Accepted Manuscript

Title: A rapid integrated bioactivity evaluation system based on near-infrared spectroscopy for quality control of Flos Chrysanthemi

Author: Guoyu Ding Baiqing Li Yanqi Han Aina Liu Jingru Zhang Jiamin Peng Min Jiang Yuanyuan Hou Gang Bai

PII: S0731-7085(16)30510-6
DOI: http://dx.doi.org/doi:10.1016/j.jpba.2016.09.008
Reference: PBA 10845

To appear in: Journal of Pharmaceutical and Biomedical Analysis

Received date: 8-6-2016
Revised date: 3-9-2016
Accepted date: 3-9-2016

Please cite this article as: Guoyu Ding, Baiqing Li, Yanqi Han, Aina Liu, Jingru Zhang, Jiamin Peng, Min Jiang, Yuanyuan Hou, Gang Bai, A rapid integrated bioactivity evaluation system based on near-infrared spectroscopy for quality control of Flos Chrysanthemi, Journal of Pharmaceutical and Biomedical Analysis http://dx.doi.org/10.1016/j.jpba.2016.09.008

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Research Articles

A rapid integrated bioactivity evaluation system based on near-infrared spectroscopy for quality control of *Flos Chrysanthemi*


*a State Key Laboratory of Medicinal Chemical Biology and College of Pharmacy, Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Tianjin 300071, People's Republic of China*

*b Department of Traditional Chinese Medicine, Tianjin Institute of Pharmaceutical Research Co., Ltd, Tianjin 300193, People's Republic of China*

*Corresponding author: Yuanyuan Hou and Gang Bai*

Tel/Fax: +86-022-23506930 (Y. Hou); +86-022-23506792 (G. Bai)

*E-mail address:* houyy@nankai.edu.cn (Y. Hou); gangbai@nankai.edu.cn (G. Bai)
GRAPHICAL ABSTRACT

Highlights

- Boxplot analysis combined with the interval LODs theory was used to screen out the Q-marker ingredients for NIR analysis.
- Machine learning approach was firstly introduced to solve the nonlinear spectrum-activity relationship.
- A rapid bioactivity-based evaluation system based on near infrared spectroscopy for quality control of Flos Chrysanthemi was developed.
ABSTRACT

For quality control of herbal medicines or functional foods, integral activity evaluation has become more popular in recent studies. The majority of researchers focus on the relationship between chromatography/mass spectroscopy and bioactivity, but the connection with spectrum-activity is easily ignored. In this paper, the near infrared reflection spectra (NIRS) of Flos Chrysanthemi samples were collected as a representative spectrum technology, and corresponding anti-inflammation activities were utilized to illustrate the spectrum-activity study. HPLC/Q-TOF-MS identification and heat map clustering were used to select the quality markers (Q-marker) from five cultivars of Flos Chrysanthemi. Using boxplot analysis and the interval limits of detection (LODs) theory, six crucial markers, namely, chlorogenic acid, 3,5-dicaffeoylquinic acid, 1,5-dicaffeoylquinic acid, luteoloside, apigenin-7-O-β-D-glucoside, and luteolin-7-O-6-malonylglucoside were screened out. Then partial least squares regression (PLSR) calibration models combined with synergy interval partial least squares (siPLS) and 12 different spectral pretreatment methods were developed for the parameters optimization of these Q-markers in Flos Chrysanthemi powder. After comparing the relationship between Q-marker contents and anti-inflammation activity via three machine learning approaches and PLSR, back-propagation neural network (BP-ANN) displayed a more excellent non-linear fitting effect, as its $R$ for new batches reached 0.89. These results indicated that the integrated NIRS and bioactive strategy was suitable for fast quality management in Flos Chrysanthemi, and also applied to other botanical food quality control.
Abbreviations used

CA, chlorogenic acid; L-7G, luteoloside; 3,5-diCQA, 3,5-dicafeoylquinic acid; 1,5-diCQA, 1,5-dicafeoylquinic acid; 3-meto-1,5-diCQA, 3-methoxyoxalyl-1,5-dicafeoylquinic acid; A-7G, apigenin-7-O-β-D-glucoside; L-7GM, luteolin-7-O-6-malonylglucoside; BJ, Boju; CJ, Chuju; GJ, Gongju; HBJ, Hangbaiju; HJ, Huaiju

Keywords

Flos Chrysanthemi, near infrared spectroscopy, anti-inflammatory activity, quality control, quality markers
1. **Introduction**

Herbal medicines (HMs) contain the most complicated chemical constituents. Accordingly, the quality evaluation of HMs is difficult. Currently, to ensure the safety and efficacy of HMs, fingerprinting combined with chemometrics [1, 2] and representative components determination [3] are two common strategies. However, people are much more interested in the holistic effects of HMs; therefore, concentrating on the numerous constituents will lead to incredible conclusions. The lack of correlation analysis with integral bioactivity evaluation is a major challenge for HM quality management [4]. To address this issue, different attempts have been made. For example, the quantitative composition-activity relationship (QCAR) study makes it possible to discover active components in HMs and to predict the holistic bioactivity by its chemical composition [5-7]. In addition, fingerprint-activity relationship analysis is an emerging approach to facilitating the bioactivity relevant to the quality control of HMs [8-10]. The benefits of HMs are generally considered to elicit effects *via a synergistic mode* by multi-compounds and multi-targets [11, 12]. Thereof, some novel strategies were provided to uncover the key compounds that contribute to certain pharmacological effects [13-15].

Unfortunately, these analytical methods were almost solely based on chromatography or mass spectrometry, but the efficiency correlation analysis with the spectral information is easily ignored. As a representative spectrum technology, near-infrared spectroscopy (NIRS), has advantages over other analytical techniques in that it is a fast, easily used, and low-cost method that records spectra of solid and liquid samples with
a huge information and allows for the quantification of multiple components. Additionally, NIRS has been widely applied to fast quality control on raw materials, active pharmaceutical ingredients, and in the food industry. Currently, the most critical concern is whether the ingredients determined by the spectral analysis can be correlated with the function. In fact, some analysts have been aware of the challenge, so they began to utilize NIRS to evaluate the total antioxidant capacity in the green tea, apples and quinoa [16-18]. It was easily found that this evaluation index mostly focused on the antioxidant capacity in vitro with PLSR, such as vitamin E, polyphenol and flavonoids, and the results indicated that antioxidant activity with these bioactive components can be predicted with linear algorithm-PLSR [5, 7]. However, the non-linear bioactivities such as immunoregulation and analgesic effects were rarely been studied [6, 19].

_Flos Chrysanthemi_ (FC) is a medicinal and edible cognate that is the flower of _Chrysanthemum morifolium_ Ramat. Modern pharmacological research shows that FC has broad pharmacological effects. During its long evolution, eight main varieties of FC have been developed, of which Hangbaiju (HBJ), Boju (BJ), Chuju (CJ) and Gongju (GJ) are the four standard varieties cited in the Chinese Pharmacopeia[20]. Caffeoylquinic acids, dicaffeoylquinic acids and flavonoids are the major ingredients of FC. Moreover, it has been reported that the extract of _Chrysanthemum indicum_ L. exerts anti-inflammatory properties by suppressing MAPK and NF-κB-dependent pathways [21]. Chlorogenic acid (CA) can suppress LPS-induced COX-2 expression by attenuating the activation of NF-κB-dependent pathways and JNK/AP-1 signaling
pathways [22].

To reveal the complex spectrum-activity relationship, the relationship between the NIRS and anti-inflammation activities of FC samples was introduced as a model. NF-κB which is involved in the early immune response and the synthesis of cytokines and chemokines, was selected as a bioactive index [23]. Heat map analysis combined with HPLC/Q-TOF-MS identification was carried out for the clustering of the fingerprints of FC samples. As seen by combining these results with our previous pharmacological study[19], some NF-κB inhibitors (including caffeoylquinic acid, dicaffeoylquinic acid and flavone derivatives) were selected for NIRS quantitative analysis. After the limits of detection (LODs) evaluation by PLS-NIRS quantification, six Q-markers were selected to construct the relationship between ingredients and anti-inflammation activity via three non-linear machine learning approaches and one linear PLSR methods. Then, the established spectrum-activity analysis strategy was used successfully to evaluate a new batch of FC samples.

2. Materials and methods

2.1. Chemicals and Reagents

Reference standards of CA, L-7G, and 3,5-diCQA were purchased from the Chinese Institute for the Control of Pharmaceutical and Biological Products (Beijing, China). The purities of all of the standards were greater than or equal to 98%. HPLC grade acetonitrile, phosphoric acid and formic acid were purchased from Merck (Darmstadt, Germany). Ultrapure water was prepared with a Milli-Q purification system (Millipore,
Bedford, MA, USA). All other reagents were of analytical grade and were purchased from Yifang S&T (Tianjin, China). A total of 55 samples from five cultivars [BJ, CJ, GJ, HBJ and Huaiju (HJ)] were purchased from AnGuo herb markets in China. All of these samples were identified as *Chrysanthemum morifolium* Ramat. by Professor Tiejun Zhang from the Tianjin Institute of Pharmaceutical Research. The voucher specimens are deposited in the Composite Drugs and Systems Biology Lab at Nankai University.

### 2.2. Sample Preparation

All sample materials were finely pulverized and filtered through a 100-mesh sieve. The dried powder (1 g) was extracted with 100 mL of a methanol-water (7:3, v/v) solution using an ultrasonic extraction apparatus (40 kHz, 500 W, Ningbo, China) for 30 min at room temperature. The extracts were centrifuged at 12,000 rpm for 10 min, and the supernatant was used for HPLC analysis and Q-TOF-MS identification. Primary stock solutions of the three reference compounds CA, L-7G and 3,5-diCQA at final concentrations of 96.8, 107.6 and 315.2 µg/mL, respondingly, were prepared by dissolving the accurately weighed reference compounds in a methanol-water (7:3, v/v) solution. All of the solutions were stored at 4°C and were brought to room temperature before use.

### 2.3. HPLC Analysis

HPLC analysis was performed using a Shimadzu 20A HPLC system (Shimadzu Co., Kyoto, Japan). Chromatographic separations were conducted using an Agilent Eclipse Plus C_{18} column (100 mm×4.6 mm, 3.5µm) operated at 25°C. The mobile phase system
was acetonitrile (A) and water with 0.1% phosphoric acid (B) at a flow rate of 1.0 mL/min, and gradient elution was completed as follows: 0-5 min, 10-28% (v/v) A; 5-12 min, 28-32% (v/v) A; 12-20 min, isocratic 32% (v/v) A; 20-35 min, 32-80% (v/v) A; 35-36 min, 80-100% (v/v) A; 36-39 min, isocratic 100% (v/v) A; 39-41 min, 100-10% (v/v) A; 41-45 min, isocratic 10% (v/v) A. The compounds were detected by UV spectroscopy at 348 nm. The sample injection volume was 10 µL.

2.4. HPLC/Q-TOF-MS Analysis

A Waters Acquity system (Waters Co., Milford, MA, USA) equipped with a photodiode array detector (PDA) and a Waters Q-TOF Premier Mass Spectrometer with an electrospray ionization system (Water MS Technologies, Manchester, UK) were used for sample analysis. Data acquisition was performed with the help of MassLynx V4.1 software (Waters Co., USA). Separations were performed using the same chromatographic column as mentioned in section 2.3. The mobile phase consisted of acetonitrile (A) and water with 1% formic acid (B) at a flow rate of 0.4 mL/min and the same gradient as described in section 2.3 were used for analysis. The compounds were detected by PDA scanning from 190 to 450 nm. The sample injection volume was 5 µL. These conditions were used to reproduce the separation achieved with the HPLC method described in section 2.3. The ESI-MS spectra were acquired in both the positive and negative ion modes. The conditions for ESI-MS analysis were as follows: the capillary voltage was set to 3.0 kV for positive mode and 2.5 kV for negative mode, the sample cone voltage was set to 30 V, the desolvation gas flow was set to 600 L/h at 350°C, the cone gas was set to 50 L/h, and the source temperature was 110°C. The Q-
TOF Premier acquisition rate was 0.1 s with a 0.02 s inter-scan delay. The MS spectra were acquired from 100 to 1500 Da. Leucine enkephalinamide acetate was used as the lock mass ($m/z$ 555.2931 in ESI$^+$; $m/z$ 553.2775 in ESI$^-$) at a concentration of 200 ng/mL and a flow rate of 0.2 µL/min. The MS/MS analyses were used to obtain the mass fractions of target ions.

2.5. NIRS Collection

NIRS were acquired using a Brucker TENSOR 37 FT-NIR spectrometer (Bruker Optik, Ettlingen, Germany) with an InGaAs detector and an integrating sphere module over the 12000 to 4000 cm$^{-1}$ spectral range. The spectra were acquired with the OPUS spectral acquisition software (Bruker Optik, Ettlingen, Germany) and collected at a resolution of 8 cm$^{-1}$ per spectrum by averaging 64 scans. Each sample was tested three times, and an average spectrum was taken to reduce the instrument noise. To reduce the error of operation, a uniform glass vessel (diameter: $d = 20$ mm, height: $h = 50$ mm) was used to load the samples (approximately 1.0 g of powder covering the bottom of the vessel) for NIRS measurement.

2.6. Multivariate statistical analysis

2.6.1. NIRS Analysis

To avoid sample selection bias, sample set partitioning based on the joint x-y distances (SPXY) algorithm was used to split the dataset into calibration and validation sets. A total of 32 samples were used for the calibration set, and the remaining 15 samples comprised the validation set. Preprocessing methods, such as non-pretreatment, autoscaling (AUTO), multiple-scatter correction (MSC), standard normal variable
transformation (SNV), de-trend (DT), DT+SNV, smooth, Savitzky-Golay first derivative (S/G 1st der), Savitzky-Golay second derivative (S/G 2nd der), S/G 1st der + SNV, S/G 1st der + MSC and S/G 1st der + DT were used to process the FC data. The synergy interval partial least squares (siPLS) method was used for variable selection in NIRS. The spectral preprocessing methods, spectra intervals and PLS components of the optimum model were applied for the remaining 15 samples. In this process, the root mean square error of prediction (RMSEP), the corresponding R (RMSEP) and residual predictive deviation (RPD) were used to validate the stability of the model. All algorithms were implemented in Matlab 2013b (MathWorks, Natick, MA, USA) using Windows 8.1.

2.6.2. Quantitative HotMap Analysis

A quantitative heat map was used to display the clustering effect using the 14 major common peak areas of the chromatographic fingerprint. The hierarchical cluster analysis of the quantitative heat map was also generated using Matlab 2013b (MathWorks, Natick, MA, USA) under Windows 8.1. The process was conducted as follows: after vector normalization of the 14 major common peak areas in the chromatographic fingerprints from 47 batches of FC samples, a hierarchical cluster tree of the peak areas for the major components was created by defining the linkage function as the ward method, and the distance between the samples was computed using the Euclidean distance formula.

2.6.3. The interval LODs theory

The theory of interval LODs used in this article was briefly described below. A more
detailed description and the process used to derive the mathematical equations was published [24].

The overall LOD formula against both type I and type II errors:

\[
\text{LOD} = \left( t_{\alpha,\gamma} + t_{\beta,\gamma} \right) \text{var}(y_0)^{1/2}
\]

Eq. (A.1)

where \( \text{var}(y_0) \) is the concentration variance for a blank sample. \( t_{\alpha,\gamma} \) and \( t_{\beta,\gamma} \) take into account the probability of making type I errors and type II errors with a probability \( \alpha, \beta \) and with the \( \gamma \) degrees of freedom. In the application, \( t_{\alpha,\gamma} + t_{\beta,\gamma} \) takes the approximate value of 3.3. The key point in regard to Eq. (A.1) is the criterion adopted for estimating the variance of the predicted concentration. The basic assumption throughout the interval LOD theory is the well-known expression:

\[
\text{var}(y) = \|b\|^2 \text{var}(x) + \left( h + \frac{1}{I} \right) \|b\|^2 \text{var}(x) + \left( h + \frac{1}{I} \right) \text{var}(y_{\text{cal}})
\]

Eq. (A.2)

To estimate the LOD, the value of \( \text{var}(y_0) \) is required. In this regard, the leverage when the analyte concentration \( (h_0) \) is zero plays a fundamental role. Even if blank samples are not included in the calibration set, the LODs interval theory aims to find the limits of \( H_0 \) in score space. In multivariate calibration, \( h_0 \) assumes different values depending on the sample background composition. Therefore, it is more reasonable to consider the existence of LOD intervals:

\[
\text{LOD}_{\text{min}} = 3.3 \left[ \|b\|^2 \text{var}(x) + \left( h_{0\text{min}} + \frac{1}{I} \right) \|b\|^2 \text{var}(x) + \left( h_{0\text{min}} + \frac{1}{I} \right) \text{var}(y_{\text{cal}}) \right]^{1/2} \text{Eq.}(A.3)
\]

\[
\text{LOD}_{\text{max}} = 3.3 \left[ \|b\|^2 \text{var}(x) + \left( h_{0\text{max}} + \frac{1}{I} \right) \|b\|^2 \text{var}(x) + \left( h_{0\text{max}} + \frac{1}{I} \right) \text{var}(y_{\text{cal}}) \right]^{1/2} \text{Eq.}(A.4)
\]

2.7. Machine learning methods

To determine the components-activity relationship of the \( FC \), the dataset was randomly divided into two subsets: a training set of 37 samples and a test set of 10 samples. The
training set was used to calculate the model parameter, while the test set was used to evaluate its generalization ability and to monitor the training process to reduce the overfitting. To determine the optimal components-activity relationship, various algorithms including RF, nu-SVR and BP-ANN were introduced into the components-activity regression analysis. The $RMSE$ for training, $RMSEP$ for testing and their respective correlation coefficient $R$ were used as the evaluation criterion.

2.7.1. Random Forest Regression (RF)

RF is an ensemble learning method that contains a large number of decision trees. In this study, the R (version 3.0.1) Random Forest package (version 4.6-7) in regression mode was used to grow the trees with the training data. Here, we operate the RF with the default parameters; ntree (number of trees to grow) is 500 as its performance does not generally improve significantly beyond this threshold. In contrast, $m_{try}$ (number of variables randomly sampled as candidates at each split) has some influences on the performance and thus constitutes the only tuning parameter of the RF algorithm.

2.7.2. nu-Support Vector Regression (nu-SVR)

The principle of SVR is based on the structural risk minimization, which minimizes an upper bound of the generalization error. The basic idea of SVR is to map the input variables into a high dimensional feature space where they are linearly correlated with the output variable. To solve the curse of dimensionality, the kernel function (KF) is introduced and realize the nonlinear transformation. In this study, the release (Version 3.21) of libsvm by Chih-Jen Lin [25] in Matlab 2013b was used to realize the nu-SVR algorithm which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. In detail, a
radial basis function such as the KF with the parameter $\gamma$ and the legalization argument $C$ were optimized with the cross-validation for the training set in terms of RMSECV. During this process, keeping the value of $\nu$ equal to 0.5, for $C$, a range of parameter settings from 0.5 to 50 with 0.2 steps and for $\gamma$, a range from 0.5 to 50 with 0.5 steps were searched in detail.

2.7.3. Back propagating - artificial neutral network (BP-ANN)

The feed-forward BP-ANN is a supervised ANN learning technique and can be realized using single-layer as well as multi-layer networks, which is particularly effective in modeling complex non-linear systems [26, 27]. In this paper, the weights and bias values in the BP-ANN topology were updated with the resilient backpropagation algorithm (Rprop) using Matlab 2013b. One hidden layer was introduced into the BP-ANN structure, and the tan-sigmoid was selected as the activation function from the input layer to the hidden layer; a linear function was selected as the activation function from the hidden layer to the output layer. Overfitting to the training data was prevented by restricting the optimum number of nodes in the hidden layer, which was optimized between 1 and 30 in this experiment. A normalization procedure was necessary for training the BP-ANN. Therefore, the input and target variables were processed by mapping the minimum and maximum values from -1 to 1. At the same time, the methodology of ‘early stopping by cross-validation’ was applied to prevent overfitting with the 10 test samples.

2.8. Relative significance of the inputs

The relative significance of each input in the BP-ANN was established by evaluating
the partitioning of the BP-ANN connection weights using the Garson’s modified algorithm[27]. The calculation sequence was the following:

\[ C_{ij} = W_{ij} K_j \]  \hspace{1cm} (Eq.B.1)

\( C_{ij} \) is the contribution from each input node \((i)\) to the output node via hidden node \((j)\), \(W_{ij}\) is the weight matrix from the input node \((i)\) to the hidden node \((j)\), and \(K_j\) is the weight matrix from the hidden node \((j)\) to the output node.

\[ r_{ij} = \frac{|C_{ij}|}{\sum_{i=1}^{n} |C_{ij}|} \]  \hspace{1cm} (Eq.B.2)

\( r_{ij} \) is the relative contribution from each input node \((i)\) to the output node via hidden node \((j)\). \(n\) is the number of input node.

\[ RI_i = M_i \sum_{j=1}^{m} r_{ij} / \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \]  \hspace{1cm} (Eq.B.3)

\( RI_i \) is the relative importance of each input variable \((i)\). \(M_i\) is the mean concentration of the \(i\)th component.

\[ RI_i = RI_i / \sum_{i=1}^{n} RI_i \times 100\% \]  \hspace{1cm} (Eq.B.4)

The sum of \(RI_i\) is normalized to 100%.

2.9. Anti-inflammation activity analysis via NF-κB inhibition

To evaluate the anti-inflammation effects on NF-κB activation, an NF-κB-dependent luciferase assay was carried out according to our previous report [19]. HEK 293 cells were co-transfected in 96-well plates with NF-κB luciferase reporter plasmid pGL4.32 and Renilla luciferase reporter vector plasmid pRL-TK (Fitchburg, WI, USA) at 100 and 9.6 ng per well, respectively. Transfection was performed for 24 h using Lipofectamine 2000 according to the manufacturer’s instructions (Carlsbad, CA, USA). Cells were pretreated with drugs or ultrasonic extraction before human TNF-α
stimulation (10 ng/mL) for 6 h. After stimulation, cells were lysed and assayed for luciferase activity using a Promega dual-luciferase reporter assay system.

3. Results and discussion

3.1. Fingerprints and HotMap Analysis of FC

As shown in Fig. 1A, the typical HPLC-UV chromatograms of five cultivars of FC extracts shared similar fingerprints. To further assess the quality characteristics between the chromatograms, a hierarchical agglomerative clustering analysis of 47 batch samples was performed based on the relative peak areas of the 14 common chromatographic peaks determined at 338 nm. Then, a 47 × 14 matrix was formed, and the data were further analyzed and presented as a heat map, as shown in Fig.1B. Three of the medicinal cultivars (HJ, CJ and BJ) showed rough differences from the two edible cultivars (HBJ and GJ) identified as the common ingredients of flavonoid, glycosides and caffeoylquinic acid derivatives. Then, these major compounds were selected as potential Q-markers and investigated in the following study.

3.2. Identification of the Q-markers

To characterize the Q-marker compounds, extracts of HBJ (HBJ-1), a representative cultivar that contained all 14 common peaks, were subjected to HPLC-integrated Q-TOF-MS analysis. The typical HPLC-UV/MS fingerprints are shown in Fig. 1C. Fourteen specific peaks were characterized using the Waters PDA detector. By comparing the retention times, UV spectra and m/z of the characteristic molecular and fragment ions with those of the three standards CA, L-7G and 3,5-diCQA, the other 11
peaks were tentatively identified. Detailed information was presented in Table 1. As shown in Fig. 1D, the identified compounds in FC extracts belonged to three categories of skeleton, namely, caffeoylquinic acids, dicaffeoylquinic acids and flavonoids, and consistent with previously published reports [28, 29]. Therefore, the standards of CA, L-7G, and 3,5-diCQA, which represented different skeleton compounds, were used for the HPLC quantitative methodological study. The method validation parameters are listed in Table 2. The calibration curves exhibited good linearity ($R^2 > 0.9997$) with wide linear ranges, and the RSD values of repeatability were below 5%. Recovery tests were conducted by addition of precise amounts of CA, L-7G and 3,5-diCQA standards to 0.1 g of FC sample powders. The obtained results showed that the recoveries of these three compounds were more than 95%, indicating that the reference values of CA, L-7G and 3,5-diCQA were reliable for use in the NIRS calibration. Additionally, they could act as the common references for multiple evaluations of these three categories of compounds.

3.3. Q-markers selection and Quantification by PLSR Algorithm

3.3.1. The interval LODs and boxplot analysis for Q-marker selection

For samples with complex chemical compositions, higher detection limits and lower sensitivities were generally observed in NIRS than with the HPLC method. We anticipated that the measurement of low abundance compounds in the FC would be difficult. Therefore, three types of compounds, CA, L-7G and 3,5-diCQA were used to calculate their corresponding LODs firstly. And the interval LODs theory[24] was introduced to calculate the $\text{LOD}_{\text{min}}$ and $\text{LOD}_{\text{max}}$ using Eq.(A.3-4). As shown in Table 3,
LOD\textsubscript{mijn} depended on the range of concentrations, and it had the potential to decrease with the more lower range of concentrations in the training test and a smaller number of calibration latent variables. However, LOD\textsubscript{max} values, which considered the variability of the background composition, were not all below 0.09%. This result implied that the NIRS for quality control of the HMs was not suitable for content lower than 0.1%. As shown in Fig. 2, the boxplot displayed the different concentration distributions of these compounds. Six compounds that were detected at higher levels among the different categories of compounds and their medians were higher than their corresponding LOD\textsubscript{max}. These compounds were selected as Q-markers. In addition, all six Q-markers had anti-inflammatory properties [19] and were fit for quality control of FC.

3.3.2. The optimum selection of preprocessing methods, wavenumber regions and LV numbers for calibration models

In general, spectral pretreatment methods, the number of latent variables (LV) and variable selection were the three critical parameters for improving PLS model performance. In this study, 12 different spectral pretreatment methods were used to reduce the noise in the raw spectra. Due to the difference in concentration or structure of the Q-markers, the preprocessing methods for the six Q-markers were dissimilar (Table 4). For example, CA and 3,5-diCQA, which had relatively higher contents, would increase their predicted power by the scatter corrections (DT + SNV, SNV). These scatter correction methods were designed to reduce the physical variability between samples due to scatter and baseline shifts between samples. For others, spectral
derivatives (S/G 1st der and smooth) were employed to remove the baseline and increase the resolution. The siPLS method was used to build the calibration models. This method was an all-possible-interval-combinations procedure test based on all possible PLS on all subsets of intervals and reported the subset of sets giving the lowest RMSECV or RMSEP. The spectrum was first divided into 20 equidistant subintervals, and the combinations of one, two or three intervals were selected to build the calibration model (Fig. 3, wavenumber selection). The most suitable wavenumber spans were within the common NIRS region of 4000-9000 cm⁻¹, where the vibration information was rich enough to determine the contents of the different compounds, especially for the natural products [30]. After the preprocessing was performed on the raw spectra, different wavenumber combinations were fitted using the PLS algorithm for each Q-marker. The maximum number of LV was set to 20. The Leave-One-Out cross-validation (LOOCV) method was used to choose the optimum number of LV. The optimum LV number was determined by the lowest RMSECV and the highest $R^2$ (RMSECV). Higher LV numbers led to a better predictive power of the model but increased the risk of over-fitting. Therefore, we balanced the RMSECV and the risk of over-fitting and chose reasonable LV to build a more robust model for prediction of the unknown sample (Fig. 3, latent variable selection). The parameters for the most suitable conditions for calibration and validation are shown in Table 4. However, the LVs of CA and L-7GM used for prediction of the samples were a little higher. This could be caused by the complex composition and trace amount in the herbs. All of the compounds exhibited a perfect predictive effect during the calibration. However, during validation in the LOOCV
method, such a predictive effect was not achieved during the calibration. Predictive power decreased with the decreased content of the compounds in the raw herb. The predictive effects for these three main compounds (CA, L-7G and 3,5-diCQA) and the other ingredients are shown in Fig. 3 and the supplementary data (Fig. S1), respectively. The relatively poor performance could be attributed to the heterogeneity of the sample set and smaller sample size. With this designed model, the remaining 15 samples acted as the external validation to test the robustness of the model and this result is listed in Table S1.

3.4. The comparison of machine learning approaches for activity fitting

The activities of the 47 batches of FC extracts were evaluated at the cellular level by using the luciferase reporter assay system. The RLU ratio of NF-κB was used to evaluate their anti-inflammation activity. To reveal the relationship between NIR spectra and anti-inflammation activity via NF-κB inhibition, three machine learning machine methods (RF, nu-SVR, BP-ANN) and PLSR were carried out for activity prediction with six Q-markers’ contents. As shown in Table 5, the capacity of BP-ANN, RF and nu-SVR were similarly excellent, and BP-ANN was distributed more closer to the parity line (y = x) (Fig S2). Due to its incapability for nonlinear approximation, the PLSR method was a poor model for anti-inflammation activity. Therefore, the BP-ANN model (three-layered configuration of 6-9-1 nodes) was selected to build the relationship between the Q-markers and holistic effect in the following test. The relative importance of the contributions for the concentrations of the various Q-markers were determined through a systematic analysis of the partitioning of the BP-ANN connection.
weights. Their contributions to the anti-inflammation activity were ranked as follows: CA (29.4%) > 3,5-diCQA (29.0%) > 1,5-diCQA (19.3%) > L-7GM (8.0%) > L-7G (7.4%) > A-7G (6.9%). CA was the most influential among these six Q-markers, although its concentration was not the highest. In terms of the categories of compounds, the CA and caffeoylquinic acid derivatives displayed better anti-inflammation activity than flavonoid glycosides.

To validate whether these six Q-markers could represent all 14 major compounds, their absolute contents, determined by HPLC, were used to build the relationship with activity. Finally, a three-layered BP-ANN configuration of 14-8-1 nodes was constructed. The relative impact value of individual components ranged from 1.3% to 33.3%, and the sum contributions of the six Q-markers were all above 76% (Fig. 4 right). Considering the excellent correlation between experimental value and predicted value, calculated by these six Q-markers, illustrated that these six Q-markers made major contributions to anti-inflammation. However, to further validate that the selection of these Q-markers was appropriate, new batches of FC samples needed to be tested to determine the robustness of the model.

3.5. Model performance and robustness

To further validate the reliability of using the spectrum-activity strategy for quality management of FC, an external validation experiment was performed finally. Another eight batches of FC samples were examined to validate the integral anti-inflammation effects via the prediction of six Q-markers’ contents by NIRS. As shown in Fig. 5, the established BP-ANN model displayed an excellent non-linear fitting effect, and their corresponding correlation coefficient was 0.89 (Fig. 5 & Table S2). This result illustrated that the integrated spectrum-activity strategy was robust and capable of being implemented for quality control of the unknown samples.
4. Conclusion

In this paper, an integrated spectrum-activity strategy was provided for fast quality management of HMs. The techniques used included HPLC/Q-TOF-MS, heat map clustering and biological evaluation to select the Q-markers from different cultivars of FC. NIRS technology which directly detected FC solid powder, was used to analyze the Q-marker content. Moreover, BP-ANN was introduced to solve the complex relationship between the Q-markers and the holistic anti-inflammation activity. The advances in technology demonstrated that an integrated approach based on NIRS not only was capable of multi-compounds quantification, but also could predict integral bioactivity for HMs quality control.

SUPPORTING INFORMATION DESCRIPTION

NIRS method parameter optimal process and reference VS predicted value of the quality ingredients; Figures of the predictive effect from the remaining 3 crucial bioactive marker ingredients and the model robustness to predict anti-inflammation activity of the new batches of Flos Chrysanthemi samples.

Notes

The authors declare no competing financial interest.

Acknowledgements

This work was supported by a Grant from the National Natural Science Foundation of China (Nos. 81430095; 81373506), the Specialized Research Fund for the Doctor Program of Higher Education of China (Nos. 20120031110042) and the Key Program of Natural Science of Foundation of Tianjin, China (Nos. 13JCZDJC31400).
References


Figure Captions

Fig 1. HPLC-UV fingerprint chromatograms (A) and heat map hierarchical clustering analysis (B) for the five cultivars of Flos Chrysanthemi; HPLC/Q-TOF-MS identification (C) and chemical structures (D) of these principal Q-markers in FC extract.

Fig 2. Boxplot analysis of Q-markers in different categories of FC. The three orange lines are the intervals LOD values for three types of compounds.

Fig 3. Predictive effect (right) of three representative Q-marker compounds with the optimal wavenumber selected numbers according to RMSECV for different pretreated methods (left) on the suitable latent variable (middle). (Panel A, CA; B, 3,5-diCQA; C, L-7G)

Fig 4. The relative importance of each input in determining the anti-inflammation activity for the six Q-markers (left) and the fourteen potential Q-markers (right), as estimated using Garson’s modified algorithm.

Fig 5. Predicted values of anti-inflammation activity plotted against the experimental values using BP-ANN
### Table 1. Identification of the common principal compounds by HPLC-PDA and HPLC-Q-Tof analysis.

<table>
<thead>
<tr>
<th>No</th>
<th>( t_R )</th>
<th>( \text{MS}^* )</th>
<th>MS/MS</th>
<th>MS/MS</th>
<th>MS/MS</th>
<th>Formula</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.49</td>
<td>355.1019</td>
<td>163</td>
<td>353.0859</td>
<td>191, 179, 173, 135</td>
<td>( \text{C}<em>{16}\text{H}</em>{18}\text{O}_9 )</td>
<td>Chlorogenic acid</td>
</tr>
<tr>
<td>2</td>
<td>14.45</td>
<td>449.1072</td>
<td>287</td>
<td>447.0944</td>
<td>285</td>
<td>( \text{C}<em>{21}\text{H}</em>{26}\text{O}_{11} )</td>
<td>Luteoloside</td>
</tr>
<tr>
<td>3</td>
<td>14.69</td>
<td>517.1367</td>
<td>499, 163</td>
<td>515.1190</td>
<td>353, 335, 191, 179, 173</td>
<td>( \text{C}<em>{25}\text{H}</em>{24}\text{O}_{12} )</td>
<td>3,4-Dicaffeoylquinic acid</td>
</tr>
<tr>
<td>4</td>
<td>15.51</td>
<td>517.1354</td>
<td>499, 163</td>
<td>515.1175</td>
<td>353, 191, 179</td>
<td>( \text{C}<em>{25}\text{H}</em>{24}\text{O}_{12} )</td>
<td>3,5-Dicaffeoylquinic acid</td>
</tr>
<tr>
<td>5</td>
<td>15.84</td>
<td>517.1357</td>
<td>499, 163</td>
<td>515.1167</td>
<td>353, 191, 179</td>
<td>( \text{C}<em>{25}\text{H}</em>{24}\text{O}_{12} )</td>
<td>1,5-Dicaffeoylquinic acid</td>
</tr>
<tr>
<td>6</td>
<td>16.67</td>
<td>603.1367</td>
<td>585, 423, 163</td>
<td>601.1184</td>
<td>557, 515, 439, 395, 233</td>
<td>( \text{C}<em>{28}\text{H}</em>{26}\text{O}_{15} )</td>
<td>3-Methoxyxalyl-1,5-dicaffeoylquinic acid</td>
</tr>
<tr>
<td>7</td>
<td>19.59</td>
<td>517.1346</td>
<td>499, 163</td>
<td>515.1167</td>
<td>353, 191, 179, 173</td>
<td>( \text{C}<em>{25}\text{H}</em>{24}\text{O}_{12} )</td>
<td>4,5-Dicaffeoylquinic acid</td>
</tr>
<tr>
<td>8</td>
<td>19.59</td>
<td>433.1116</td>
<td>271</td>
<td>431.0958</td>
<td>269</td>
<td>( \text{C}<em>{21}\text{H}</em>{20}\text{O}_{10} )</td>
<td>Apigenin-7-O-β-D-glucoside</td>
</tr>
<tr>
<td>9</td>
<td>22.49</td>
<td>535.1089</td>
<td>449, 287</td>
<td>533.0943</td>
<td>489, 285</td>
<td>( \text{C}<em>{26}\text{H}</em>{22}\text{O}_{14} )</td>
<td>Luteolin-7-O-6-malonylg glucoside</td>
</tr>
<tr>
<td>10</td>
<td>26.74</td>
<td>519.1126</td>
<td>271</td>
<td>517.1005</td>
<td>473, 269</td>
<td>( \text{C}<em>{24}\text{H}</em>{22}\text{O}_{13} )</td>
<td>Apigenin-7-O-6-malonylg glucoside</td>
</tr>
<tr>
<td>11</td>
<td>27.37</td>
<td>549.1238</td>
<td>301</td>
<td>547.1100</td>
<td>503, 299</td>
<td>( \text{C}<em>{25}\text{H}</em>{24}\text{O}_{14} )</td>
<td>Diosmetin-7-O-6-malonylg glucoside</td>
</tr>
<tr>
<td>12</td>
<td>28.33</td>
<td>593.1877</td>
<td>447, 285</td>
<td>591.1705</td>
<td>283</td>
<td>( \text{C}<em>{26}\text{H}</em>{22}\text{O}_{14} )</td>
<td>Linarin</td>
</tr>
<tr>
<td>13</td>
<td>29.27</td>
<td>447.1273</td>
<td>285</td>
<td>491.1192</td>
<td>283</td>
<td>( \text{C}<em>{22}\text{H}</em>{22}\text{O}_{10} )</td>
<td>Acacetin-7-O-β-D-glucoside</td>
</tr>
<tr>
<td>14</td>
<td>31.05</td>
<td>533.1271</td>
<td>285</td>
<td>531.1168</td>
<td>487, 283</td>
<td>( \text{C}<em>{25}\text{H}</em>{24}\text{O}_{13} )</td>
<td>Acacetin-7-O-6-malonylgactoside</td>
</tr>
</tbody>
</table>
Table 2. The methodology parameters and the calibration curves of the HPLC analysis

<table>
<thead>
<tr>
<th>Compounds</th>
<th>$t_R$(min)</th>
<th>Linear ranges (µg/mL)</th>
<th>Calibration curves</th>
<th>$R^2$</th>
<th>LOD (µg/mL)</th>
<th>Repeatability (RSD%, $n=6$)</th>
<th>Recovery (%,$n=3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>8.6</td>
<td>4.8-96.8</td>
<td>$y=9373x+2319$</td>
<td>0.9998</td>
<td>0.32</td>
<td>0.6</td>
<td>99.5</td>
</tr>
<tr>
<td>L-7G</td>
<td>15.6</td>
<td>5.4-107.6</td>
<td>$y=15434x+3638$</td>
<td>0.9997</td>
<td>0.53</td>
<td>0.9</td>
<td>101.6</td>
</tr>
<tr>
<td>3,5-diCQA</td>
<td>17.6</td>
<td>15.8-315.2</td>
<td>$y=12705x+15423$</td>
<td>0.9998</td>
<td>1.22</td>
<td>1.2</td>
<td>96.2</td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>3,5-diCQA</td>
<td>L-7G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concentration range (%, 10 mg/g)</td>
<td>0.21-1.07</td>
<td>0.39-2.43</td>
<td>0.01-0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LV</td>
<td>14</td>
<td>11</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(x)½</td>
<td>3.9 × 10⁻³</td>
<td>1.4 × 10⁻³</td>
<td>0.1 × 10⁻³</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(ycal)</td>
<td>0.01</td>
<td>0.04</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOD&lt;sub&gt;min&lt;/sub&gt; (%, 10 mg/g)</td>
<td>0.08</td>
<td>0.24</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOD&lt;sub&gt;max&lt;/sub&gt; (%, 10 mg/g)</td>
<td>0.10</td>
<td>0.31</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOD&lt;sub&gt;ref&lt;/sub&gt; (%, 10 mg/g)</td>
<td>3.2 × 10⁻³</td>
<td>1.2 × 10⁻³</td>
<td>5.3 × 10⁻³</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# LOD<sub>ref</sub> (%, 10 mg/g) is the LOD from the HPLC results.
Table 4. Most suitable conditions for the calibration and validation of the six Q-marker ingredients

<table>
<thead>
<tr>
<th>Compounds</th>
<th>Pretreated methods</th>
<th>Interval number</th>
<th>LV</th>
<th>$R_{cal}$</th>
<th>RMSEC</th>
<th>$R_{val}$</th>
<th>RMSECV</th>
<th>Mean concentration (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>DT+SNV</td>
<td>15 16 19</td>
<td>14</td>
<td>0.99</td>
<td>0.02</td>
<td>0.94</td>
<td>0.04</td>
<td>0.45</td>
</tr>
<tr>
<td>3,5-diCQA</td>
<td>SNV</td>
<td>3 15 17</td>
<td>11</td>
<td>0.99</td>
<td>0.02</td>
<td>0.96</td>
<td>0.08</td>
<td>0.87</td>
</tr>
<tr>
<td>1,5-diCQA</td>
<td>one-DC</td>
<td>15 17</td>
<td>10</td>
<td>0.99</td>
<td>0.02</td>
<td>0.95</td>
<td>0.07</td>
<td>0.41</td>
</tr>
<tr>
<td>L-7G</td>
<td>Smoothing</td>
<td>4 17 18</td>
<td>11</td>
<td>0.99</td>
<td>0.02</td>
<td>0.94</td>
<td>0.05</td>
<td>0.30</td>
</tr>
<tr>
<td>A-7G</td>
<td>DT</td>
<td>2 6 15</td>
<td>7</td>
<td>0.99</td>
<td>0.01</td>
<td>0.91</td>
<td>0.04</td>
<td>0.15</td>
</tr>
<tr>
<td>L-7GM</td>
<td>one-DC+MSC</td>
<td>5 20</td>
<td>15</td>
<td>0.99</td>
<td>0.01</td>
<td>0.92</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>Machine learning Algorithm</td>
<td>parameters</td>
<td>$RMSE$</td>
<td>$R_{RMSE}$</td>
<td>$RMSEP$</td>
<td>$R_{RMSEP}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------</td>
<td>--------</td>
<td>------------</td>
<td>---------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLSR</td>
<td>$l_v=4$</td>
<td>23</td>
<td>25%</td>
<td>15</td>
<td>32%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nu-SVR</td>
<td>$C=16.7$; $g=28$</td>
<td>17</td>
<td>87%</td>
<td>12</td>
<td>84%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP-ANN</td>
<td>node=9</td>
<td>10</td>
<td>91%</td>
<td>8.9</td>
<td>87%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>$m_{to}=5$</td>
<td>11</td>
<td>96%</td>
<td>10</td>
<td>87%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，
提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。
图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：
图书馆首页 文献云下载 图书馆入口 外文数据库大全 疑难文献辅助工具