Computational Prediction of Origins of *Glycyrrhizae Radix*
From Liquid Chromatography Mass Spectra

Momeen Abrar Fahad
47-096916
Arita Laobratory

A method of analyzing Liquid Chromatography Mass Spectrometry (LC-MS) data has been proposed and a system has been developed to investigate the relationship between the chemical constituents and the location of origin along with biological activity of *Glycyrrhizae Radix*. *Glycyrrhizae Radix*, derived from a species of the genus *Glycyrrhiza*, is one of the most popular medicinal plants used in crude drugs. It has been used in many Asian and European countries for over 2000 years [1]. It shows a variety of pharmacological activities and used as a remedy in Japanese traditional medicine ("Kampo" medicine) [2]. The biological activities of *Glycyrrhizae Radix* are attributed to its chemical constituents [3]. The relationship between the biological activity and the chemical constituent is complex as the biological activities do not depend on few discrete constituents; instead it depends on the combination of several related chemical constituents. It is also well known that the level of active compounds varies widely depending on the plant species, geographic source, harvesting and processing [4]. 100 samples of radix of *Glycyrrhiza uralensis* species have been collected from several places around China and Mongolia cultivated with different farming methods and from wild. For the initial investigation 33 samples have been selected. The biological activities of these hot-water extracts of the 33 samples have been measured. Even though the samples have been derived from same species they demonstrate high variance in the level of the biological activity level. The biological activity is measured as lipase inhibitory activity, ranges 1.1 to 64 in relative scale for the 33 samples. To understand the reason of the variance, the chemical constituents and their quantity has to be known for each sample extract. For that purpose, liquid chromatography mass spectrometry (LC-MS) technique has been used for these sample extracts, using Shimadzu Liquid Chromatography – Ion Trap – Time of Flight mass spectrometer. Because of the non-targeted nature of the investigation, it is required to detect every possible unique chemical constituent in the sample extract. LC column for wide-target metabolomics research has been used, which is expected to separate all kinds of metabolites. Sample preparation and LC-MS measurement were performed by Prof. Tanaka, Institute of Natural Medicine, University of Toyama.

For sample comparison analysis, it is required to detect peaks within the Total Ion Current (TIC) chromatogram generated from the LC-MS data and to align TIC chromatograms. Raw LC-MS data contain noises and disparity in retention time. It is essential to improve the signal-to-noise ratio using different signal enhancement techniques before using the TIC chromatogram for sample comparison. In TIC chromatogram the peaks contain significant information. In the case of a fine separation, different peaks within the TIC chromatogram correspond to different compounds of the separated mixture. Different signal enhancement techniques have been tested to determine the optimum enhancement required for peak detection process. These are Simple Moving Average [5], Median Filtering [6], Savitzky-Golay filtering [7] and Triangular Moving Average [5]. Each enhancement technique has its own strengths and limitations. After testing these signal enhancement techniques, Triangular Moving Average technique using 11 points, has been observed to work optimum for the peak detection process. After the signal enhancement process has been completed, the enhanced signal is used for peak detection. The detection of peaks in a chromatogram is crucial for qualitative and quantitative analysis. The minimum height threshold for a possible peak is considered as 600,000 TIC. The shape of the peak is also taken into account to distinguish possible peaks from noises. For positive scan the average number of peaks detected from a LC-MS TIC chromatogram is 40. The maximum number of detected peaks is 32 and the minimum is 51. For negative scan LC-MS TIC chromatogram, the maximum number of detected peaks is 54, the minimum is 33 and the average is 40. Even though all the samples are from same species and the same method has been used to generate LC-MS, the peak of the same constituent in different sample may occur in different retention times based on column performance and column

overloading with sample [8, 9]. Peak specific mass spectra has been analyzed to thoroughly compare the extracted samples and to align the LC-MS TIC chromatogram. Peaks with significant heights those have correlation co-efficiency [10] value more than 0.80 are considered common among the samples. Based on these common significant peaks or the Landmark peaks, required retention time corrections have been calculated to align all TIC chromatograms. After the peak identification and peak alignment process is completed, to extract features for support vector machine classification and decision tree learning, m/z chromatograms have been analyzed. Raw m/z chromatograms have been enhanced using Median Filtering technique with 11 points. From the enhanced m/z chromatogram signal the peak centroids have been detected. The similar centroids of all samples have been grouped into features.

A set of features has been selected, with minimum 7 common centroids and within retention time of 8 minutes to 28 minutes for better accuracy. Under this condition the number of features for positive ions is 562 and for negative ions 392. Using these sets of features, Support Vector Machine [11, 12] has been used with both linear kernel and radial basis function kernel, to classify the samples based on their location of origin and biological activity level. Using radial basis function kernel for support vector machine, achieved accuracy for location of origin based classification is 90%. Using the same procedure, achieved accuracy for biological activity based classification ranges from 67% to 78%.

To investigate more about the details of this relationship decision tree learning [13], which is able to process real values has been used. Decision tree learning system maps observation about an item to the conclusions of the item's target value using a predictive model. Because of its white box nature, decision tree learning is an ideal tool to determine the internal dependency of classification. Using decision tree learning sets of metabolites significant for location of origin based classification and biological activity based classifications have been detected. To identify the metabolites, the sets of metabolites have been compared to information found in KNApSAcK [14] database.

References:

1. Zhang Q, Ye M, (2009), Chemical analysis of the Chinese herbal medicine Gan-Cao (licorice), J. Chromatogr. A, 1216, 1954-1969.
2. Society of Japanese Pharmacopoeia, (2007), The Japanese Pharmacopoeia, 15th ed., The Ministry of Health, Labour and Welfare (English edition), Tokyo, 1026.
3. Asl M N, Hosseinzadeh H, (2008), Review of pharmacological effects of Glycyrrhiza sp. and its bioactive compounds, Phytother. Res., 22, 709-724.
4. Liang Y Z, Xie P, Chan K, (2004), Quality control of herbal medicines. J. Chromatogr. B, 812, 53-70.
5. Felinger A, (1998), Data Analysis and Signal Processing in Chromatography, Elsevier, 145-159.
6. Moore A W, Jorgenson J W, (1993), Median Filtering for Removal of Low-Frequency Background Drift, Anal. Chem. 65, 188-191.
7. Savitzky A, Golay M J E, (1964), Smoothing and differentiation of data by simplified least squares procedures Anal. Chem. 36, 1627–163.
8. Etxebarris N, Zuloaga O, Olivares M, Batolome L J, Navarro P, (2009), Retention-time locked methods in gas chromatography, J. Chromatogr. A, 1216, 1624.
9. Van Nderkassel A M, Daszykowski M, Eilers P H, Heyden Y V, (2006), A comparison of three algorithms for chromatograms alignment, J. Chromatogr., A, 1118, 199.
10. Pesyna G M, Venkataraghavan R, Dayrnger H E, McLafferty F W, (1976), Probability based matching system using a large collection of reference mass spectra, Anal. Chem., vol. 48, 1362-1368.
11. SVMlight: http://svmlight.joachims.org/
12. LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/
13. Decision tree applet: http://www.montefiore.ulg.ac.be/~geurts/dtapplet/dtexplication.html
14. KNApSAcK database: http://kanaya.naist.jp/knapsack_jsp/top.html