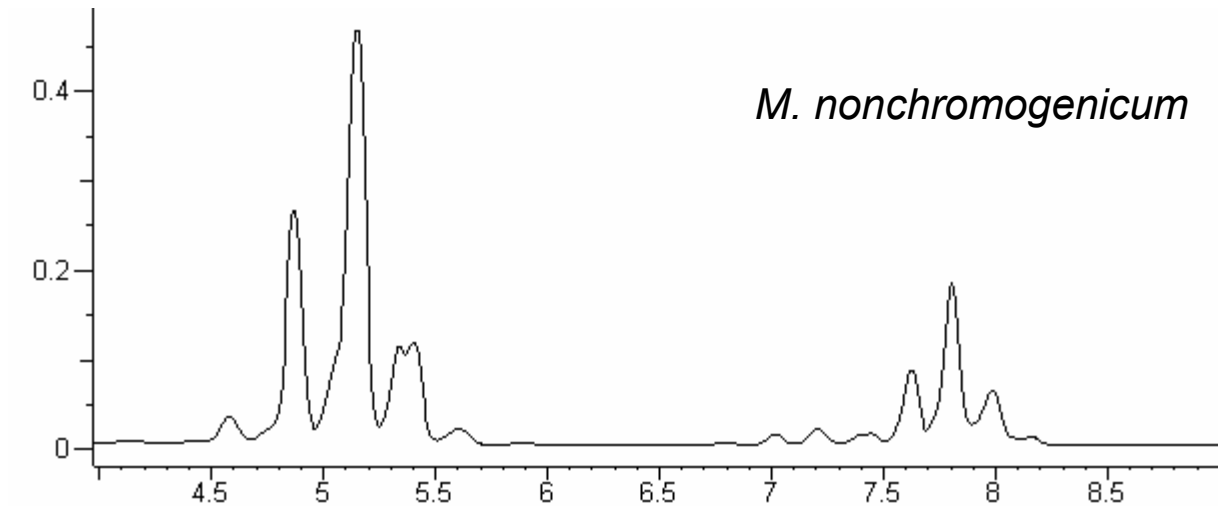
Diagnostic Region of Mycolic Acids



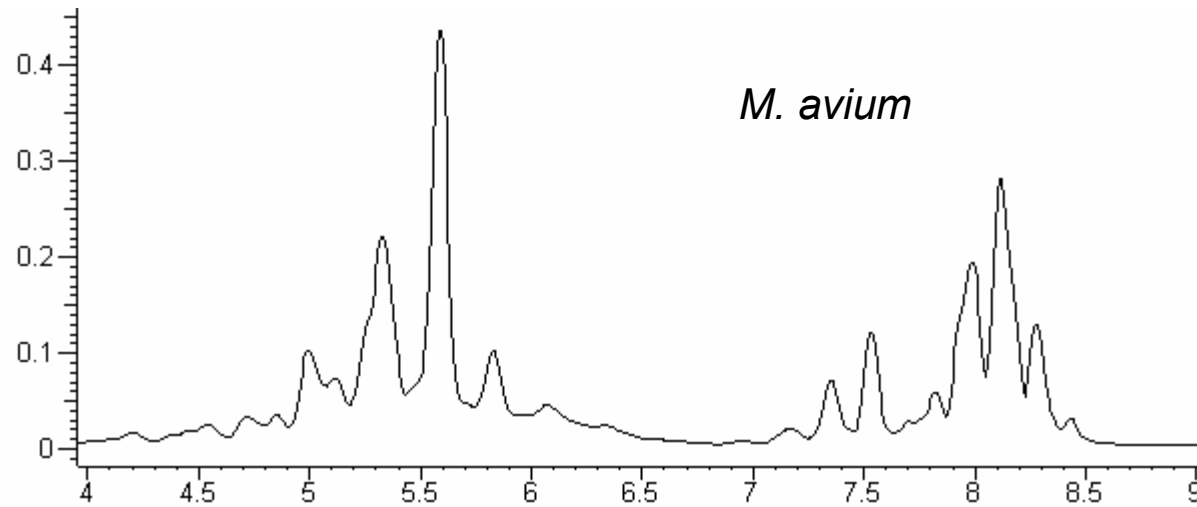
Single Cluster, Late Eluting



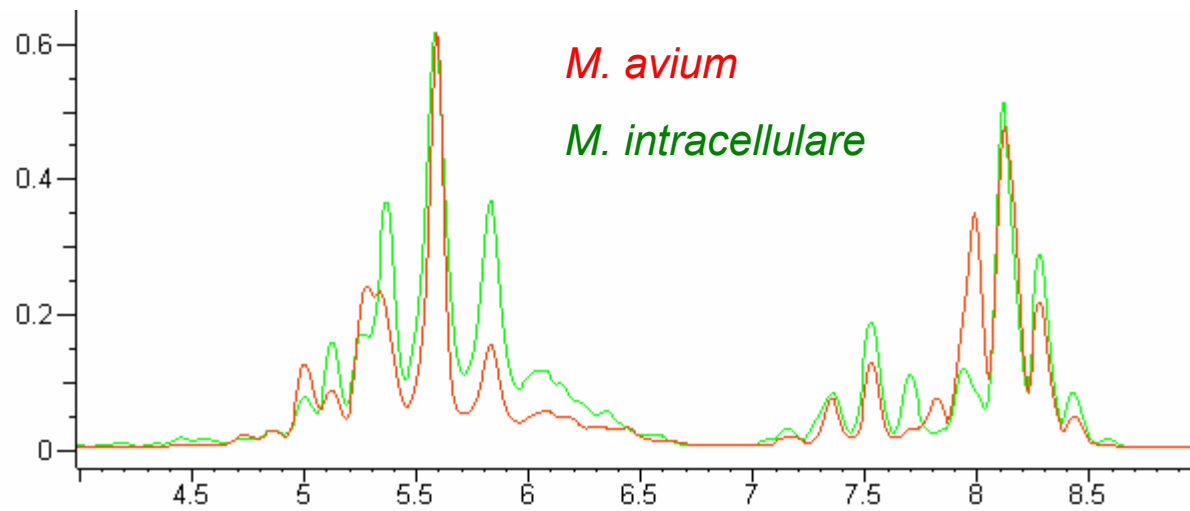
Double Cluster, Early Eluting



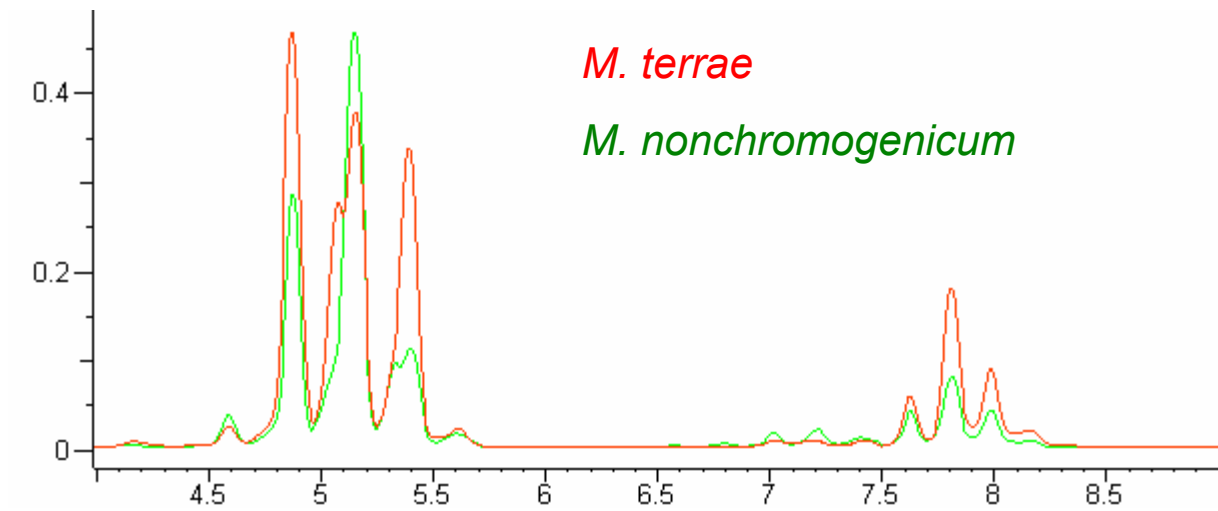
Double Cluster, Mid and Late Eluting



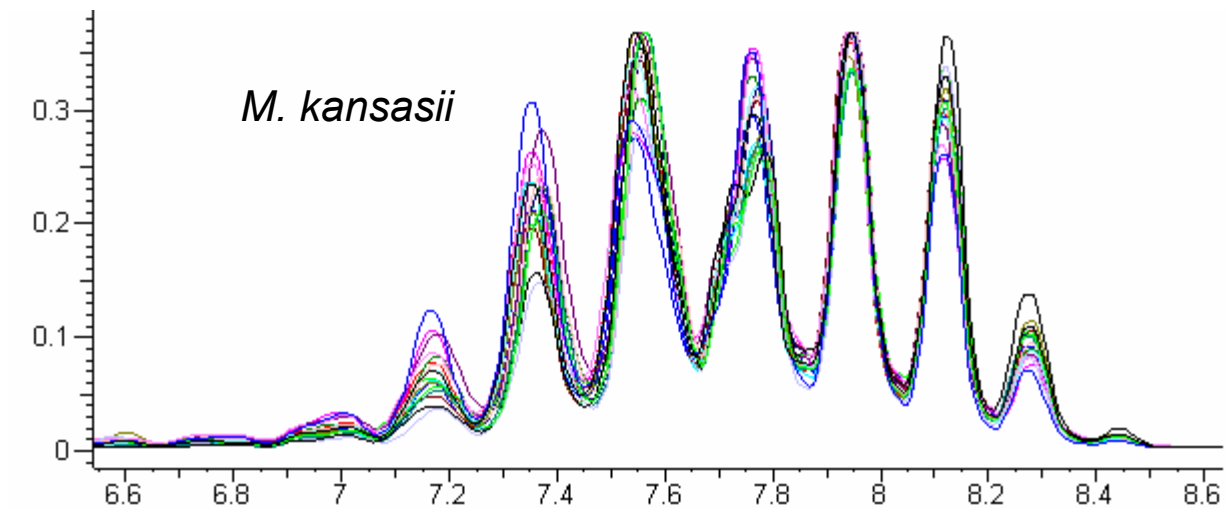
Similar Species



Similar Species

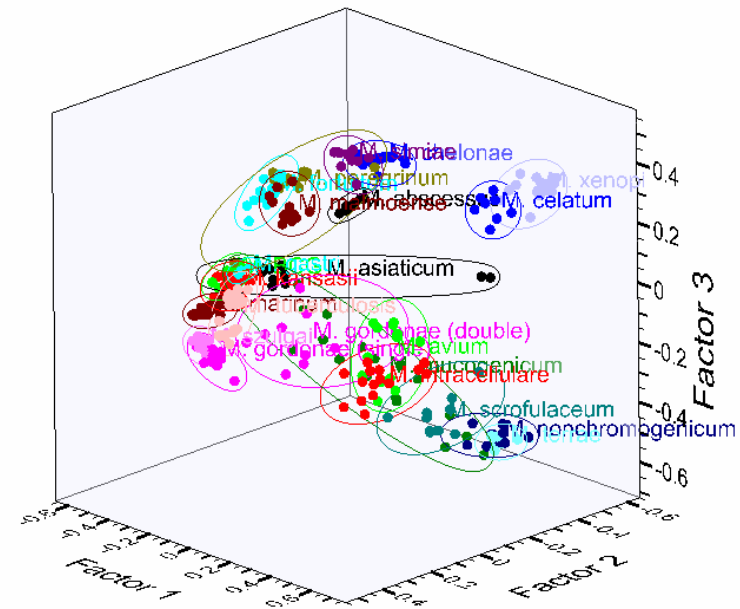


Species Variation



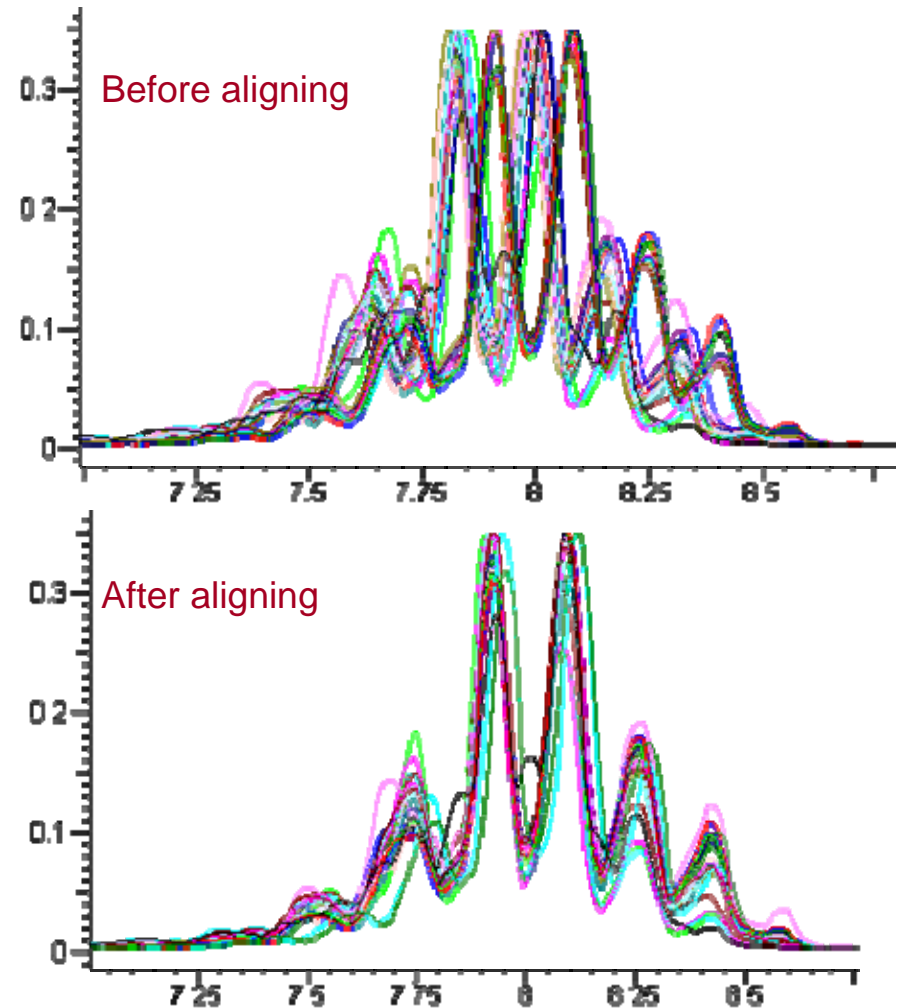
Principal Components Analysis

- Analyzelt™ MVP (KnowItAll) parameters
 - Time range restricted to 4.0 to 9.0 minutes
 - Vector-length normalization
 - Mean centering
- Results
 - Some clusters have large spread
 - Similar species pairs too far apart



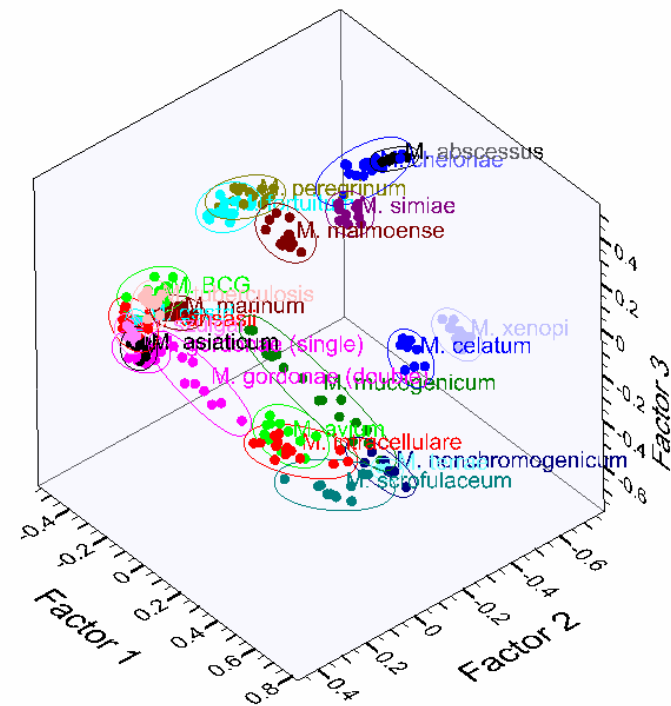
Principal Components Analysis

- Alignment
 - Alignment targets are category medians
 - Medians aligned first using internal standards
 - Correlation Optimized Warping (LineUp™) for final alignment
 - Example: *M. tuberculosis* profiles



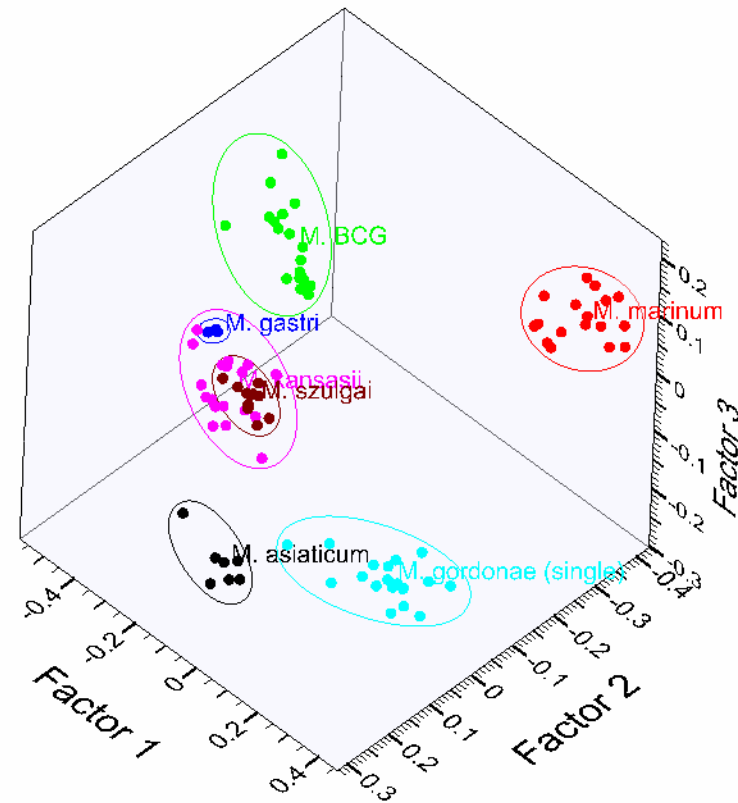
Principal Components Analysis

- Aligned profiles
- 11 outliers removed
- Processing
 - Time range restricted to 4.0 to 9.0 minutes
 - Vector-length normalization
 - Mean centering
- Similar species
 - *M. chelonae*, *M. abscessus*
 - *M. malmoense*, *M. simiae*
 - *M. fortuitum*, *M. peregrinum*
 - *M. xenopi*, *M. celatum*
 - *M. terrae*, *M. nonchromogenicum*
 - *M. avium*, *M. intracellulare*, *M. scrofulaceum*



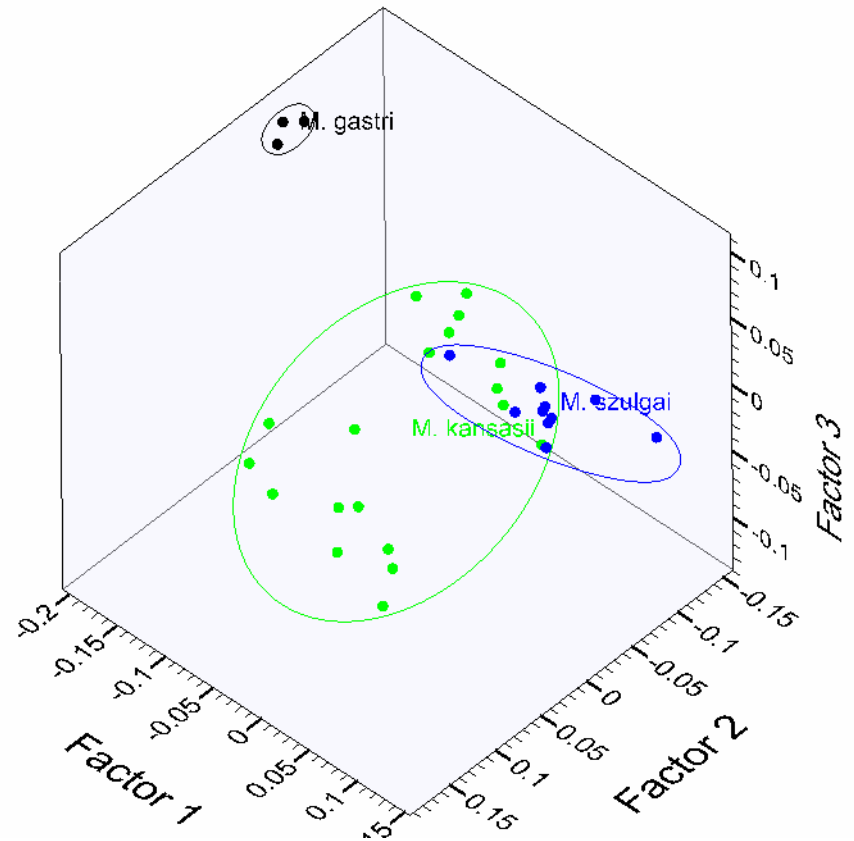
PCA on Subgroup

- 7 species appear in one cluster
- New subset created
 - Highlight samples of interest
 - Convert to Hit List
 - Pass to Analyzelt MVP
- PCA conditions same as prior analysis

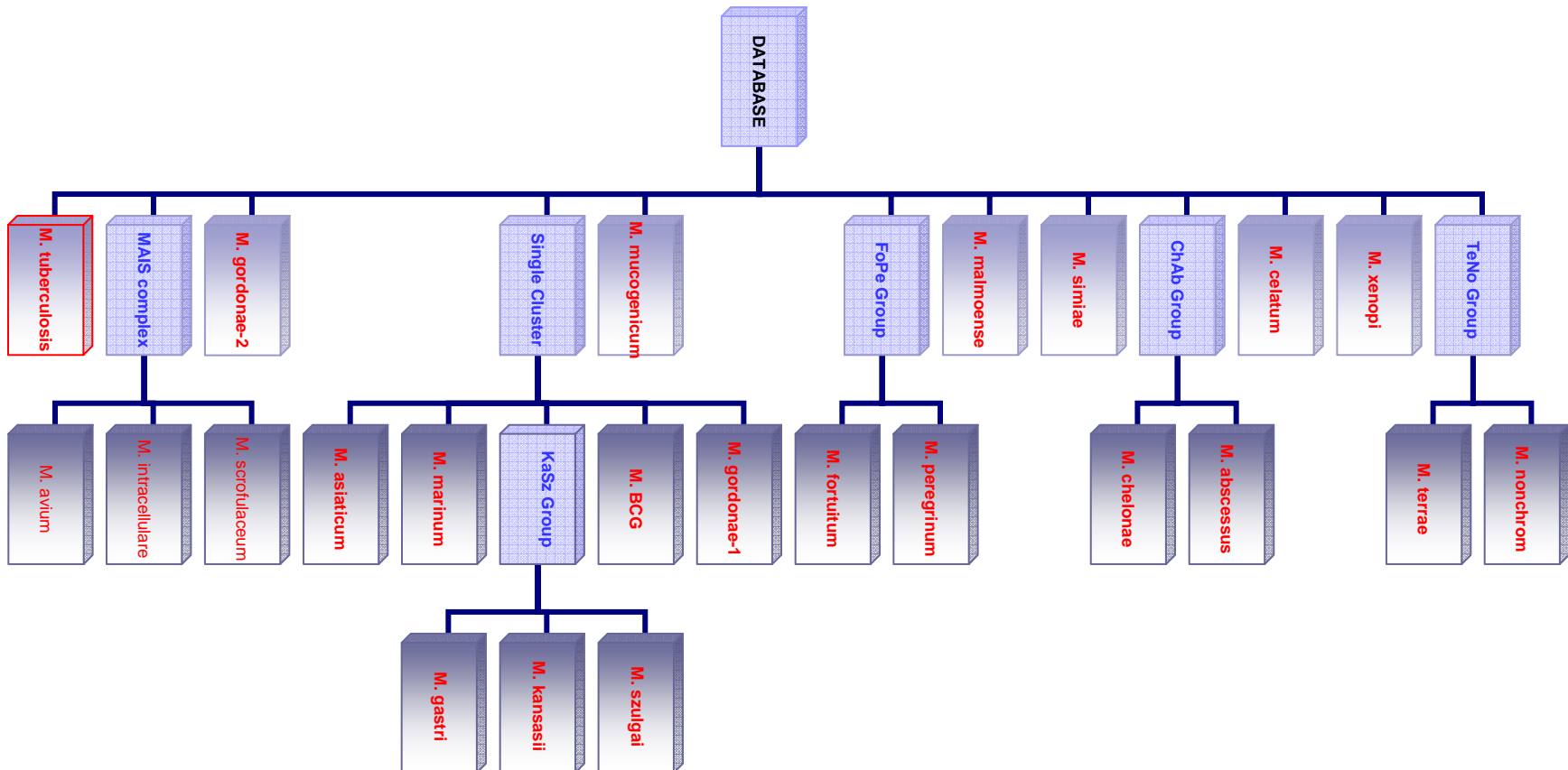


PCA on Subgroup of Subgroup

- 3 strains in one cluster
- New subset created and analyzed
- PCA, same conditions



Multivariate Decision Tree

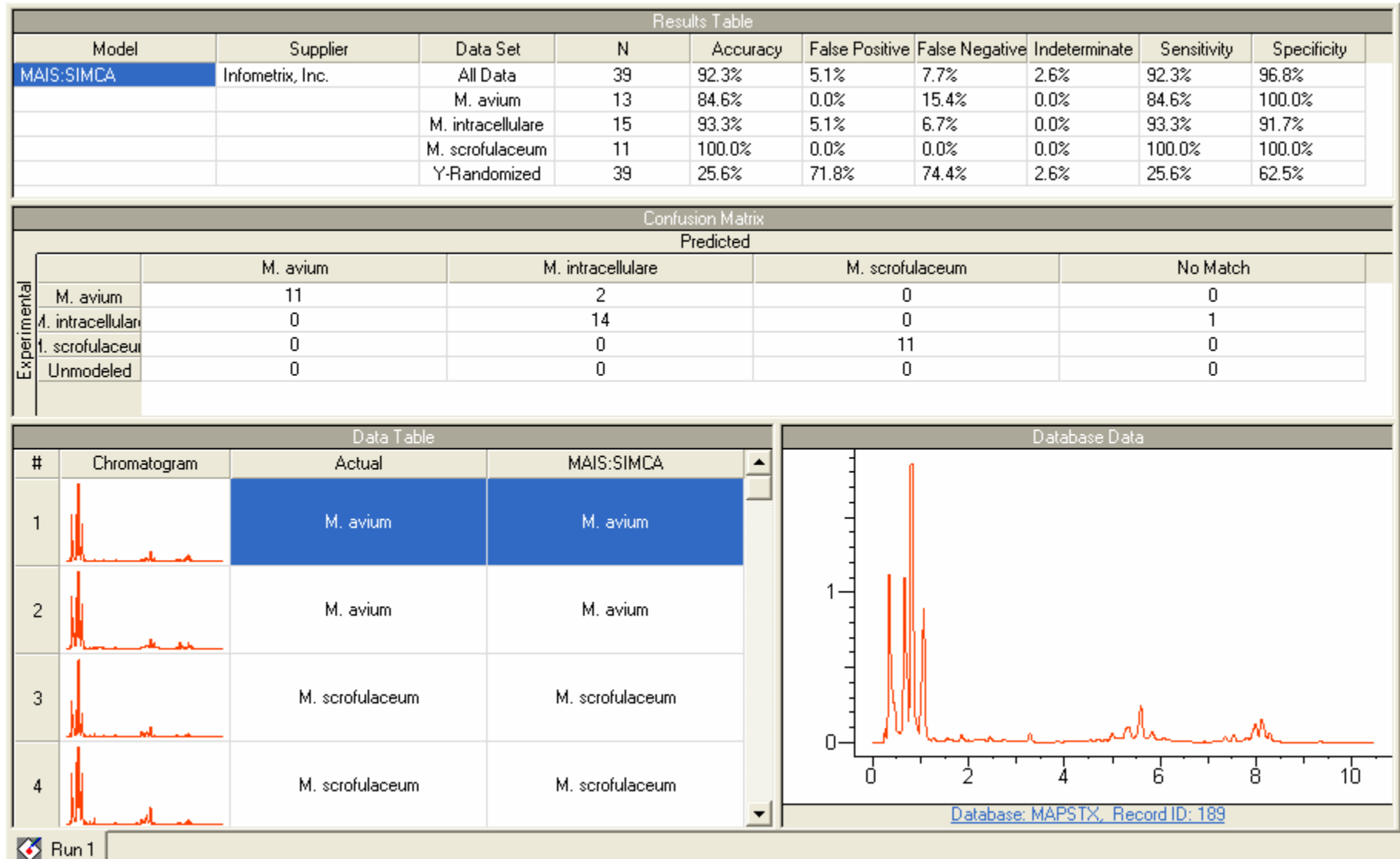




Classification Methodology

- KNN, SIMCA, PLS-DA models evaluated
 - KNN – evaluate up to 10 neighbors
 - SIMCA – one PCA model for each category
 - PLS-DA – one PLS model on a binary Y for each category
- One model (per algorithm) created at each node in decision tree
- Predictions run in Pirouette or in KnowItAll
 - KNN – consensus of nearest neighbors
 - SIMCA – class residual < threshold
 - PLS-DA – predicted Y > 0.5 and X residual probability < threshold

Model Validation in KnowItAll



Classification Results

- Divide each category into training (167 samples) and evaluation (143 samples) subsets*
- Make classification models on set of all training samples and on the 6 subgroups in the decision tree
- Predict on the corresponding evaluation sets
- Qualify results at $p < 0.05$
- Success rate, as fraction correct:

Modeling	TB	AV	IN	SC	GO	G2	KA	TE	NO	XE	MR	MU	ML	SI	SZ	BC	AS	GA	CE	FO	PE	CH	AB	Total
KNN	1.00	0.63	0.75	1.00	1.00	1.00	0.75	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93
SIMCA	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
PLS-DA	1.00	0.88	1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Prediction	TB	AV	IN	SC	GO	G2	KA	TE	NO	XE	MR	MU	ML	SI	SZ	BC	AS	GA	CE	FO	PE	CH	AB	Total
KNN	1.00	0.60	1.00	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				1.00	1.00	0.92	1.00	0.98
SIMCA	1.00	1.00	0.86	1.00	1.00	1.00	1.00		0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00				1.00	1.00	0.92	1.00	0.97
PLS-DA	1.00	0.80	1.00	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				1.00	1.00	0.85	1.00	0.98

* Kennard, R.W. and Stone, L.A. Computer aided design of experiments. Technometrics. 11(1):137-148 (1969).



Species Complexes

- With PCA, could not totally separate samples from strains in so-called MAIS cluster, containing
 - *M. avium*
 - *M. intracellulare*
 - *M. scrofulaceum*
- Pathology and treatment do not differ
- Maintain as a single group

- Other complexes
 - *M. chelonae*, *M. abscessus*
 - *M. fortuitum*, *M. peregrinum*
 - *M. terrae*, *M. nonchromogenicum*



Results

- With exception of *M. terrae* samples in KNN, modeling success was > 99% across all samples
- All algorithms produced the same success rate in prediction, > 97%
- If instead of ID to species, use complexes (where relevant), success rate was 100%



Considerations

- Previous work with Peak Heights

- 👍 Processing essentially the same among vendors, relatively easy for technician
- 👎 Complexity in establishing reliable peak ID tables among different laboratories
- 👎 Different peak finding and integration among software vendors
- 👎 May miss diagnostic peaks

- Current work with whole Profiles

- 👍 No peak ID table needed
- 👎 Chromatographic alignment mandatory; may require external software
- 👎 Variability in solvent programs among laboratories
- 👍 Captures nuances in profiles that do not qualify as peaks



Conclusions

- Combination of informatics database and chemometrics toolkit offers several advantages
 - Simple storage and mining of chromatographic profiles in a single database
 - Quick PCA tool for characterizing differences among samples in a data subset
 - Easy transfer of hit lists to external program for intensive multivariate modeling
 - Whole profile analysis as reliable as and can replace analysis of peak height data



Acknowledgments

- HPLC Mycobacteria User's Group
 - Standardized Method for HPLC Identification of Mycobacteria
 - Mycolic Acid Pattern Standards for HPLC Identification of Mycobacteria

http://www.cdc.gov/nchstp/tb/Laboratory_Services/Liquid_Chroma.htm