Journal of Chromatography A, 1518 (2017) 34–45

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Full length article

A modification on the vector cosine algorithm of Similarity Analysis for improved discriminative capacity and its application to the quality control of Magnoliae Flos

Guang Yang[a,b,1], Xin Zhao[c,d,1], Guorong Fan[c,e,f,*]

[a] Department of Pharmaceutical Analysis, School of Pharmacy, Second Military Medical University, No. 325 Guohe Road, Shanghai 200433, PR China
[b] No.208 Hospital of Chinese People's Liberation Army, Changchun 130062, PR China
[c] Department of Clinical Pharmacy, Shanghai General Hospital, School of Medicine, Shanghai Jiaotong University, No. 100 Haining Road, Shanghai 200025, PR China
[d] Department of Pharmaceutical Analysis, School of Pharmacy, China Pharmaceutical University, No. 24 Tong Jia Xiang, Nanjing 210009, PR China
[e] Shanghai Key Laboratory for Pharmaceutical Metabolite Research, School of Pharmacy, Second Military Medical University, No. 325 Guohe Road, Shanghai 200433, PR China
[f] School of Medicine, Tongji University, No. 1239 Siping Road, Shanghai 200092, PR China

ARTICLE INFO

Article history:
Received 2 March 2017
Received in revised form 1 August 2017
Accepted 10 August 2017
Available online 16 August 2017

Keywords:
Similarity Analysis
Hierarchical Cluster Analysis
Principal Component Analysis
Chemometrics
Magnolia biondii Pamp.
Fingerprint

ABSTRACT

Chromatographic fingerprint analysis has been widely used in quality control of herbal medicines, and Similarity Analysis (SA) as a well-established method has been applied in the quality control practice as well as publications related to the study of herbal medicines and preparations. However, in some cases the results of SA do not fit well with those of other chemometric approaches and quantitative analysis, and the problem remains unsolved. In this study, a modified SA algorithm has been proposed, with its advantages discussed in theory. The extract of dried flower bud of *Magnolia biondii* Pamp. obtained by pressurized solvent extraction was then selected as a case to verify the modified algorithm. After identification of the components, fingerprint analysis was performed using different chemometric methods including Hierarchical Cluster Analysis (HCA), Principal Component Analysis (PCA) as well as original and modified SA methods, and the improved discriminative capacity of modified SA algorithm was illustrated. Characteristic chemical markers were then identified using the modified SA approach and then confirmed using PCA method. The quantitative results were then utilized to confirm the advantage of modified SA approach over the original one. The study made a modification to the widely applied SA algorithm, which was possibly a beneficial improvement in fingerprint analysis and quality control practice of herbal medicines.

1. Introduction

As a holistic approach to the treatment of diseases, Traditional Herbal Medicines are believed to exert their therapeutic or preventive effects through the interaction of multiple components rather than the bioactivity of single chemicals. The conventional quality control methods which mainly focus on major constituent are thus insufficient when applied to the evaluation of complex mixtures such as herbal preparations.

Chromatographic fingerprint analysis has been widely applied in the quality control of herbal medicines. For Instance, In the latest version of Chinese Pharmacopeia, fingerprint similarity has been applied as one of the requirements in the assessment of Traditional Chinese Medicines [1]. Generally, the peak-integrated files of test and reference samples produced by chromatogram workstation are converted to AIA files and inputted to the Similarity Evaluation System for chromatographic fingerprints of Traditional Chinese Medicines software released by Chinese Pharmacopoeia Commission in 2004 and 2012 [1]. After alignment and identification of mutual peaks, the congruence coefficient or Similarity Index (SI) based on vector cosine algorithm [2] was calculated, and a value

Abbreviations: SA, Similarity Analysis; SI, Similarity Index; HCA, Hierarchical Cluster Analysis; PCA, Principal Component Analysis; *M. biondii*, *Magnolia biondii* Pamp; PSE, pressurized solvent extraction; CAD, charged aerosol detection.

* Corresponding author at: Department of Clinical Pharmacy, Shanghai General Hospital, School of Medicine, Shanghai Jiaotong University, No. 100 Haining Road, Shanghai 200025, PR China.
E-mail addresses: diabloyg@126.com (G. Yang), yoyo0132@163.com (X. Zhao), guorfan@outlook.com (G. Fan).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.chroma.2017.08.033
0021-9673/

higher than 0.85-0.95 compared with reference chromatogram is generally accepted as the threshold of qualification. Compared with other established approaches such as Hierarchical Clustering Analysis (HCA) and Principal Component Analysis (PCA), it could provide a concrete value ranged from 0 to 1 to reflect the similarity between two samples, which is suitable for direct comparison and decision-making in the practice of quality control. Meanwhile, Similarity Analysis (SA) has been selected as a routine method in many publications related to the study of herbal medicines and preparations [3–9].

Although widely applied, a potential problem may exist in the vector cosine method-based SA. It is observed that compared with HCA and PCA, the results of SA were less discriminative in some cases [5,6,10–13]. Generally, it happened that the results from SA and HCA& PCA were in contradiction, where samples who shared high SI value between each other, or those who showed close SI values against the reference, were not clustered into the same groups, which means the SI matrix was not informative and discriminative enough to explain the classification in HCA and PCA. Or, although two samples were similar according to the calculated SI values, the contents of quantified components were quite different. Such situations may be an indication that the SI, in some cases is not sufficient for the reflection of difference between samples. Unfortunately, to the best of our knowledge, the problem remains unnoticed since in more cases, the full SI matrix was not reported, and the difference between the results of SA and other chemometric approaches were roughly attributed to the difference between algorithms but not well explained [3]. Since such cases may happen in the quality control practice of herbal medicines, modification should be done to improve the discriminative capacity of SA.

The dried flower buds of *Magnolia biondii* Pamp., known as 'Xin-yi' in China, is one of the most widely used medicinal plants officially listed in the Chinese Pharmacopeia, with its clinical application recorded as treating chill, headache, nasal congestion, allergic rhinitis and nasosinusitis [1]. Both the essential oil and solvent extract from *M. biondii* are considered to be responsible for its biological effects [14–17]. However, in previous study more attention have been paid to the quality study of the essential oil from *M. biondii*, and the quality assessment of the solvent extract attracted less interest. In Consideration of these, the solvent extract of *M. biondii* was selected as an object of our study to verify the modification to the SA.

In this study, a modified SA algorithm was proposed based on the widely applied vector cosine method-based SA appoach in fingerprint analysis of herbal medicines. The solvent extract of *M. biondii* was then selected as a case to verify the modified SA approach. The original and modified SA approaches were both applied as well as other well-established methods such as HCA and PCA in comparison of *M. biondii* samples from different origins and batches to illustrate the improvement of modified algorithm. The quantitative results of identified characteristic chemical markers as well as external data from literature were then utilized to comfirm the advantage of modified SA approach over the original one. The study made a modification to the widely applied SA algorithm, which was possibly a beneficial improvement in fingerprint analysis and quality control practice of herbal medicines.

2. Theory

Define **x** and **v** as two different chromatograms to be compared, and ***n*** as the number of mutual peaks. The definition 'mutual peaks' does not necessarily means the peak must be detected in every sample; generally, it is detected in more than 50% or 80% of the samples analyzed. In other words, the peak area can be zero in some samples. Then both **x** and **v** contain n measurements of peak area as $x_1, x_2, \ldots x_n$ and $v_1, v_2, \ldots v_n$. In original vector cosine algorithm for SA, the congruence coefficient c, or SI, is then calculated as follows [2]:

$$c(x,\ v) = \frac{\sum_{i=1}^{n} x_i v_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2 \sum_{i=1}^{n}(v_i)^2}} (i = 1,\ 2,\ 3, \ldots n) \quad (1)$$

where x_i and v_i are the *i*th in **x** and **v**, representing the peak area of the *i*th mutual peaks.

According to the formula, the essence of vector cosine algorithm can be concluded as follows: the chromatogram of each sample is converted to a *n*-dimensional vector based on the integrated areas of n mutual peaks, and the SI value is the cosine of the angle between two vectors be compared. When the peak areas of two samples are similar, the angle between those converted vectors is small, and the cosine value is close to 1, indicating high similarity.

However, a drawback exists in this algorithm. It is generally reasonable to convert peak-integrated chromatograms to vectors, but the calculation only utilizes the angular information of the vectors and fails to take magnitude information into calculation. Consider such situation: the area of each peak in **v** is just two times as large as those of the corresponding peak in **c,** where the value of SI is thus calculated as 1. It is obvious that from the perspective of quantitative analysis these two sample cannot be regarded as similar samples, but it cannot be figured out from the result of SA. One of the reason, from our view, is that the original algorithm failed to make full used the information from the vectors generated from samples. Thereby, a modified formula for the calculation of SI is proposed as follows:

$$c(x,\ v) = \frac{\sum_{i=1}^{n} x_i v_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2 \sum_{i=1}^{n}(v_i)^2}} \times \min\left(\sqrt{\frac{\sum_{i=1}^{n}(x_i)^2}{\sum_{i=1}^{n}(v_i)^2}},\ \sqrt{\frac{\sum_{i=1}^{n}(v_i)^2}{\sum_{i=1}^{n}(x_i)^2}} \right)$$
$$(i = 1,\ 2,\ 3, \ldots n) \quad (2)$$

where min() means the smaller value of two calculated in the parenthesis. The essence of the modification is that comparison of the magnitude between two vectors was included and utilized as a factor in the calculation, with the range of the final SI unchanged. For instance, the SI in the above mentioned situation is calculated as 0.5 and 1 respectively, using the modified and original algorithm.

3. Experimental

3.1. Chemicals and reagents

17 different batches of *M. biondii* in China were purchased from local Traditional Chinese Medicine pharmacies, and further confirmed by Dr. Luping Qin (Department of Pharmacognosy, School of Pharmacy, Second Military Medical University, Shanghai, China). Reference substances including magnoflorine, denudatone, tiliroside, phillygenin, eudesmin, magnolin, fargesin and veraguensin were purchased from Shanghai yuanye Bio-Technology Co., Ltd. (Shanghai, China). *epi*-Magnolin A was prepared using preparative liquid chromatography [18]. The HPLC grade organic solvent acetonitrile was purchased from Merck (Darmstadt, Germany). Analytical grade ethanol and formic acid were obtained from China