6-1-2021

Linear discriminant analysis based on gas chromatographic measurements for geographical prediction of USA medical domestic cannabis

Ramia Z. Al Bakain
The University of Jordan

Yahya S. Al-Degs
Hashemite University

James V. Cizdziel
University of Mississippi

Mahmoud A. Elsohly
University of Mississippi School of Pharmacy

Recommended Citation

Al Bakain, R. Z., Al-Degs, Y. S., Cizdziel, J. V., & Elsohly, M. A. (2021). Linear discriminant analysis based on gas chromatographic measurements for geographical prediction of USA medical domestic cannabis. Acta Chromatographica, 33(2), 179–187. https://doi.org/10.1556/1326.2020.00782

AKADÉMIAI KIADÓ

Acta Chromatographica

33 (2021) 2, 179–187

DOI:
10.1556/1326.2020.00782

ORIGINAL RESEARCH PAPER

Linear discriminant analysis based on gas chromatographic measurements for geographical prediction of USA medical domestic cannabis

RAMIA Z. AL BAKAIN[1*], YAHYA S. AL-DEGS[2], JAMES V. CIZDZIEL[3] and MAHMOUD A. ELSOHLY[4,5]

[1] Department of Chemistry, School of Science, The University of Jordan, P.O. Box 11942, Amman, Jordan

[2] Chemistry Department, The Hashemite University, P.O. Box 150459, Zarqa, Jordan

[3] Department of Chemistry and Biochemistry, University of Mississippi, University, MS, 38677, USA

[4] National Center for Natural Products Research, University, MS, 38677-1848, USA

[5] Department of Pharmaceutics and Drug Delivery, School of Pharmacy, University of Mississippi, University, MS, 38677, USA

Received: April 06, 2020 • Accepted: April 24, 2020
Published online: June 5, 2020

ABSTRACT

Fifty four domestically produced cannabis samples obtained from different USA states were quantitatively assayed by GC–FID to detect 22 active components: 15 terpenoids and 7 cannabinoids. The profiles of the selected compounds were used as inputs for samples grouping to their geographical origins and for building a geographical prediction model using Linear Discriminant Analysis. The proposed sample extraction and chromatographic separation was satisfactory to select 22 active ingredients with a wide analytical range between 5.0 and 1,000 μg/mL. Analysis of GC-profiles by Principle Component Analysis retained three significant variables for grouping job (Δ^9-THC, CBN, and CBC) and the modest discrimination of samples based on their geographical origin was reported. PCA was able to separate many samples of Oregon and Vermont while a mixed classification was observed for the rest of samples. By using LDA as a supervised classification method, excellent separation of cannabis samples was attained leading to a classification of new samples not being included in the model. Using two principal components and LDA with GC–FID profiles correctly predict the geographical of 100% Washington cannabis, 86% of both Oregon and Vermont samples, and finally, 71% of Ohio samples.

KEYWORDS

geographical prediction, medical cannabis, linear discriminant analysis, gas chromatography, principal component analysis

INTRODUCTION

Cannabis and other by-products are gaining popularity due to their reported medicinal applications [1, 2]. The first medicinal usage originated in the Middle East and Asia was back in the 6th century BC [2, 3]. Later, cannabis was introduced to Western medicine during the 19th century [1, 2]. Interestingly, cannabis is the most widely cultivated and yet publicly illegal drug worldwide [1]. Chemically, cannabis is made up of complex blend of constituents including cannabinoids, terpenoids, flavonoids, carbohydrates, and hydrocarbons [3, 4]. Among the reported constituents of cannabis, cannabinoids and terpenoids are the most dominant with more than 200 compounds isolated from both classes [5–10]. Moreover, the

*Corresponding author.
Tel.: +962 6 535 5000 ext. 22138;
Fax: +962 65300253
E-mail: r.bakain@ju.edu.jo

AKJournals

total number of detected or isolated constituents in cannabis (*Cannabis sativa* L.) has steadily increased over the last decades [7–11].

Many advanced chromatographic methods were adopted for the analysis of as many active ingredients as possible in cannabis samples [11–13]. Among adopted methods, Gas Chromatography–Flame Ionization Detector (GC–FID), and Gas Chromatography–Mass Spectrometry (GC–MS) were the most common methods, and this is likely attributed to the volatility of the active components in cannabis [10, 11]. Beside GC, classical Liquid Chromatography (HPLC) and Supercritical Fluid Chromatography (SFC) have been used for cannabis analysis [3]. In our laboratory, more than 50 solutes, both identified and unidentified, were separated by GC–FID in a reasonable run time (around 52 min) [12]. The broad and comprehensive chromatographic responses of a variety of chemical classes of cannabis constituents have prompted many investigators to study the clustering of cannabis samples based on their chemical profiles (i.e., according to different cultivars and geographical origins) [3, 12, 14]. Chemical profile monitoring is an essential aspect in samples' classification and hence consumer's protection, because blending of cannabis of different varieties may not be declared. As mentioned earlier, conventional analytical methods that achieve maximum cannabis identification mainly involve chromatographic detection and quantitation of the different components (i.e., fingerprint) [3, 6, 14].

Multivariate analysis including Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) were mostly employed in cannabis clustering [12, 14]. Many published reports indicate that the characteristic chemical classes for cannabis clustering are the cannabinoids and the terpenoids [3, 12, 14]. Other organic constituents such as flavonoids, carbohydrates, and hydrocarbons were of limited applicability toward cannabis clustering. Among the cannabinoids, Δ^9-THC is the main compound responsible for the psychoactive nature of cannabis, while CBN can reflect the age and storage condition of a sample [15]. CBG was the first cannabinoid identified and the first biogenic cannabinoid formed in cannabis [15]. On the other hand, the typical scent of cannabis is attributed to specific terpenoids including myrcene and limonene [12, 15]. Jin et al. proposed a chemometric classification of Canadian cannabis samples based on LC and GC measurements using 10 cannabinoids and 14 terpenoids [14]. Al Bakain and co-workers outlined a convenient clustering of 23 USA-cannabis samples using GC–FID measurements and PCA [12]. The results indicated fair separation of cannabis samples from California and Oregon, where CBN/Δ^9-THC ratio was the most dominant variable for clustering [12]. In an interesting investigation, the application of ultra-high-performance SFC was reported to detect nine cannabinoids CBD, Δ^8-THC, THCV, Δ^9-THC, CBN, CBG, THCA-A, CBDA, and CBGA in several USA-domestic cannabis samples but was not used for clustering purpose [3].

Although researchers have recognized the practicality of using cannabinoids and terpenoids as informative indices toward cannabis clustering, to the best of our knowledge, there is no published work aimed at predicting the classification of USA-domestic cannabis samples based on GC-chromatographic profiles of cannabinoids and terpenoids using Linear Discriminant Analysis (LDA). In this work, quantitative classification of 54 collected cannabis samples according to geographical origins based on GC-profiles of 22 components were carried out with the aid of PCA and LDA. Then, LDA model was used to predict the possible origin of unknown medical cannabis.

EXPERIMENTAL

Chemicals and reagents

Seven cannabinoids (Δ^8-THC, Δ^9-THC, THCV, CBC, CBG, CBL and CBN) and fifteen terpenoids (α-pinene, β-pinene, α-humulene, β-caryphyllene, α-terpinol, myrcene, limonene, caryophylleneoxide, fenchol, linalool, carveol, terpinolene, cineol, guaiol, and α-bisabolol) were purchased from Sigma-Aldrich® (St. Louis, MO) and used as markers for samples classification. The chemical structures of the isolated organics are provided in Table 1. Phenanthrene (>99% purity) supplied from Sigma-Aldrich® (St. Louis, MO) was used as internal standard (IS). Ultra pure water (18 MΩ cm^{-1}) generated by Milli-Q Plus water purification system (Millipore, Billerica, MA) was used to prepare aqueous solutions and for dilution purposes.

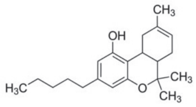
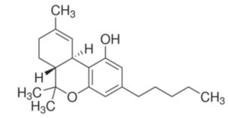
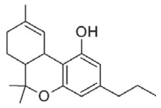
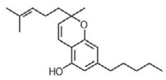
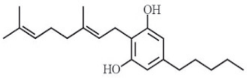
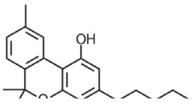
Collection of cannabis samples

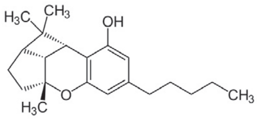
Fifty-four cannabis samples seized by law enforcement from the states of Washington, Vermont, Ohio, and Oregon were used in this study. The samples were split as following: 26 (five samples from Washington, six from Vermont, six from Ohio, and nine from Oregon) and those samples were used to train the LDA model. Another 28 samples (seven from each state) were used to test the efficiency of LDA model for unknown sample classification. The samples were obtained from seizures by the The Drug Enforcement Administration (DEA) and/or state agents representing the Domestic Cannabis Eradication/Suppression Program (DCE/SP) and submitted to the National Institute on Drug Abuse (NIDA) for analysis under a national potency monitoring program. The samples were received in sealed plastic bags and stored in temperature controlled vault in the Coy Waller Complex at the University of Mississippi until analyzed.

Extraction of active ingredients and GC chromatographic separation

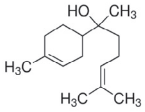
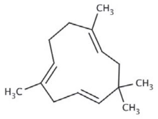
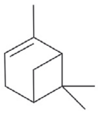
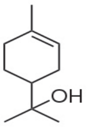
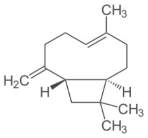
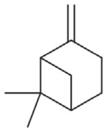
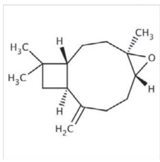
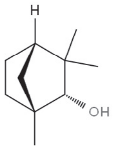
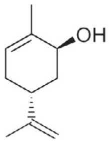
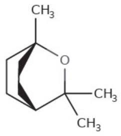
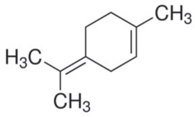
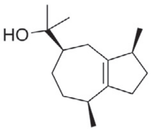
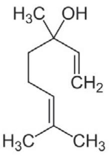
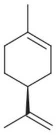
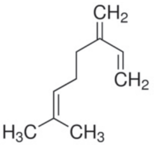
Initially, the samples were mixed and ground to get a uniform and homogenized matrix. One hundred milligrams plant samples were each extracted using 3.0 mL methanol–chloroform (9:1 v/v) containing the internal standard (0.2 mg/mL phenanthrene). The mixtures were then sonicated for 15 min at room temperature and finally centrifuged. Phenanthrene was used as a retention time marker and as an internal standard (IS). All samples were analyzed in duplicate extractions and injections, and the average results of the

Table 1. Structural formula of assayed cannabinoids and terpenoids

 Δ^8-THC/Δ^8-Tetrahydrocannabinol	 Δ^9-THC /Δ^9-Tetrahydrocannabinol	 THCV/ Tetrahydrocannabivarin
 CBC/Cannabichromene	 CBG/Cannabigerol	 CBN/Cannabinol



CBL/ Cannabicyclol

Terpenoids		
 α-bisabolol	 α -humulene	 α – pinene
 α –terpinol	 β – caryphyllene	 β –pinene
 Caryophylleneoxide	 Fenchol	 Carveol
 Cineol	 Terpinolene	 Guaiol
 Linalool	 Limonene	 Myrcene

two injections were registered. Before GC analysis of samples, standard solutions of cannabinoids and terpenoids (0.1 mg/mL in methanol) were injected for qualitative identification. Splitless mode injections on the GC were used using Agilent GC 6890. The column was (DB-5) 30 m length, 0.25 mm internal diameter with a film thickness of 0.25 μm (J&W Scientific Inc., Folsom, CA). Helium was used as the carrier gas at a flow rate of 25 cm/s. Detection was achieved using a Flame Ionization Detector (FID) and the injection volume was 1.0 μL. The run-to-run repeatability (n = 12) and intraday reproducibility (n = 10) of peak area and retention time were measured, and RSD values were less than 1% in all cases. More experimental information is available in the literature [12].

Multivariate data analysis for samples classification

A data matrix was created from the GC-profiles with rows representing cannabis samples from different states while columns containing the contents of the analyzed components. Three quantities were created to run PCA and LDA; $\mathbf{X}_{26\text{samples}\times 22\text{solutes}}$, $\mathbf{X}_{28\text{samples}\times 22\text{solutes}}$, and $\mathbf{y}_{26\times 1}$. The main matrix $\mathbf{X}_{26\text{samples}\times 22\text{solutes}}$ containing the chromatographic profiles of the 26 samples, received from the four states, was analyzed by PCA. PCA was applied to mean-centered data to reduce the dimensionality of the data and to gain visual insight into natural grouping of cannabis samples according to the geographic origin. Moreover, the outputs of PCA are helpful to indicate the best component (solutes) contributing to samples grouping. PCA was created using two factors (to build the model) as estimated by leave-one-out cross-validation methodology [16, 17]. For better sample separation, LDA assumes class data following a multivariate normal distribution and seeks to maximize the ratio among class variance and minimize the ratio within class variance, which ensures maximum separation among groups [18, 19]. Initially, discriminant classifiers are created using $\mathbf{X}_{26\text{samples}\times 22\text{solutes}}$ and $\mathbf{y}_{26\times 1}$ (containing the class membership of states) with the optimum number of factors (i.e., the number needed to create the LDA model). Classification of the new samples was achieved by subjecting directly the $\mathbf{X}_{28\text{samples}\times 22\text{solutes}}$ matrix (i.e., new samples from the four states) to LDA classifier and then assessing the accuracy of the model. LDA was performed on the first two factors and the classification model was created and validated using the leave-one-out cross-validation methodology [18–20]. PCA was performed using Chemoface 1.61 software [12, 21] under Matlab® (Mathworks, 8.6, USA), while LDA was performed using XLSTAT software (Excel, Microsoft®).

RESULTS AND DISCUSSION

Quantification of cannabinoids and terpenoids in different samples

Simultaneous quantification of various cannabinoids and terpenoids in cannabis by GC has been well documented [15]. Based on the literature methods, GC–FID or MS detection are the most appropriate methods for cannabis analysis [15–22]. Derivatization (usually silylation) is necessary when information about cannabinoid acids is needed [3]. The total cannabinoid content, i.e., the amount of total cannabinoids (neutral and decarboxylation of the acidic forms) is also determined using GC–FID [1, 2]. Fig. 1 displays the chromatographic profiles of some cannabis samples analyzed by GC–FID without derivatization (total cannabinoids and terpenoids).

As shown in both chromatograms, the separated components were 15 terpenoids (10–30 min) and seven cannabinoids which appeared over the interval (40–52 min). The late retention of the cannabinoids is attributed to their polar nature as indicated from the phenolic group in cannabinoids (Table 1). Moreover, in all samples, the contents of the cannabinoids were much higher than the terpenoids. Among the cannabinoids, Δ^9-THC was present in high levels (13.25%). In fact, all samples exhibited comparable quantitative GC profile of the fifteen-terpenoids and the seven-cannabinoids as fingerprint, which permit the application of numerical classification methods [3, 12]. The adopted GC–FID methodology presented good linearities (R^2 > 0.99) toward all analytes and over a wide dynamic range 5–1,000 μg/mL. Moreover, low detection limits (1.0–3.0 μg/mL) were reported, which allowed for accurate quantification of cannabinoids and terpenoids in the extracts of cannabis samples. The concentrations (wt%) of the individual components are provided in Table 2.

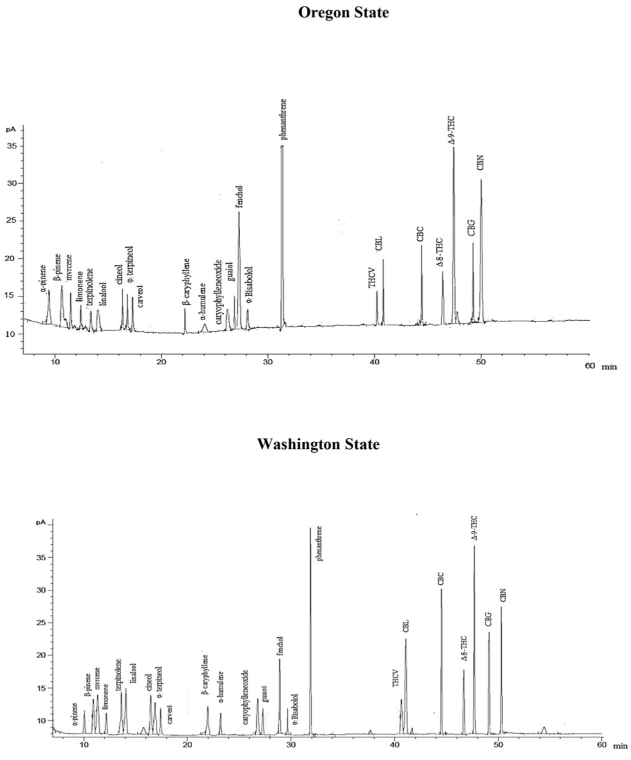


Fig. 1. GC–FID chromatograms of cannabis samples of Washington and Oregon states

Table 2. Chemical profiles (mass%, $n = 3$) of terpenoids and cannabinoids in cannabis samples used for PCA and LDA measurements (Cannabis samples collected from different cultivars sites in four USA states: Washington (WA1–WA5), Vermont (VT1–VT6), Ohio (OH1–OH6), and Oregon (OR1–OR9))

Solute	WA1	WA2	WA3	WA4	WA5	VT1	VT2	VT3	VT4	VT5	VT6	OH1	OH2
α-Bisabolol	0.038	ND	0.076	0.152	0.014	0.017	0.057	0.011	0.217	0.004	0.009	0.223	0.165
α-Humulene	0.008	0.017	0.005	0.009	0.024	0.151	0.021	0.095	0.097	0.033	0.027	0.092	0.121
α-Pinene	0.016	ND	0.008	0.055	0.041	0.392	0.006	ND	0.019	0.065	ND	0.031	ND
α-Terpinol	0.012	ND	0.016	0.026	0.008	0.015	ND	ND	0.015	0.004	0.027	0.004	ND
β-Caryphyllene	0.011	ND	0.662	0.007	0.015	0.556	0.065	0.289	0.345	0.136	0.076	0.322	0.334
β-Pinene	0.023	ND	0.049	0.031	0.075	0.169	ND	0.017	ND	0.034	ND	ND	ND
Caryophylleneoxide	0.029	ND	0.013	0.008	0.018	0.053	0.011	0.012	0.065	0.012	0.024	0.082	0.061
Carveol	0.012	ND	0.007	0.027	0.016	0.021	0.063	0.038	0.024	0.024	0.027	0.019	0.012
Cineol	0.012	ND	0.023	0.006	ND	0.012	ND	ND	0.009	0.004	ND	0.005	ND
Fenchol	0.038	0.051	0.012	0.152	0.014	0.102	0.014	0.076	0.217	0.004	0.009	0.223	0.01
Guaiol	0.029	ND	0.09	0.008	0.018	0.013	0.044	0.082	0.023	0.004	0.013	0.015	0.112
Limonene	0.009	0.007	0.031	0.009	0.011	0.021	0.019	0.013	0.047	0.012	ND	0.011	ND
Linalool	0.029	0.033	0.067	0.029	0.026	0.034	0.041	0.023	0.066	0.019	0.015	0.004	ND
Myrcene	0.023	ND	0.049	0.03	0.075	0.169	ND	0.017	ND	0.033	ND	ND	ND
Terpinolene	0.014	ND	0.067	0.008	0.017	0.034	0.041	0.023	0.066	0.019	0.015	0.023	0.015
Δ^8-THC	0.019	ND	0.021	0.013	0.009	0.034	0.005	0.021	0.017	0.017	ND	0.006	0.012
Δ^9-THC	0.009	ND	12.15	0.007	2.441	11.706	0.026	5.561	7.403	3.392	0.159	0.85	1.801
THCV	0.222	0.006	0.111	0.022	0.012	0.049	0.009	0.041	0.025	0.032	0.008	0.019	0.014
CBC	0.046	ND	0.068	0.038	0.014	0.289	0.175	0.126	0.156	0.015	0.048	0.117	0.117
CBG	0.006	ND	0.232	0.056	0.057	0.256	0.061	0.405	0.736	0.166	0.023	0.051	0.183
CBL	0.014	ND	0.016	0.014	0.012	0.022	0.008	0.006	0.009	ND	0.004	0.012	0.008
CBN	1.539	0.008	0.917	0.313	0.824	0.204	0.159	0.654	0.856	0.166	0.243	0.859	1.297

As shown in Table 2, virtually all the 22 selected components were found in the samples, except for one sample (WA2) which was found to be free of most terpenoids and cannabinoids, especially Δ^9-THC. In general, Δ^9-THC, CBG, and CBN were the three major components found in all samples, while the rest of solutes existed in relatively low contents.

From distribution point of view, the contents of Δ^9-THC, CBG, and CBN exhibited the largest variability among all cannabis samples. Moreover, the profiles of separated components displayed different patterns in terms of the contents of the terpenoids/cannabinoids (Fig. 1). Hence, chemometric classification of samples is expected to be possible. In a similar study, Δ^9-THC, CBG, and CBN were found necessary for cannabis classification [3, 12]. However, both cannabinoids and terpenoids were reported to be necessary for USA-domestic produced cannabis samples using k-means algorithm [3, 12].

Cannabis characterization by PCA

Natural clustering of cannabis produced in different USA states in the space of the most significance principal components can be viewed by PCA. The obtained chromatographic data matrix (26 samples × 22 solutes) was subjected to PCA. Two principal components explaining 92.43% of the total variance in the data were obtained. As shown in Fig. 2A, cannabis samples of different origin were not properly grouped from each other. As shown, six Oregon samples (OR4–OR9) were grouped together but with four samples of Vermont origin (VT1, VT3–VT5). The first PC, which explained 75.43% of the total variance, was mainly contributed by Δ^9-THC (loadings > -0.85). The second PC (17.00% of the total variance) correlated positively with CBN (loading $> +0.8$). Comparing score and loading plots (Fig. 2A and B), the dominant variables that separate a sample(s) from the others can be viewed. The separation of Vermont samples (VT1, VT3–VT5) and Oregon samples (OR4–OR9) would be attributed to the higher contents of Δ^9-THC. However, the higher content of CBN was the reason for isolation of the samples: OR1, WA1, WA3, and VT2. The unique isolation of WA2 was attributed to the absence of many terpenoids and cannabinoids, especially Δ^9-THC. It is rather interesting to see the limited contribution of the 15 terpenoids for samples clustering.

As shown in Fig. 2B, the main contributing variables to the first two PCs were Δ^9-THC/CBN/CBC and the viability of using these cannabinoids as variables to discriminate cannabis samples of different origins was further validated. Unfortunately, cannabis classification was not fully accomplished using only three cannabinoids as indices. The PCA outputs (Fig. 2A) indicated that the scores of VT6, OH1, OR3, OH2 samples mixed together and not isolated according to their origins. Although Δ^9-THC was essential to separate six Oregon samples (OR4–OR9) and four Vermont samples (VT1, VT3–VT5) from the rest of samples, but the model was not capable to widen the distance among groups. The same performance was observed for CBN as it separates two Washington samples but mixed with two samples obtained from Oregon and Vermont states. With a small

Table 2. Continued

OH3	OH4	OH5	OH6	OR1	OR2	OR3	OR4	OR5	OR6	OR7	OR8	OR9
0.105	0.101	0.031	0.059	0.007	0.039	0.043	0.039	0.047	0.008	0.176	0.078	0.068
0.037	0.103	0.111	0.238	0.062	0.038	0.019	0.038	0.055	0.008	0.087	0.119	0.112
0.168	0.068	0.044	0.013	0.173	0.074	0.006	0.071	0.009	0.014	0.046	0.023	0.033
0.011	0.016	0.005	0.011	0.008	0.006	ND	0.006	0.005	0.009	0.007	0.006	0.004
0.122	0.013	0.281	0.595	0.148	0.131	0.043	0.131	0.179	0.261	0.247	0.315	0.359
0.026	0.050	0.019	0.021	0.045	0.036	ND	0.035	0.014	0.028	0.018	0.019	0.044
0.038	0.049	0.051	0.086	0.034	0.044	0.011	0.044	0.031	0.058	0.021	0.029	0.037
0.007	0.005	0.003	0.015	0.034	0.024	0.064	0.024	0.031	0.017	0.006	0.039	0.008
0.008	0.037	0.004	0.019	0.017	0.006	0.005	0.006	0.004	0.013	0.006	0.007	0.022
0.004	0.101	0.031	0.019	0.019	0.039	0.043	0.039	0.008	0.071	0.176	0.005	0.068
0.008	0.019	0.015	0.132	0.062	0.007	0.014	0.007	0.015	0.078	0.014	0.034	0.031
0.009	0.047	ND	0.024	0.033	0.011	ND	0.011	0.014	0.018	0.011	0.015	0.035
0.018	0.053	0.017	0.039	0.031	0.02	0.025	0.021	0.014	0.061	0.022	0.032	0.075
0.026	0.049	0.019	0.021	0.045	0.021	ND	0.02	0.014	0.028	0.018	0.019	0.044
0.018	0.052	0.004	0.039	0.016	0.021	0.025	0.02	0.004	0.061	0.021	0.032	0.044
0.013	0.036	0.039	0.039	0.025	0.01	0.006	0.01	0.043	0.047	0.024	0.022	0.066
2.966	13.254	8.162	9.172	0.124	2.84	0.293	2.84	8.411	10.82	5.889	4.419	13.36
0.023	0.101	0.091	0.111	0.055	0.011	0.023	0.011	0.078	0.052	0.032	0.061	0.077
0.175	0.289	0.294	0.399	0.531	0.053	0.123	0.053	0.214	0.017	0.14	0.132	0.225
0.126	0.256	0.406	0.349	0.375	0.176	0.011	0.176	0.264	0.017	0.101	0.415	0.403
0.014	0.08	0.021	0.042	0.042	0.005	0.006	0.005	0.006	0.005	0.006	0.01	0.029
1.211	0.895	1.091	1.258	1.159	0.871	0.251	0.871	0.561	0.916	0.258	0.178	0.438

loading value (+0.3), CBC was helpful to separate one cannabis sample from Washington state (WA2).

It should be stressed that PCA is used for sample or variable grouping within a previous knowledge on the class membership of studied samples. Efficient discrimination among cannabis samples is achieved by LDA as will be discussed in the following section.

Quantitative cannabis classification by LDA

As an unsupervised clustering method, PCA could not afford complete geographical classification of cannabis samples collected from the four USA states. The failure of PCA would be attributed to the fact that no previous information on class membership of the samples involved in the model. Hence, PCA grouped the samples in a lower dimension space of the original data. Due to the modest performance of PCA toward geographical classification of cannabis samples, LDA was selected. LDA, unlike PCA, is a supervised classification method in which class memberships of samples should be known before modeling. With the aid of class information, LDA selects a direction that achieves minimum within-class distance and maximum separation among the classes [18–20]. In LDA, the first job is to create a discriminant function which is created from $\mathbf{X}_{26\text{samples}\times22\text{solutes}}$ and $\mathbf{y}_{26\times1}$ (class membership codes of samples from states). The next step is to check the classes of new samples ($\mathbf{X}_{28\text{samples}\times22\text{solutes}}$) not involved in the model. The

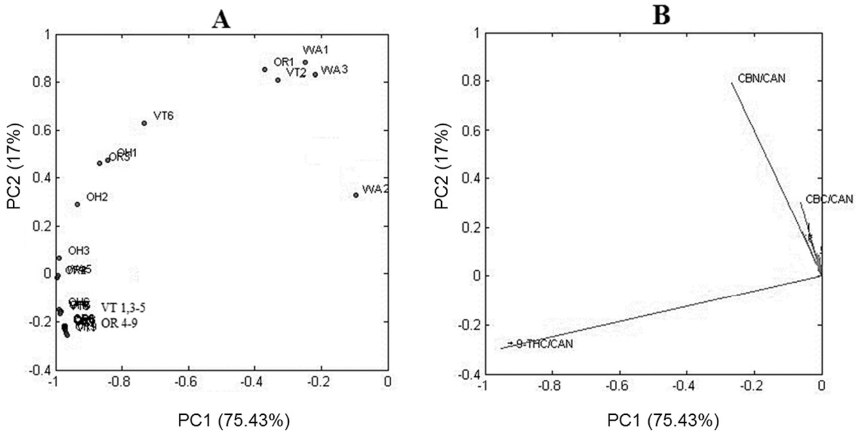


Fig. 2. PCA outputs for measured GC profiles, A: score plot, and B: loading plot of the 26 cannabis samples and the 22 solute at four USA states

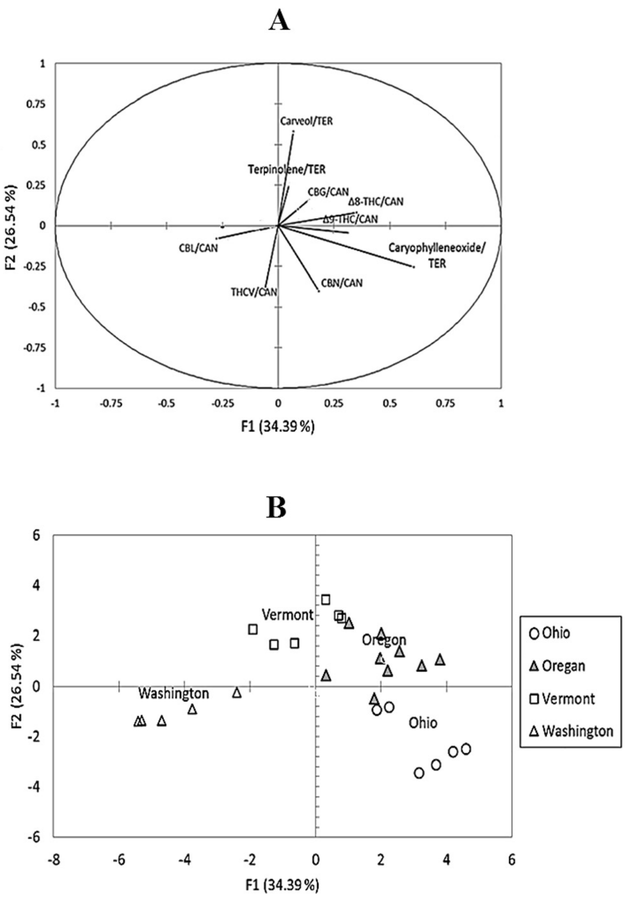


Fig. 3. LDA plots for the discrimination of 26 cannabis samples based on GC–FID measurements, A: variables selection and B: samples classification

outputs of LDA are presented in Fig. 3, where Fig. 4 showed the level (i.e., relative abundance) of the nine cannabinoids and terpenoids selected by the LDA model in the four USA states samples.

It seems that LDA considered more indices to build classification functions (F1 and F2). As shown in Fig. 3A, two LDA factors (explained around 61% of total variance) were used to create the necessary classification functions. The interesting point in Fig. 3A was the larger number of constituents (nine components) that were involved in the LDA classification of the samples in comparison to PCA (three components) (see Fig. 2B). The larger number of variables was attributed to the extra information on class membership involved while building LDA. Fig. 3A showed that all cannabinoids were significant in LDA prediction except CBC because it had low loading value in PCA, which is shown previously in Fig. 2B. Among measured terpenoids, three terpenes were selected by LDA: caryophylleneoxide, carveol, and terpinolene. Both caryophylleneoxide and caravel were highly significant with loading values > +0.5 for F1 and F2, respectively. It was interesting to notice the high anticorrelation between carveol and THCV.

The feasibility of using LDA to discriminate the geographical origin of cannabis samples was assessed. The LDA outputs of samples collected from Washington, Vermont, Ohio, and Oregon are displayed in Fig. 3B. Compared with the outputs of PCA, LDA exhibited more sharp classes among the 26 samples. LDA was also convenient to isolate Washington samples from others. Ohio samples were also isolated away from Washington and Vermont samples. Only few samples from Oregon seem to be close to Vermont and Ohio samples. Variables selection by forward stepwise-criterion was adopted to find the most appropriate variables for LDA. Analysis indicated that Δ^9-THC is the most significant variable for samples discrimination regarding geographical origins. The earlier results were in agreement with LDA and PCA results as Δ^9-THC was highly related to the first classification function F1 and first principal component PC1, respectively. Internal validation of LDA using two factors indicated a 100% accuracy of classification, i.e., all samples were correctly classified to their geographical origins. Therefore, the classification performance of LDA was further assessed in the next section by classification of 28 new cannabis samples not involved in building the model in order to check the power of this model.

LDA-model validation

To validate our LDA model, new 28 cannabis samples (seven from each state) not involved in building the model were applied and the classification accuracy is provided in Table 3.

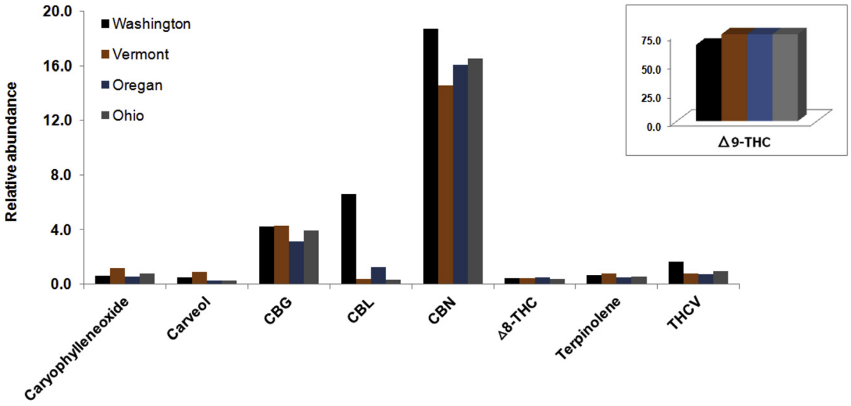


Fig. 4. Relative abundance of the nine cannabinoids and terpenoids selected by LDA model in the four USA states samples

Table 3. Results of LDA classification of cannabis samples from Washington, Vermont, Oregon, and Ohio states using chemical profiles of terpenoids and cannabinoids obtained by GC–FID

USA states	Number of samples (*n*)	LDA prediction				Classification accuracy (%)
		Washington	Vermont	Oregon	Ohio	
Washington	7	7	0	0	0	100%
Ohio	7	0	0	1	6	86%
Vermont	7	0	6	1	0	86%
Oregon	7	0	0	5	2	71%

Classification accuracy was evaluated as the percentage of correctly classified samples by LDA in a given class compared to the total number of samples in that class. As indicated in Table 3, LDA has acceptable classification accuracy for new samples according to their geographical origins. The best result was reported for Washington samples as all samples were correctly classified with 100% accuracy. For seven new samples from Ohio, the accuracy of the classification was rather acceptable (86%) as only one sample out of seven was incorrectly classified with Oregon samples (Fig. 3B). The same discussion holds true for LDA classification for Vermont samples, one sample was misclassified with Oregon samples. For Oregon samples, two samples out of seven were incorrectly classified with a final accuracy of 71%. These misclassified samples were specified to Ohio and this was expected considering the high closeness between Oregon and Ohio samples as indicated in Fig. 3B. The comparable chemical profiles of some samples obtained from Oregon/Vermont/Ohio would retard the classification efficiency of LDA.

But in general, LDA along with GC–FID profiles of cannabis samples was found practical for classification and prediction of samples according to their geographical origins.

CONCLUSION

Chemical profiles obtained by GC–FID along with LDA were found efficient for geographical classification of USA-domestic cannabis samples and for predicting the possible origin of unknown cannabis samples. As an unsupervised classification method, PCA was of limited application for classification of 26 cannabis samples to their geographical sites due to the absence of classification markers while building the model. As a supervised method, LDA was created with a previous knowledge for class member of samples. LDA was efficient for selecting more components (Δ^8-THC, Δ^9-THC, THCV, CBG, CBL, CBN, terpinolene, carveol, and caryophylleneoxide). Reasonable classification of 28 new samples according to their geographical origins was achieved using GC–FID–LDA with accuracies of 100, 86, and 71% for Washington, Ohio/Vermont, and Oregon states, respectively.

Conflicts of interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

ACKNOWLEDGMENT

Dr Ramia Al Bakain gratefully acknowledges the financial support for her Fulbright post doctorate year 2017/2018 at The University of Mississippi, USA, provided by (1) The Binational Fulbright Commission in Jordan, and (2) the University of Jordan. Indeed, great appreciation goes to Dr Omar Marzouk for his experimental help and technical support. Many thanks also go to Dr Mohamed Radwan, Elsayed Ibrahim, Avery Claire Jones, and Chandrani Gon Majumdar for their assistance.

REFERENCES

1. Gunn, R.; Jackson, K.; Borsari, B.; Metrik, J. A longitudinal examination of daily patterns of cannabis and alcohol co-use among medicinal and recreational veteran cannabis users. *Drug. Alcohol. Depen.* **2019**, *205*, 107661, 1–8.
2. O'Brien, K. Medicinal cannabis: Issues of evidence. *Eur. J. Integr. Med.* **2019**, *28*, 114–20.
3. Wang, M.; Wang, Y. -H.; Avula, B.; Radwan, M. M.; Wanas, A. S.; Mehmedic, Z.; Antwerp, J.; ElSohly, M. A.; Khan, I. A. Quantitative determination of cannabinoids in cannabis and cannabis products using ultra-high-performance supercritical fluid chromatography and diode array/mass spectrometric detection. *J. Forensic. Sci.* **2017**, *62*, 602–11.
4. Radwan, M. M.; Elsohly, M. A.; Slade, D.; Ahmed, S. A.; Khan, I. A.; Ross, S. A. Biologically active cannabinoids from high-potency Cannabis sativa. *J. Nat. Prod.* **2009**, *72*(5)**,** 906–11.
5. Ahmed, S. A.; Ross, S. A.; Slade, D.; Radwan, M. M.; Khan, I. A.; Elsohly M. A. Structure determination and absolute configuration of cannabichromanone derivatives from high potency Cannabis sativa. *Tetrahedron. Lett.* **2008**, *49*(42), 6050–3.
6. Radwan, M. M.; Ross, S. A.; Slade, D.; Ahmed, S. A.; Zulfiqar, F.; Elsohly, M. A. Isolation and characterization of new cannabis constituents from a high potency variety. *Planta. Med.* **2008**, *74*(3), 267–72.
7. Ahmed, S. A.; Ross, S. A.; Slade, D.; Radwan, M. M.; Zulfiqar, F.; Matsumoto, R. R.; Xu, Y. -T.; Viard, E.; Speth, R. C.; Karamyan, V. T.; ElSohly, M. A. Cannabinoid ester constituents from high-potency Cannabis sativa. *J. Nat. Prod.* **2008**, *71*(4), 536–42.
8. Zulfiqar, F.; Ross, S. A.; Slade, D.; Ahmed, S. A.; Radwan, M. M.; Ali, Z.; Khan, I. A.; ElSohly, M. A. Cannabisol, a novel D9-THC dimer possessing a unique methylene bridge, isolated from Cannabis sativa. *Tetrahedron. Lett.* **2012**, *53*(28), 3560–2.

9. Ibrahim, A. K.; Radwan, M. M.; Ahmed, S. A.; Slade, D.; Ross, S. A.; ElSohly, M. A.; Khan, I. A. Microbial metabolism of cannflavin A and B isolated from Cannabis sativa. *Phytochemistry* **2010**, *71*, 1014–9.
10. Radwan, M. M.; Elsohly, M. A.; Slade, D.; Ahmed, S. A.; Wilson, L.; El-Alfy, A. T.; Khan, I. A.; Ross, S. A. Non-cannabinoid constituents from a high potency Cannabis sativa variety. *Phytochemistry* **2008**, *69*(14), 2627–33.
11. Brenneisen, R.; Elsohly, M. A. Chromatographic and spectroscopic profiles cannabis of different origins: Part 1. *J. Forensic Sci.* **1988**, *33*, 1385–404.
12. Al Bakain, R. Z.; Al-Degs, Y. S.; Cizdziel, J. V.; Elsohly, M. A. Comprehensive classification of USA cannabis samples based on chemical profiles of major cannabinoids and terpenoids. *J. Liq. Chrom. Rel. Technol.* **2019**, *43*, 172–84.
13. Elkins, A. C.; Deseo, M. A.; Rochfort, S.; Ezernieks, V.; Spangenberg, G. Development of a validated method for the qualitative and quantitative analysis of cannabinoids in plant biomass and medicinal cannabis resin extracts obtained by super-critical fluid extraction. *J. Chromatogr. B* **2019**, *1109*, 76–83.
14. Jin, D.; Jin, S.; Yu, Y.; Lee, C.; Chen, J. Classification of Cannabis Cultivars Marketed in Canada for Medical Purposes by Quantification of Cannabinoids and Terpenes Using HPLCDAD and GC-MS. *J. Anal. Bioanal. Tech.* **2017**, *9*, 1–9.
15. Brenneisen, R.; Elsohly, M. A. Chemistry and Analysis of Phytocannabinoids and Other Cannabis Constituents. Forensic Science Medicine: Marijuana and the Cannabinoids; Humana Press: Totowa, **2007**, p 17–49.
16. Brereton, R. G. Chemometrics: Statistics and Computer Application in Analytical Chemistry; John Wiley & Sons: Chichester, **2007**, UK.
17. Otto, M. Applied Chemometrics for Scientists, 3rd ed.; Wiley-VCH: New York, NY, **2016**.
18. Arvanitoyannis, I. S.; Katsota, M. N.; Psarra, E. P.; Soufleros, E. H.; Kallithraka, S. Application of quality control methods for assessing wine authenticity: use of multivariate analysis (chemometrics). *Trends. Food. Technol. Sci.* **1999**, *10*, 321–36.
19. Berrueta, L. A.; Alonso-Salces, R. M.; Héberger, K. Supervised pattern recognition in food analysis. *J. Chromatogr. A* **2007**, *1158*, 196–214.
20. Bellomarino, S. A.; Conlan, X. A.; Parker, R. M.; Barnett, N. W.; MJ, A. Geographical classification of some Australian wines by discriminant analysis using HPLC with UV and chemiluminescence detection. *Talanta* **2009**, *80*, 833–8.
21. Nunes, C. A.; Freitas, M. P.; Pinheiro, A. C. M.; Bastos, S. C. Chemoface: A novel free user-friendly interface for chemometrics. *J. Braz. Chem. Soc.* **2012**, *23*, 2003–10.
22. Al Bakain, R. Z.; Al-Degs, Y. S.; Cizdziel, J. V.; Elsohly, M. A. Comprehensive chromatographic profiling of cannabis from 23 USA States marketed for medical purposes. *Acta. Chromatogr.* **2020**, (in press).