Processing and Classification of Proteomics Mass Spectra (MS) data in R with caMassClass package

By Jarek Tuszynski

jaroslaw.w.tuszynski@saic.com

(703) 676-4192

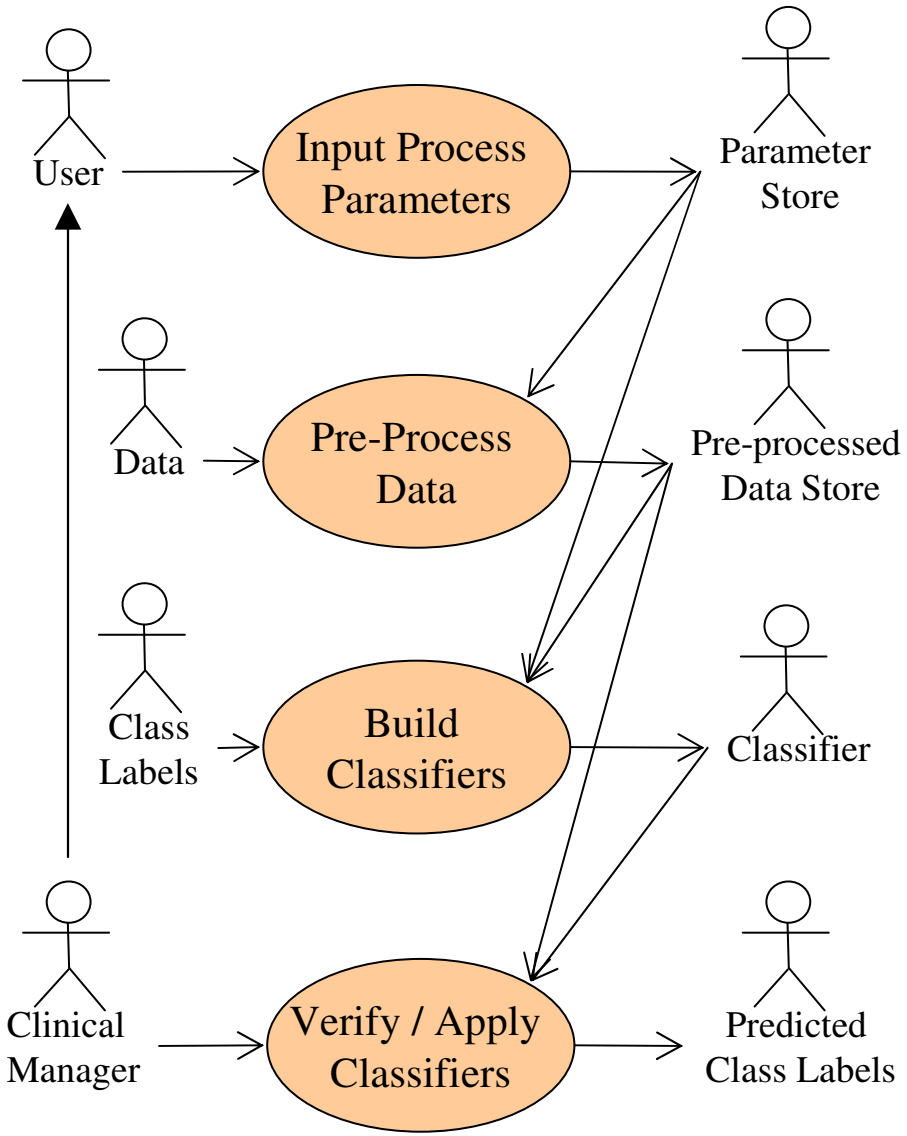
caMassClass Package

- Package of functions for processing and classification of protein mass spectra data.
- Released as "open source" through CRAN website, together with its companion package "caTools"
- Functions range:
  - from generic (moved to caTools) to specific
  - from low level (easily used in other codes; IO using R structures) to high level (one-function pipelines with file IO)

caTools Package

- This presentation will focus on functions specialized to narrow task of analyzing MS data
- However, specialized functions required development of various generic tools which were placed in a separate package “caTools”:
  - fast moving window statistic functions (mean, minimum, maximum, MAD, quantile) needed for peak finding.
  - fast calculation of Area Under ROC Curve (AUC) , aka. Wilcoxon test needed for feature selection
  - base64 encoder/decoder needed for mzXML support
  - round-off error free sum and ‘cumsum’

Use Cases

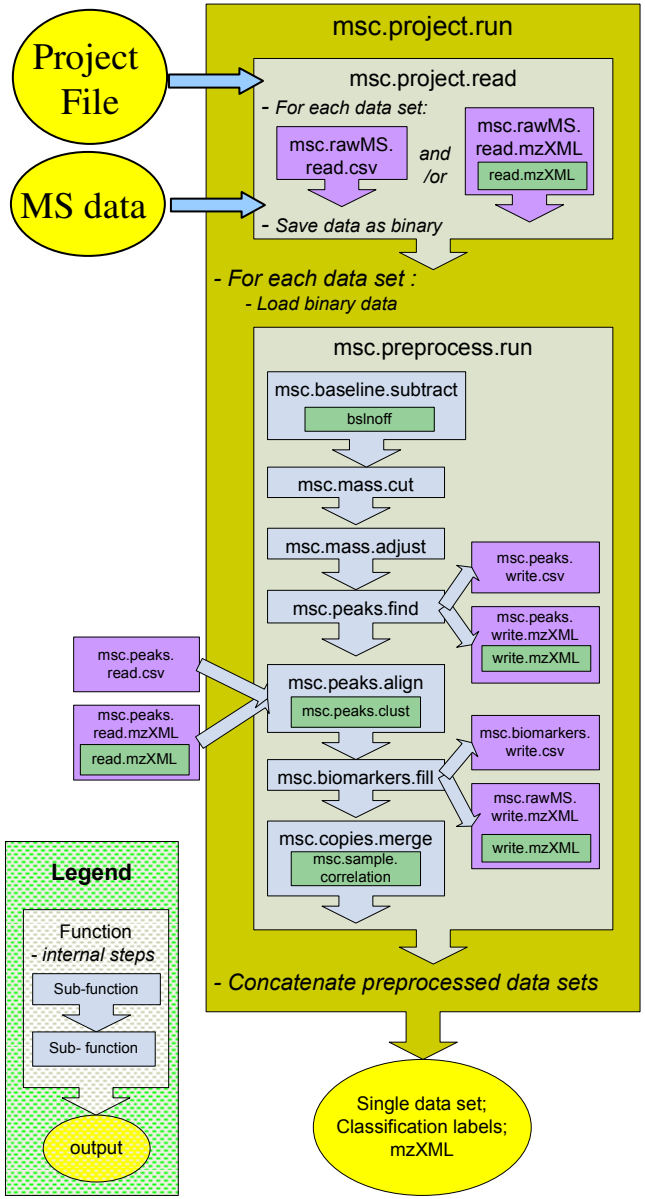


- User inputs Process Parameters, which will uniquely describe the rest of the flow. The parameters are saved into *Parameter Store*, which will be retrieved by remaining processes.
- *Data* is pre-processed according to user specifications retrieved from *Parameter Store,* and then stored in *Pre-processed Data Store.*
- Classifiers are built using *pre-processed data* and *class labels*. The algorithms used and steps of the process are specified by *Parameter Store.*
- *Classifier* is verified by a *User* or applied by a *Clinical Manager.* That is done by running the *classifier* on unlabeled *pre-processed data* in order to predict the class labels.

Terms Used

- **data set** – *features by samples data where each sample has one or more MS spectra (***copies***). All MS spectra were taken under the same conditions.*
- **data sets** – *data sets taken under different conditions for the same samples (example SELDI data using* IMAC3-Cu & WCX2 chips)
- **class labels** – *describe samples (for example "cancer", "normal", "benign")*
- **preprocessing** – *steps used to improve and lower dimensionality of the data, performed without use of class labels*
- **biomarkers** – *aligned peaks. We might or might not know what they are.*

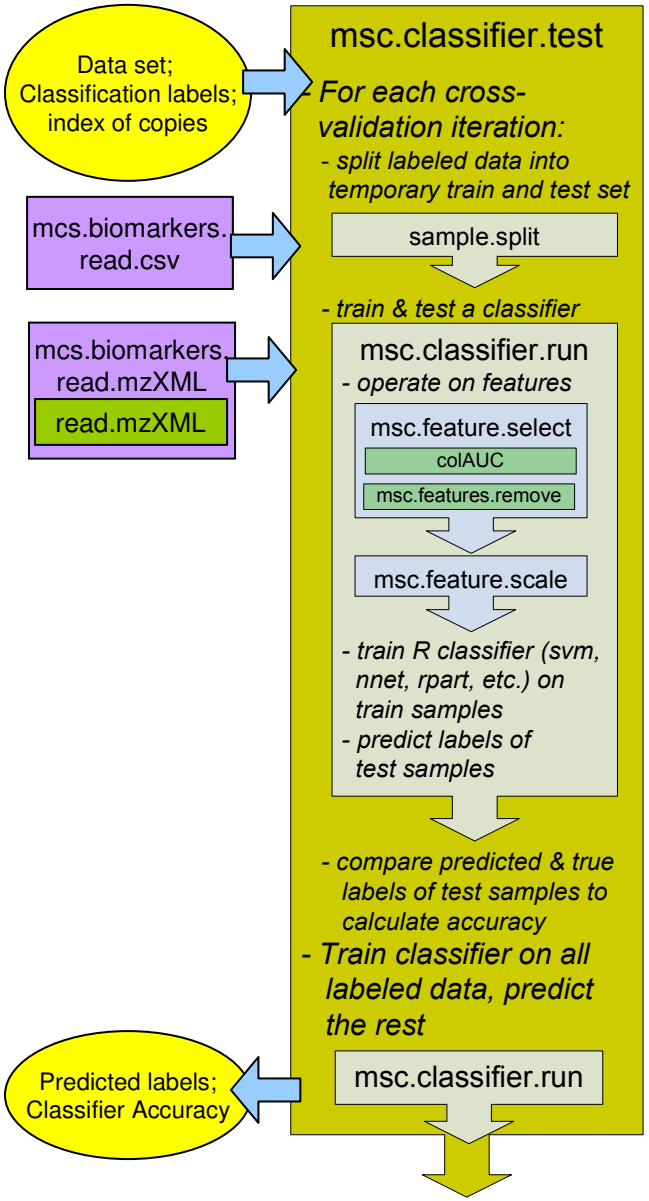
Data Input and Pre-Processing



Project Run:

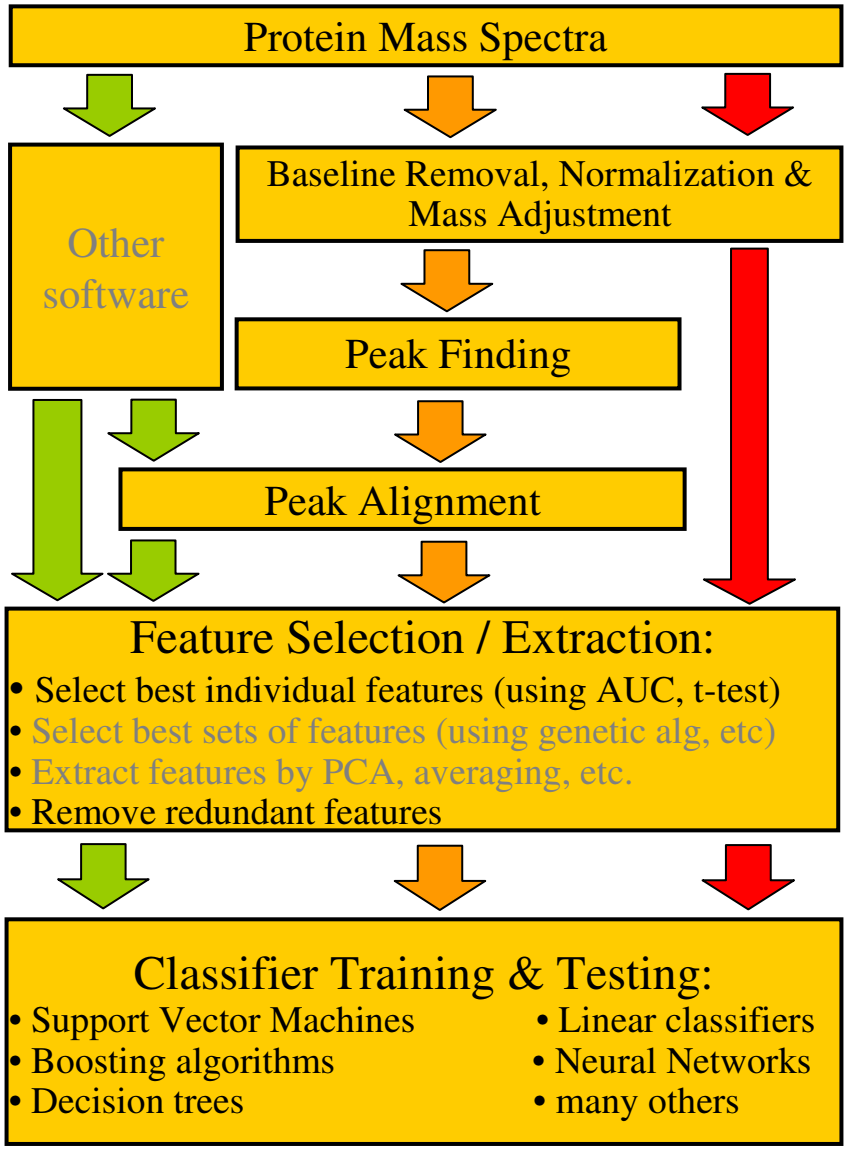
- **Read input files** and save than in R binary format
- **Preprocess Pipeline**:
  - **Base-line subtraction** – optional step since it is usually performed as part of data collection.
  - **Trimming low & high m/z values**
  - **Normalization** – match means and/or mediums of all samples. (performed by mcs.mass.adjust)
  - **Mass Drift Adjustment** – shift each row to the right or the left if it improves its correlation with the rest of the samples.
  - **Peak Finding and Alignment** – steps designed to reduce dimensionality of the data by extracting common peaks (aka biomarkers) from the data.
  - **"Filling"** of biomarker matrix fills gaps caused by lack of a peak in given sample in given range.
  - **Merging** of copies of each sample:
    - Average copies in order to reduce noise
    - Keep all copies
    - Throw out the outliers
- **Concatenate data sets** increasing number of features

Classification



- For each step of **cross-validation** :
  - **Split samples** of *Pre-processed Data* into temporary test and train sets.
  - Perform **feature selection** on train set:
    - Individual feature selection using: AUC, T-test, etc.
    - Individual feature removal: for highly correlated features remove sub-optimal features.
  - Perform **classification** on train set using:
    - Support Vector Machine (svm)
    - Neural Networks (nnet)
    - CART - Classification And Regression Trees (rpart)
    - Boosting algorithms (LogitBoost)
  - **Test the classifier** on test data set, and keep track of its performance
- **Build final classifier** using all *Pre-processed Data with* labels, by following feature selection and classification steps above.
- **Predict labels** of all un-labeled samples

Algorithm families supported by caMassClass



Different approaches for classification of Protein MS :

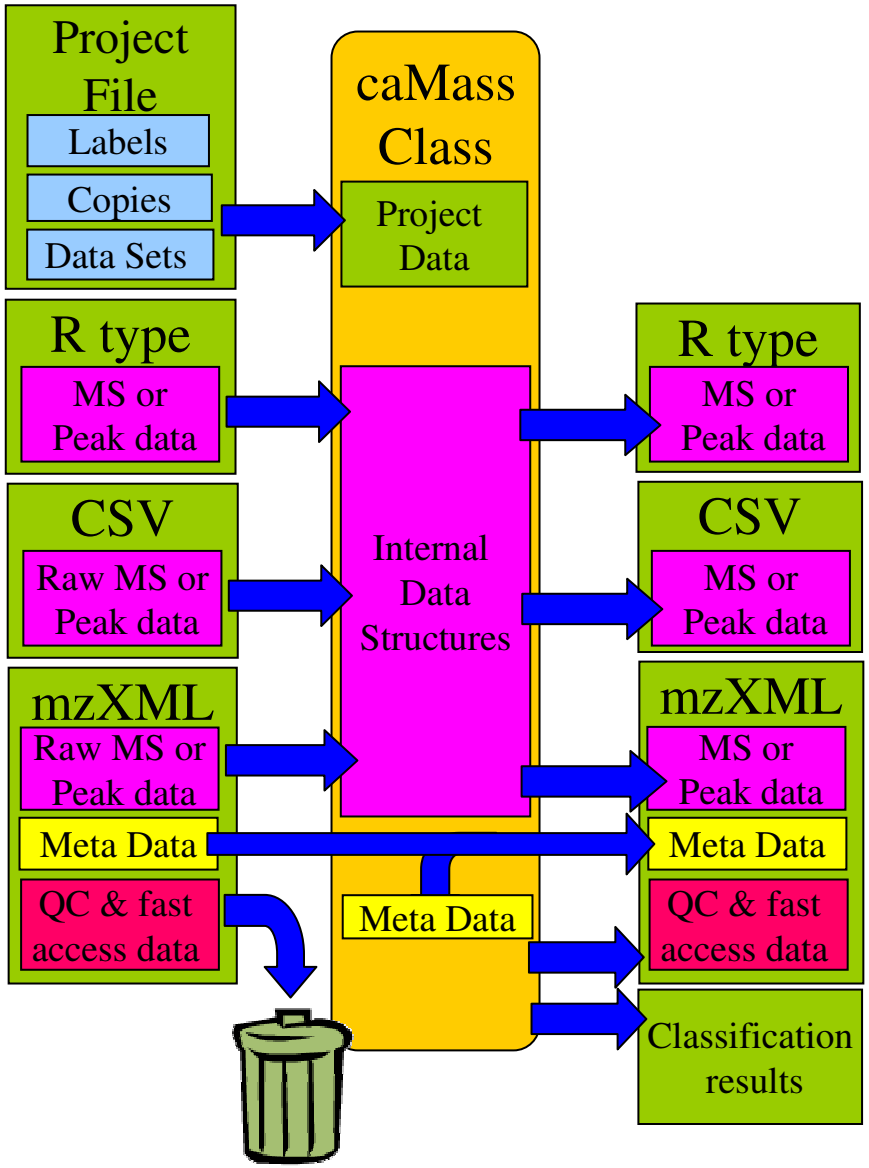
- Green : method used in analysis of EVMS data as described by Bao-Ling Adam
- Orange : same as green but without use of proprietary software. Similar to method described by K. Baggerly
- Red : method used in Petricoin/Liotta study where feature selection was done by genetic algorithm and Kohonnen SOM's were used for classification

Input Data Types

- Input data can be in form of:
  - Raw MS spectra (all have to have the same length and m/z values)
  - Baseline subtracted MS spectra
  - Uneven list of peaks for each spectrum
  - Biomarker matrix (sample by biomarker table) with or without missing values.
- Input data can have:
  - Multiple copies of each sample
  - Multiple data sets
  - Two or more class labels
- Input/Output files can be in form of:
  - CSV files (multiple directories, compressed & uncompressed)
  - mzXML files
  - “Project File” is in the form of CSV file

IO Data Formats

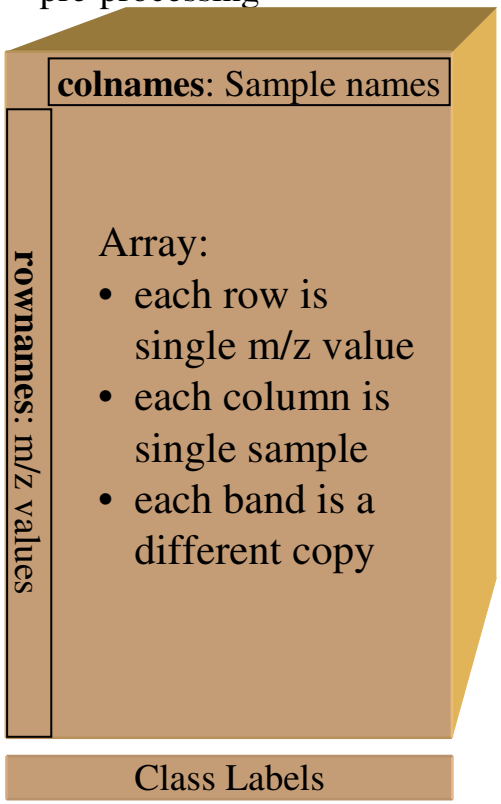
- Raw MS or Peak data:
  - scan (meta-data) - copied
  - peaks - replaced
- Meta Data:
  - parentFile - appended
  - msInstrument - copied
  - dataProcessing - appended
  - separation - copied
  - spotting - copied
- Quality Control (QC) & fast access data:
  - offset – replaced
  - indexOffset - replaced
  - sha1 – replaced
- Project File
  - Sample Class Labels (i.e. “cancer”, “normal”)
  - Sample copies (multiple copies of scans of the same sample)
  - Data sets (multiple experiments performed on the same samples)
  - Sample names (CSV file names)



Internal Data Structures

- Simple types designed to be fast and extensible
- Three main data structures are:

3D format used during pre-processing



Uneven Peak List used in peak finding section

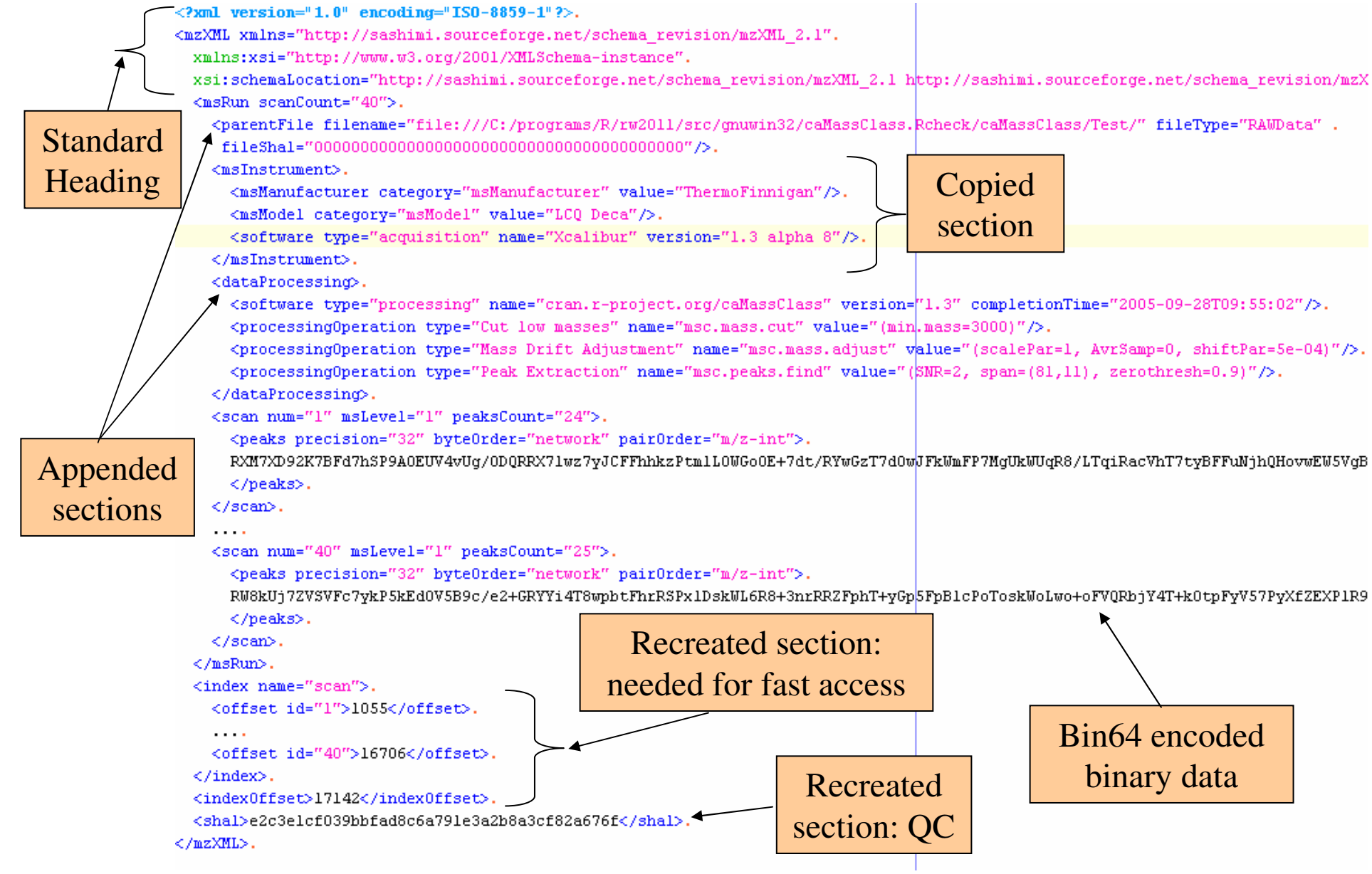
Spectrum.Tag	Spectrum.	Intensity	Substance.Mass
cancer_01(1)	1	0.517369	2960.36
cancer_01(1)	1	0.98591	3894.02
cancer_01(1)	1	1.667703	3965.85
cancer_01(1)	1	1.667703	3982.16
cancer_01(1)	1	0.435958	4293.57
cancer_01(1)	1	0.476308	4310.54
cancer_01(1)	1	0.444201	4483.32
cancer_01(1)	1	1.434796	4655.72
cancer_01(1)	1	0.69378	4759.69
cancer_01(1)	1	0.476156	5349.39
cancer_01(1)	1	4.007973	5917.95
cancer_01(1)	1	4.318063	5933.6
cancer_01(1)	1	1.193908	6124.45
cancer_01(1)	1	0.534523	6955.01
cancer_01(1)	1	4.064739	7779.58
cancer_01(1)	1	0.798553	8155.69
cancer_01(1)	1	0.312816	8615.99
cancer_01(1)	1	1.135725	8946.77
cancer_01(1)	1	5.005366	9301.59
cancer_01(1)	1	2.001326	9509.51
cancer_01(1)	1	0.276836	10277.8
cancer_01(1)	1	0.255963	11745.1
cancer_01(1)	1	0.784887	13894.4
cancer_02(1)	2	0.500555	2959.36
cancer_02(1)	2	0.941613	3892.87
cancer_02(1)	2	1.287152	3965.85
cancer_02(1)	2	0.208383	4292.36
cancer_02(1)	2	1.117763	4654.45
cancer_02(1)	2	0.665819	4759.69
cancer_02(1)	2	0.48962	5348.04

Biomarkers matrix used during classification

Same Sample Index / Class Labels

	M3894.6	M3965.85	M3982.16	M4079.54	M4284.5
cancer_01(1)	0.9616	1.6266	1.6266	0.4729	0.4252
cancer_02(1)	0.9451	1.2919	0.0000	0.3405	0.2092
cancer_03(1)	0.8889	1.1636	0.0000	0.3666	0.2097
cancer_04(1)	1.2880	1.5457	0.0000	0.3153	0.7957
cancer_05(1)	0.9964	1.5826	0.0000	0.4046	0.3315
cancer_06(1)	0.9052	0.0000	0.0000	0.2531	0.1969
normal_01(1)	1.2410	0.0000	2.0169	0.0000	0.4520
normal_02(1)	1.1391	0.0000	0.0000	0.0000	0.0000
normal_03(1)	1.0525	1.3626	0.0000	0.0000	0.2630
normal_04(1)	1.1320	0.0000	0.0000	0.0000	0.0000
normal_05(1)	1.4636	1.6314	1.1889	0.0000	0.4756
normal_06(1)	0.7593	0.0000	0.0000	0.0000	0.0000
cancer_01(2)	0.9185	1.3135	1.3135	0.0000	0.3508
cancer_02(2)	1.0493	1.2660	0.0000	0.5149	0.3132
cancer_03(2)	0.9893	1.1365	0.0000	0.0000	0.0000
cancer_04(2)	1.3401	1.4539	0.0000	0.0000	0.5812
cancer_05(2)	1.5399	2.3200	0.0000	0.0000	0.5812
cancer_06(2)	1.1330	0.0000	0.0000	0.3799	0.3172
normal_01(2)	1.4561	0.0000	2.1135	0.4678	0.6964

mzXML File Example




```
<?xml version="1.0" encoding="ISO-8859-1"?>.
<mzXML xmlns="http://sashimi.sourceforge.net/schema_revision/mzXML_2.1".
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance".
  xsi:schemaLocation="http://sashimi.sourceforge.net/schema_revision/mzXML_2.1 http://sashimi.sourceforge.net/schema_revision/mzX
  <msRun scanCount="40">.
    <parentFile filename="file:///C:/programs/R/rw2011/src/gnuwin32/caMassClass.Rcheck/caMassClass/Test/" fileType="RAWData" .
     fileSha1="0000000000000000000000000000000000000000"/>.
    <msInstrument>.
      <msManufacturer category="msManufacturer" value="ThermoFinnigan"/>.
      <msModel category="msModel" value="LCQ Deca"/>.
      <software type="acquisition" name="Xcalibur" version="1.3 alpha 8"/>.
    </msInstrument>.
    <dataProcessing>.
      <software type="processing" name="cran.r-project.org/caMassClass" version="1.3" completionTime="2005-09-28T09:55:02"/>.
      <processingOperation type="Cut low masses" name="msc.mass.cut" value="(min.mass=3000)"/>.
      <processingOperation type="Mass Drift Adjustment" name="msc.mass.adjust" value="(scalePar=1, AvrSamp=0, shiftPar=5e-04)"/>.
      <processingOperation type="Peak Extraction" name="msc.peaks.find" value="(SNR=2, span=(81,11), zerothresh=0.9)"/>.
    </dataProcessing>.
    <scan num="1" msLevel="1" peaksCount="24">.
      <peaks precision="32" byteOrder="network" pairOrder="m/z-int">.
      RXM7XD92K7BFd7hSP9A0EUV4vUg/0DQRRX71wz7yJCFFhhkzPtm1LOWGo0E+7dt/RYwGzT7d0wJFkWmFP7MgUkWUqR8/LTqiRacVhT7tyBFFuNjhQHovwEW5VgB
      </peaks>.
    </scan>.
    ....
    <scan num="40" msLevel="1" peaksCount="25">.
      <peaks precision="32" byteOrder="network" pairOrder="m/z-int">.
      RW8kUj7ZV5VFc7ykP5kEd0V5B9c/e2+GRYYi4T8wpbtFhrRSPx1DskWL6R8+3nrRRZFphT+yGp5FpB1cPoToskWoLwo+oFVQRbjY4T+k0tpFyV57PyXfZEXP1R9
      </peaks>.
    </scan>.
  </msRun>.
  <index name="scan">.
    <offset id="1">1055</offset>.
    ....
    <offset id="40">16706</offset>.
  </index>.
  <indexOffset>17142</indexOffset>.
  <sha1>e2c3e1cf039bbfad8c6a791e3a2b8a3cf82a676f</sha1>.
</mzXML>.
```

Example Run (1)

- A small data set was provided by Center for Prostate Disease Research containing SELDI Data in form of CSV files:
  - train set contained 41 cancerous and 40 normal samples
  - blinded test set contained 79 samples
- Project file was created:

Two copies

name	label	IMAC1	IMAC2
p0003	1	cpdr_data/p0003.csv	cpdr_data/p0003(2).csv
p0004	1	cpdr_data/p0004.csv	cpdr_data/p0004(2).csv
p0009	1	cpdr_data/p0009.csv	cpdr_data/p0009(2).csv
pb001	0	cpdr_data/pb001.csv	cpdr_data/pb001(2).csv
pb002	0	cpdr_data/pb002.csv	cpdr_data/pb002(2).csv
pb003	0	cpdr_data/pb003.csv	cpdr_data/pb003(2).csv
pn0002	2	cpdr_data/pn0002.csv	cpdr_data/pn0002(2).csv
pn0003	2	cpdr_data/pn0003.csv	cpdr_data/pn0003(2).csv
pn0061	2	cpdr_data/pn0061.csv	cpdr_data/pn0061(2).csv
pn0064	2	cpdr_data/pn0064.csv	cpdr_data/pn0064(2).csv

Name to be used in classification output

Files in csv format. Other formats allowed:
- individually compressed csv
- csv extracted from zip'ed file
- sample extracted from mzXML file

Example Run (2)

- Data Input and Pre-Processing was done by:

```
fname = "F:/projects/NCI/plasma-1/InputFiles.csv";.
ddump = "F:/projects/NCI/plasma-1/data.Rdata";.
.
msc.project.run(fname,.
  baseline.removal = 0,.
  min.mass = 3000,                                                    # msc.mass.cut.
  mass.drift.adjustment = 1, shiftPar=0.0005,                         # msc.mass.adjust.
  peak.extraction = 1, .
   PeakFile="F:/projects/NCI/plasma-1/PeakFile.csv", SNR=2, span=c(81,11), zerothresh=0.9,  # msc.peaks.find.
   BmrkFile="F:/projects/NCI/plasma-1/BmrkFile.csv", BinSize=c(0.002, 0.008), tol=0.97,     # msc.peaks.align.
   FlBmFile="F:/projects/NCI/plasma-1/FlBmFile.csv", FillType=0.9                           # msc.biomarkers.fill.
).
X=msc.project.run(fname,.
    baseline.removal = 0,.
    min.mass = 3000,                                  # msc.mass.cut.
    mass.drift.adjustment = 1, shiftPar=0.0005,       # msc.mass.adjust.
    peak.extraction = 0,                              # no peak extraction.
    merge.copies = 1+4)                               # msc.copies.merge.
save(X, file=ddump).
```

Project File

Preprocessing with peak extraction

Output files

Preprocessing without peak extraction

- The code above created three output files that will be used during classification:
  - BmrkFile.csv – Biomarker Matrix (Aligned peaks) with NA's when there were no peaks
  - FlBmFile.csv – "Filled" Biomarker Matrix without NA's
  - Data.rdata - MS spectra

Example Run (3)

- In case of 'BmrkFile.csv' file 'R's function 'tune.svm' was used to find optimal values for SVM parameters "cost", and "gamma".
- Training and running a classifier was done by :

No feature selection

```
out = msc.classifier.test ( X, Y, iters=100, SplitRatio=3/4, .
       RemCorrCol=0, KeepCol=0, prior=1, same.sample=SameSamples, ScaleType="none", .
       method="svm", cost = 32, gamma = 0.062) .
```

- Cross validation gave following results for train set:

(other data sets, usually larger, gave results up to 94% correct)

	True	
Predicted	1	2
1	0.791	0.267
2	0.209	0.733

- Predicted labels for the whole blinded test set were also calculated.

Example Run (4)

- In case of raw data file ‘data.Rdata’ file ‘tune.svm’ was used again to find optimal parameters
- Training and running a classifier was done by :

```
out = msc.classifier.test ( X, Y, iters=100, SplitRatio=3/4, prior=1,
        RemCorrCol=0.95, KeepCol=200, ScaleType="none",.
        same.sample=SameSamples, method=method, cost=2, gamma=2^-10)
```

Heave feature selection

- In this approach reduction of number of featured was mostly accomplished by feature selection performed during cross-validation.
- The results of this approach were worse than in case of algorithm with peak-finding.

Planned Extensions

- Implement or translate other established algorithms for different pre-processing steps to R
- Add other standard R classification algorithms to "mcs.classifier.run" function
- Improve mzXML reader to be faster and use less memory
- Add Quality Control functions, actively testing for specific problems with data

Other Related Codes in R

Name	Author / Group	Affiliation	Package released on	Description
PROcess	Xiaochun Li	Harvard	BioConductor	“A package for processing protein mass spectrometry data”
ppc	R. Tibshirani, T. Hastie & B. Narasimhan	Stanford	CRAN	"Sample classification of protein mass spectra by peak probability contrasts"
msBase & msCalib	Witold Wolski	Max Planck Institute (Germany)	BioConductor	“visualization & storage of mass spectrometric mass lists ”
RProtiomics		Duke	Not in form of a package	
msInspect	Computational Proteomics Analysis System	Fred Hutchinson Cancer Research Center	Uses some R functions. Not in form of a package	R used for “alignment and registration steps ”
Q5	R. Lilien, H. Farid, & B. Donald	Dartmouth	Matlab code was released. Any R code?	“Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum. ”