Supporting material

1. Theory

1.1 Fuzzy Rule-Building Expert Systems (FuRES)

FuRES is a pattern recognition method for complex datasets based on multivariate rule-building expert systems.[1] Different from other previous rule-building expert systems, FuRES uses fuzzy logic instead of crisp logic to classify different classes. FuRES does not have any adjustable parameter, such as the number of latent variables in (partial least squares (PLS) regression, so it does not require a separate set of data to optimize the model. As a robust classification method, FuRES has been used in many fields such as proteomics,[2] jet fuel classification,[3] bacterium classification,[4] etc. FuRES classification produces a classification tree that is easy to interpret and understand. The process of training is a process to minimize classification entropies by applying multivariate rules from the root of the tree to the branches. The classification of logic is manifested by a tree structure. At the root of the tree, the classes have the greatest entropy for which a general rule is used to first split the data into two subsets that exhibit smaller entropies. As one proceeds from the root of tree, the rules become more precise until all the data are grouped into classes at the leaves of the tree.

1.2 Projected Difference Resolution (PDR)

For complex datasets, the distribution of objects and classes cannot be accurately assessed by looking at the principal component scores, especially when the variance spanned by the first two components is less than 90%. Visually assessing plots of principal component scores is usually not ideal because it is not quantitative. As a powerful tool, the PDR quantitative metric was devised that measures the separations of data clusters in a multi-dimensional space in the context of

chromatographic resolution.[5] The stepwise calculations are given as follows. First, the difference vector between two class means is calculated.

$$d_{a,b} = \bar{x}_a - \bar{x}_b \quad (1)$$

for which $\bar{x}_a$ and $\bar{x}_b$ are the class means and $d_{a,b}$ is the difference vector between $\bar{x}_a$ and $\bar{x}_b$. Objects are row vectors. Each data object is projected to the difference vector $d_{a,b}$ as described in the following equation:

$$p_i = x_i d^T \quad (2)$$

for which p_i is the inner product of each data object x_i and the class average difference vector $d_{a,b}$. The resolution of two classes then can be calculated according to the equation below:

$$Rs_{a,b} - \frac{|\bar{p}_a - \bar{p}_b|}{2 \times (s_a + s_b)} \quad (3)$$

for which $\bar{p}_a$ and $\bar{p}_b$ are the averages of the inner products; s_a and s_b are the standard deviations of the two classes. As with chromatographic resolution, when the *Rs* value is larger than 1.5, the objects of the two classes are considered baseline resolved in the multivariate data space. Typically all possible pairs of classes are measured and either the minimum or geometric average PDR is reported.

1.3 Baseline correction

When a mass spectrometer is coupled with an HPLC as a detector, mobile phase programs in liquid chromatography may cause drift in the baselines of the total ion current (TIC) chromatograms. Common background components in TIC chromatograms of different samples may have an adverse effect on pattern recognition because TIC chromatograms will appear similar to one another because of the common

baseline components. As a result, classification or pattern recognition may result in error. Baseline correction therefore is important for pattern recognition.

In this work, an in-house baseline correction algorithm that prevents negative peaks in TIC caused by overfitting was used.[6] This algorithm is described by equations (4-6). A basis (V) is constructed from mass spectra collected from chromatographic regions that have no analytical peaks. The background spectrum is estimated by projecting a spectrum with analytical signal onto this basis and subtracted from the spectrum to accomplish the correction. Given as

$$x_c = x - (xV)V^T \tag{4}$$

for which an uncorrected mass spectrum *x* is a row vector*;* V is the orthogonal basis obtained by singular value decomposition; and x_c is the unconstrained corrected mass spectrum that may have negative peaks.

Negative peaks in the total ion current from chance correlations are avoided by introducing a regularization parameter.

$$\lambda = \frac{\bar{x} - e}{\bar{x} - \bar{x}_c} \tag{5}$$

for which $\bar{x}_c$ is the average peak intensity of the corrected mass spectrum; $\bar{x}$ is the mean peak intensity of the uncorrected mass spectrum *x*; *e* denotes for an error threshold that defines the smallest negative intensity peak allowed in the corrected ion chromatogram; λ is the regularization parameter that prevents overfitting of the background components.

Lastly, a constrained background correction is accomplished with the addition of the regularization parameter which is

$$x_{c_\lambda} = x - \lambda(xV)V^T \tag{6}$$

to yield the corrected mass spectrum x_{c*} whose sum of mass peaks will not be less than zero.

2. Experimental details

2.1 Samples, Reagents and Sample Pretreatment

P. quinquefolius L samples from China and the United States used in this study are listed in Table 1. The information of age, size, and specific locations in each country was not completely available, and was believed to vary. For the samples from the United States, information about suppliers was the only additional information available while for the samples from China, some information about suppliers, sub-locations, shape and size was also available.

Table 1. Ginseng samples grown in two countries used in this work. AM and CH denote the samples grown in the United States and China, respectively

Class	Sample ID	Source	Additional Information
AM	WBQ31	US	Heil[a]
AM	WBQ32	US	Drath[a]
AM	WBQ33	US	Weege[a]
AM	WBQ34	US	Bauman[a]
AM	WBQ35	US	Heier[a]
AM	WBQ36	US	Untiedt[a]
AM	WSQ1	US	Schumacher[a] Pure Wisconsin[b]
AM	WSQ2	US	Schumacher[a] Pure Wisconsin[b]
AM	WSQ3	US	Schumacher[a] Wisconsin[b] Ginseng
AM	WSQ4	US	Schumacher[a] Wisconsin[b] Ginseng
AM	WSQ5	US	Schumacher[a] Wisconsin[b] Ginseng
AM	WSQ6	US	Schumacher[a] Pure Wisconsin[b]
CH	ZQ01	China	Tong RenTang[a], Huairong, Beijing[b]/Slices
CH	ZQ03	China	Huairong, Beijing[b]/ Short Head
CH	ZQ04	China	Huairong, Beijing[b]/Round Head
CH	ZQ07	China	Huairong, Beijing[b]/Short Head
CH	ZQ08	China	Huairong, Beijing[b]/Medium Slices
CH	ZQ16	China	Wendeng, Shandong[b]/Large Slices
CH	ZQ17	China	Wendeng, Shandong[b]/Medium Slices
CH	ZQ18	China	Qingdao, Shandong[b]/Short Head
CH	ZQ19	China	Xinbin, Liaoning[b]/Long Head
CH	ZQ20	China	Tonghua, Jilin[b]/ Long Head
CH	ZQ22	China	Fusong, Baishan, Jilin[b]/ Small Slices
CH	ZQ23	China	Fusong, Baishan, Jilin[b]/Long Head

a. Suppliers of Ginseng.

b. Sub-locations.

HPLC-grade acetonitrile (EMD Chemicals Inc, Gibbstown, NJ) and HPLC-grade methanol (PHARMCO-AAPER, Brookfield, CT, US) were used for the mobile phase components and the extracting solvent, respectively. De-ionized water (18 MΩ) for sample preparation and as one component of the mobile phase was obtained using a water purification system (Nanopure Diamond Barnstead, Thermo Scientific).

Ginseng root samples were ground into fine powder and stored at room temperature. For each ground sample, about 30 mg was weighed and mixed with 1 mL of methanol-water (60:40, volume ratio) in a 2-mL plastic vial and sonicated for 60 min at room temperature. The extracted samples were centrifuged at 8000 rpm for 15 min. Then the supernatant was filtered by using 13-mm (pore size of 0.45 μm) polyvinylidene fluoride (PVDF) syringe filters (General Separation Technologies, Inc., US). The extracts were then stored in a refrigerator and brought to room temperature prior to analysis.

2.2 Instrumentation

An Agilent 1100 HPLC equipped with a G1322A online degasser, a G1312A binary pump, a G1313A autosampler, and a G1316A temperature controlled column compartment was used for separation. A Thermo Finnigan PolarisQ mass spectrometer modified with a Thermo Finnigan Deca XP electrospray ionization (ESI) source and ion optics was used to couple with the HPLC for the online mass spectra collection, as described previously.[7] Data was collected using the XCalibur development kit (XDK) provided by Thermo and was custom-modified in Visual Basic 6.0 (Redmond, WA, USA). A high voltage power supply (Stanford Research System, Inc., Sunnyvale, CA) was used for the ionization of samples in the Deca XP continuous ESI source.

2.3 HPLC-MS Data Collection Conditions

The mobile phase was composed of A: water and B: acetonitrile. Gradient elution was used as follows: from 0 to 3 min, 80% A and 20% B; mobile phase B was increased from 20% to 25% between 3 and 7 minutes, from 25% to 45% between 7 and 20 minutes, from 45% to 75% between 20 and 25 minutes, from 75% to 80% between 25 and 28 minutes, and 80% to 95% between 28 and 40 minutes. The column temperature was controlled at 35 °C.

For the mass spectra collection, a mass range extension (MRE) program was written to eject ions at a q_z value of 0.45 to double the mass scan range (the highest mass was extended from 1000 Th to 2000 Th) to cover the possible mass range of ginsenosides. The mass spectrometer was optimized and calibrated with an ESI tune mix (Agilent Technology, US). The mass spectra were collected in a negative ion mode with voltage of -4.5 kV. The capillary temperature was set at about 350 °C and the electron-multiplier was set at 1.375 kV.

3. Results

3.1 Sample two-way data image

The separation of components can be seen in a two-way data image (the TIC chromatogram and mass spectra) in Figure 1. In this figure, a complex and rich pattern of peaks can be seen.

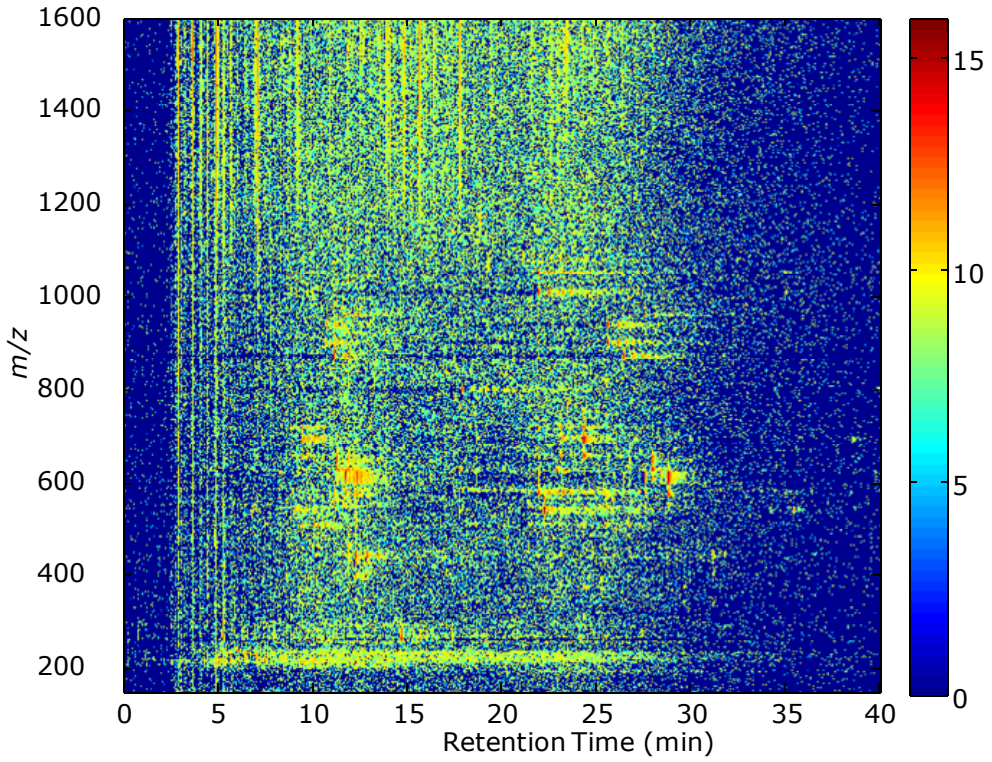

Figure 1. Two-way image (log values) of sample WSQ2 from one HPLC—MS run.

3.2 Alignment Effect

The enlarged total ion chromatograms before and after retention time alignment with an in-house alignment program are given in Figure 2 to demonstrate the effect of the retention time alignment. It can be seen obviously that most peaks in the chromatogram are more aligned compared with those before the treatment.

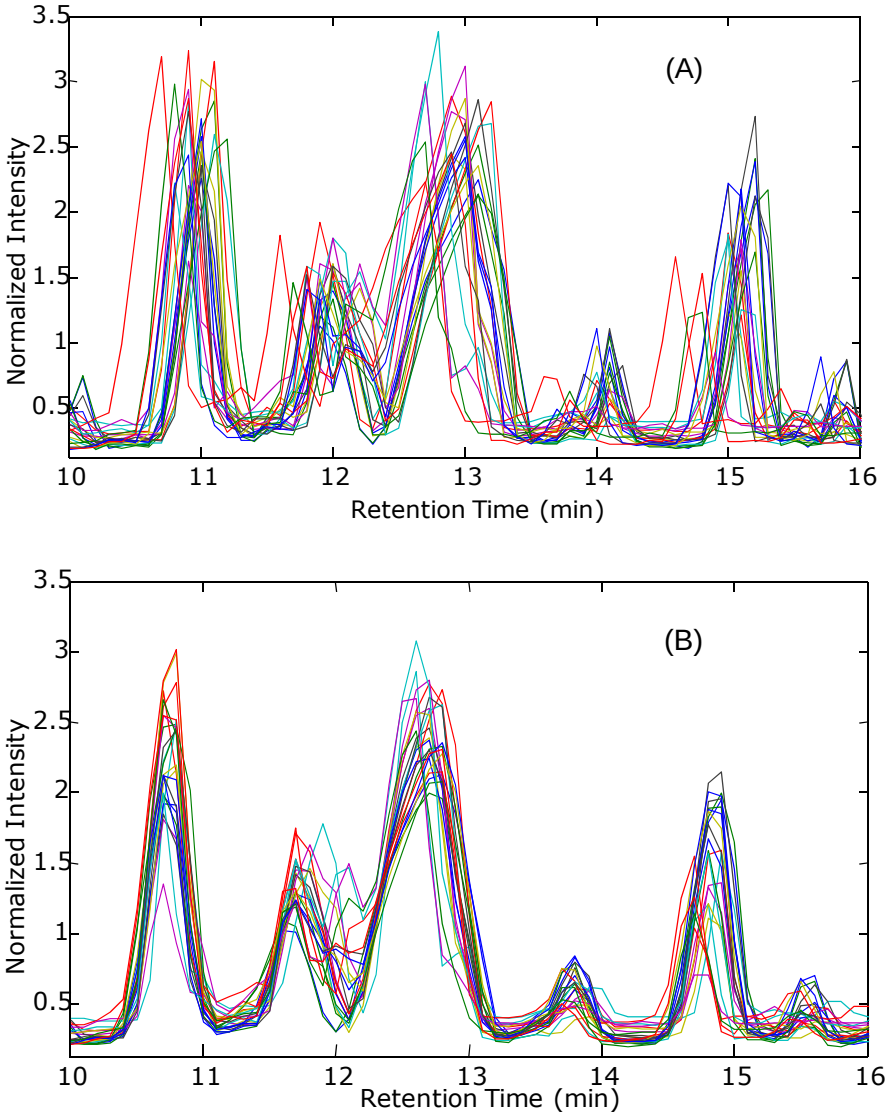

Figure 2. Enlarged TIC chromatogram (from RT=10 min to RT=16 min) before RT alignment (A) and after RT alignment (B).

3.3 Mass spectra of the samples

The mass spectra of the *P. quinquefolius* L samples grown in the United States and 12 *P. quinquefolius* L samples grown in China are displayed in Figure 3. The peak ratios of *m/z* 574.3 to peak *m/z* 1031 are significantly different between the two classes. The intensity ratio of peak *m/z* 574.3 to peak *m/z* 1031 grown in China is about 2.8:1 for the 12 samples and that in the samples grown in the United States is about 1.1:1 for the 12 samples.

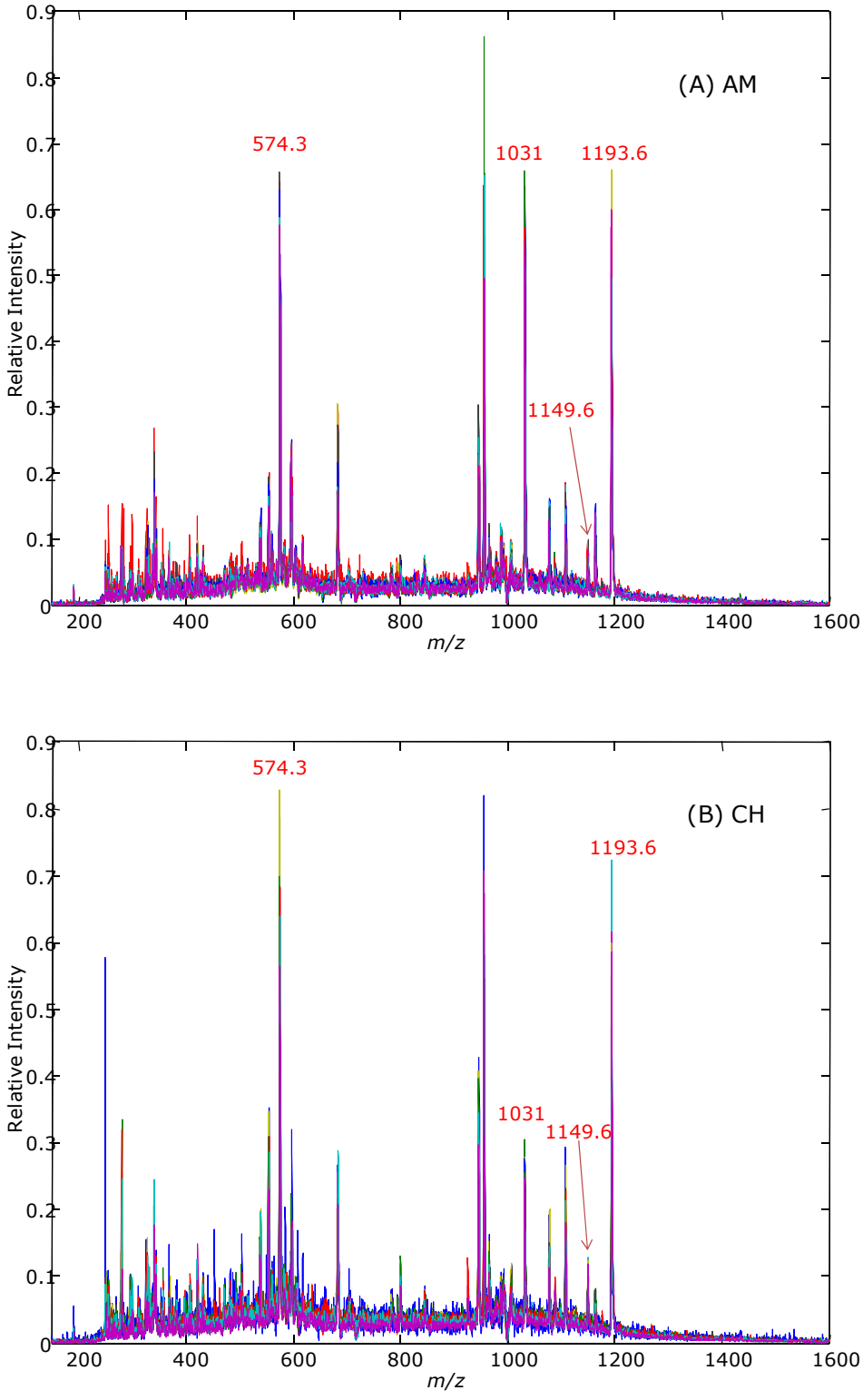

Figure 3. Mass spectra of all 12 *P. quinquefolius* L samples grown in the United States (A) and all 12 *P. quinquefolius* L samples grown in China (B).

3.4 The extracted ion chromatogram of ion *m/z* 574.3 and the corresponding mass spectra of the major peaks

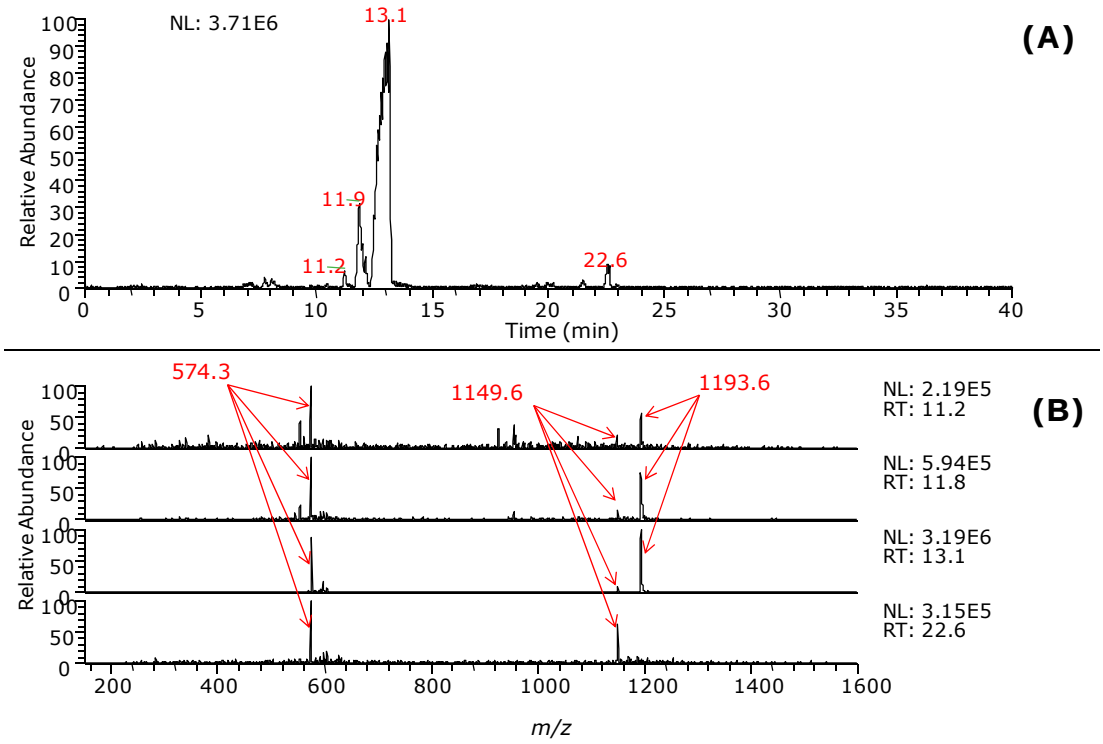

Figure 4. The extracted ion chromatogram (A) of ion *m/z* 574.3 and the corresponding mass spectra of the major peaks (B).

Sample: WBQ36.

3.5 PCA score plotting using first two most prominent rules

The PCA plotting by using the two most prominent peaks (14.9 min, *m/z* 1031) and (12.6 min, *m/z* 574.3) is displayed in Figure 5. The first and second PCs span 69% of the variation of the data set.

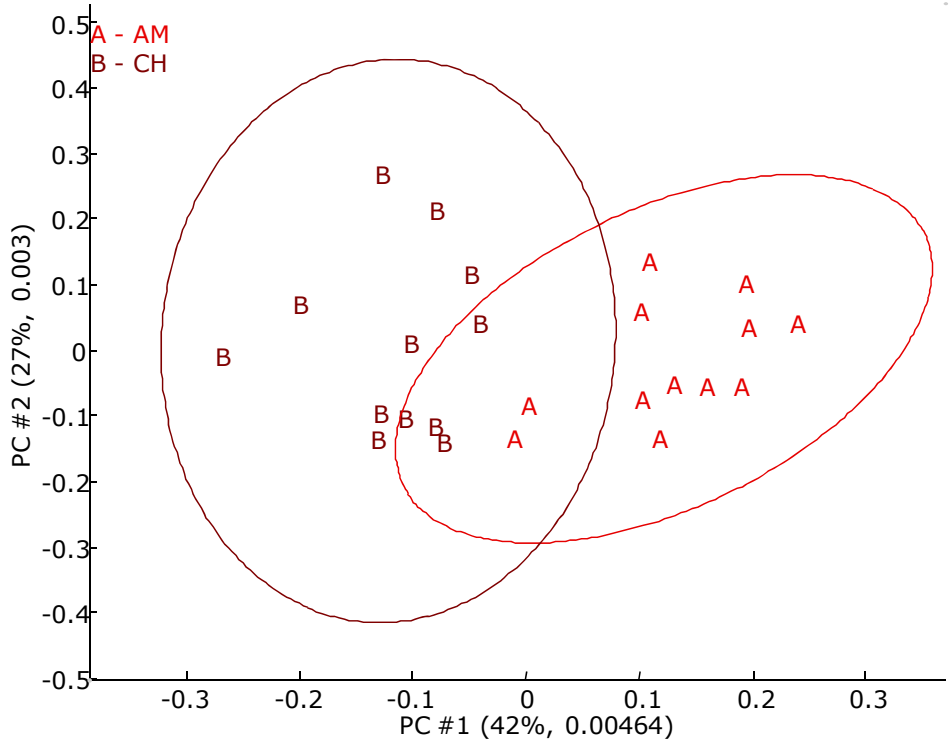

Figure 5. PCA score plot with only first two largest magnitude peaks in the two-way FuRES rule.

AM: *P. quinquefolius* L samples grown in the United States; CH: *P. quinquefolius* L samples grown in China.

(1) Harrington, P. B. *Journal of Chemometrics* **1991**, *5*, 467-486.

(2) Harrington, P. B.; Laurent, C.; Levinson, D. F.; Levitt, P.; Markey, S. P. *Anal. Chim. Acta* **2007**, *599*, 219–231.

(3) Lu, Y.; Harrington, P. B. *Analytical Chemistry* **2007**, *79*, 6752-6759.

(4) Lu, Y.; Harrington, P. B. *Analytical and Bioanalytical Chemistry* **2010**, *397*, 2959-2966.

(5) Cao, L. *Dissertation OHIO Athens* **2004**, 180.

(6) Xu, Z.; Sun, X.; Harrington, P. d. B. *Analytical Chemistry* **2011**, *83*, 7464-7471.

(7) Jackson, G. P.; Laskay, U. A.; Collin, O. L.; Hylanda, J. J.; Nichol, B.; Pasilis, S. P.; Duckworth, D. C. *Journal of the American Society for Mass Spectrometry* **2007**, *18*, 2017-2025.