Pharmacogn. Mag.
A multifaceted peer reviewed journal in the field of Pharmacognosy and Natural Products
www.phcog.com | www.phcog.net

ORIGINAL ARTICLE

Identification of *Polygonatum odoratum* Based on Support Vector Machine

Zhong Li, Jie Zheng[1], Qin Long, Yi Li, Huaying Zhou[2], Tasi Liu[3], Bin Han

Department of Traditional Chinese Medicine Resources, College of Traditional Chinese Medicine, Guangdong Pharmaceutical University, [1]Department of Pharmaceutical Engineering, College of Chemical Engineering and Light Industry, Guangdong University of Technology, [2]Department of Computer Science, College of Medical Information Engineering, Guangdong Pharmaceutical University,Guangzhou, [3]Department of Traditional Chinese Medicine Resources, College of Traditional Chinese Medicine, Hunan University of Chinese Medicine, Changsha, China

Submitted: 27-Sep-2019 Revised: 31-Oct-2019 Accepted: 21-Apr-2020 Published: 20-Oct-2020

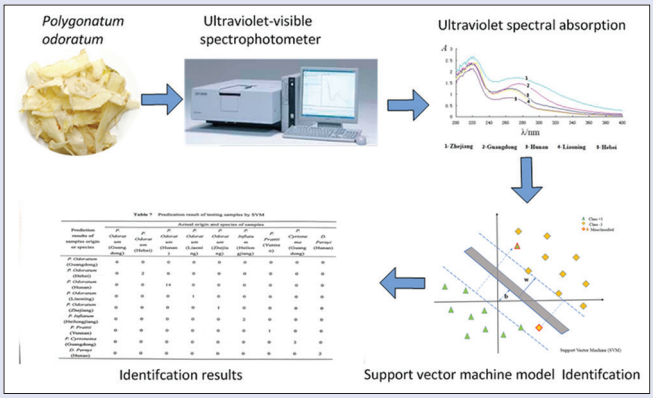
ABSTRACT

Background: The dried rhizome of *Polygonatum odoratum* (Mill.) Druce has been widely used in traditional medicinal preparations in China, Japan, and Korea. In China, it is distributed in Hunan, Guangdong, and Liaoning provinces, and its quality differs from habitat to habitat. In addition, *P. odoratum* has some adulterants, such as *Polygonatum inflatum* Kom, *Polygonatum prattii* Baker, and *Polygonatum cyrtonema* Hua. The morphological traits and chemical composition of the aforementioned adulterants have many similarities with those of *P. odoratum*. Therefore, it is possible that people often use adulterants instead of *P. odoratum* for clinical treatment. **Objectives:** We aimed to establish a reliable and accurate classification model of *P. odoratum* based on the support vector machine (SVM) and identify it from different habitats; we also aimed to identify its adulterants. **Materials and Methods:** In this study, we first determined the ultraviolet (UV) absorption spectrum of the water extract of the rhizome from 162 samples (including *P. odoratum* from Hunan, Guangdong, Heilongjiang, Yunnan, and Liaoning Provinces and adulterant species including *P. inflatum*, *P. prattii*, *P. cyrtonema*, and *Disporopsis pernyi* (Hua) Diels) by UV-visible spectrophotometry. The UV absorption data were preprocessed with the SVM model before establishing the habitat and other details. **Results:** According to our results, the SVM model showed a prediction accuracy of 100%. The model accurately identified five different habitats and four different adulterants of *P. odoratum*. Pretreatment of samples with UV spectrum might be useful in the accurate identification of *P. odoratum*. **Conclusion:** The SVM model seems very prospective in identifying herbs with multiple habitats and its adulterants.

Key words: Adulterants, identification, *Polygonatum odoratum*, support vector machine, ultraviolet

SUMMARY

- In this study, R language optimized the ultraviolet (UV) spectral data of water extracts of *Polygonatum odoratum* and helped to establish the support vector machine (SVM) to identify and classify *P. odoratum* from different habitats and its various adulterants. Our results showed that the prediction accuracy of the SVM model was 100%, and the method of SVM pretreatment UV spectrum could be used to identify *P. odoratum*.



Abbreviations used: *P.*: *Polygonatum*; UV: Ultraviolet; SVM: Support vector machine; TCM: Traditional Chinese medicine; Fig: Figure.

Correspondence:

Ph.D. Huaying Zhou,
College of Medical Information Engineering,
Guangdong Pharmaceutical University,
Guangzhou,510006,China.
E-mail: 287059250@qq.com

DOI: 10.4103/pm.pm_410_19

Access this article online
Website: www.phcog.com
Quick Response Code:

INTRODUCTION

Polygonatum odoratum (Mill.) Druce, native to many parts of the world, belongs to the Liliaceae family. It has been widely used in the preparations of traditional Chinese medicine (TCM) as a component of medications intended to treat diabetes, Qi-tonify, and clear the heat. *P. odoratum* is distributed in many provinces of China, such as Hunan, Guangdong, and Liaoning, and its quality differs based on its habitat. In addition, the morphological traits and chemical composition of the adulterants of *P. odoratum* have many similarities with *P. odoratum*. For example, the rhizome of *Polygonatum inflatum* is used as *P. odoratum* in Northeast China and *Polygonatum prattii* is mistaken for *P. odoratum* in Sichuan and Yunnan Provinces. *Polygonatum cyrtonema* is mixed with *P. odoratum* to be used as medicine in other parts of China. Therefore, the adulterants of *P. odoratum* have been wrongly used for thousands of years in China. However, only *P. odoratum* has been included in the "*Pharmacopoeia of The People's Republic of China*" as an authentic Chinese medical herb.[1] It is a great challenge to distinguish different species under the same genus – *Polygonatum* – as their dried and sliced rhizomes are very much similar looking. So far, there is no effective method to identify the difference between *P. odoratum* and its adulterants. However, it is highly essential to identify each one of them with a reliable and accurate method when applied as a medicine;

Cite this article as: Li Z, Zheng J, Long Q, Li Y, Zhou H, Liu T, *et al*. Identification of *Polygonatum odoratum* based on support vector machine. Phcog Mag 2020;16:538-42.

the medicinal ingredients of the misnamed *P. odoratum* might cause severe consequences to the patients.

In recent years, many new technologies and methods have been established to identify and characterize Chinese herbal medicines; support vector machine (SVM) is a well-known classification paradigm in machine learning. It is a supervised learning model with associated learning algorithms that analyze data for classification and regression analysis. It has become a hot spot for new research due to its inherent outstand learning ability[2,3] and has been widely used to classify TCM,[4] elucidate structure–activity relationships,[5,6]

We combined ultraviolet (UV) spectrometry with SVM to identify *P. odoratum* samples from different habitats such as Hunan, Guangdong, Heilongjiang, Yunnan, and Liaoning Provinces and from different neighboring species such as *P. inflatum*, *Polygonatum prattii*, *P. cyrtonema*, and *Disporopsis pernyi* (Hua) Diels. In this study, we aimed to develop a fast-identifying model based on SVM.

MATERIALS AND METHODS

Samples

Samples were collected from Hunan, Guangdong, Heilongjiang, Yunnan, and Liaoning Provinces in China. Table 1 shows information regarding each sample. All these samples were authenticated by Associate Professor Zhong Li (College of TCM, Guangdong Pharmaceutical University, Guangzhou, China).

Preparation of water extract

The coarse powder of each sample was accurately weighed (2.0 g) and placed in a 50 mL volumetric flask by adding 20 mL distilled water, and then, the samples were ultrasonicated for 20 min at room temperature and filtered through a 20um quantitative fitler paper. From this stock solution, 5 mL of the filtrate was taken into a 100 mL volumetric flask and was diluted up to the mark with distilled water and mixed well. This working solution was used to test the absorbance.

Ultraviolet absorption spectroscopy

The absorbance of the samples was measured at 200–400 nm using a UV–visible spectrophotometer, and the sampling interval was set to 1 nm. During SVM modeling, the absorbance of each sample will affect the classification in a different way; therefore, it is necessary to centralize and standardize the absorbance of different ranges of wavelengths. This will ensure that all data participate in the construction of the classifier model on the same scale. The spectral data are centralized and standardized using the scale function of R software.[7] In addition, the data also include redundant information which might lead to errors in the SVM modeling; therefore, it is necessary to optimize the wavelength range. In this study, based on the SVM variable selection method,[8] the absorbance data in the wavelength range of 200–400 nm are sorted according to the closeness of the SVM classification index and the optimal wavelength absorbance data are selected to establish the classifier model.

Table 1: Samples information of *Polygonatum odoratum* and its adulterants

n	Species	Place of origin
1	*P. odoratum* (Mill.) Druce	Guangdong
2	*P. odoratum* (Mill.) Druce	Hebei
3	*P. odoratum* (Mill.) Druce	Hunan
4	*P. odoratum* (Mill.) Druce	Liaoning
5	*P. odoratum* (Mill.) Druce	Zhejiang
6	*P. inflatum* Kom.	Heilongjiang
7	*P. prattii* Baker	Yunnan
8	*P. cyrtonema* Hua	Guangdong
9	*D. pernyi* (Hua) Diels	Hunan

P. odoratum: *Polygonatum odoratum*; *P. inflatum*: *Polygonatum inflatum*; *P. prattii*: *Polygonatum prattii*; *P. cyrtonema*: *Polygonatum cyrtonema*; *D. pernyi*: *Disporopsis pernyi*

Creating the support vector machine model

From the total sample, we randomly selected 80% of the samples as the training set and the remaining 20% of the samples as the testing set. Data on the training set were used to build the SVM model, and then, the data on the testing set were used to validate the model. The program interface of Lib SVM in e1071 provided by R Software was used to create SVM classifier modeling.[9]

RESULTS

Ultraviolet spectral absorption of *Polygonatum odoratum* from different habitats and different species

As shown in Figure 1, the UV spectral absorption of water extracts of *P. odoratum* from different regions is very similar; therefore, it is difficult to find the region of origin of *P. odoratum* only by the spectral analysis. Furthermore, the UV spectral data of *P. odoratum* and other neighboring species are also very similar [Figure 2]. The traditional method of tasting or visualizing the product was therefore not helping to differentiate between *P. odoratum* and its adulterants, which has caused major confusion for thousands of years. Therefore, it is important to develop an advanced method to identify *P. odoratum* and improve its production standard and quality.

Selection of ultraviolet wavelength

The variable selection algorithm is performed by SVM, and the absorbances of all wavelengths are sorted according to the importance degree. Table 2 shows the sorting results. AvgRank is the sorting index – the smaller the value of AvgRank, the closer the relation to its classification. In this study, the top 40 wavelengths of absorbance data were used for SVM modeling analysis.

A total of 162 samples from 9 producing places were randomly divided into the training set and testing set by "Sample" function in Software R [Table 3].

Selection of radial basis function kernel γ and error warning factor C

We screened the radial basis function kernel γ while creating the SVM model. According to the results, when γ value changes from 0.125 to 16,

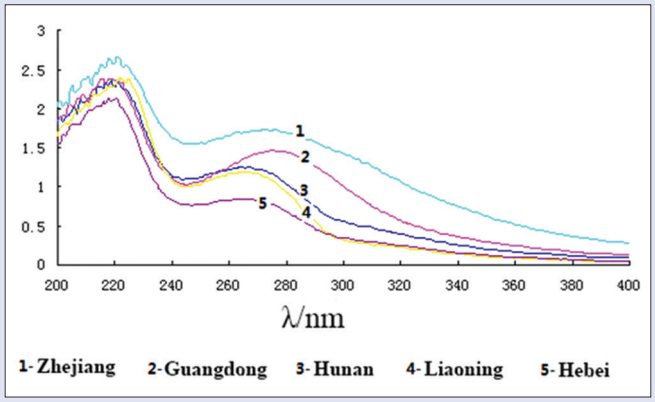


Figure 1: Ultraviolet spectral absorption of *Polygonatum odoratum* from different producing areas

it has no impact on the results of the training set, whereas the accuracy for the testing set decreases [Table 4]. This is because γ controls the amplitude of the radial base function, which controls the generalizability of SVM. Based on this, γ = 0.125 was selected.

Based on the same philosophy, error warning factor C was also screened to optimize the SVM model. Table 5 shows the results. The smaller the value of C, the smaller will be the penalty, which makes the training error larger. The structural risk to the system is confined by empiric risk and confidence level; therefore, a large training error may cause an increase in the structural risk and worsen the generalizability of the system. Therefore, the value of C has a tremendous influence on the system's generalizability. Based on the data presented in Table 5, it is obvious that when C is between 2 and 16, the classification accuracy is stabilized. Therefore, we selected the value of C = 2 for this model.

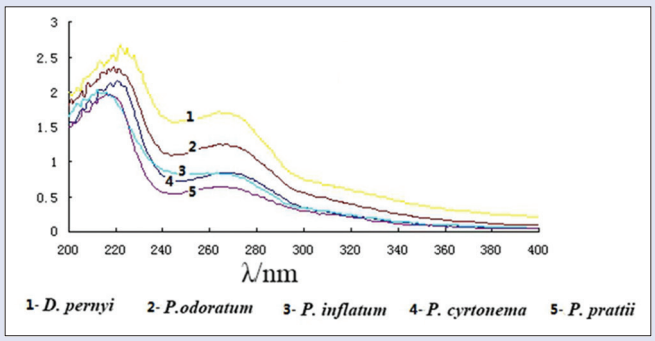


Figure 2: Ultraviolet spectral absorption of different *Polygonatum* species

Identification results by support vector machine

The optimized SVM classifier was built with the optimized parameters obtained from the previous analysis on the radial base kernel γ and error warning factor C. Table 6 shows the prediction result of 130 training samples by SVM, and the prediction accuracy rate was 100%. With the optimized factors, all 32 samples of the test set were validated. The identification accuracy by SVM was 100% [Table 7].

DISCUSSION

UV absorption spectroscopy is commonly used to identify the structure of compounds or in the determination of the composition. Due to the different saturation values of each of the chemical components contained in TCM, the peak of the absorption curve, the shape of the peak, and

Table 2: The importance sorting of support vector machine variables

λ/nm	AvgRank	λ/nm	AvgRank	λ/nm	AvgRank	λ/nm	AvgRank	λ/nm	AvgRank
382	8.8	355	39.0	326	82.8	282	121.6	292	160.0
384	9.8	367	39.5	325	83.0	208	122.1	313	160.5
383	9.8	399	41.0	252	83.2	274	123.3	293	162.4
378	10.7	398	41.7	324	83.9	202	123.4	232	163.3
381	10.9	348	42.2	251	86.0	262	123.8	294	165.1
385	11.0	358	43.4	250	87.5	283	124.7	312	166.0
377	11.1	350	44.5	323	87.6	263	125.0	295	166.6
374	11.2	349	46.6	259	87.8	207	127.6	231	167.9
380	11.2	357	48.4	249	89.4	273	127.7	296	168.3
376	11.6	344	50.6	248	90.9	237	128.2	311	169.8
386	12.5	343	50.9	247	91.8	211	129.6	230	170.5
375	12.5	356	51.3	246	93.3	272	129.8	297	171.7
373	12.6	342	53.1	260	94.5	317	130.6	225	173.7
387	14.7	345	53.9	245	94.9	264	130.9	298	174.0
370	14.7	351	56.8	322	95.4	284	131.6	228	174.9
369	14.9	353	57.5	244	95.6	271	133.4	229	177.3
388	15.7	347	58.7	243	97.4	236	133.5	310	177.4
372	16.9	333	58.9	242	98.4	285	135.2	299	177.9
391	17.4	336	59.7	241	100.5	265	135.8	300	180.3
389	18.2	337	60.6	321	101.7	266	136.9	214	181.1
390	19.3	338	61.1	240	103.2	269	138.3	309	181.3
368	19.3	335	62.6	277	105.1	286	138.5	221	181.5
392	21.2	346	65.3	279	105.6	270	138.6	227	184.4
371	22.1	332	65.7	278	105.6	212	139.0	301	184.8
360	24.4	334	67.6	201	106.8	268	139.2	224	185.1
393	24.5	340	67.8	320	106.9	316	140.1	226	185.1
364	29.1	339	67.9	239	108.6	287	143.0	220	185.5
394	30.7	331	71.5	276	110.0	267	144.1	219	186.7
362	31.5	341	72.3	200	110.4	288	146.3	217	189.4
396	31.7	352	72.4	206	110.4	210	146.4	302	189.5
363	31.9	330	73.2	280	112.0	235	147.5	308	189.9
395	33.2	328	74.7	281	112.6	203	150.2	307	191.0
361	33.4	329	76.6	319	112.8	289	150.6	306	192.2
354	34.9	255	77.7	204	113.0	315	150.8	303	192.5
379	35.0	258	77.9	261	115.7	290	154.5	218	193.0
397	35.3	254	78.1	205	117.6	234	156.4	222	193.8
366	35.6	256	79.2	275	118.5	291	157.6	304	194.2
400	36.9	327	79.4	238	119.9	314	158.3	305	194.5
359	38.5	257	80.2	209	121.0	233	159.4	216	194.8
365	38.6	253	80.2	318	121.0	213	159.7	223	196.8

Table 3: The information of training set and test set samples

Species (producing place)	Training set (130)	Testing set (32)
P. odoratum (Guangdong)	24	6
P. odoratum (Hebei)	10	2
P. odoratum (Hunan)	46	14
P. odoratum (Liaoning)	5	1
P. odoratum (Zhejiang)	5	1
P. inflatum (Heilongjiang)	4	2
P. prattii (Yunnan)	11	1
P. cyrtonema (Guangdong)	15	3
D. pernyi (Hunan)	10	2

P. odoratum: *Polygonatum odoratum*; *P. inflatum*: *Polygonatum inflatum*; *P. prattii*: *Polygonatum prattii*; *P. cyrtonema*: *Polygonatum cyrtonema*; *D. pernyi*: *Disporopsis pernyi*

Table 4: Support vector machine predicating ability on different radial basis function kernel γ

Predication accuracy (%)	γ value							
	0.125	0.25	0.5	1	2	4	8	16
Training set	100	100	100	100	100	100	100	100
Testing set	100	96.88	90.63	75	43.75	43.75	43.75	43.75

Table 5: Support vector machine predicating ability on different error warning factor C

Predication accuracy (%)	Error warning factor C							
	0.125	0.25	0.5	1	2	4	8	16
Training set	45.38	84.61	98.46	100	100	100	100	100
Testing set	53.12	75.00	96.87	96.87	100	100	100	100

the strength of the peak are different. In addition, UV absorption spectroscopy is a simple and effective method used to identify distantly related to traditional Chinese herbal medicines, but it is not effective when identifying adulterants. SVM can solve this problem more effectively. Based on the UV spectral data of *P. odoratum* samples, SVM successfully identified and differentiated between the *P. odoratum* samples from different habitats and its adulterants.

Proper identification and classification of TCM is a common problem. The majority of the data obtained are unlabeled which lead to problems in identification and classification. Therefore, the question of how to use these data effectively and improve the accuracy of modeling techniques needs an urgent answer.

In this study, we used the R language to screen the best spectral pretreatment. It is found that the absorbance data of the first 40 wavelengths can effectively be used in the modeling and analysis of herbal medicines. The predicted accuracy of the established SVM classifier is 100%, which is, in turn, based on the predicted accuracy of the whole model. This shows that SVM has the advantages of a faster learning rate, high accuracy, and high generalizability. These features can help to solve the quality problem of TCM originating from different habitats and provide a new method for the effective identification of complex TCM.

Financial support and sponsorship

This work was supported by Guangdong Science and Technology Department Project (Grant No. 2016A020226018), Ministry of National Science and Technology Support Program Project of China (No. 2011BA101B09), and Central support for a local college project (Grant No. 51348000).

Table 6: Predication result of training samples by support vector machine

Prediction results of samples origin or species	Actual origin and species of samples								
	P. odoratum (Guangdong)	*P. odoratum* (Hebei)	*P. odoratum* (Hunan)	*P. odoratum* (Liaoning)	*P. odoratum* (Zhejiang)	*P. inflatum* (Heilongjiang)	*P. prattii* (Yunnan)	*P. cyrtonema* (Guangdong)	*D. pernyi* (Hunan)
P. odoratum (Guangdong)	24	0	0	0	0	0	0	0	0
P. odoratum (Hebei)	0	10	0	0	0	0	0	0	0
P. odoratum (Hunan)	0	0	46	0	0	0	0	0	0
P. odoratum (Liaoning)	0	0	0	5	0	0	0	0	0
P. odoratum (Zhejiang)	0	0	0	0	5	0	0	0	0
P. inflatum (Heilongjiang)	0	0	0	0	0	4	0	0	0
P. prattii (Yunnan)	0	0	0	0	0	0	11	0	0
P. cyrtonema (Guangdong)	0	0	0	0	0	0	0	15	0
D. pernyi (Hunan)	0	0	0	0	0	0	0	0	10

P. odoratum: *Polygonatum odoratum*; *P. inflatum*: *Polygonatum inflatum*; *P. prattii*: *Polygonatum prattii*; *P. cyrtonema*: *Polygonatum cyrtonema*; *D. pernyi*: *Disporopsis pernyi*

Table 7: Predication result of testing samples by support vector machine

Prediction results of samples origin or species	Actual origin and species of samples								
	P. odoratum (Guangdong)	*P. odoratum* (Hebei)	*P. odoratum* (Hunan)	*P. odoratum* (Liaoning)	*P. odoratum* (Zhejiang)	*P. inflatum* (Heilongjiang)	*P. prattii* (Yunnan)	*P. cyrtonema* (Guangdong)	*D. pernyi* (Hunan)
P. odoratum (Guangdong)	6	0	0	0	0	0	0	0	0
P. odoratum (Hebei)	0	2	0	0	0	0	0	0	0
P. odoratum (Hunan)	0	0	14	0	0	0	0	0	0
P. odoratum (Liaoning)	0	0	0	1	0	0	0	0	0
P. odoratum (Zhejiang)	0	0	0	0	1	0	0	0	0
P. inflatum (Heilongjiang)	0	0	0	0	0	2	0	0	0
P. prattii (Yunnan)	0	0	0	0	0	0	1	0	0
P. cyrtonema (Guangdong)	0	0	0	0	0	0	0	3	0
D. pernyi (Hunan)	0	0	0	0	0	0	0	0	2

P. odoratum: Polygonatum odoratum; P. inflatum: Polygonatum inflatum; P. prattii: Polygonatum prattii; P. cyrtonema: Polygonatum cyrtonema; D. pernyi: Disporopsis pernyi

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Chinese Pharmacopoeia Commission. Pharmacopoeia of The People's Republic of China (Part 2). Beijing: Chemical Industry Press; 2015.
2. Tahir M, Jan B, Hayat M, Shah SU, Amin M. Efficient computational model for classification of protein localization images using extended threshold adjacency statistics and support vector machines. Comput Methods Programs Biomed 2018;157:205-15.
3. Chen C, Chen LX, Zou XY, Cai PX. Predicting protein structural class based on multi-features fusion. J Theor Biol 2008;253:388-92.
4. Jun WY, Yue Y, Yu ZJ, Song LX, Jiang WY, Tao ZW. Geographical origin discrimination of herba epimedii by near infrared spectroscopy. Lishizhen Med Mater Medica Res 2017;28:1902-5.
5. Ruiz IL, Gómez-Nieto MÁ. Advantages of Relative Versus Absolute Data for the Development of Quantitative Structure-Activity Relationship Classification Models.J Chem Inf Model;2017:2776-88.
6. Luque Ruiz I, Gómez-Nieto MÁ. Robust QSAR prediction models for volume of distribution at steady state in humans using relative distance measurements. SAR QSAR Environ Res 2018;29:529-50.
7. Tierney L. The R Statistical Computing Environment. In: Statistical Challenges in Modern Astronomy V. New York: Springer; 2012.
8. Zhang C, Shen T, Liu F, He Y. Identification of coffee varieties using laser-induced breakdown spectroscopy and chemometrics. Sensors (Basel) 2017;18:95.
9. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071),Tu wien.UTC; 2019.