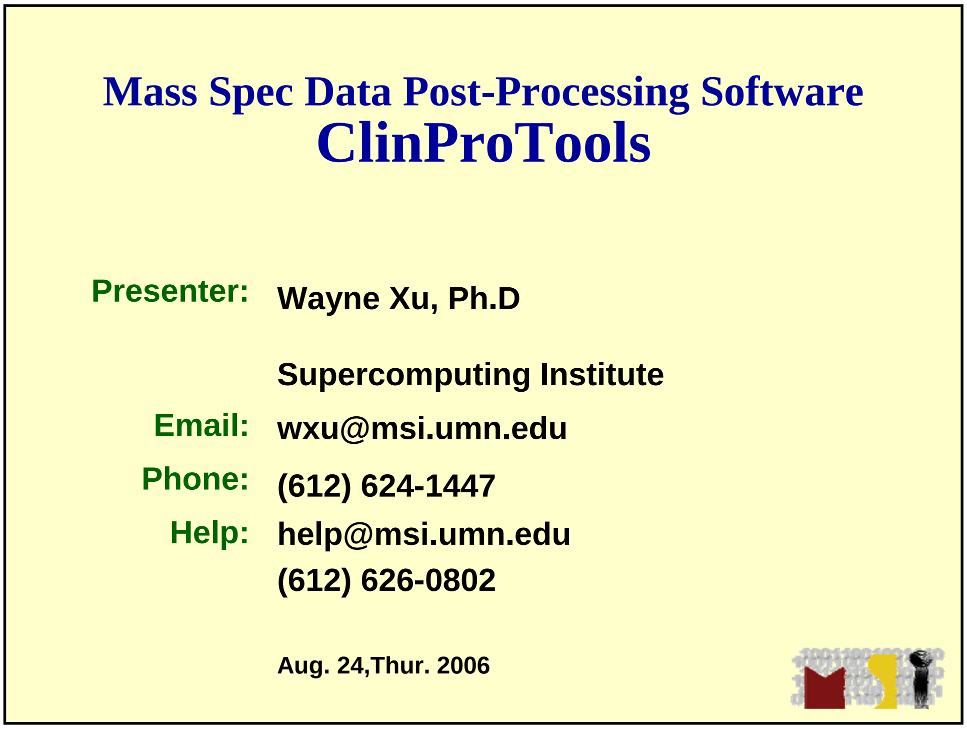


Mass Spec Data Post-Processing Software
ClinProTools

Presenter: Wayne Xu, Ph.D

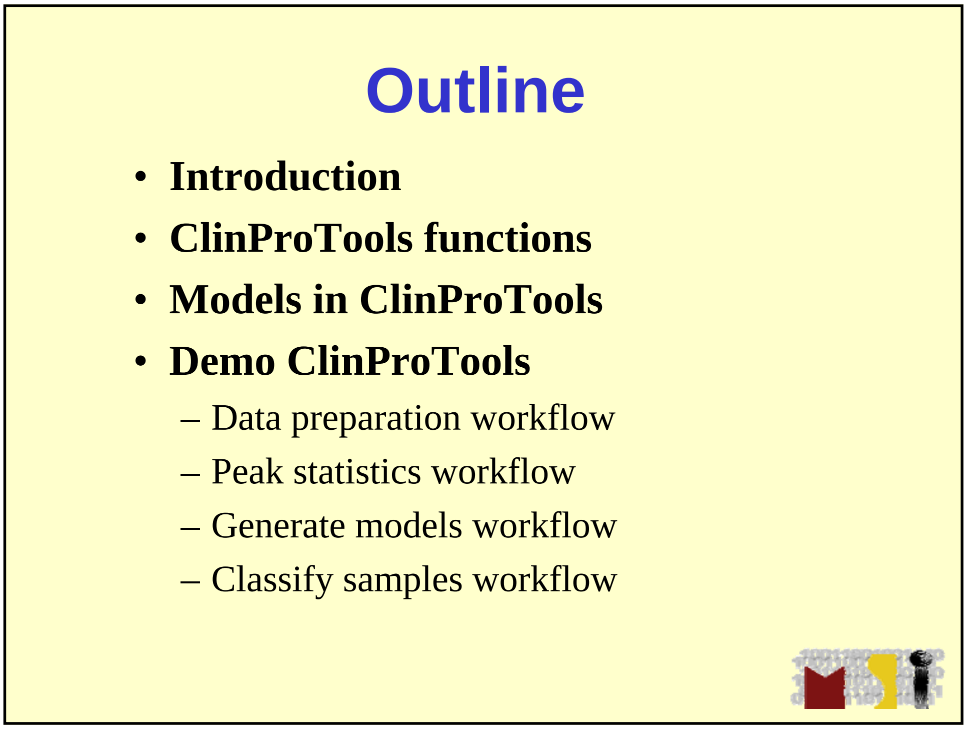
Supercomputing Institute

Email: wxu@msi.umn.edu

Phone: (612) 624-1447

Help: help@msi.umn.edu
(612) 626-0802

Aug. 24,Thur. 2006

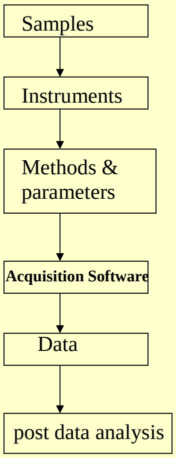


Outline

- **Introduction**
- **ClinProTools functions**
- **Models in ClinProTools**
- **Demo ClinProTools**
  - – Data preparation workflow
  - – Peak statistics workflow
  - – Generate models workflow
  - – Classify samples workflow

Introduction

Mass Data Acquisition & Post Processing



Samples	ClinProt kits
Instruments	Bruker's mass spectrometer
Methods & parameters	MALDI-TOF protocol_1
Acquisition Software	FlexControl
Data	Spectra data
post data analysis	ClinProTools, ….

Center for Mass Spectrometry and Proteomics, UMN,

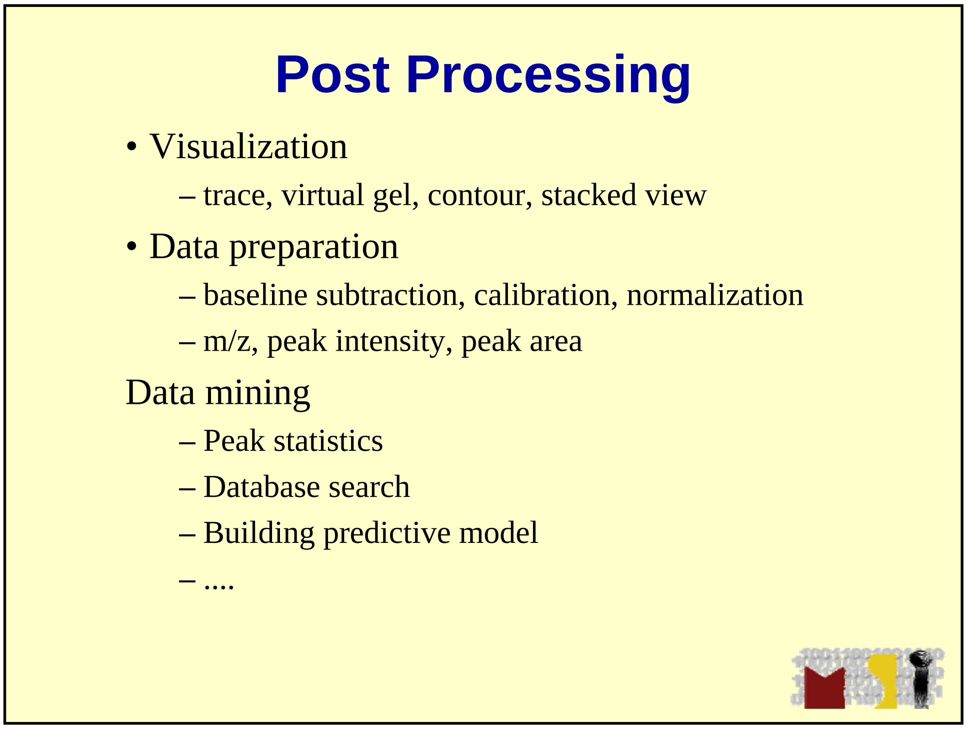
"Proteomics Workshop", 5-6 times / year

Post Data Analysis Software in Supercomputing Institute

- Post data analysis software: Analyze data that comes off from the mass spec instruments
  - Analyst QS
  - BioAnalysis
  - ClinProTools
  - Mascot
  - Sequest
  - Peaks Online
  - Pro ID
  - Pro QuanT
  - Pro TS Data
  - Scaffold

Sample Preparation

- Sample types
  - For peptide mass fingerprint
  - For tandem mass (ms/ms)
  - For quantitative Traq labeling
  - For intact mass (biological fluids)
- Biological fluids preparation (blood serum, blood plasma,..):
  - Highly concentrated components, similar mass to charge (m/z) ratio, may result in overlapping peaks.
  - A selective enrichment of specific proteins according to their biological, chemical or physical properties can improve spectra quality significantly.
  - Bruker’s enrichment/prefraction magnetic microbead system with different functionalized surfaces are provided as different profiling kits (ClinProt Kits) and each kit contains a detailed protocol for sample preparation (optimized on blood serum).
  - Mass analysis reproducibility is significantly depending on reproducibility of sample preparation



Post Processing

- Visualization
  - trace, virtual gel, contour, stacked view
- Data preparation
  - baseline subtraction, calibration, normalization
  - m/z, peak intensity, peak area

Data mining
  - Peak statistics
  - Database search
  - Building predictive model
  - ....

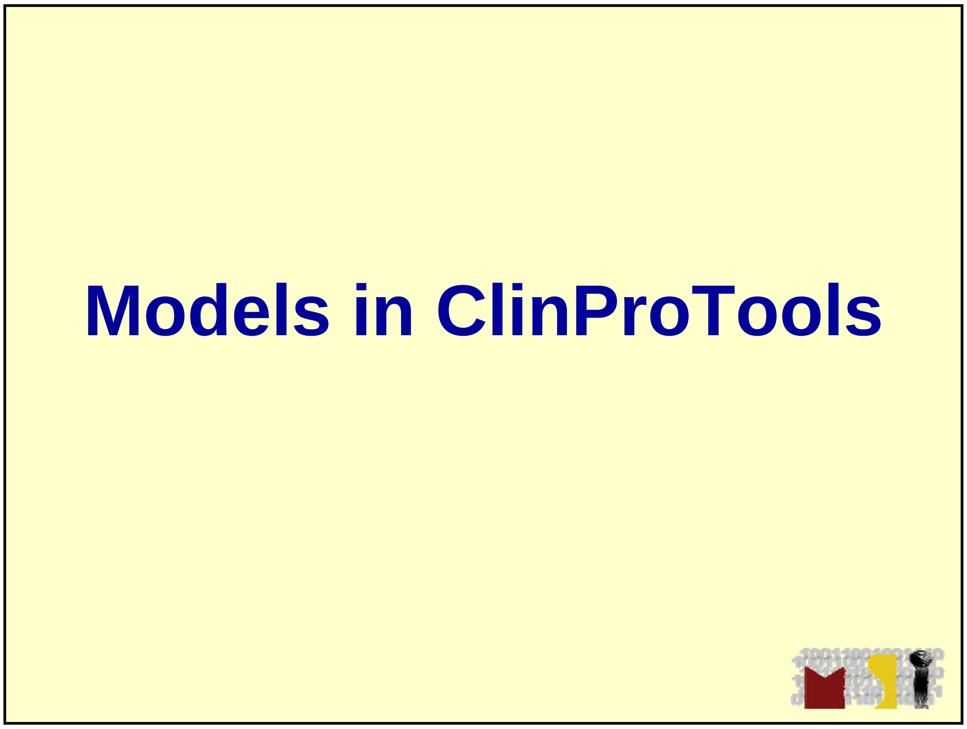
ClinProTools Functions

ClinProTools Functions

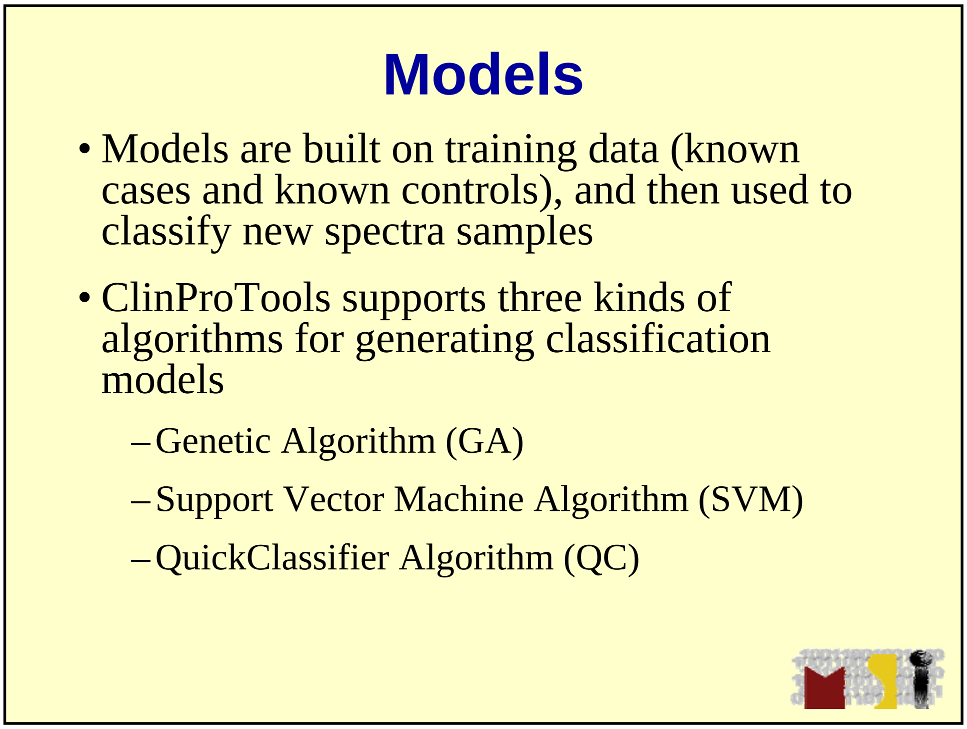
- Main:
  - Combine intuitive visualization and multiple mathematical algorithms to generate pattern recognition **models**
- Use:
  - Detect intact protein differential levels between cases and controls in order to discover **biomarkers to predict or diagnose** diseases

Features

- Visualization:
  - averaged spectra, compared spectra, and single spectrum with intuitive visualization features such as trace, virtual gel, contour and stacked views
- Data normalization:
  - processing parameters for baseline substraction, peak definition, calibration, normalization
- Data mining:
  - Averages and compares peaks from different spectra
  - Generates and validates pattern recognition models using different sophisticated mathematical and bioinformatic algorithms
- Biomarkers:
  - Highlights the locations of the biomarkers and allows users to visually inspect individual spectrum to verify their results
- Results:
  - Stores the detailed results for each analysis

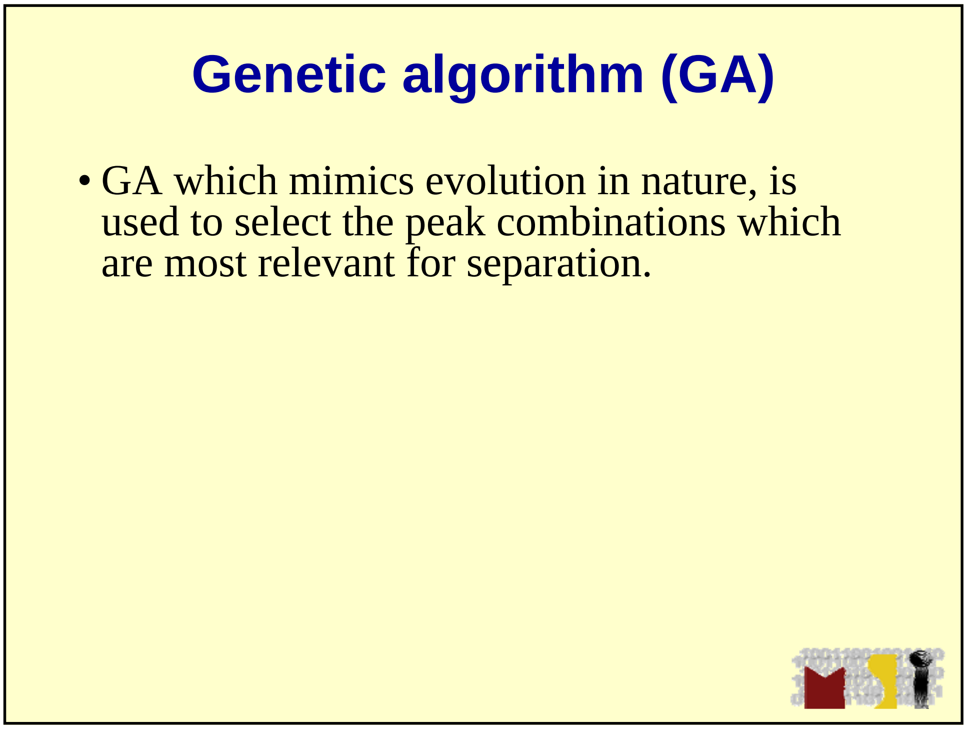


Models in ClinProTools



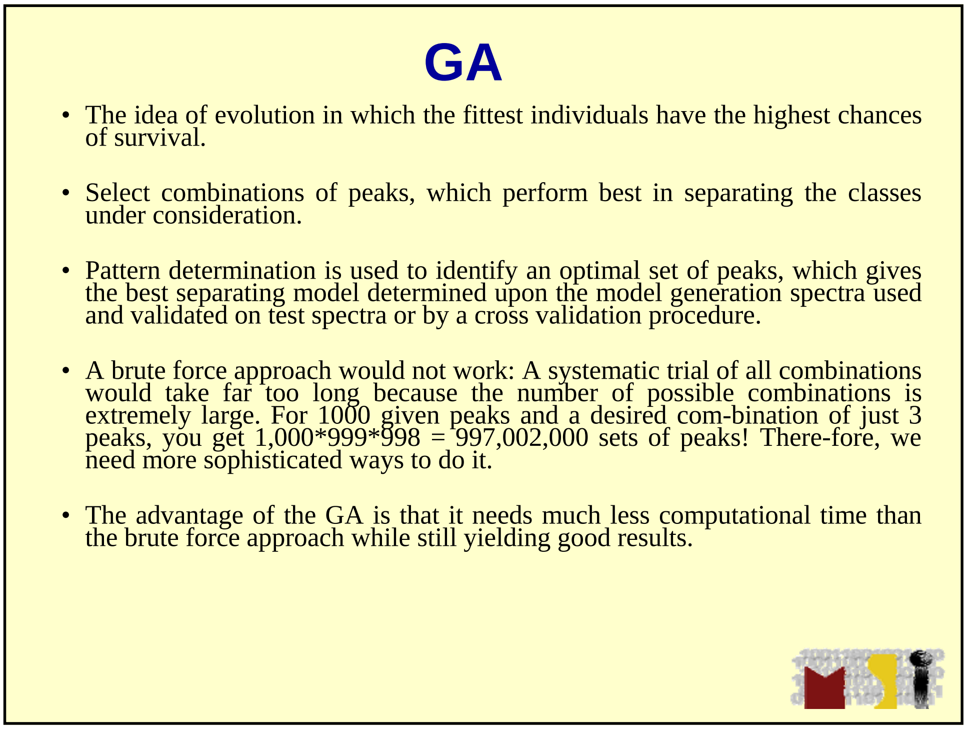
Models

- Models are built on training data (known cases and known controls), and then used to classify new spectra samples
- ClinProTools supports three kinds of algorithms for generating classification models
  - Genetic Algorithm (GA)
  - Support Vector Machine Algorithm (SVM)
  - QuickClassifier Algorithm (QC)



Genetic algorithm (GA)

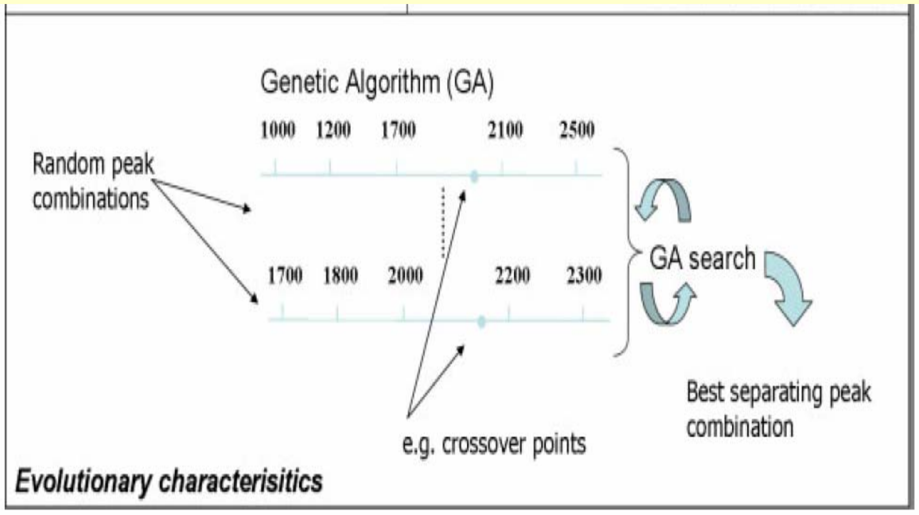
- GA which mimics evolution in nature, is used to select the peak combinations which are most relevant for separation.



GA

- The idea of evolution in which the fittest individuals have the highest chances of survival.
- Select combinations of peaks, which perform best in separating the classes under consideration.
- Pattern determination is used to identify an optimal set of peaks, which gives the best separating model determined upon the model generation spectra used and validated on test spectra or by a cross validation procedure.
- A brute force approach would not work: A systematic trial of all combinations would take far too long because the number of possible combinations is extremely large. For 1000 given peaks and a desired com-bination of just 3 peaks, you get 1,000*999*998 = 997,002,000 sets of peaks! There-fore, we need more sophisticated ways to do it.
- The advantage of the GA is that it needs much less computational time than the brute force approach while still yielding good results.

GA



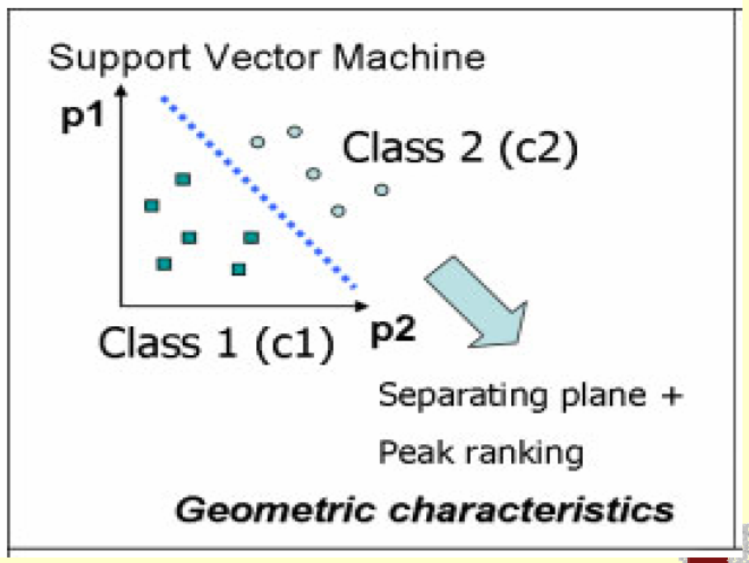
GA

- This is done by optimizing a cost function, which aims on optimal class separation with variance high between classes. Using the cost function each peak combination is rated by an expense factor, which is used as a measure for the fitness.
- The crossover combines randomly selected pairs of peak combinations to produce child peak combinations, which replace their parent peak combination. The intention here is to combine two fairly good peak combinations to form even better ones.
- The expectation is that the average fitness of all peak combinations rises and the best fitness observed will improve.
- The result of the GA is the peak combination which is proved to separate best between the different classes.

Support Vector Machine Algorithm (SVM)

- SVM is motivated from statistical learning theory and is at first used to determine separation planes between the different data classes. Upon the obtained planes, a peak ranking can be calculated in a second step.

Support Vector Machine Algorithm (SVM)



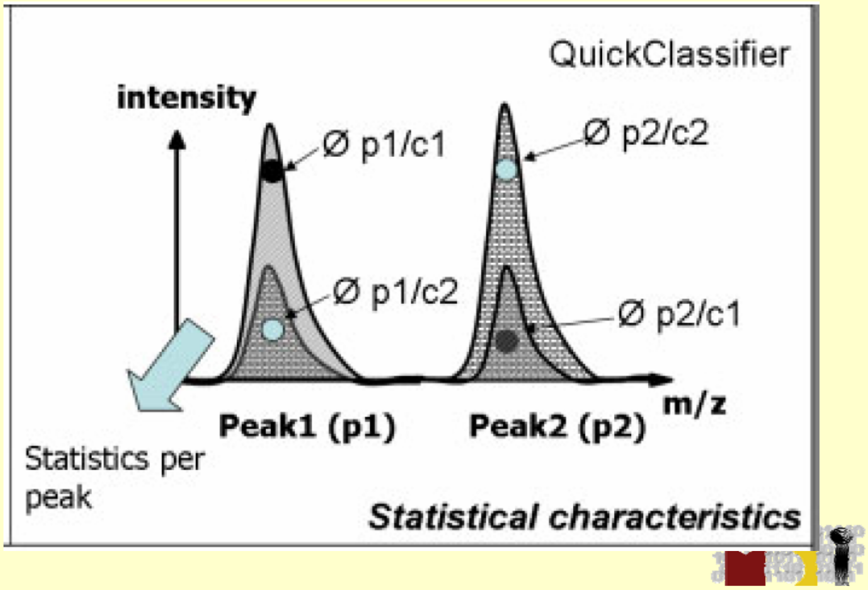
VSM

- A peak ranking is derived from the obtained hyperplane solution. The procedure is iterated until for each class a classifier (class vs rest) is obtained.
- Upon the obtained SVM model the best number of peaks is determined by a clustering in the subspace taken from the k best peaks and the (best) solution is stored as the final model.

QuickClassifier Algorithm (QC)

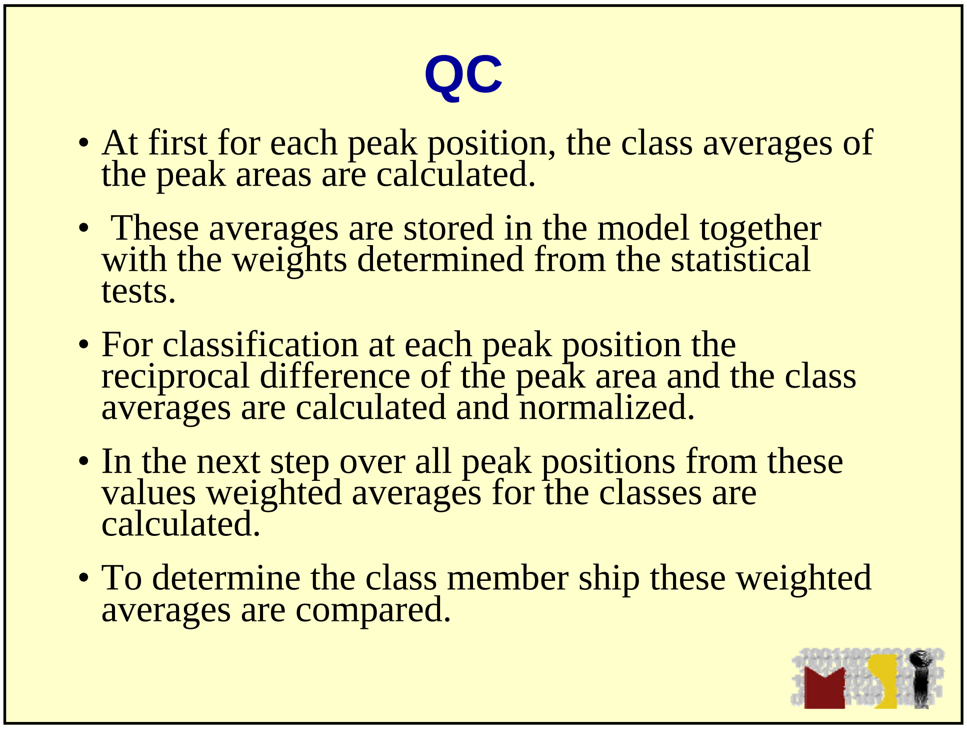
- a univariate sorting algorithm. The class averages of the peak areas are stored in the model together with some statistical data like the p-values at certain peak positions.
- For classification, the peak areas are sorted per peak and a weighted average over all peaks is calculated.

QuickClassifier Algorithm (QC)



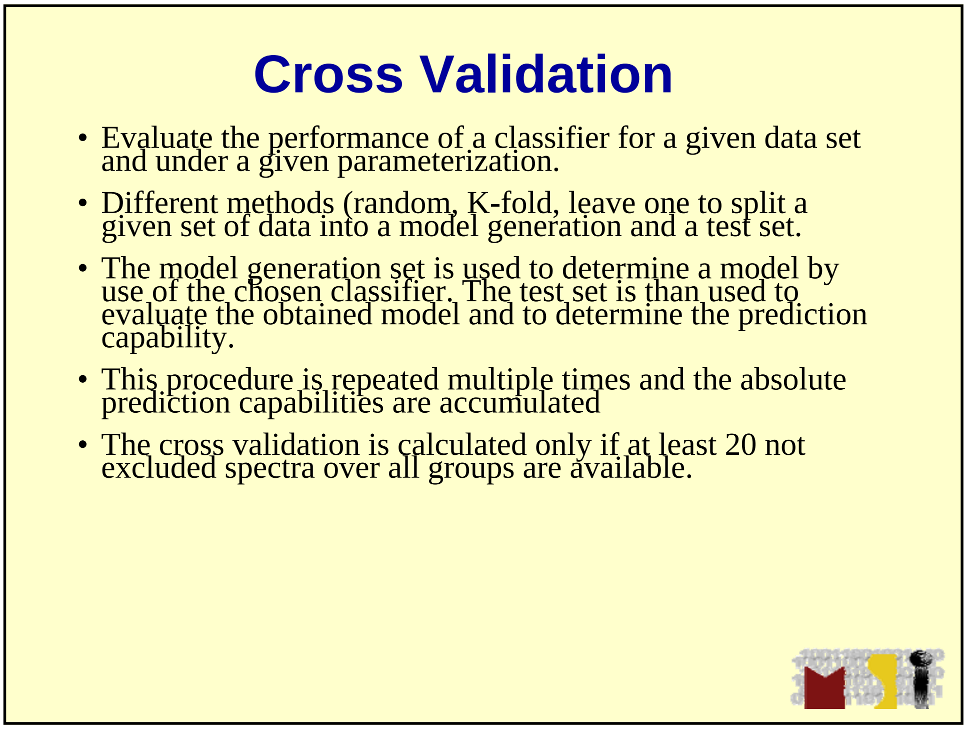
QC

- The QC algorithm has a good performance.
- The classification allows not only determining the class membership but calculates also a likeliness for each class.
- If there are only few samples available for the model generation, the validity of the classification seems to be better in comparison to other algorithms in many cases.



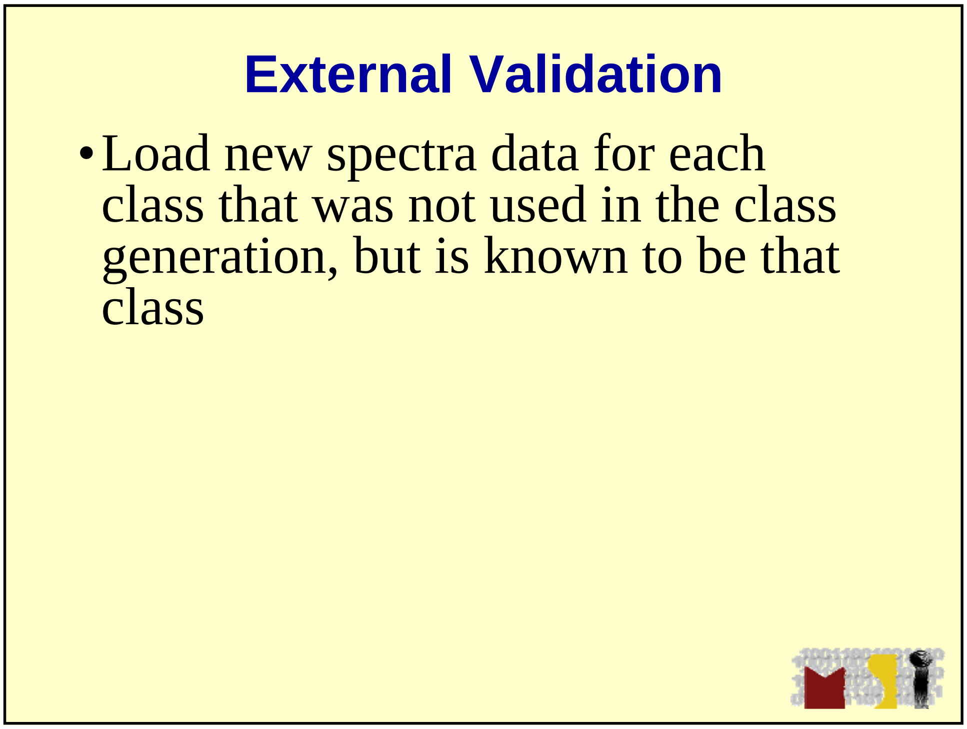
QC

- At first for each peak position, the class averages of the peak areas are calculated.
- These averages are stored in the model together with the weights determined from the statistical tests.
- For classification at each peak position the reciprocal difference of the peak area and the class averages are calculated and normalized.
- In the next step over all peak positions from these values weighted averages for the classes are calculated.
- To determine the class member ship these weighted averages are compared.



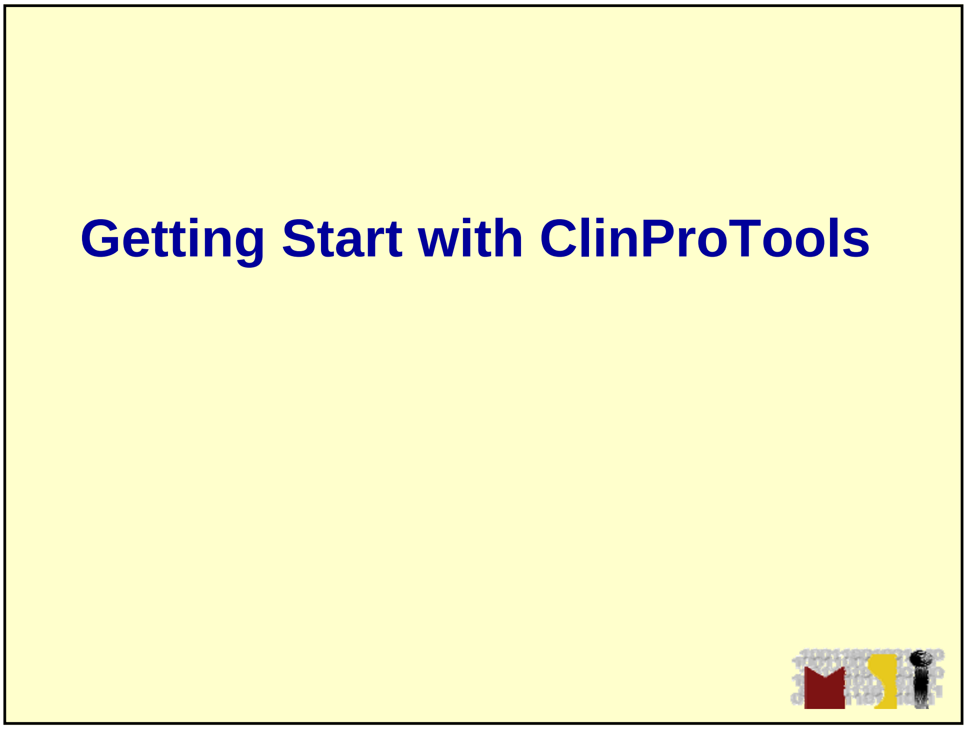
Cross Validation

- Evaluate the performance of a classifier for a given data set and under a given parameterization.
- Different methods (random, K-fold, leave one to split a given set of data into a model generation and a test set.
- The model generation set is used to determine a model by use of the chosen classifier. The test set is than used to evaluate the obtained model and to determine the prediction capability.
- This procedure is repeated multiple times and the absolute prediction capabilities are accumulated
- The cross validation is calculated only if at least 20 not excluded spectra over all groups are available.

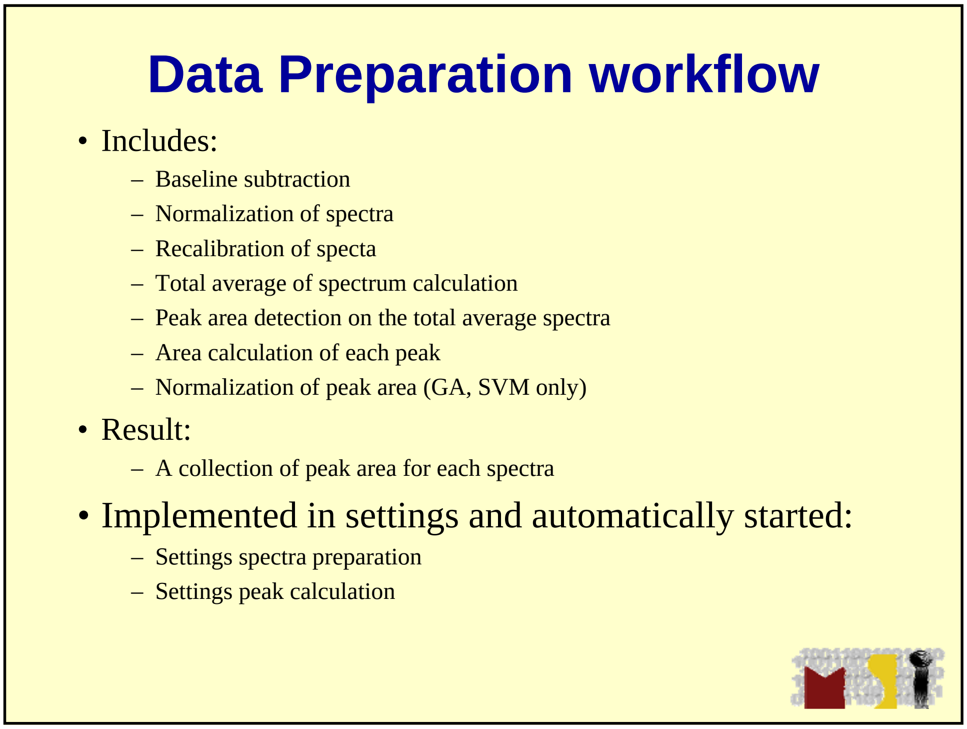


External Validation

- Load new spectra data for each class that was not used in the class generation, but is known to be that class

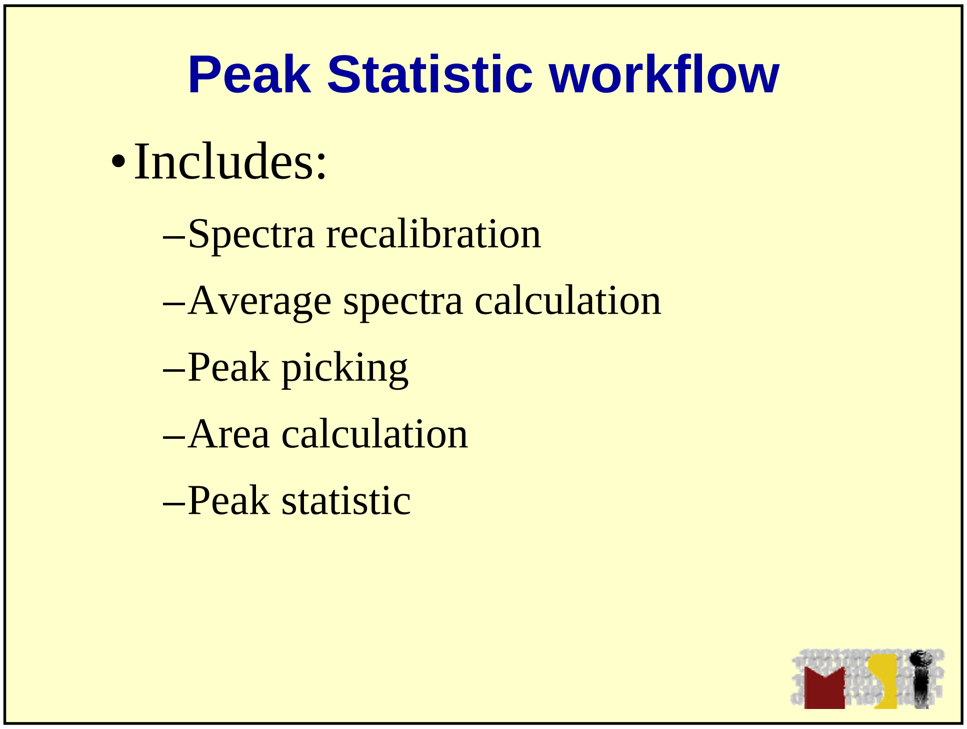


Getting Start with ClinProTools



Data Preparation workflow

- Includes:
  - Baseline subtraction
  - Normalization of spectra
  - Recalibration of specta
  - Total average of spectrum calculation
  - Peak area detection on the total average spectra
  - Area calculation of each peak
  - Normalization of peak area (GA, SVM only)
- Result:
  - A collection of peak area for each spectra
- Implemented in settings and automatically started:
  - Settings spectra preparation
  - Settings peak calculation



Peak Statistic workflow

- Includes:
  - Spectra recalibration
  - Average spectra calculation
  - Peak picking
  - Area calculation
  - Peak statistic

Peak Statistic workflow

- Input files (more than 1):

  Open Model Generation Class

- Start peak statistic:

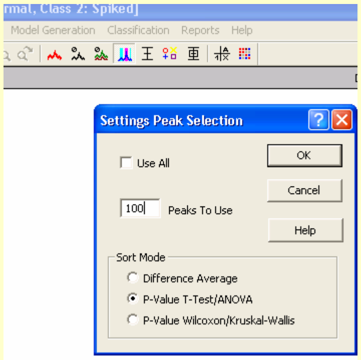
  Reports | Peak Statistic

  (spectra recalibration, averaging, calculation)

Peak Selection Settings

The picked Peak number will strongly influence the quality of model algorithms. In many cases, a reasonable reduction of peaks improves the classification.

Default: use all, or 100



Settings Peak Selection

□ Use All

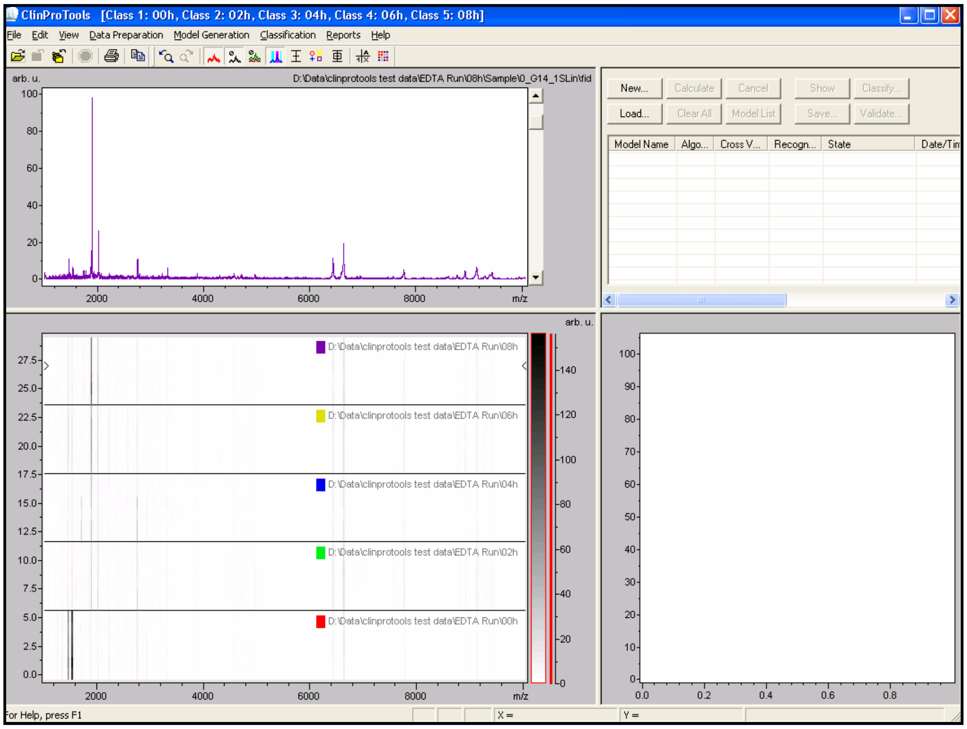
100 Peaks To Use

Sort Mode

○ Difference Average

● P-Value T-Test/ANOVA

○ P-Value Wilcoxon/Kruskal-Wallis


ClinProTools [Class 1: 00h, Class 2: 02h, Class 3: 04h, Class 4: 06h, Class 5: 08h]
File Edit View Data Preparation Model Generation Classification Reports Help
arb. u.
D:\Data\clinprotools test data\EDTA Run\08h\Sample\0_G14_1SLin\fid
100
80
60
40
20
0
2000
4000
6000
8000
m/z
New...
Calculate
Cancel
Show
Classify...
Load...
Clear All
Model List
Save...
Validate...
Model Name
Algo...
Cross V...
Recogn...
State
Date/Tim
arb. u.
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
D:\Data\clinprotools test data\EDTA Run\08h
D:\Data\clinprotools test data\EDTA Run\06h
D:\Data\clinprotools test data\EDTA Run\04h
D:\Data\clinprotools test data\EDTA Run\02h
D:\Data\clinprotools test data\EDTA Run\00h
140
120
100
80
60
40
20
0
2000
4000
6000
8000
m/z
100
90
80
70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
For Help, press F1
X =
Y =

Peak Statistic Calculation

- View Report:
  - –Peak Statistic
  - –Spectra View
  - –Gel View
  - –2D Peak Distribution View

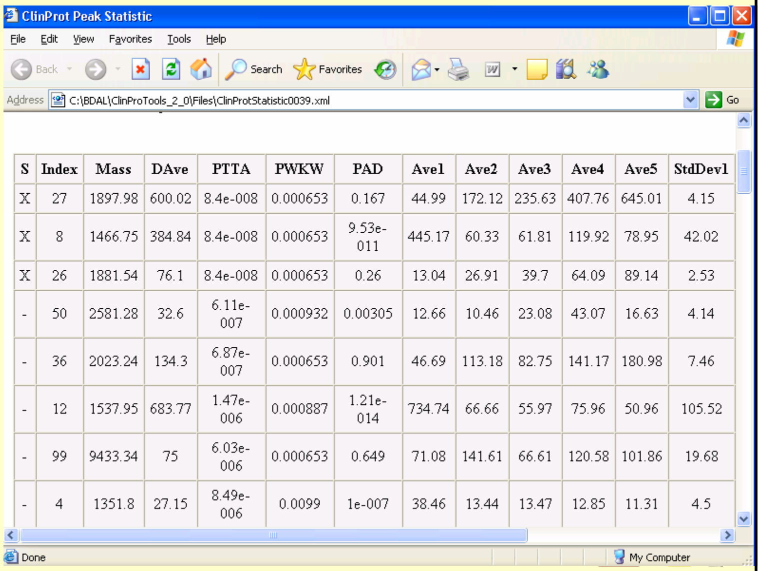
Peak Statistic

S: Used in model generation

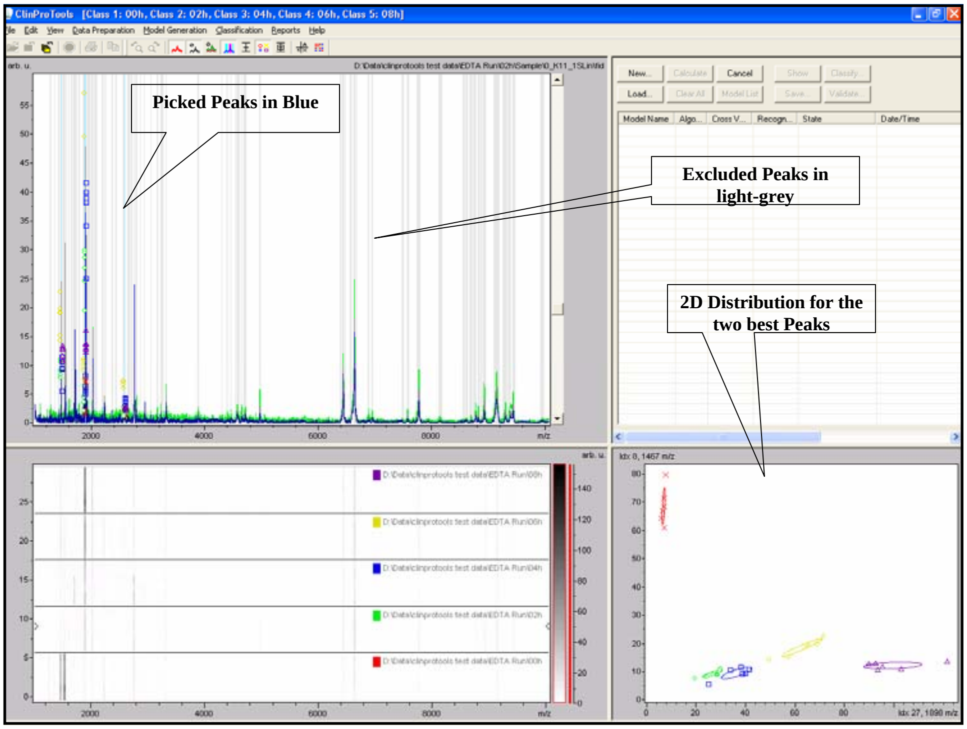
DAve: Difference between the Max and Min average peak area of all classes

PTTA: P-value of T test or ANOVA

PWKW: P-value of Wilcoxon or Kruskal-Wallis test


ClinProt Peak Statistic
File Edit View Favorites Tools Help
Back Search Favorites
Address C:\BDAL\ClinProTools_2_0\Files\ClinProtStatistic0039.xml
Go
Done
My Computer

S	Index	Mass	DAve	PTTA	PWKW	PAD	Ave1	Ave2	Ave3	Ave4	Ave5	StdDev1
X	27	1897.98	600.02	8.4e-008	0.000653	0.167	44.99	172.12	235.63	407.76	645.01	4.15
X	8	1466.75	384.84	8.4e-008	0.000653	9.53e-011	445.17	60.33	61.81	119.92	78.95	42.02
X	26	1881.54	76.1	8.4e-008	0.000653	0.26	13.04	26.91	39.7	64.09	89.14	2.53
-	50	2581.28	32.6	6.11e-007	0.000932	0.00305	12.66	10.46	23.08	43.07	16.63	4.14
-	36	2023.24	134.3	6.87e-007	0.000653	0.901	46.69	113.18	82.75	141.17	180.98	7.46
-	12	1537.95	683.77	1.47e-006	0.000887	1.21e-014	734.74	66.66	55.97	75.96	50.96	105.52
-	99	9433.34	75	6.03e-006	0.000653	0.649	71.08	141.61	66.61	120.58	101.86	19.68
-	4	1351.8	27.15	8.49e-006	0.0099	1e-007	38.46	13.44	13.47	12.85	11.31	4.5


ClinProTools [Class 1: 00h, Class 2: 02h, Class 3: 04h, Class 4: 06h, Class 5: 08h]
File Edit View Data Preparation Model Generation Classification Reports Help
arb. u.
New... Calculate Cancel Show Classify
Load... Clear All Model List Save... Validate
Model Name Algo... Cross V... Recogn... State Date/Time
2000 4000 6000 8000 m/z
Idx: 8, 1467 m/z
Idx: 27, 1898 m/z

Picked Peaks in Blue

Excluded Peaks in light-grey

2D Distribution for the two best Peaks

Model Generation workflow

- Includes:
  - Spectra recalibration
  - Average spectra calculation
  - Peak calculation
  - Model generation

Model Generation workflow

- Input files:

  File | Opene Model generation Class (normal vs spiked)

- Select models:

  Model Generation | New Model

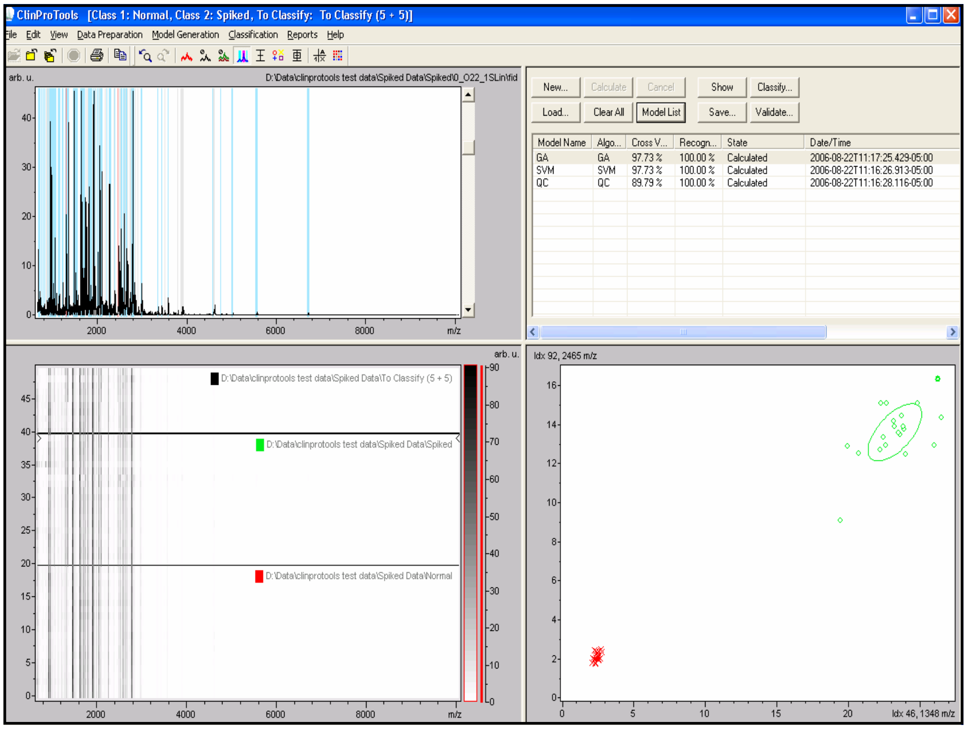
  Use default settings

- Generate models:

  Model Generation | Calculate

- Save models:

  Model Generation | Save Model As


ClinProTools [Class 1: Normal, Class 2: Spiked, To Classify: To Classify (5 + 5)]
File Edit View Data Preparation Model Generation Classification Reports Help
arb. u.
D:\Data\clinprotools test data\Spiked Data\Spiked\0_O22_1SLin\fid
40
30
20
10
0
2000
4000
6000
8000
m/z
New...
Calculate
Cancel
Show
Classify...
Load...
Clear All
Model List
Save...
Validate...
arb. u.
D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)
D:\Data\clinprotools test data\Spiked Data\Spiked
D:\Data\clinprotools test data\Spiked Data\Normal
45
40
35
30
25
20
15
10
5
0
2000
4000
6000
8000
m/z
90
80
70
60
50
40
30
20
10
0
Idx 92, 2465 m/z
16
14
12
10
8
6
4
2
0
0
5
10
15
20
Idx 46, 1348 m/z

Model Name	Algo...	Cross V...	Recogn...	State	Date/Time
GA	GA	97.73 %	100.00 %	Calculated	2006-08-22T11:17:25.429-05:00
SVM	SVM	97.73 %	100.00 %	Calculated	2006-08-22T11:16:26.913-05:00
QC	QC	89.79 %	100.00 %	Calculated	2006-08-22T11:16:28.116-05:00

View Models: Model list

ClinProt Model List

BRUKE

ClinProTools Version: 2.0 build 365

Name	Algo	Validation				GA Param							SVM Param		QC Param	KNN Param	X Val Param				Date/Time
		XVal	X1	X2	Rec Cap	Max NBP	Max Gen	Auto NPC	Num PCs	Mut Rate	Cross Rate	Var RS	Auto BPD	Num BstP	Sort Mode	Num OfNN	Mode	Part LOut	Num Iter	Num K	
GA	GA	97.7 %	100 %	95.5 %	100 %	5	50	true		0.2	0.5	false				3	random	20 %	10		2006-08-22T11:17:25
SVM	SVM	97.7 %	100 %	95.5 %	100 %								true			3	random	20 %	10		2006-08-22T11:16:26
QC	QC	89.8 %	90 %	89.6 %	100 %										p value wkw		random	20 %	10		2006-08-22T11:16:28

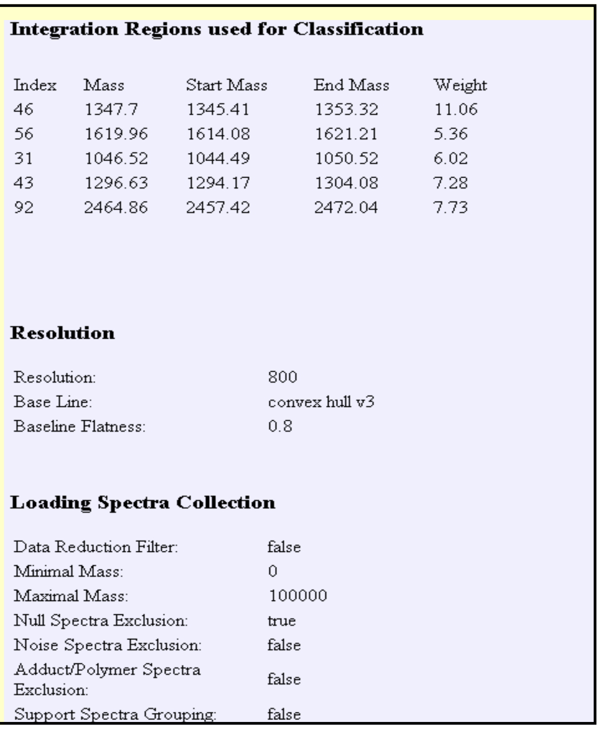
Validation: overall, each class, recognition capacity

GA: max nm peaks; max nm of generations;

SVM: auto detect peaks;

Cross valid param: percent of data points to leave out per iteration

View detailed info of a model



Integration Regions used for Classification

Index	Mass	Start Mass	End Mass	Weight
46	1347.7	1345.41	1353.32	11.06
56	1619.96	1614.08	1621.21	5.36
31	1046.52	1044.49	1050.52	6.02
43	1296.63	1294.17	1304.08	7.28
92	2464.86	2457.42	2472.04	7.73

Resolution

Resolution:	800
Base Line:	convex hull v3
Baseline Flatness:	0.8

Loading Spectra Collection

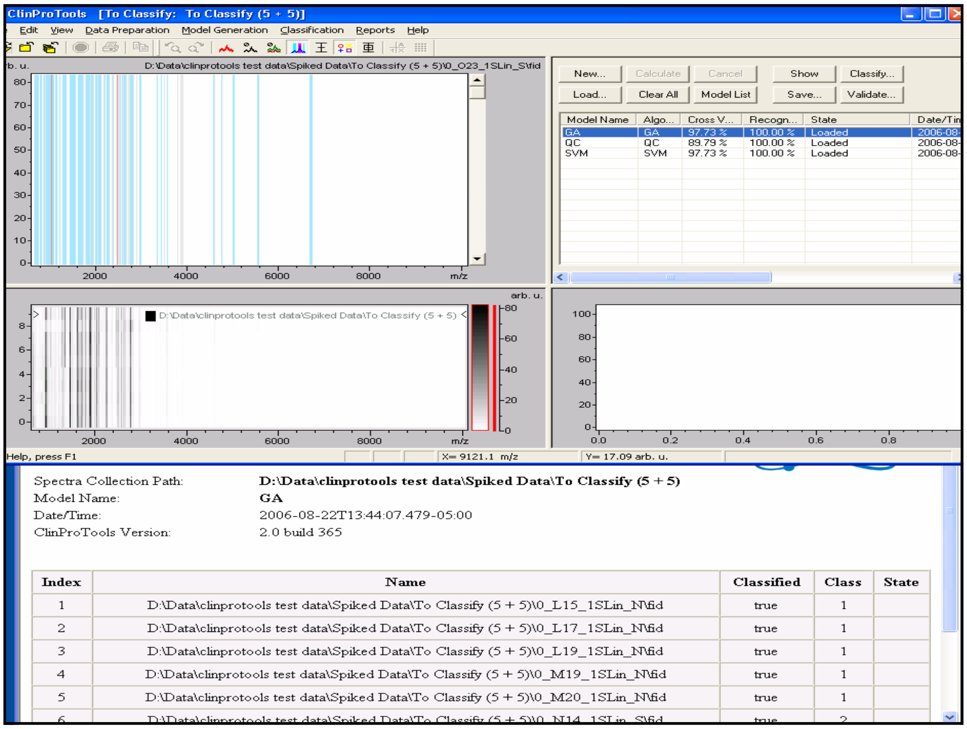
Data Reduction Filter:	false
Minimal Mass:	0
Maximal Mass:	100000
Null Spectra Exclusion:	true
Noise Spectra Exclusion:	false
Adduct/Polymer Spectra Exclusion:	false
Support Spectra Grouping:	false

Classify Samples workflow

- Includes:
  - –Selection of model
  - –Selection of spectra to be classified
  - –Data preparation of the new spectra
  - –Classification

Classify Samples workflow

- Load models:
- Select model:
- Classification | classify
- View:
  - –Classification report
  - –View gel
  - –View spectra



Model Name	Algo...	Cross V...	Recogn...	State	Date/Tim
GA	GA	97.73 %	100.00 %	Loaded	2006-08-
QC	QC	89.79 %	100.00 %	Loaded	2006-08-
SVM	SVM	97.73 %	100.00 %	Loaded	2006-08-

Spectra Collection Path: **D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)**
Model Name: **GA**
Date/Time: 2006-08-22T13:44:07.479-05:00
ClinProTools Version: 2.0 build 365

Index	Name	Classified	Class	State
1	D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)\0_L15_1SLin_N\fid	true	1	
2	D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)\0_L17_1SLin_N\fid	true	1	
3	D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)\0_L19_1SLin_N\fid	true	1	
4	D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)\0_M19_1SLin_N\fid	true	1	
5	D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)\0_M20_1SLin_N\fid	true	1	
6	D:\Data\clinprotools test data\Spiked Data\To Classify (5 + 5)\0_N14_1SLin_S\fid	true	2	