Summary of Doctoral Thesis

Enrollment year: 2018 y 10 m
UGAS Specialty: Bioresources Science
Major of Food Science
Name Son Hae Reon

Title	**Efficient profiling base on the database to elucidate chemical structure and biosynthetic pathway of soyasaponins**

Introduction and purpose

In Asia, soybeans have been used for generations as soy based foods, such as soy paste (miso), soy source, tofu, and fermented soybeans (natto and tempe); in addition soybeans are processed as forage, cooking oil, and biofuel. The worldwide soybean production is about 300 Mt, behind maize 1000 Mt, rice 740 Mt, and wheat 730 Mt. Over 80% of wheat and rice are used for food and the remaining 20% are used for other non-food products. By contrast, less than 20% of soybeans are consumed as foods and over 75% are used for processed products, such as soybean oil or flour for livestock feed, and the remaining soybeans are used for other industrial purposes.

It is unfortunate that only small portions of the worldwide soybean production are used for foods because soybeans have high nutritional values and health benefits. Soybeans not only contain high amounts of protein, fat and dietary fiber, but also additional health benefit components, such as isoflavones and saponins, which are not abundant in wheat or rice. Daily consumption of soybean based foods is considered to be a contributing factor in improving human health and longevity.

Soyasaponins or soybean saponins are found in various legumes, but they are particularly abundant in soybean seeds. Soyasaponins are complex molecules consisting of more than 100 structural variations that show various health benefits and taste characteristics depending on the chemical structures. Soyasaponins contain sugar chains in the molecule but sugar chain sequences having the same molecular mass like a hexose-hexose sequence cannot be elucidated by LC-MS analysis. Therefore, time consuming purification steps and instrumental analysis, such as HI-MS, IR, NMR, have always been required to analyze the chemical structures.

This research has demonstrated that the chemical structures and biosynthetic pathway of soyasaponins can be elucidated by using a new method called "profiling base on the database (PBOD)" without purification and instrumental analysis. We believe that this new method would improve efficiency in analyzing the structural variations showing various health benefits and taste characteristics. Overview of this new PBOD method is summarized below:

First, we created a set of database (TLC, LC-PDA/MS/MS and genetic information) of major 36 soyasaponins. Among the 36 soyasaponins, chemical structures of 20 components have been reported by purification and instrumental analysis. It is also reported that the sugar chain sequences of soyasaponins are strictly regulated by the combination of some genes on independent loci relating to soyasaponin biosynthesis. By using LC-PDA/MS/MS and genetic information without purification and instrumental analysis, we were able to detect the sugar chain sequences containing a glucose and a galactose residue of the remaining 16 soyasaponins.

Second, we evaluated whether the PBOD can elucidate the chemical structures of unknown soyasaponins produced by additional enzymatic modification to express genes of aglycone parts, except sugars. The estimated chemical structures of newly discovered Sg-6 saponins found in the

wild soybeans by the PBOD were identical with the chemical structures elucidated by purification and instrumental analysis afterwards.

These results strongly suggest that this newly developed PBOD method in this research is a very effective tool to elucidate the chemical structure and biosynthetic pathway of soyasaponins. Furthermore, the database can be expanded to include new soyasaponin information, and the expanded database may enhance the profiling procedure and efficiency.

Materials and methods

We obtained soybean (*Glycine max*) seeds of five varieties; Enrei, Ryuho, Norin No. 3, Ibarakimame No. 7, and Kinusayaka, and two accessions of wild soybeans (*Glycine soja*); GD50029-2 and GD50326-2 from the National Institute of Crop Science, National Agriculture and Food Research Organization (NARO), Tsukuba, Japan. These wild soybeans contain Sg-6 saponins in the seeds. An inbred progeny of the F8 line accumulating Sg-6 saponins in the seed was obtained from the cross between Shirosennari (*Glycine max*) and GD50326-2, which were grown and harvested in the field at Iwate University, Morioka, Japan.

Basic procedure of purification was followed by the method of Itabashi *et. al.* [1]. A thin layer chromatography (TLC) and a liquid chromatography-photo diode arrey-tandem mass spectrometry (LC-PDA/MS/MS) analysis was carried out for the preliminary analysis assay of saponin composition according to Krishnamurthy *et. al.* [2, 3]. Fourier transform infrared spectoscopy (FT-IR) (SPECTRUM2000, Parkin Elmer, Japan) with a range of 4000-800 cm^{-1} were observed (0.2 cm^{-1} measurement accuracy). Purified soyasaponin powder (2 mg) was placed on an attenuated total reflectance (ATR) accessory followed by a few drops of methanol to spread the powder to achieve close contact on the ATR cell, then methanol was air-dried. The purified soyasaponin powder (40 mg) was dissolved in 0.75 mL of pyridine-d_5 for a nuclear magnetic resonance (NMR) analysis. ^{1}H-NMR (500 MHz), ^{13}C-NMR (125 MHz), DEPT and various 2D-NMR data (COSY, NOESY, HMQC, HMBC) were measured by using NMR equipment (AVANCE III 500, Bruker, United States).

Results

1. Construction of Soyasaponin Database

The nomenclatures of 36 types of soyasaponins were based on the combination of the three aglycone structures and the sugar moiety composition of each aglycone. The conjugated sugar moiety at the C-3 position of the aglycone varies depending on the sugar sequence and has been specifically categorized as the αg (Glc-Gal-GlcUA-), βg (Rham-Gal-GlcUA-), γg (Gal-GlcUA-), αa (Glc-Ara-GlcUA-), βa (Rham-Ara-GlcUA-), and γa (Ara-GlcUA-) types. Different soyasaponin chemical structures demonstrated different health functionalities.

The main 36-soyasaponin database was constructed based on TLC analysis, LC-PDA/MS/MS analysis, and the results of reported genetic analysis to date.

The information available from TLC analysis can predict the composition of saponin aglycone based on the color. Recently, TLC analysis has been widely studied as a method of primary screening for phenotypes in large samples using plant alcohol extracts. In this analysis, soyasaponins show different colors depending on the aglycone group as follows: group A saponins containing soyasapogenol A are purple, group B and group DDMP saponins containing

soyasapogenol B are brown, and group E saponins containing soyasapogenol E are blue.

LC-PDA/MS/MS analysis provide more detailed information on soyasaponins than TLC. Under the analysis condition, retention time (RT) of each soyasaponin varies depending on the physicochemical characteristics of the molecule such as hydrophilicity, hydrophobicity, and polarity. Group A saponins are relatively hydrophilic and elute earlier in the process followed by group B, group E, and group DDMP saponins. Precise MS information (within 5 ppm tolerance) on the molecular ion peak $[M+H]^+$, which is the parent MS, can provide the molecular formula of the component.

The soyasaponin compositions of the seeds of five different varieties were classified by LC-PDA/MS/MS analysis. By using MS data, all βg type saponins showed the "rhamnose desorbed form" from the C-3 sugar chain of the molecular ion peak as the base peak. The base peak of all other saponins were the "glucuronic acid conjugated form" at the C-3 position of aglycone. Thus, the base peak of MS data varies depending only on the C-3 sugar chain sequence. If the C-3 sugar chain is the βg type, the base peak is always the "rhamnose desorbed form". However, if the C-3 sugar chain is not the βg type, the base peak is always the "glucuronic acid conjugated form", irrespective of different aglycones.

Sugar sequences of soyasaponins were genetically controlled and the combination of soyasaponin biosynthetic genes on some loci is decided by the soyasaponin phenotype. Thus, the MS/MS fragmentation spectra can confirm the sugar chain sequence and the molecular formula of aglycone. The chemical structure and biosynthetic pathway of new soyasaponins can be predicted by using a flowchart of the profile from the database.

2. Utilization of Database to Characterize New Soyasaponins (Group Sg-6 Saponin)

The chemical structure of group Sg-6 saponins were elucidated to estimate tentative chemical structures in some lines of wild soybean by using the PBOD. Group Sg-6 saponins were isolated to obtain a highly purified substance, and chemical structures of three aglycones (soyasapogenols H, I, and J) were confirmed via instrumental analysis. The conjugated sugar moiety at the C-3 position in the group Sg-6 saponins was the same as in common soyasaponins. However, LC-PDA/MS/MS, IR, and 2D-NMR analyses showed the chemical structures of the aglycone of soyasapogenol H, I, and J consisted of $-CH_2OH$, -COOH, or $-CH_2$-O-CO-CH_2-COOH (malonyl) moieties, respectively, at the C-29 position of soyasapogenol E (3β,24-dihydroxyolean-12-en-22-one). Elucidated chemical structures were identical with those estimated by the PBOD.

Conclusion and consideration

Soyasaponins are complex chemical components comprising of a triterpenoid aglycone and sugar moieties. Prediction of the complex biosynthetic pathway and chemical structure of soyasaponins has been studied for over the past 20 years. Predicting the chemical structure of soyasaponins is particularly important when the biosynthesis pathway has not yet been revealed. If we predict the chemical structure, experiments can be designed based on know or unknown molecular weights, such as target analysis or non-target analysis, respectively. Furthermore, it is difficult to obtain the exact mass data of secondary metabolites in plants because it is necessary to provide a broad overview on the experimental design as well as theoretical structure elucidation techniques for high molecules by using mass spectrometry.

The MS and MS/MS fragmentation database developed in this study efficiently provides

information on the chemical structure of soyasaponins by using a simple method that needs only extraction, hydrolysis, and LC-PDA/MS/MS analysis without time-consuming purification. The database can also provide useful information of soyasaponins detected in soy foods and chemical structures of some components in other plants, which are difficult to elucidate without purification [4].

Therefore, this study focused on examining the possibility of the chemical structure and biosynthesis pathways of new soyasaponins (group Sg-6 saponin) by using the PBOD, a new method that does not require the purification of soyasaponins. The NMR analysis results and the prediction results of the PBOD showed the same chemical structure for the group Sg-6 saponin, which validated the efficacy and accuracy of the proposed approach. Also, when new soyasaponins are discovered, accuracy can be improved by adding more new database information for profiling. Thus, the PBOD method has considerable potential.

Furthermore, we can use the PBOD (profiling base on the database) method for the forward genetic screening of ethyl methanesulfonate (EMS)-treated soybeans. In addition, new soyasaponins (A-series saponins; Y3 and Y4) were found in Category 6 mutants. The Y3 and Y4 saponins ware purified in small amounts and they were used as arabinosyl acceptor *in vitro* and *in vivo* experiments to estimate unknown genes encoding the enzyme having arabinosyltransferase activity. These facts help tremendously in estimating the unknown biosynthetic pathways of soyasaponins. Also, the rapid screening of phenotypes using the PBOD is the key in overcoming the limitations of materials required for estimating and verifying unknown biosynthetic pathways in EMS-treated soybeans.

In conclusion, the results of this study can provide multiple insights in the field of research on saponin-containing plants, secondary metabolites, and pharmaceuticals.

Reference

1. M. Itabashi, C. Tsukamoto, A. Kurosaka, P. Krishnamurthy, T. S. Shin, S. H. Yang, E. Son and G. Chung (**2016**). Efficient method for large-scale preparation of two components H and I of Sg-6 saponins from whole seeds of wild soybean (Glycine soja Sieb. and Zucc.). ***Journal of Liquid Chromatography and Related Technologies***, 39 (14), 640-646.
2. P. Krishnamurthy, C. Tsukamoto, S. H. Yang, J. D. Lee and G. Chung (**2012**). An improved method to resolve plant saponins and sugars by TLC. ***Chromatographia***, 75, 1445-1449.
3. P. Krishnamurthy, Y. Fujisawa, Y. Takahashi, H. Abe, K. Yamane, K. Mukaiyama, H. R. Son, S. Hiraga, A. Kaga, T. Anai, Tsukamoto and M. Ishimoto (**2019**). High-throughput screening and characterization of a high-density soybean mutant library elucidate the biosynthesis pathway of triterpenoid saponins. ***Plant and Cell Physiology***, 60 (5), 1082-1097.
4. H. Son, K. Mukaiyama, Y. Omizu, and C. Tsukamoto (**2021**). A fragmentation database of soyasaponins by liquid chromatography with a photodiode array detector and tandem mass spectrometry. ***Analytical Science and Technology***, 34 (4), 172-179.