ORIGINAL PAPER

An ^{1}H-NMR-based Metabolomics approach for understanding the influence of geographical origin differentiation on the major diarylheptanoid phytoestrogens in *Curcuma comosa* rhizome

Worawut Chaiyasaeng[a], Jamrearn Buaruang[b], Apichart Suksamrarn[a], Waraluck Chaichompoo[c,d], Wachirachai Pabuprapap[a], Wang Nguitragool[e], and Boon-ek Yingyongnarongkul[a,*]

[a] Department of Chemistry and Center of Excellence for Innovation in Chemistry, Faculty of Science, Ramkhamhaeng University, Bangkapi, Bangkok, Thailand, 10240
[b] Division of Environmental Science, Faculty of Science, Ramkhamhaeng University, Bangkapi, Bangkok, Thailand, 10240
[c] Department of Food and Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok, Thailand, 10330
[d] Center of Excellence in Natural Products for Ageing and Chronic Diseases, Chulalongkorn University, Bangkok, Thailand, 10330
[e] Department of Molecular Tropical Medicine and Genetics, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand, 10400

*Corresponding author: *boonek@ru.ac.th (B-e. Yingyongnarongkul)*

Received: 03 April 2024 / Revised: 23 April 2024 / Accepted: 26 April 2024

Abstract. *Curcuma comosa* has been widely used for gynecological herbal or traditional medicinal purposes. It has been reported that this turmeric plant offers several biological activities, such as anti-inflammatory, anti-lipidemic, and estrogen-like effects. Authentication of *C. comosa* is crucial to ensure its authenticity and prevent adulteration. Plants from different origins will have distinct metabolite compositions due to the influence of soil nutrition, climate, temperature, and humidity. We used ^{1}H-NMR spectroscopy in combination with chemometrics (Principal Component Analysis (PCA) and Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA)) analysis to unveil metabolic differences among *C. comosa* samples from six different provinces of Thailand (Kalasin, Nakhon Phanom, Samut Sakhon, Nakhon Pathom, Phetchabun, and Chiang Mai). The ^{1}H-NMR analysis revealed the presence of twelve metabolites, including DA1-DA5, sucrose, α-glucose, β-glucose, formate, fatty acids, and two DA-OAc, which were identified as potential metabolites for differentiating geographical origins. Additionally, a higher normalized abundance of DA1-DA5 was observed in *C. comosa* from Nakhon Pathom, Phetchabun, Samut Sakhon, and Nakhon Phanom compared to Chiang Mai and Kalasin. Both the optimal PCA and OPLS-DA models demonstrated a good fit ($R^2 > 0.8$) and strong predictivity ($Q^2 > 0.5$). The optimal OPLS-DA model was validated through permutation tests, yielding high values for the original R^2 and Q^2. In conclusion, metabolite fingerprinting using ^{1}H-NMR spectroscopy and chemometrics provides a powerful tool for authenticating *C. comosa* rhizomes.

Keywords: Chemometrics, Fingerprinting, Food authentication, *Curcuma comosa*, Wan Chak Motluk, OPLS-DA, PCA

1. Introduction

Curcuma comosa Roxb. is a member of the Zingiberaceae family and commonly found in tropical and subtropical regions of Asia, with notable presence in Thailand, Myanmar, Malaysia, and Indonesia. In Thailand, it is referred to as 'Wan Chak Motluk.' This plant's rhizome is edible and has a long history of traditional use in gynecological herbal medicine. It is known for its various therapeutic properties, including potent estrogen-like activity (Suksamrarn et al., 2008; Winuthayanon et al., 2009), anti-inflammatory effects (DeFilipps et al., 2018; Sodsai et al., 2007; Suksen et al., 2016), anti-lipidemic benefits (Piyachaturawat et al., 1999), and an ability to effectively enhance human osteoblast function (Tantikanlayaporn et al., 2013a). Previous studies reported that active constituents include diarylheptanoids such as (3*S*)-1-(3,4-dihydroxy-phenyl)-7-phenyl-(6*E*)-6-hepten-3-ol (DA1), (3*R*)-1,7-diphenyl-(4*E*,6*E*)-4,6-heptadien-3-ol (DA2), (3*S*)-1,7-diphenyl-(6*E*)-6-hepten-3-ol (DA3), 1-(4hydroxyphenyl)-7-phenyl-(4*E*,6*E*)-4,6-heptadien-3-one (DA4), and 1,7-Diphenyl-(4*E*,6*E*)-4,6-heptadiene-3-one (DA5) (Figure 1) (Suksamrarn et al., 2008; Yingngam et al., 2018). These diarylheptanoids

are of interest because they exhibit hypolipidemic and estrogenic-like activities. Specifically, DA1-DA3 are most likely the major diarylheptanoids in this plant's rhizome, responsible for many pharmacological effects both *in vitro* and *in vivo* (Piyachaturawat et al., 1995; Tabboon et al., 2019; Tantikanlayaporn et al., 2013b). Traditional uses and scientific studies suggest that the rhizomes of *C. comosa* have the potential in applications such as dietary supplements, alternative remedies and estrogen replacement (Keeratinijakal et al., 2010).

In recent years, the analysis and control of *C. comosa* quality have emerged as crucial research areas in food safety supervision. Determining the geographical origin of *C. comosa* is essential to maintain its high quality and advance the *C. comosa* industry (Petchprayoon et al., 2022; Tabboon et al., 2019). Based on our knowledge, the quality of *C. comosa* can be influenced by environmental conditions in the cultivation area, resulting in a distinctive metabolic profile. It is evident that the content of diarylheptanoids in the plant material varies depending on the collection location. Food authentication is of great interest to a wide range of individuals and organizations, including the scientific community, law enforcement, food producers, importers, exporters, and consumers. This is because food fraud, driven by economic motives, such as misbranding and false geographical origin, has become a growing worldwide concern in recent years (Danezis et al., 2016). The geographical origin is one of the most crucial aspects of food authenticity, and consumers have the right to access complete information about the food they are consuming (Macready et al., 2020). Therefore, determining the origin of food is essential in managing the food supply chain. Recently, metabolomics technology has been widely applied in the field of agricultural food analysis. Utilizing Nuclear Magnetic Resonance (NMR) spectroscopy technology, in combination with multivariate statistical methods such as Principal Component Analysis (PCA) and Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA), has shown excellent geographical discrimination capability (Zhou et al., 2021). This technique offers several advantages, including high throughput, good reproducibility, ease of operation, rich structural information, and environmental friendliness. These advantages have led to its applications in food composition analysis, adulteration testing, and geographical traceability (Pontes et al., 2017; Sobolev et al., 2019). However, to the best of the authors' knowledge, no previous studies have been conducted for *C. comosa* authentication using a ^{1}H-NMR-based metabolomics approach to establish a relationship with the geographical origin.

Therefore, the objective of this study was to assess the potential of ^{1}H-NMR spectroscopy combined with chemometrics (PCA and OPLS-DA) analysis as an approach to identify metabolic differences within *C. comosa* samples originating from various geographical locations (Kalasin (KSN), Nakhon Phanom (NPM), Samut Sakhon (SKN), Nakhon Pathom (NPT), Phetchabun (PNB), and Chiang Mai (CMI)). Furthermore, in comparison to previous *C. comosa* fingerprinting studies, unsupervised PCA was initially applied to provide an overview of the trends in *C. comosa* samples and form clusters, while supervised OPLS-DA models were developed for classification purposes. OPLS-DA, followed by Variable Importance in Projection (VIP) analysis and Analysis of Variance (ANOVA) with post-hoc Tukey's Honest Significant Difference (HSD) test, was further applied to identify potential chemical markers responsible for reliable group separation. The findings demonstrate that the combination of NMR and chemometrics can provide discrimination biomarkers for the geographic origin of *C. comosa* industry.

2. Materials and Methods

2.1 Materials and chemicals

The rhizomes of 32 *C. comosa* individuals were collected from various locations in Thailand, including KSN and NPM in the northeastern region, SKN and NPT in the central region, and CMI and PNB in the northern region. Detailed sampling information can be found in Table S1. Metabolites, including DA1-DA4, were isolated and purified from *C. comosa*.

Methanol (CH_3OH, 99%) was obtained from Apex Chemicals Co., Ltd. (Huai Khwang, Bangkok, Thailand). The internal standard reference for NMR analysis, 3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid sodium salt (TSP-d_4) (98.0% D), was sourced from Sigma-Aldrich (St. Louis, MO, USA). Additionally, 99.8% CD_3OD was purchased from Cambridge Isotope Laboratories (Andover, MA, USA), and NMR tubes with a borosilicate composition (tube diameter 5 mm) were procured from Merck (Darmstadt, Hessen, DEU).

2.2 Preparation and drying of samples

Fresh *C. comosa* rhizomes were thoroughly washed to remove all physical contaminants, including soil and gravel. Subsequently, the rhizomes were chopped into uniformly sized small pieces and left to air dry at room temperature until a constant weight was reached. Afterward, the dried samples were ground and passed through a 40-mesh sieve before being stored at -20 °C in preparation for extraction.

2.3 Crude extraction of samples

For each sample, 5 g of dried rhizome powder was macerated in 250 ml of 99% CH_3OH for three days at room temperature. Afterward, the supernatant was filtered using Whatman filter paper and then evaporated to dryness under vacuum with a rotary evaporator (Meierseggstrasse, Flawil, Switzerland). The crude rhizome extracts were obtained separately and stored at -20 ˚C for further analys

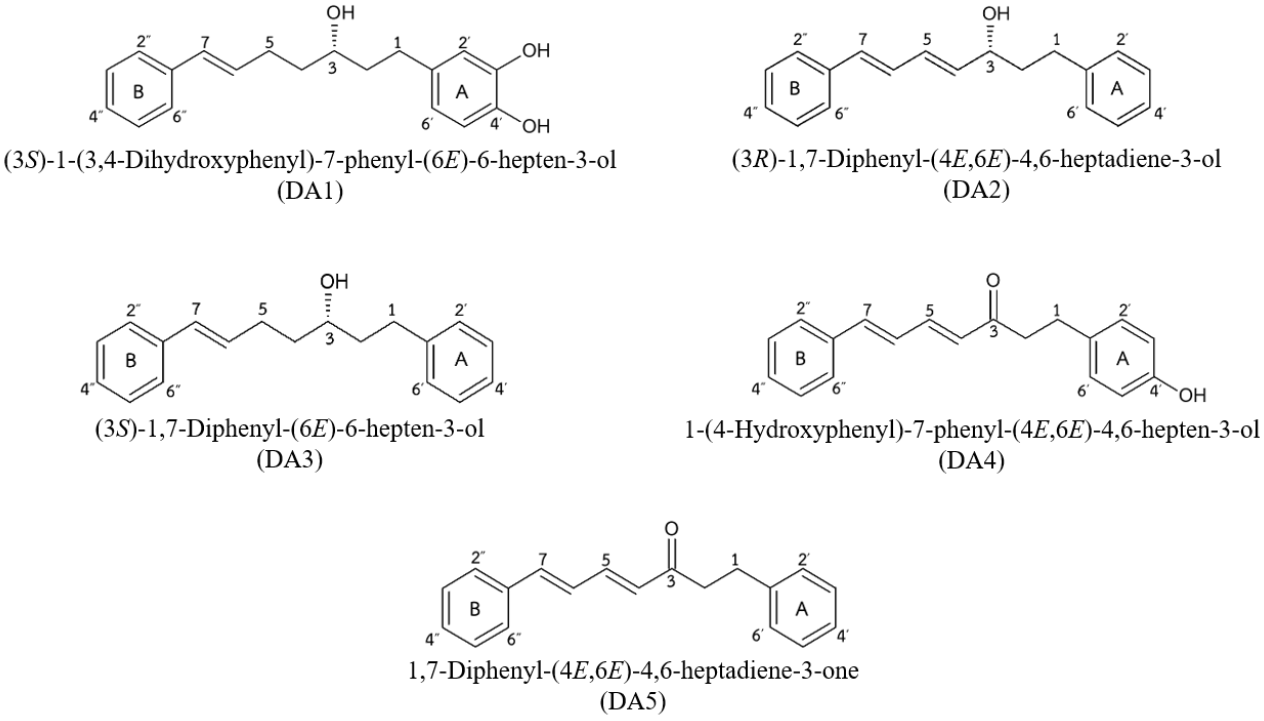

Figure 1. Chemical structures of the major (DA1-DA3) and some minor diarylheptanoid (DA4-DA5) composition in *C. comosa* rhizomes.

2.4 Sample preparation for analysis

For each sample, 20 mg of rhizome extract was accurately weighed and placed into a 1.5 ml microcentrifuge tube. Subsequently, 650 µl of CD_3OD containing 0.01% TSP-d_4 was added, and the mixture was vortexed for 1 minute at room temperature. After 10 minutes of ultrasonication at room temperature, the mixture was centrifuged at 8000 rpm for 10 min. Finally, 600 µl of the supernatant was transferred into a 5 mm NMR tube for analysis by ^{1}H-NMR. Quality control (QC) samples were prepared by mixing 2 µl of each *C. comosa* extract with 3.6 ml of CD_3OD containing 0.01% TSP-d_4. Six QC samples were also analyzed to ensure instrumental conditions and data quality during the analyses. ^{1}H-NMR spectra were acquired using a 400 MHz NMR spectrometer (Bruker, USA).

2.5 Pre-processing ^{1}H-NMR data

The analysis of the *C. comosa* extracts involved recording all ^{1}H-NMR spectra using the standard Bruker pulse program zg30, with an acquisition time of 4.0894465 seconds. The one-dimensional NMR experiment was conducted at 298.0 K, including 2 dummy scans, and a receiver gain set to 71.8. A total of 60 scans were recorded, with a Free Induction Decay (FID) size of 65K, FID resolution of 0.24 Hz, and a spectral width of 20.0254 ppm. The ^{1}H-NMR spectra were subjected to Fourier transformation using Bruker Biospin Topspin software (Version 4.1.4, Bruker Biospin GmbH, Rheinstetten, Germany). Subsequently, phase and baseline correction of the ^{1}H-NMR spectra were performed using MestReNova software (Version 14.3.2, Mestrelab Research, Santiago de Compostella, Spain). TSP-d_4, with a chemical shift at δ 0.00, was used as a reference for the crude extract spectrum. Regions containing residual water signals (δ 4.55-4.65, 4.75-4.94 ppm) and residual methanol signals (δ 3.28-3.335 ppm) were removed before data normalization. Two normalization methods (total and standardized area) were applied and their performance was compared. To simplify the spectral data for statistical analysis, spectral binning was conducted. In our study, the spectral regions from 0.5 to 10.0 ppm were divided into bins with equal widths of 0.005 ppm. Relative intensities of the binned spectral data in both total and standardized area normalization were obtained by dividing the spectral data by the total area of all bins and the area of the reference peak, respectively. The binned datasets were then converted into a data matrix or feature table using Microsoft Office Excel (version 2019, Microsoft, Redmond, WA, USA) to reformat the file for multivariate analysis. Metabolite confirmation was carried out by comparing chemical shifts with those of standard compounds from parallel experiments and by using correlation resonances, particularly statistical total correlation spectroscopy (STOCSY). Additionally, we conducted 2D ^{1}H-^{1}H correlation spectroscopy (COSY), ^{1}H-^{13}C heteronuclear single quantum correlation (HSQC), and heteronuclear multiple bond correlation (HMBC) experiments to aid in metabolite identification. Furthermore, the resonances of interest were cross-referenced with online metabolite databases such as the Biological Magnetic Resonance Data Bank (BMRD; http://www.bmrb.wisc.edu) and the Human Metabolome Database (HMDB; http://www.hmdb.ca/) (Dona et al., 2016; Hernández-Guerrero et al., 2021; Lodge et al., 2021; Montoya-García et al., 2023).

2.6 Multivariate statistical analysis of metabolomics data

To establish the optimal model, we applied two normalization methods (total and standardized area) and two scaling methods (unit variance (UV) and Pareto). Subsequently, the resulting NMR datasets were imported into SIMCA-P version 18.0 (Umetrics, Umea, Sweden) for Principal Component Analysis (PCA) and Orthogonal Partial Least Squares-Discriminant Analysis (OPLS-DA) to discriminate the geographic origin of *C. comosa* rhizome. The number of components was determined using the autofit function to select the significant number of components. The effectiveness of the established models was assessed using the cumulative (cum) R^2 and Q^2, where values close to 1 indicate an excellent model, and

values greater than 0.5 indicate good model quality (Rocchetti et al., 2021; Wang et al., 2023). The OPLS-DA model was further validated through a permutation test (200 permutations) and analysis of variance of the cross-validated residuals (CV-ANOVA, *p*-value < 0.05) to detect and prevent overfitting (Chao et al., 2017). Additionally, the areas under the curve (AUC) of the receiver operating characteristic (ROC) curves were used to evaluate the overall predictive power of the OPLS-DA models. The Variable Importance in Projection (VIP) approach was utilized to select variables with the highest discrimination potential in the OPLS-DA model, referred to as markers (considering a VIP score > 1). Subsequently, an Analysis of Variance (ANOVA) with a post-hoc Tukey's HSD test was conducted using the MetaboAnalyst 5.0 platform (http://www.metaboanalyst.ca) to investigate significant differences in the concentrations of VIP-selected metabolites for geographical differentiation (Kim et al., 2020; Rivera-Pérez et al., 2021). A false discovery rate (FDR) analysis was also performed on the data, and all significant metabolites predicted by ANOVA were found to be within the FDR cutoff of 0.05 (Singh et al., 2022; Yousf et al., 2019).

3. Results

3.1 ^{1}H-NMR-based metabolomic profile analysis

In order to assign each signal to the corresponding metabolites, 1D ^{1}H-NMR spectra, 2D correlation experiments (COSY and HSQC) and HMBC were used. Connectivity information was obtained from 1D spectra together with the chemical shift data reported in the literature (Suksamrarn et al., 2008). In addition, spectra of standard compounds (Figure S1), STOCSY plot (Figure S2, S3, S4, S5), and HMDB and BMRB databases were utilized. In this study, 12 metabolites including DA1-DA5, 2 acetate-group diarylheptanoids (DA-OAc), sucrose, α-glucose, β-glucose, formate, and fatty acid were identified in *C. comosa*. Representative ^{1}H-NMR spectra of *C. comosa* from SKN province are displayed in Figure 2 with metabolites labeled. Table 1 summarizes the experimental information obtained from the NMR spectra from the *C. comosa* extracts.

3.2 Multivariate statistical analysis of metabolomics

3.2.1 Sample clustering using unsupervised PCA

In the first phase of multivariate statistical analysis, unsupervised PCA was conducted to assess sample clustering from different geographical origins without providing any prior information about sample classes. The optimal PCA model was established by applying Standardized area and Pareto scaling, as listed in Table S2. A tight clustering of quality control samples (QC) was observed (Figure 3A), indicating that acceptable stability of the metabolomics platform throughout the run. However, only NPT samples clearly separate from the other groups (Figure 3A). When the PCA score plot was constructed excluding QC samples (Figure 3B), clear differentiation of NPT, SKN, and PNB samples from the others was evident; NPT *C. comosa* samples were clustered in the second quadrant whereas SKN and PNB *C. comosa* samples were in the fourth quadrant. Distinguishing CMI, NPM, and KSN *C. comosa* samples proved challenging using the unsupervised method.

3.2.2 Geographical discrimination using supervised OPLS-DA

In the second phase of the statistical analysis, a highly satisfactory OPLS-DA model was established to discriminate the geographical origin of *C. comosa* and identify potential biomarkers responsible for separating *C. comosa* by region. The optimal OPLS-DA model was established by applying standardized area normalization, UV scaling, and utilizing 14 components (5 predictive + 9 orthogonal), resulting in the highest R^2Y (0.985) and Q^2 (0.792) values, as listed in Table 2. R^2Y and Q^2 parameters exceeding 0.5 and approaching 1 indicated a strong correlation (R^2Y) and predictive power (Q^2) of the model for

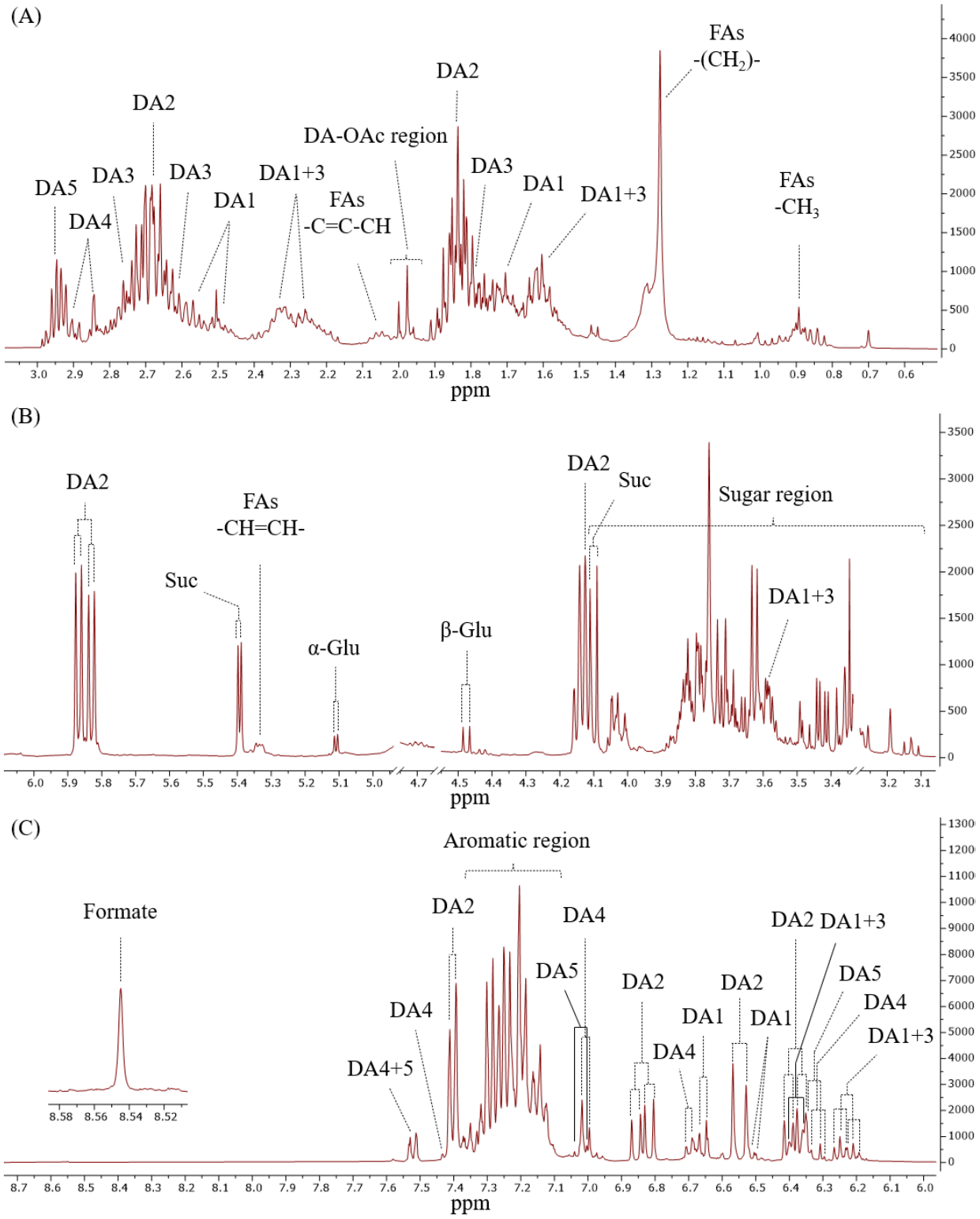

Figure 2. Representative 400 MHz ^{1}H-NMR spectra of a *C. comosa* extract from SKN province with expansions from δ 0.6-3.0 ppm (A), δ 3.1-6.0 ppm (B) and δ 6.0-8.7 ppm (C), recorded in CD_3OD containing 0.01% TSP-d_4.

Table 1. Peak assignment of ^{1}H-NMR spectra in *C. comosa* extracts, recorded in CD_3OD containing 0.01% TSP-d_4.

Metabolites	Abbreviation	Position	Chemical Shift (Multiplicity)	Assignment Method and Ref.
(3*S*)-1-(3,4-dihydroxy-phenyl)-7-phenyl-(6*E*)-6-hepten-3-ol	DA1	1a 1b 2 3 4 5a 5b 6 7 2′ 5′ 6′ 2″-6″	2.47-2.55 (*m*) 2.59-2.68 (*m*) 1.66-1.75 (*m*) 3.56-3.61 (*m*) 1.55-1.66 (*m*) 2.20-2.29 (*m*) 2.29-2.38 (*m*) 6.22 (*dt*) 6.39 (*d*) 6.64 (*br s*) 6.66 (*d*) 6.50 (*dd*) 7.13-7.34 (*m*)	1D, 2D NMR, STOCSY, Suksamrarn et al., 2008
(3*R*)-1,7-diphenyl-(4*E*,6*E*)-4,6-heptadien-3-ol	DA2	1 2 3 4 5 6 7 2′-6′ 2″-6″	2.66-2.75 (*m*) 1.81-1.90 (*m*) 4.13 (*q*) 5.84 (*dd*) 6.38 (*dd*) 6.83 (*dd*) 6.54 (*d*) 7.13-7.42* 7.13-7.42*	1D, 2D NMR, STOCSY, Suksamrarn et al., 2008
(3*S*)-1,7-diphenyl-(6*E*)-6-hepten-3-ol	DA3	1a 1b 2 3 4 5a 5b 6 7 aromatics-H	2.60-2.65 (*m*) 2.75-2.82 (*m*) 1.66-1.79 (*m*) 3.56-3.61 (*m*) 1.55-1.66 (*m*) 2.20-2.29 (*m*) 2.29-2.38 (*m*) 6.22 (*dt*) 6.39 (*d*) 7.11-7.32 (*m*)	1D, 2D NMR, STOCSY, Suksamrarn et al., 2008
1-(4hydroxyphenyl)-7-phenyl-(4*E*,6*E*)-4,6-heptadien-3-one	DA4	1 2 4 5 2′,6′ 3′,5′ 2″,6″ 3″-5″	2.87-2.90 (*m*) 2.81-2.85 (*m*) 6.31 (*d*) 7.41 (*dd*) 7.00 (*d*) 6.69 (*d*) 7.52 (*dd*) 7.26-7.37 (*m*)	1D, 2D NMR, STOCSY, Suksamrarn et al., 2008
1,7-Diphenyl-(4*E*,6*E*)-4,6-heptadiene-3-one	DA5	1, 2 4 7 2″,6″	2.93 (*m*) 6.32 (*d*) 7.02 (*d*) 7.52 (*dd*)	1D, 2D NMR, STOCSY, Suksamrarn et al., 2008
Sucrose	Suc	Glu-1 Fructose-3′	5.39 (*d*) 4.09 (*d*)	1D, STOCSY, BMRB, HMDB, Navarro et al., 2020, Gauthier et al., 2023

Table 1. (*continued*)

Metabolites	Abbreviation	Position	Chemical Shift (Multiplicity)	Assignment Method and Ref.
α-Glucose	α-Glu	1	5.10 (*d*)	1D, STOCSY, BMRB, HMDB, Navarro et al., 2020, Gauthier et al., 2023
β-Glucose	β-Glu	1	4.47 (*d*)	1D, STOCSY, BMRB, HMDB, Navarro et al., 2020, Gauthier et al., 2023
Fatty acid	FAs	Terminal methyl group Backbone-CH_2 Allylic-H Olefinic-H	0.87-0.90 (*m*) 1.24-1.36 (*m*) 2.00-2.06 (*m*) 5.31-5.33 (*m*)	1D, 2D NMR, STOCSY, BMRB, HMDB, Abreu et al., 2022, Knothe et al., 2004
Formate	Formate	1-H	8.54 (*s*)	1D, 2D NMR, 1D, STOCSY, BMRB, HMDB, Montoya-García et al., 2023
DA-OAc region	DA-OAc region	CH_3COO-	1.95-2.00	1D, STOCSY, Suksamrarn et al., 2008

s, singlet; *d*, doublet; *dd*, double of doublet; *dt*, doublet of triplets; *br d*, broad doublet; *q*, quartet and *m*, multiplet
*Overlapping signals

geographical discrimination of samples. Model validation was performed through a significant CV-ANOVA *p*-value of 1.29×10^{-3}. Furthermore, potential overfitting of the model was ruled out by permutation tests (200 permutations) with satisfactory R^2 and Q^2 values for each original class (Table 3). The permuted models of R^2 and Q^2 are presented on the left side, while the original R^2 and Q^2 models are displayed on the right side. The results indicated that all permutation tests confirmed the validity of the OPLS-DA model, as evidenced by the R^2 and Q^2 values in all permuted models being lower than those in the original R^2 and Q^2 models (Figure S6). Additionally, the intersection points of the Q^2 regression lines for all OPLS-DA models had a negative intercept (Table 3). This suggests that OPLS-DA can serve as a powerful statistical tool for classifying *C. comosa* from different origins with a high level of validity. Furthermore, in terms of geographical authentication using OPLS-DA, a correct classification rate of 100% was achieved for the samples, highlighting the excellent predictive capability of the OPLS-DA model for distinguishing the origin of *C. comosa* (Table 3). ROC curves were also examined for the tested classes of geographical origin (KSN, CMI, SKN, NPM, NPT, PNB), and AUC values of 1 were obtained for all the investigated classes, indicating the strong performance of these classification models (Figure 4B, Table 3). In Figure 4A, the OPLS-DA scores plot clearly illustrates the differentiation of the six *C. comosa* provinces, suggesting the unique metabolite profiles of each. SKN *C. comosa* samples were predominantly clustered in the first quadrant, CMI *C. comosa* samples in the second quadrant, NPT *C. comosa* samples in the third quadrant, while KSN and NPM *C. comosa* samples were primarily in the fourth quadrant, and PNB *C. comosa* samples were located near the center of the quadrant.

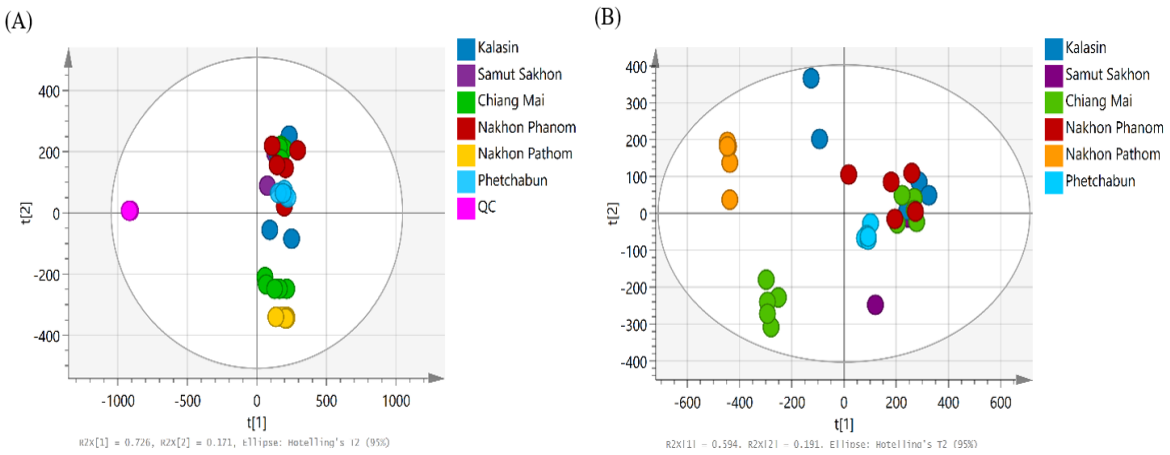

Figure 3. Principal component analysis (PCA) score plot of ^{1}H-NMR data for discriminating the six different origin of *C. comosa* samples from Thailand, including Kalasin, Nakhon Phanom, Samut Sakhon, Nakhon Pathom, Chiang Mai, Phetchabun province, and QC samples.

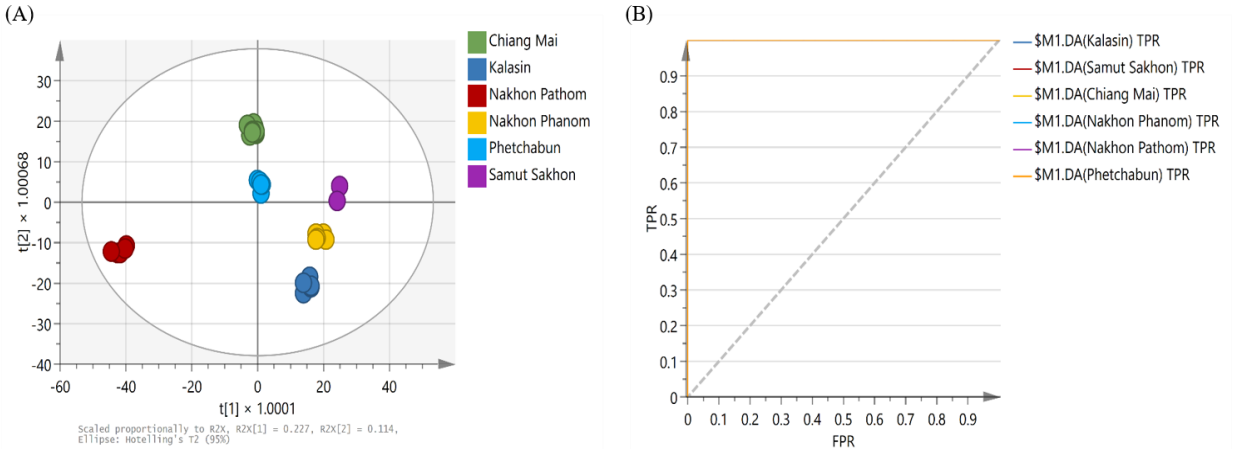

Figure 4. Orthogonal partial least squares-discriminant analysis (OPLS-DA) score plot of ^{1}H-NMR data for discriminating the six different origins of *C. comosa* samples from Thailand, including Chiang Mai, Kalasin, Nakhon Pathom, Nakhon Phanom, Phetchabun, and Samut Sakhon province (A); ROC curves for each OPLS-DA model (B).

Table 2. Selection of the optimal OPLS-DA model based on various normalization and scaling methods for discriminating the geographical origin of *C. comosa*.

Group No.	Normalization method	Scaling method	Component number[a]	R^2Y (cum)	Q^2 (cum)
1	Standardized area	UV	5+9	0.985	0.792
2		Par	5+1	0.767	0.556
3	Total area	UV	5+7	0.973	0.782
4		Par	5+3	0.898	0.716

UV, unit variance; Par, pareto

[a]No. of components expressed as "No. of predictive components + No. of orthogonal components"

3.2.2 Potential metabolites differentiating geographical origins

As a result, from the initially detected features in *C. comosa* fingerprints, 12 compounds were identified, including DA1-DA5, sucrose, α-glucose, β-glucose, formate, fatty acid, and two acetate groups of diarylheptanoids (DA-OAc). These metabolites were putatively identified as discriminant metabolites capable of differentiating samples from different geographical origins. The metabolites selected as markers, along with VIP scores > 1 and *p*-values < 0.05 (ANOVA with post-hoc Tukey's HSD test and FDR analysis < 0.05), were considered potential markers for distinguishing the origin of *C. comosa* (Table 4). The normalized abundance of selected metabolites, according to the assessed origin, is summarized in box plots (Figure 5).

4. Discussion

For geographical differentiation of *C. comosa* samples, several metabolites were considered. Among these, DA1 (*p*-value of 5.40×10^{-7}, VIP score = 1.12), DA3 (*p*-value of 4.43×10^{-7}, VIP score = 1.00), DA-OAc at 1.98999 ppm (*p*-value of 6.99×10^{-4}, VIP score = 1.31), and Fatty Acids (FAs) (*p*-value of 7.18×10^{-6}, VIP score = 1.01) exhibited significantly higher levels in *C. comosa* samples from NPT, while significantly lower levels were found in samples from SKN (Figure 5) compared to the other investigated origins. On the other hand, DA2 (*p*-value of 1.10×10^{-4}, VIP score = 1.11) and Sucrose (*p*-value of 7.18×10^{-5}, VIP score = 1.29) were found at significantly higher levels in *C. comosa* samples from SKN, with significantly lower levels in samples from NPT (Figure 5) compared to the other investigated origins. Additionally, DA4 (*p*-value of 2.06×10^{-5}, VIP score = 1.64) was found at significantly higher levels in *C. comosa* samples from NPM, with significantly lower levels in samples from SKN (Figure 5) compared to the other investigated origins. Lastly, DA5 (*p*-value of 1.81×10^{-3}, VIP score = 1.16) was found at significantly higher levels in *C. comosa* samples from PNB, with significantly lower levels in samples from KSN (Figure 5) compared to the other investigated origins. DA-OAc at 1.96999 ppm (*p*-value of 6.00×10^{-4}, VIP score = 1.14) was found at significantly higher levels in *C. comosa* samples from KSN, while significantly lower levels were found in samples from SKN (Figure 5) compared to the other investigated origins. α-glucose (*p*-value of 1.02×10^{-3}, VIP score = 1.14) was found at significantly higher levels in *C. comosa* samples from NPT, with significantly lower levels in samples from NPM (Figure 5) compared to the other investigated origins. Similarly, β-glucose (*p*-value of 2.17×10^{-3}, VIP score = 1.45) was found at significantly higher levels in *C. comosa* samples from SKN, with significantly lower levels in samples from NPM (Figure 5) compared to the other investigated origins. Formate (*p*-value of 3.06×10^{-8}, VIP score = 1.42) was found at significantly higher levels in *C. comosa* samples from PNB, while significantly lower levels were found in samples from NPT (Figure 5) compared to the other investigated origins. Furthermore, it was discovered that *C. comosa* sourced from CMI province exhibited the second-highest levels of DA4, DA5, sucrose, and formate among all provinces examined. These results highlight the significant influence of geographical origins, particularly the province of production, on the diarylheptanoid phytoestrogen composition of *C. comosa*.

Table 3. Performance and validation results of OPLS-DA models for the geographical discrimination of *C. comosa* samples.

Model parameters	Geographical origin
R^2X(cum)	0.962
R^2Y(cum)	0.985
Q^2(cum)	0.792
NC[b]	5+9
NxK	32x1837
Permutation tests (R^2 ; Q^2)	0.860; -0.837 (KSN) 0.846; -0.794 (CMI) 0.874; -0.688 (SKN) 0.865; -0.753 (NPM) 0.853; -0.856 (NPT) 0.850; -0.796 (PNB)
CV-ANOVA *p*-value	1.29×10^{-3}
ROC[c]	1 (KSN) 1 (CMI) 1 (SKN) 1 (NPM) 1 (NPT) 1 (PNB)
Fisher's probability	1.4×10^{-19}
CCR (%)[d]	100

[b] NC: No. of components expressed as "No. of predictive components + No. of orthogonal components"
N: No. of observations
K: No. of variables
[c] ROC: Receiver operating characteristic curve findings indicated as areas under the ROC curves (AUC)
[d] CCR: Correct classification rate of prediction samples indicated as a percentage (%)

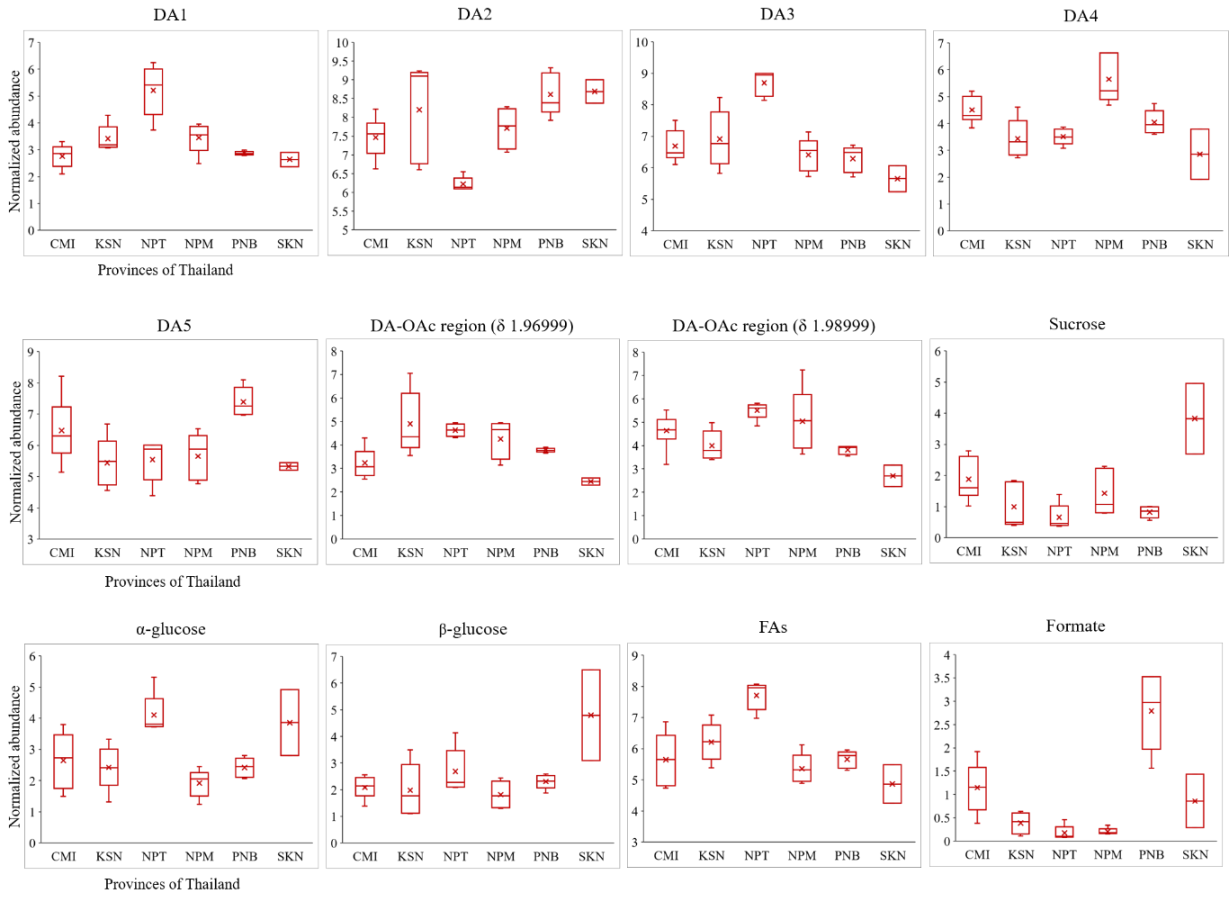

Figure 5. Box and whisker plots showing normalized abundance of metabolites (DA1-DA5 and other metabolites) under different geographical origins of *C. comosa* rhizome; Kalasin (KSN), Nakhon Phanom (NPM), Samut Sakhon (SKN), Nakhon Pathom (NPT), Phetchabun (PNB), and Chiang Mai (CMI).

Table 4. One-way ANOVA followed by post-hoc analysis, showed twelve metabolites under geographical origin with significance of the metabolites change.

Metabolites	Chemical shift (ppm)	*p*-value	FDR	VIP score	Tukey's HSD
DA1	2.51499	5.40×10^{-7}	9.71×10^{-6}	1.12	NPT-CMI; NPT-KSN; NPM-NPT; PNB-NPT; SKN-NPT
DA2	6.52997	1.10×10^{-4}	5.81×10^{-4}	1.11	NPT-CMI; NPT-KSN; NPM-NPT; PNB-NPT; SKN-NPT
DA3	2.75999	4.43×10^{-7}	6.39×10^{-6}	1.00	NPT-CMI; NPT-KSN; NPM-NPT; PNB-NPT; SKN-NPT
DA4	6.28497	2.06×10^{-5}	1.33×10^{-4}	1.64	NPM-CMI; SKN-CMI; NPM-KSN; NPM-NPT; PNB-NPM; SKN-NPM
DA5	2.93999	1.81×10^{-3}	4.88×10^{-3}	1.16	PNB-KSN; PNB-NPT; PNB-NPM; SKN-PNB
DA-OAc region	1.96999	6.00×10^{-4}	2.05×10^{-3}	1.14	KSN-CMI; NPT-CMI; SKN-KSN; SKN-NPT
DA-OAc region	1.98999	6.99×10^{-4}	2.30×10^{-3}	1.31	SKN-CMI; NPT-KSN; PNB-NPT; SKN-NPT; SKN-NPM
Sucrose	5.39497	7.18×10^{-5}	4.16×10^{-4}	1.29	NPT-CMI; SKN-CMI; SKN-KSN; SKN-NPT; SKN-NPM; SKN-PNB
α-glucose	5.09998	1.02×10^{-3}	3.09×10^{-3}	1.14	NPT-CMI; NPT-KSN; NPM-NPT; PNB-NPT; SKN-NPM
β-glucose	4.45998	2.17×10^{-3}	5.57×10^{-3}	1.45	SKN-CMI; SKN-KSN; SKN-NPT; SKN-NPM; SKN-PNB
FAs	0.87499	7.18×10^{-6}	6.81×10^{-5}	1.01	NPT-CMI; NPT-KSN; NPM-NPT; PNB-NPT; SKN-NPT
Formate	8.54496	3.06×10^{-8}	1.4×10^{-6}	1.42	NPT-CMI; NPM-CMI; PNB-CMI; PNB-KSN; PNB-NPT; PNB-NPM; SKN-PNB

This study analyzed *C. comosa* rhizome samples from various locations in Thailand, including KSN, NPM, SKN, NPT, PNB, and CMI using metabolomics technology based on NMR combined with chemometrics (PCA and OPLS-DA) analysis. The metabolic differences between the different origins of *C. comosa* samples were significant and could be identified by OPLS-DA models. The ^{1}H-NMR analysis revealed twelve metabolites, including DA1-DA5, sucrose, α-glucose, β-glucose, formate, fatty acid, and two DA-OAc compounds, which were responsible as potential metabolites differentiating geographical origins. Additionally, a higher normalized abundance of DA1-DA5 was found in *C. comosa* from NPT, PNB, SKN, and NPM compared to CMI and KSN. The analysis of different metabolites has, for the first time, highlighted the potential of applying a ^{1}H-NMR metabolomics approach to assess the metabolic changes in *C. comosa* samples resulting from cultivation in different areas. Our findings confirm that the major diarylheptanoid composition of *C. comosa* is highly dependent on the region, i.e. the environmental conditions related to the cultivation area, resulting in distinctive metabolic profiles. In conclusion, this study demonstrated a successful method to differentiate *C. comosa* from different sources. The method could lead to the establishment of discrimination biomarkers for the origin for the *C. comosa* and may provide a quality control and authenticating method for the industry.

Acknowledgements

This work was supported by the Center of Excellence for Innovation in Chemistry (PERCH-CIC) and Facility at Department of Chemistry, Faculty of Science, Ramkhamhaeng University.

Appendix. Supporting information

Supplementary data to this article can be found online at https://drive.google.com/file/d/ 1iv6UmF4Oi4_NA12sRaHQ65VUutgJAg39/view?usp=sharing

References

Abreu AC, Mora S, Tristán AI, Martín GE, Prados PA, Moreno M, and Fernández I (2022) NMR-based Metabolomics and Fatty Acid Profiles to Unravel Biomarkers in Preclinical Animal Models of Compulsive Behavior. J. Proteome Res. 21(3), 612-622

Chao J, Dai Y, Cheng HY, Lam W, Cheng Y C, Li K, Peng WH, Pao LH, Hsieh M T, Qin XM, and Lee MS (2017) Improving the concentrations of the active components in the herbal tea ingredient, Uraria crinita: The effect of post-harvest oven-drying processing. Scientific reports, 7(1), 38763

Danezis GP, Tsagkaris AS, Brusic V, and Georgiou CA (2016) Food authentication: state of the artprospects." Curr. Opin. Food Sci., 10, 22-31

DeFilipps RA, and Krupnick GA (2018) The medicinal plants of Myanmar. PhytoKeys, 102, 1-341

Dona AC, Kyriakides M, Scott F, Shephard E A, Varshavi D, Veselkov K, and Everett JR (2016) A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. Comput. Struct. Biotechnol. J, 14, 135-153

Gauthier JR, Burns D, Sheng J, and D'eon JC (2022) Exploring the composition and authenticity of honey and syrup samples using quantitative NMR spectroscopy and principal component analysis in an upper-year undergraduate analytical environmental course. J. Chem. Educ., 100(1), 161-169

Hernández GCJ, Villa-Ruano N, Zepeda VLG, Hernández FAD, Ramirez EK, Zamudio LS, Hidalgo MD, and Becerra ME (2021) Bean cultivars (*Phaseolus vulgaris* L.) under the spotlight of NMR metabolomics. Int. Food Res., 150, 110805

Keeratinijakal V, Kladmook M, and Laosatit K (2010) Identification and characterization of *Curcuma comosa* Roxb., phytoestrogens-producing plant, using AFLP markers and morphological characteristics. J. Med. Plant Res., 4(24), 2651-2657

Kim T J, Park JG, Kim HY, Ha SH, Lee B, Park SU, Seo WD, and Kim JK (2020) Metabolite profiling and chemometric study for the discrimination analyses of geographic origin of perilla (*Perilla frutescens*) and sesame (*Sesamum indicum*) seeds. Foods, *9*(8), 989

Knothe G, and Kenar JA (2004) Determination of the fatty acid profile by 1H-NMR spectroscopy. Eur. J. Lipid Sci. Technol., 106(2), 88-96

Lodge S, Nitschke P, Kimhofer T, Wist J, Bong SH, Loo RL, Masuda R, Begum S, Richards T, Lindon JC, and Bermel W (2021) Diffusion and relaxation edited proton NMR spectroscopy of plasma reveals a high-fidelity supramolecular biomarker signature of SARS-CoV-2 infection. Anal. Chem., 93(8), 3976-3986

Macready AL, Hieke S, Klimczuk KM, Szumiał S, Vranken L, and Grunert KG (2020) Consumer trust in the food value chain and its impact on consumer confidence: A model for assessing consumer trust and evidence from a 5-country study in Europe. Food Policy, 92, 101880

Montoya GCO, García MR, Magdaleno VJJ, Volke HVH, Villa RN, Zepeda VLG, and Becerra ME (2023) NMR-based metabolomics to determine the fluctuation of metabolites in hydroponic purslane crops at different harvesting times. Int. Food Res., 166, 112489

Navarro Y, Soengas R, Iglesias MJ, and Ortiz FL (2020) Use of NMR for the analysis and quantification of the sugar composition in fresh and store-bought

fruit juices. J. Chem. Educ., 97(3), 831-837

Petchprayoon C, Charupant K, Krabesri S, Kitisripanya T, Aneklaphakij C, Bunsupa S, Anantachoke N, Satitpatipan V, Chatsumpun N, and Sithisarn P (2022) Pharmacognostic characteristics, physical properties and chromatographic fingerprints of *Curcuma comosa* Roxb. rhizome samples collected in Thailand. Agr. Nat. Resour., 56(4), 781-790

Piyachaturawat P, Charoenpiboonsin J, Toskulkao C, and Suksamrarn A (1999) Reduction of plasma cholesterol by *Curcuma comosa* extract in hypercholesterolaemic hamsters. J. Ethnopharmacol. 66(2), 199–204

Piyachaturawat P, Ercharuporn S, and Suksamrarn A (1995) Uterotrophic effect of *Curcuma comosa* in rats. Int. J. Pharmacogn., 33(4), 334-338

Pontes JGM., Brasil AJM, Cruz GC, de Souza RN, and Tasic L (2017) NMR-based metabolomics strategies: plants, animals and humans. Anal. Methods, 9(7), 1078-1096

Rivera P, Romero GR, and Frenich A (2021) Application of an innovative metabolomics approach to discriminate geographical origin and processing of black pepper by untargeted UHPLC-Q-Orbitrap-HRMS analysis and mid-level data fusion. Int. Food Res. 150, 110722

Rocchetti G, Michelini S, Pizzamiglio V, Masoero F, and Lucini L (2021) A combined metabolomics and peptidomics approach to discriminate anomalous rind inclusion levels in Parmigiano Reggiano PDO grated hard cheese from different ripening stages. Int. Food Res.149, 110654

Singh P, Kumar P, Pande V, Kumar V, and Dhiman RC (2022) Untargeted metabolomics-based response analysis of temperature and insecticide exposure in Aedes aegypti. Scientific Reports, 12(1), 2066

Sobolev AP, Thomas F, Donarski J, Ingallina C, Circi S, Marincola FC, Capitani D, and Mannina L (2019) Use of NMR applications to tackle future food fraud issues. Trends Food Sci. Technol., 91, 347-353

Sodsai A, Piyachaturawat P, Sophasan S, Suksamrarn A, and Vongsakul M (2007) Suppression by *Curcuma comosa* Roxb. of pro-inflammatory cytokine secretion in phorbol-12-myristate-13-acetate stimulated human mononuclear cells. Int. Immunopharmacol., 7(4), 524-531

Suksamrarn A, Ponglikitmongkol M, Wongkrajang K, Chindaduang A, Kittidanairak S, Jankam A, Yingyongnarongkul BE, Kittipanumat N, Chokchaisiri R, Khetkam P, and Piyachaturawat P (2008) Diarylheptanoids, new phytoestrogens from the rhizomes of *Curcuma comosa*: Isolation, chemical modification and estrogenic activity evaluation. Bioorg. Med. Chem., *16*(14), 6891-6902

Suksen K, Charaslertrangsi T, Noonin C, Jariyawat S, Devakul W. Na Ayutthaya A, Suksamrarn, Tuchinda P, and Piyachaturawat P (2016) Protective effect of diarylheptanoids from *Curcuma comosa* on primary rat hepatocytes against t-butyl hydro-peroxide-induced toxicity. Pharm. Biol., 54(5), 853-862

Tabboon P, Tuntiyasawasdikul S, and Sripanidkulchai B (2019) Quality and stability assessment of commercial products containing phytoestrogen diaryheptanoids from *Curcuma comosa*. Ind. Crops Prod., 134, 216-224

Tantikanlayaporn D, Robinson LJ, Suksamrarn A, Piyachaturawat P, and Blair HC (2013a) A diarylheptanoid

phytoestrogen from Curcuma comosa, 1,7- diphenyl-4,6-heptadien-3-ol, accelerates human osteoblast proliferation and differentiation. Phytomedicine, 20(8-9), 676-82

Tantikanlayaporn D, Wichit P, Weerachayaphorn J, Chairoungdua A, Chuncharunee A, Suksamrarn A, and Piyachaturawat P (2013b). Bone sparing effect of a novel phytoestrogen diarylheptanoid from *Curcuma comosa* Roxb. in ovariectomized rats. PloS one, 8(11), 78739

Wang Z, Chen X, Liu Q, Zhang L, Liu S, Su Y, Ren Y, Yuan C (2023) Untargeted metabolomics analysis based on LC-IM-QTOF-MS for discriminating geographical origin and vintage of Chinese red wine. Int. Food Res., 165, 112547

Winuthayanon W, Suksen K, Boonchird C, Chuncharunee A, Ponglikitmongkol M, Suksamrarn A, and Piyachaturawat P (2009) Estrogenic activity of diarylheptanoids from *Curcuma comosa* Roxb. requires metabolic activation. J. Agric. Food Chem., 57(3), 840-845

Yingngam B, Brantner A, Jinarat D, Kaewamatawong R, Rungseevijitprapa W, Suksamrarn A, Piyachaturawat P, Chokchaisiri R, (2018) Determination of the marker diarylheptanoid phytoestrogens in *Curcuma comosa* rhizomes and selected herbal medicinal products by HPLC-DAD. Chem. Pharm. Bull., 66(1), 65-70

Yousf S, Sardesai DM, Mathew AB, Khandelwal R, Acharya JD, Sharma S, Chugh J (2019) Metabolic signatures suggest o-phosphocholine to UDP-N-acetylglucosamine ratio as a potential biomarker for high-glucose and/or palmitate exposure in pancreatic β-cells. Metabolomics, 15, 1-16

Zhou Y, Kim SY, and Lee JS (2021) Discrimination of the geographical origin of soybeans using NMR-based metabolomics. Foods, 10, 1-16