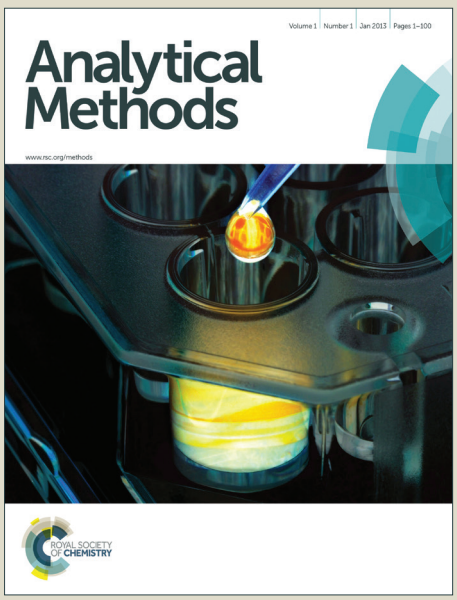
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Identification of heavy metal-contaminated *Tegillarca granosa* using infrared spectroscopy

Xiaojing Chen[a], KeLiu[a], JingboCai[c], DehuaZhu[b], Huiling Chen[a*]

[a] College of Physics and Electronic Engineering Information, Wenzhou University

[b] College of Mechanical& Electrical Engineering, Wenzhou University

[c] Zhejiang Key Laboratory of Exploitation and Preservation of Coastal Bio-resource, Zhejiang Mariculture Research Institute

* Corresponding author. Tel.: +86 577 86689027; fax +86 577 86689027

E-mail: Huiling@wzu.edu.cn (Huiling Chen*)

Abstract: This study explored the feasibility of using infrared spectroscopy for the rapid detection of heavy metal contamination in *Tegillarca granosa*. Generally, there is no specific characteristic peak of heavy metals in the infrared range. Nevertheless, these are some changes in the structure and concentration of relevant biological molecules induced by heavy metal contamination produce remarkably weak spectral information on heavy metals. In this study, we selected characteristic infrared spectral variables to obtain heavy metal information using the Competitive Adaptive Reweighted Sampling method, Successive Projection Algorithm and Genetic Algorithm. The selected variables served as inputs for classification algorithm to construct two classification models. One model was designed to classify *Tegillarca granosa* samples that were uncontaminated (healthy) and contaminated by a certain heavy metal (Cu, Cd, Pb, or Zn) (Design I). The other model was designed to classify all sample varieties, including the samples that were uncontaminated and contaminated by the four heavy metals (Design II). The two models

Corresponding author: Huiling@wzu.edu.cn (Huiling Chen*), Fax: +86 577 86689027 Tel: +86 577 86689027,Wenzhou University Chashan University Town, Wenzhou, Zhejiang Province, People's Republic of China

were validated using10-fold cross validation. The prediction accuracy by combination of Competitive Adaptive Reweighted Sampling method and Support Vector Machine algorithm reached 95% for Design I and 92% for Design II. The results of this study indicated the potential of infrared spectroscopy in evaluating heavy metal contamination in *Tegillarca granosa*.

Keywords: Heavy metal; Infrared spectroscopy; Aquatic product; *Tegillarca granosa*; CARS; Support vector machine;

Introduction

Tegillarca granosa is a nutritious food source because of its low cholesterol content and high protein, ferrum, calcium, carbohydrate, riboflavin, and various trace element contents [1-2]. However, *Tegillarca granosa* principally thrives in tidelands close to sewages, which are highly exposed to heavy metal contamination. Given its non-selective filter-feeding behavior and low mobility, *Tegillarca granosa* can accumulate heavy metals at concentrations ten to thousand times higher than other aquatic species. The accumulated heavy metals remain in *Tegillarca granosa* for a long time, making this seafood hazardous to human health. Soluble heavy metal ions are absorbed through the gills of *Tegillarca granosa* and then distributed throughout the body through blood circulation. These ions can be accumulated in specific body parts or on surface cells; they can also be absorbed through the digestive tract during feeding. *Tegillarca granosa* absorbs soluble and particulate heavy metals. Soluble heavy metals are principally absorbed by the body surface, whereas particulate heavy metals are absorbed through the ingestion and digestion of heavy metal-contaminated food [3]. Upon entering the food chain, these heavy metals accumulate to high concentrations through bio-concentration and bio-magnification. Some metals can be easily discharged, whereas others may accumulate and affect *Tegillarca granosa* tissues. When

ingested, *Tegillarca granosa* containing toxic levels of heavy metals can cause serious harm to human health. Therefore, it is important and necessary to detect heavy metal pollution in *Tegillarca granosa*is.

At present, common methods of heavy metal detection include graphite furnace atomic absorption spectrometry, flame atomic absorption spectrometry, atomic fluorescence spectrometry, and inductively coupled plasma mass spectrometry [4-5]. However, these methods are expensive, labor intensive, complex, time consuming, and require a large sample size. Therefore, a fast, simple and reliable method that overcomes all these drawbacks is necessary to be developed to detect heavy metal contamination in *Tegillarca granosa*. In addition, *Tegillarca granosa* can reflect the level of heavy metal contamination in its surrounding environment. *Tegillarca granosa*, oyster, and other shellfish are widely used as indicator organisms of heavy metal contamination for the biological monitoring of marine pollution in numerous countries. Therefore, rapid detection of heavy metal contamination in *Tegillarca granosa*is urgently needed.

Infrared (IR) spectroscopy has been widely used in food safety and quality control because of its rapidity, simplicity, high precision, low maintenance cost and small sample size requirement [6-10]. IR spectroscopy detects the vibrational and rotational energies of molecules within the IR spectrum by measuring the absorbance spectrum of hydrogen bonding. This technology also gathers useful information that reflects organic molecules such as proteins, lipids and sugar in biological tissues. IR spectroscopy is an important strategy for the structural analysis of organic compounds [11-12]. However, heavy metals generally do not any show IR activity and barely have any characteristic peak in the IR spectrum. Heavy metals only indirectly change the vibrational spectrum by inhibiting antioxidant enzymes or by inducing the synthesis of

detoxification proteins in large quantities. These phenomena change the structures and concentrations of relevant biological molecules [11, 13-17]. Therefore, IR spectral information on heavy metal contamination can be indirectly obtained through the interactions between heavy metal ions and enzymes. However, this information is remarkably weak. An in-depth analysis of these weak signals is therefore crucial to detect heavy metal accumulation in *Tegillarca granosa*.

This study aims to select characteristic spectral variables that differentiate *Tegillarca granosa* samples that were uncontaminated (healthy) from those that were contaminated by copper (Cu), cadmium (Cd), zinc (Zn), and lead (Pb) using the Competitive Adaptive Reweighted Sampling(CARS) method，Successive Projection Algorithm (SPA) and Genetic Algorithm (GA). Discrimination models that classify healthy samples and heavy metal-contaminated samples were constructed using the selected spectral variables. Cd and Pb are not essential and toxic, whereas Zn and Cu are crucial to the human body. However, excessive Zn and Cu can also harm human health. Hence, detecting heavy metal contamination in *Tegillarca granosa*is important to ensure food safety. The specific objectives of the current work were to (1) spectral variables that differentiate *Tegillarca granosa* samples that were uncontaminated (healthy) and contaminated by a certain heavy metal were selected using CARS，SPA and GA (Design I). classification models were then constructed using these selected spectral variables; (2) Spectral variables that differentiate all sample varieties, including the samples that were uncontaminated and contaminated by any of the four heavy metals, were selected using CARS，SPA and GA (Design II). Classification model was then constructed with these selected spectral variables.

Materials and methods

Sample Preparation and Chemicals

Tegillarca granosa samples were purchased from Zhejiang Mariculture Research Institute at Wenzhou, China in May 2014, and were acclimatized to laboratory conditions for approximately 10 d in plastic pools with a size of 60.0 cm×40.0 cm×30.0 cm. Analytical-grade $PbCH_3COO{\cdot}3H_2O$, $CuSO_4{\cdot}5H_2O$, $CdC1_2$, and $ZnSO_4{\cdot}7H_2O$ were purchased from Chemical Reagent Co. Ltd., Shanghai, China.

Seawater prepared by over 24h of sedimentation and then sand filtration was used to maintain *Tegillarca granosa* in the tanks. The prepared seawater had a pH of 8.05±0.1, a temperature of 20.8±2.6 °C, a dissolved oxygen content of > 6 mg/L, and a salinity level of 21‰. The water was changed every 24h throughout the experiment. The containers were refilled and re-dosed with the metal toxicant.

The *Tegillarca granosa* samples in Groups I, II, III, and IV were exposed to high concentrations of $PbCH_3COO{\cdot}3H_2O$ (1.833 mg/L), $CuSO_4{\cdot}5H_2O$ (5.589mg/L), $CdC1_2$ (1.634 mg/L), and ZnSO4·$7H_2O$ (4.424 mg/L) in water, respectively. Group V (control)was reared in seawater without adding heavy metals. The *Tegillarca granosa* samples from all groups were reared for 10 d to allow heavy metal accumulation. After the rearing period, the *Tegillarca granosa* samples were sacrificed and then stored in a refrigerator at -4 °C for 15 min. The samples were freeze-dried, ground into powder, and then used for spectral analyses.

A total of 150 samples (30 samples for each variety) were prepared for further treatment, in which 30 samples from the healthy group and 30 samples from the contaminated group were used to establish the models for Designs I and II. Therefore, 60 samples in four sample sets were used for Design I, and these four sample sets were used for the intoxication analysis of Zn, Cu, Cd, and Pb. Meanwhile, 150 samples (30 samples×5 groups) were used for Design II. A column vector (Y_C)

containing the integer numbers of intoxication status of samples from the training set was concatenated to matrix X_C. Model training was based on X_C and Y_C. For the two-class problem (Design I), the integer numbers were set to 0 and 1, which represent the contaminated and healthy samples, respectively. For the five-class problem (Design II), the integer numbers were set to 0 and 1 to 4, which represent the healthy samples and the samples contaminated by Cd, Cu, Pb, and Zn, respectively.

Spectral Collection and Reference Methods for Heavy Metal Content

IR (4000−400 cm^{-1}) spectra were obtained using a Tensor 27 spectrophotometer (Bruker, Inc., Germany) equipped with a Golden Gate Diamond ATR sampling accessory. The collection of all samples was completed in an airtight collection box. The instrument need warm up about 30 minutes before measuring the spectrum. All samples were scanned 15 times, in which a scan time is about 5 seconds, and the results were averaged using OMINIC software (Version 5.2, Bruker, Inc.). The SNR of spectrophotometer is superior to 55000:1 (peak-to-peak value), resolution ration is $4cm^{-1}$, KBr beam splitter, DTGS detector and 100-micrometer diaphragm were adopted.

The reference value of the heavy metal concentration was measured using a NexION 300X ICP-MS (NexION 300X, Perkin Elmer, Inc., U.S.). The gas flow rate, auxiliary gas flow rate, RF power, and peristaltic pump of the atomized sample were set to 0.90L/min, 14L/min, 1100W, and 20rpm, respectively. The detailed test procedure was as follows. First, exactly 0.3g of sample (deviation < 0.001 g) was prepared, digested with 6mL of highly pure nitric acid, and then filled to 50mL with distilled water. The sample was subjected to microwave digestion, heated to 120 °C within 5min, and then maintained at 120 °C for 10min. The sample was heated from 120 °C to 180 °Cfor 10min and then maintained at 180 °C for 20min. Afterward, the sample was analyzed

using the instrument tested. The statistical values of the heavy metal contents of the samples are shown in Table 1. The table shows that the four heavy metals were mildly enriched in the healthy samples but were highly enriched in the contaminated samples.

Variable Selection method

CARS method

Competitive Adaptive Reweighted Sampling (CARS) is a newly proposed spectral variable selection method whose algorithm is based on the "survival of the fittest" principle of Darwin's theory of evolution [18-21]. The key concept is to consider each spectral variable as an individual and to remove unfit individuals. Spectral variables with large absolute coefficients in partial least squares (PLS) regression were selected and those with small values were excluded using CARS. Spectral variable subsets were obtained, and each subset was used to construct a model for CV. The model with the lowest root mean square error of CV (RMSECV) was selected as the optimal subset of wavelengths. The model with the lowest RMSECV effectively selected the optimized subset of spectral variables relevant to properties of interest. The algorithm was as follows:

(1) The population was subjected to N rounds of Monte Carlo (MC) sampling. In each sampling, a PLS regression model was constructed with a random partition of samples selected for validation. The absolute value of the regression coefficient $|b_i|$ was calculated, where i represents the i^{th} model.

(2) Spectral variables with relatively small $|b_i|$ were forcibly removed using the exponential decreasing function:

$$r_i = ae^{-ki} \quad (1)$$

where the constants a and k were calculated as follows:

$$a = (p/2)^{1/(N-1)} \quad (2)$$

$$k = \ln(p/2)/(N-1) \quad (3)$$

In the first round of MC sampling, all p variables were used for modeling; thus, $r_i = 1$, In the N^{th} MC sampling, only two spectral variables were used; thus, $r_N = p/2$.

(3) N rounds of CARS selection were performed. In each round, the spectral variables with large absolute values of PLS regression coefficients were selected and used to construct a PLS regression model. Then, the spectral variables with the lowest RMSECV values were selected.

SPA method

Successive projections algorithm (SPA) is a forward variable selection method in the area of spectral analysis. It uses simple operations in a vector space to select a subset of variables whose information content is minimally redundant for multivariate linear regression (MLR) [22-23]. In SPA calculation, candidate subsets of variables are constructed according to a sequence of projection operations involving the columns of the instrumental response matrix. These candidate subsets are evaluated according to the prediction performance of the resulting MLR model. In SPA, such prediction performance is assessed by an independent validation set or cross-validation of training set. With the advantages of simple, stability and good predictive performance, SPA has been successfully employed for variable selection in UV-VIS, ICP-AES and NIR spectrometry, as well as for coefficient selection in wavelet regression model.[24-27].

GA method

The ultimate goal of genetic algorithm (GAs) is the optimization of a given response function. GAs are inspired by evolution theory: in a living environment, the “best” individuals have a

greater chance to survive and a greater probability to spread their genomes by reproduction, namely the 'struggle for life'. In wavelength variables selection, GAs have five basic steps, variable coding, population initiation, response evaluation, reproductions, and mutations. Steps 3-5 alternate until a termination criterion is reached. This criterion can be based on a lack of improvement in the response or simply on a maximum number of generations or on the total time allowed for elaboration. [28-29].

Classification models

PLS-DA method

Partial least-squares discriminant analysis (PLS-DA) is the classification version of partial least-squares regression (PLSR). Differing from Partial least-squares regression (PLSR) that is PLSR of a set Y of binary variables describing the categories of a categorical variable on a set X of predictor variables. It is a compromise between the usual discriminant analysis and a discriminant analysis on the significant principal components of the predictor variables PLS-DA encodes the dependent variable with dummy variables describing the classes for the optimum separation of classes. After encoding, PLS-DA is implemented in the usual way of PLSR [30].

SIMCA method

Soft independent modelling of class analogy (SIMCA) is an established method for multivariate classification. In SIMCA classification, the residuals of several disjoint PCA models are utilized to assign an observation to one or several of the available classes. During the training of each class-specific PCA model, a distribution of the residuals for each class is generated. Given this class-specific residual distribution, any given observation can subsequently be assigned a probability of equal variance compared to the model residuals according to a *F*-test. The

probability assignment is then ultimately used to accept or reject the observation to or from each class, which is essentially a tool for detecting model outliers [31-32].

LDA method

Linear discriminant analysis (LDA) is one commonly used technique for data classification and dimensionality reduction. LDA is a mathematical transformation method from multidimensional space to one-dimensional space by maximizing the Fisher criterion, namely the ratio of between-class scatter dispersion and within-class scatter dispersion, to find the best and most easily categorized projection direction. LDA easily handles the case where the within-class frequencies are unequal and their performance has been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. The use of Linear Discriminant Analysis for data classification is applied to classification problem in speech recognition. When transformed to a different space, LDA doesn't change the location of the original data sets but only tries to provide more class separability and draw a decision region between the given classes. This method also helps to better understand the distribution of the feature data. In LDA, data sets can be transformed and test vectors can be classified in the transformed space by two different approaches: class-dependent transformation and class-independent transformation [33].

SVM method

This section briefly describes Support Vector Machine (SVM). For further details, one can refer to [34-35], which provides a complete description of the SVM theory. The linear SVM finds an optimal separating margin by solving the following optimization task:

$$\begin{aligned} &\text{Minimize } g(\mathrm{w},\xi) = \frac{1}{2}\|\mathbf{w}\|^2 + \kappa\sum_{i=1}^{n}\xi_i \\ &st: y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0 \end{aligned} \quad (4)$$

where κ is a penalty value and ξ_i is the positive slack variables. This primal problem can be reduced to the Lagrangian dual problem by introducing Lagrangian multipliers (α_i). The optimal solution α_i can be obtained under the Karush–Kuhn–Tucker conditions. If α_i>0, the corresponding data points are called SVs. Afterward, the optimal hyperplane can be constructed using the optimal parameters *w* and *b*. The linear classification function can then be given by

$$g(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b)\right) \quad (5)$$

The original input space was mapped into a high-dimensional feature space via a mapping function to maximize the function of the linear learning machine in non-linear cases. Through this mapping function, x^Tx in the input space is a form of $\phi(\mathbf{x}_i)^T\phi(\mathbf{x})$ in the feature space. The Gaussian kernel is generally utilized to detect the optimal parameter values of the radial basis function (RBF) kernel (i.e., *C* and γ). Therefore, the decision function can be expressed as follows:

$$g(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \quad (6)$$

In the SVM algorithm, the RBF kernel is utilized to map the input space into a high-dimensional feature space. Thus, the parameters *C* and γ of RBF kernel are important in designing an effective SVM model. The penalty parameter *C* determines the trade-off between the fitting error minimization and model complexity, while the kernel width γ defines the non-linear mapping from the input space to a high-dimensional feature space [36].

Model Evaluation and Algorithm Parameters

The sensitivity, specificity, and accuracy of the classification models for training and

prediction were evaluated. Sensitivity is defined as the number of positives (contaminated samples) correctly classified by the model divided by the number of all positives. Specificity is defined as the number of negatives (healthy samples) correctly identified by the model divided by the number of all negatives. Accuracy is defined as the number of correctly distinguished samples divided by the number of all samples. All models were validated through 10-fold CV. All computations and chemometric analyses were operated in Matlab 2011a (The Mathworks, Inc., Natick, MA, USA).

Results and Discussion

There were large of noise interference existed in 623−400 $cm^{-1}$and 4000−3517cm^{-1} bands, therefore these two bands were removed, resulting in only the band of 3516−624cm^{-1} was used for further analysis. Figure 1 shows a typical IR spectrum of *Tegillarca granosa*. From Figure 1, healthy samples and samples contaminated by different heavy metals show similar IR spectra. Each variety only has one sample, the IR spectra in the Figure 1 don't represent the true differences among all samples. Considering all samples produces overlapping spectral profiles. As a result, evident difference among all samples can hardly be detected by naked eye. In addition, the IR spectrum of *Tegillarca granosa*is complicated and contains numerous absorption peaks contributed by different functional groups. The wavenumber assignment mainly concentrates in two distinct ranges, namely, 3700−2800 and 1800–650 cm^{-1}. Comprehensively, the broad band centered at around 3300 $cm^{-1}$is attributed to the N–H stretching mode of Amide A [11, 14-15]. In the region between 3000 and 2800 cm^{-1}, three weak absorption peaks are found at around 2960, 2925and 2865 cm^{-1}. The absorption peaks at 2960 and 2865 cm^{-1} are attributed to the asymmetric and symmetric stretching vibrations of CH_3. These two peaks are often used to determine the lipid

structure and monitor proteins in biological systems. Another absorption band at 2925 cm^{-1} corresponds to the anti symmetric and symmetric stretching vibrations of $-CH_2$, which is mainly found in lipids [11, 14-15]. Several absorbance peaks are located in the band region range of 1800–650 cm^{-1}. This region is dominated by amide groups. Two strong absorption bands at around 1650 and 1540 cm^{-1} correspond to the amide I and amide II vibrations of structural proteins, respectively. Amide I absorption is principally related to the C=O stretching vibration of amides. Meanwhile, amide II absorption originates from amide N–H bending vibration (60%) coupled with the C–N stretching vibration (40%) mode of the polypeptide and protein backbones [13].These two absorption bands are sensitive and can be used to determine the secondary structure of proteins. The absorption bands at around 1400 cm^{-1} are assigned to COO^- symmetric stretching modes, which are mainly associated with fatty acids and amino acids [13-14]. The bands observed at 1230 cm^{-1}correspond to the PO_2 asymmetric stretching of nucleic acids in phospholipids. The band observed at 1070 cm^{-1} corresponds to the symmetric stretching of PO_2^- in nucleic acids, HO−C−H stretch, and carbohydrates. The band observed at 1040 cm^{-1} is assigned to the C−O stretching vibrations in polysaccharides. In addition, the peaks at 1000−623 cm^{-1} correspond to a fingerprint region mostly of nucleic acids [15]. We also observed the spectra of the chemical molecules and functional groups of the constituents found in *Tegillarca granosa*. This spectral information is important to detect heavy metal contamination. However, the *Tegillarca granosa* samples have numerous absorption peaks that overlap between different samples. This overlapping complicates the selection of useful spectral variables for the detection of heavy metal contamination. Therefore, variable feature selection and pattern recognition methods were utilized. In our study, the preprocessed method of standard normal variate (SNV) was employed to achieve

a centering and scaling effect. The processed spectra were used for further data analysis.

Analysis of classification Results based on different variable selection methods

Because there are lots of irrelevant information and redundant information in infrared spectra, the common variable selection methods of CARS, SPA and GA were employed to select the characteristic spectral variables. The selected variables by CARS, SPA and GA were served as input variables to construct classification model including PLS-DA, SIMCA, LDA and SVM, so 48 models for Design I (four heavy metals × four calibration methods × three variable selection methods) and 12 models for Design II (four calibration methods × three variable selection methods) was obtained. For Design I, classification results of PLS-DA, LDA and SIMCA models are below than 90% except SVM model, and classification results obtained by these linear models (PLS-DA, LDA and SIMCA) are only close to 70% for Design II, which are unexceptionally inferior to the results of SVM which classification results exceed 90% . So, in our manuscript, the SVM model as main model was employed to further analyze.

As mentioned above, the parameters C and γ of RBF kernel are important in designing an effective SVM model. In this study, grid search was utilized using 10-fold CV to optimize the two key parameters of the RBF kernel-based SVM. The values of C and γ vary as $C = \{2^{-5}, 2^{-3},\ldots,2^{15}\}$ and $\gamma = \{2^{-15},2^{-13},\ldots,2^{1}\}$. The optimal parameter pair (C, γ) was used to construct a model for training. The classification results for Design I are shown in Table 2 and those of Design II in Table 3. From Tables 2 and 3, it can be found that the variable number selected by SPA is less than those of CARS and GA, meanwhile the results of GA and SPA are slightly inferior to that of CARS, that might be because it is easy for GA to get partially optimized, in most cases the selected variables are not the optimal. As the number of variables selected by SPA must be less

than that of samples involved in the modeling and there were a few samples considered in this study, only some useful information (variable number was less than sample number) can be selected by SPA. As to results, it can find that Design I had a better identifying result than Design II, especially based on GA method. For Design II, the classification rates of both Cu and Zn were less than 90%, that was inferior to those of Cd and Pb.

As the improvement effect using CARS is a little better than those of using GA and SPA, an in-depth analysis of characteristic variables was carried out based on CARS selection. SVM model was established based on the characteristic variables selected by CARS algorithm. Figure 3 shows the classification accuracy surface in a run of 10-fold CV, where the x- and y-axes represent $\log_2 C$ and $\log_2 \gamma$, respectively. Each mesh node in the (x, y) plane of the validation accuracy represents a parameter combination, and the z-axis denotes the obtained classification accuracy value with each parameter combination. Figure 3 shows that the classification rate with the optimal parameters was over 90%. Figures 3 (a) to (e) show that the highest classification rate depended on the optimization of the two parameters. In these figures, $C>0$ and $\gamma<0$ generally yielded high classification rates. Meanwhile, the classification accuracy remained unchanged when the C value exceeded a certain threshold. For Design I, with the optimum C and γ values, the variables selected by CARS (33 for healthy samples vs. Cu-contaminated samples, 31 for healthy samples vs. Zn-contaminated, 22 for healthy samples vs. Pb-contaminated samples, and 10 for healthy samples vs. Cd-contaminated samples) were used as inputs of SVM to construct a classification model. The model was validated by 10-fold CV using all samples. The classification rates during training reached 90% (Figure 3). The corresponding prediction results are shown in Table 2. Table 2 shows that the effectiveness of the two-category classification (healthy samples vs. samples

contaminated by a single heavy metal) was highly satisfactory. This finding is evidenced by the over 95% sensitivity, specificity and accuracy of the method. Compared with Design I, Design II had slightly lower classification rate because greater complexity was required to identify more sample varieties. Detailed classification results have been listed in Table 4, including misclassifications of different samples, the prediction value for healthy samples was relatively high (over 95%). However, several varieties of the heavy metal-contaminated samples were recognized as healthy and most of them were Cu- and Zn-contaminated samples (four and five samples, respectively). This observation may be attributed to the fact that unlike Pb and Cd, Cu and Zn are essential to organisms. Little amounts of Cu and Zn may be insufficient to change the protein and lipid structures of the samples. Slight changes in the IR spectra may increase false classification rate. This phenomenon might be the reason for the lower classification rate for Cu and Zn contaminations than for Pb and Cd contaminations.

Analysis of Differences between Healthy and Contaminated Samples

The spectral variables used in Designs I and II were selected using CARS which was utilized 10 times because a different model was constructed each time and the selected characteristic spectral variables were slightly different. Figure 2 shows the characteristic spectral variables that were selected five times or more by CARS. The IR spectrum of a biological system is derived from the vibration of various functional groups; hence, the characteristic spectral information can be attributed to the interactions between these groups. Such characteristic spectral information reflects the specific characteristics of certain molecular structures. Figure 2 also shows that the spectral variables selected by both Designs I and II are concentrated within the 624−1700 cm^{-1} and 3000−3516 cm^{-1} regions.

An in-depth analysis shows that Design II selected more spectral variables than Design I. This observation can be attributed to the fact that Design II requires more variables for model construction because it must classify all five sample varieties. The variables selected by Design II included almost all characteristic peaks within the full IR spectral range and distributed in two main ranges, 1350−1540cm^{-1} reflecting the COO-symmetric stretch (fatty acids and amino acids, CH_2 bending: mainly lipids, and Amide II: N−H bending and C−N stretching of proteins) and 3500−3100cm^{-1} reflecting Amides A and B. Besides, the 1680cm^{-1} band that principally reflects the Amide I and C=O stretching of proteins was also selected [11,14-15].

For Design I, the regions and numbers of the selected spectral variables differ between the healthy samples and the samples contaminated by a single heavy metal. To differentiate healthy from Cu-contaminated samples, the highest number of spectral variables (up to 33) was selected using CARS method. These variables are concentrated at around 1500, 3100−3500, and 600cm^{-1}. In total, 31 spectral variables were selected by CARS for the model that classifies healthy from Zn-contaminated samples. These variables are concentrated at around 3000, 1600 and 1100cm^{-1}, which are close to the bands of the variables for Cu contamination. Relatively fewer variables were selected for Cd and Pb contaminations as compared with those for Cu and Zn contaminations. For Pb contamination, 22 variables were selected, which mainly concentrate at the 2900−3100, 1500−1700, 1300 and 700−1100cm^{-1} regions. For Cd contamination, only 10 variables were selected, which are distributed at around 3300, 2800, 1600, 1100, 900 and 600cm^{-1}. The different numbers of variables selected for the four heavy metal contaminants were due to the proposal that Pb and Cd as highly toxic heavy metals can significantly change the structures of *Tegillarca granosa* components (proteins, lipids and others). Therefore, less variable information is needed to

identify the samples contaminated by these two metals. Meanwhile, Cu and Zn are required at low concentrations by most organisms and are unlikely to cause significant structural changes in *Tegillarca granosa* because of their low toxicity. Hence, a more variable information is required to identify samples contaminated by these heavy metals. From another perspective, the region at around 1500 cm^{-1} was selected using CARS for the four sample varieties of heavy metal contaminations. This region principally reflects the protein structure. This selection can be explained by the mechanism of heavy metal poisoning. In other words, heavy metal poisoning stimulates *Tegillarca granosa* to synthesize metallothionein. This compound causes peroxidation reactions in the cell membrane and induces the synthesis of antioxidants, such as glutathione and superoxide dismutase. These antioxidants lead to the generation of free radicals and H_2O_2, which are highly toxic to cells. The whole defense system collapses when the generation of H_2O_2 exceeds the antioxidant capacity of an organism. Most of the structural changes caused by heavy metal contamination are related to proteins [37-40]. This observation explains why the variables selected for all four heavy metal contaminations include the region that reflects protein structure. A similar explanation is also applicable to the variable selection using Design II. If more varieties of contamination need to be classified, then more variables are required for model construction.

Conclusion

A rapid method based on IR spectroscopy and pattern recognition was proposed to differentiate healthy samples from Cu-, Cd-, Z-n, or Pb-contaminated *Tegillarca granosa* samples. The variables in the IR spectra that classified healthy and heavy metal-contaminated samples were selected using CARS. These variables were used as inputs for SVM to construct a classification model. The models that were classified between the healthy samples and the samples

contaminated by a single heavy metal showed accuracies of up to 95%. The method also classified samples contaminated by different heavy metals with an accuracy of over 90%. This study provides a new and convenient method for the rapid detection of heavy metal-contaminated aquatic products.

Acknowledgments

The authors would like to acknowledge the financial support provided by National Natural Science Foundation of China (NO.31201355).

References

[1] Institute for Nutrition and Food Safety of the Chinese Center for Disease Control and Prevention, Food composition table: National representative value, Beijing, China, 1992, pp: 32-48.

[2] J. Xu, H. Zhou, X. Yan, X. Yan, C. Zhou, P. Zhu, and B. Ma, Effect of unialgal diets on the composition of fatty acids and sterols in juvenile ark shell *tegillarca granos*a Linnaeus, *J. Agric. Food Chem.*, 2012, **60**, 3973–3980.

[3] M. Barron, G. Stehly, and W. Hayton, Pharmacokinetic modeling in aquatic animals Ⅰ. models and concepts,*Aquat.Toxicol.*,1990,**18**, 61-85.

[4] M. Soylak, and A. Aydin, Determination of some heavy metals in food and environmental samples by flame atomic absorption spectrometry after coprecipitation, *Food Chem.Toxicol.*, 2011, **49**, 1242-1248.

[5] M. Tüzen, Determination of heavy metals in fish samples of the middle Black Sea (Turkey) by graphite furnace atomic absorption spectrometry, *Food Chem.*, 2003, **80**, 119-123.

[6] R. Karoui, G. Downey, and C. Blecker, Mid-infrared spectroscopy coupled with chemometrics: a tool for the analysis of intact food systems and the exploration of their molecular structure structure-quality relationships - a review. *Chem. Rev.*, 2010, **110**, 6144-6168.

[7] X.J. Chen, D. Wu, X.C. Guan, B. Liu, G. Liu, M.C, Yan, H.L. Chen, Feasibility of Infrared and Raman Spectroscopies for Identification of Juvenile Black Seabream (Sparus macrocephalus) Intoxicated by Heavy Metals, *J. Agr. Food Chem.* 2013, 61, 12429-12435.

[8] D. Cozzolino, Recent trends on the use of infrared spectroscopy to trace and authenticate natural and agricultural food products, *Appl. Spectrosc. Rev.*, 2012, **47**: 518-530.

[9] D. Wu, J. Chen, B. Lu, L. Xiong, Y. He, Y. Zhang, Application of near infrared spectroscopy for the rapid determination of antioxidant activity of bamboo leaf extract. *Food Chem.*, 2012, **135**, 2147-2156.

[10] D. Wu, P.C. Nie, Y. He, Y.D. Bao, Determination of calcium content in powdered milk using near and mid-infrared spectroscopy with variable selection and chemometrics. *Food Bioprocess Tech.*, **5**, 1402-1410.

[11] Z. Movasaghi, S. Rehman, and I. Rehman, Fourier transform infrared (FT-IR) spectroscopy of biological tissues, *Appl. Spectrosc. Rev.*, 2008, **43**，134–179.

[12] X. Lu, M. Webb, M. Talbott, J. Van Eenennaam, S. Doroshov and B. Rasco, A study of biochemical parameters associated with ovarian atresia and quality of caviar in farmed white sturgeon (Acipenser transmontanus) by Fourier Transform Infrared (FT-IR) Spectroscopy, *Aquaculture*, 2011,**315**, 298-305.

[13] S. Akkas, M. Severcan, O. Yilmaz,and F. Severcan, Effects of lipoicacid supplementation on

rat brain tissue: an FT-IR spectroscopic and neural network study, *Food Chem*., 2007, **105**, 1281-1288.

[14] P. Palaniappan, and V. Vijayasundaram, Fourier transform infrared study of protein secondary structural changes in the muscle of Labeorohita due to arsenic intoxication, *Food Chem. Toxicol*, 2008, **46**:3534-3539.

[15] P. Palaniappan, and V. Vijayasundaram, The FT-IR study of the brain tissue of Labeorohita due to arsenic intoxication. *Microchem. J*., 2009, **91**,118-124.

[16] P. Palaniappan, and V. Vijayasundaram, The effect of arsenic exposure and the efficacy of DMSA on the proteins and lipids of the gill tissues of Labeorohita. *Food Chem.Toxicol.*, 2009,**47**, 1752-1759.

[17] P. Palaniappan, and K. Pramod, FTIR study of the effect of $nTiO_2$ on the biochemical constituents of gill tissues of Zebrafish (Daniorerio). *Food Chem. Toxicol.*, 2010, **48**: 2337-2343.

[18] H. Li, Y. Liang, Q. Xu, and D. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta*, 2009, **648**, 77–84.

[19] K. Zheng, Q. Li, J. Wang, J. Cheng, P. Cao, T. Sui, X. Wang and Y. Du, Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra. *Chemometr. Intell. Lab.*, 2012, **112**, 48-54.

[20] W. Fan, Y. Shan, and G. Li, Application of competitive adaptive reweighted sampling method to determine effective wavelengths for prediction of total acid of vinegar. *Food Anal.Method*, 2012, **5**,585-590.

[21] X. Zhang, W. Li, and B. Yin, Improvement of near infrared spectroscopic (NIRS) analysis of caffeine in roasted Arabica coffee by variable selection method of stability competitive adaptive reweighted sampling (SCARS). *Spectrochim. Acta A*, 2013, **114**, 350-356.

[22] D. Wu, Y. He, Potential of spectroscopic techniques and chemometric analysis for rapid measurement of docosahexaenoic acid and eicosapentaenoic acid in algal oil. *Food Chem*., 2014, **158**, 93-10.

[23] D. Wu, H. Shi, Y. He, X. Yu, Y. Bao, Potential of hyperspectral imaging and multivariate analysis for rapid and non-invasive detection of gelatin adulteration in prawn. *J. Food Eng.*, 2013, **119**, 680-686.

[24] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemometr. Intell. Lab., 2001, **57**, 65-73.

[25] R.K.H. Galvão, M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares, H.M. Paiva, A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemometr. Intell. Lab., 2008, **92**, 83-91.

[26] K. Liu, X.J. Chen, L.M. Li, H.L. Chen, R.X. Ruan, W.B. Liu, A consensus successive projections algorithm -multiple linear regression method for analyzing near infrared spectra, *Anal. Chim. Acta*, 2015, doi:10.1016/j.aca.2014.12.033.

[27] D Wu, X.J. Chen, X.O., Zhu, X.C. Guan, G.C. Wu, Uninformative variable elimination for improvement of successive projections algorithm on spectral multivariable selection with different calibration algorithms for the rapid and non-destructive determination of protein content in dried laver, *Ana. Method*, 2011, **3**, 1790-1796.

[28] R. Leardi, R. Boggia, M. Terrile, Genetic algorithm as a strategy for feature selection, J. Chemometr. 1992, **6**, 267-281.

[29] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. de Noord, Genetic algorithms as a tool for wavelength selection in multivariate calibration, Anal. Chim., 1995, **67**, 4295-4301.

[30] M. Pérez-Enciso, T. Tenenhaus, Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach, *Hum. Genet.*, 2003, **112**, 581-592.

[31] S. Wold, Pattern recognition by means of disjoint principal components models, *Pattern Recogn.*, 1976, **8**,127–139.

[32] E. Smidt, K. Meissi, M. Schwanninger, P. Lechner, Classification of waste materials using Fourier transfrom infrared spectroscopy and soft independent modeling of class analogy, *Waste Manage.*, 2008, **28**,1699-1710.

[33] S. Balakrishnama, A. Ganapathiraju, Linear discriminant analysis - a brief tutorial, Institute for Signal and information Processing, 1998, 1-8.

[34] C. Hsu, C. Chang, and C. Lin, A practical guide to support vector classification. In: Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2003, available at http://www.csie.ntu.edu.tw/cjlin/libsvm/.

[35] H. Li, Y. Liang,and Q. Xu, Support vector machines and its applications in chemistry. *Chemometr. Intell. Lab.*, 2009, **95**, 188-198.

[36] H. Chen, C. Huang, and X. Yu, An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. *Expert. Syst. Appl.*, 2013, **40**, 263-271.

[37] G. Winston, and R. Di Giulio, Prooxidant and antioxidant mechanisms in aquatic organisms. *Aquat Toxicol.*, 1991, **19**, 137-161.

[38] A. Viarengo, L. Canesi, M. Pertica, G. Poli, M. Moore and M. Orunesu, Heavy metal effects on lipid peroxidation in the tissues of mytilus-galloprovincialis lam. *Comp. Biochem. Phys. C*, 1990, **97**, 37-42.

[39] R. Pipe, J. Coles, and F. Carissan, Copper induced immunomodulation in the marine mussel, *Mytilusedulis*, *Aquat. Toxicol.*, 1999,**46**, 43-54.

[40] F. Regoli, G. Principato, Glutathione, glutathione-dependent and antioxidant enzymes in mussel, *Mytilus-galloprovincialis*, exposed to metals under field and laboratory conditions: implications for the use of biochemical biomarkers, *Aquat. Toxicol.*, 1995,**31**, 143-164.

Figure caption

Figure 1 Representative infrared spectra of healthy and heavy metal-contaminated samples of *Tegillarca granosa*.

Figure 2 Variables selected using CARS for Designs I and II.

Figure 3 Training accuracy surface with parameters for the SVM model.

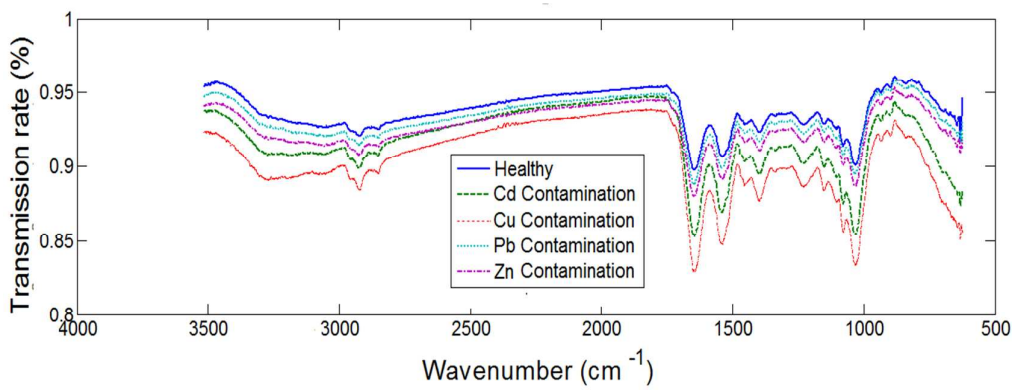


Figure 1 Representative infrared spectra of healthy and heavy metal-contaminated samples of *Tegillarca granosa*.

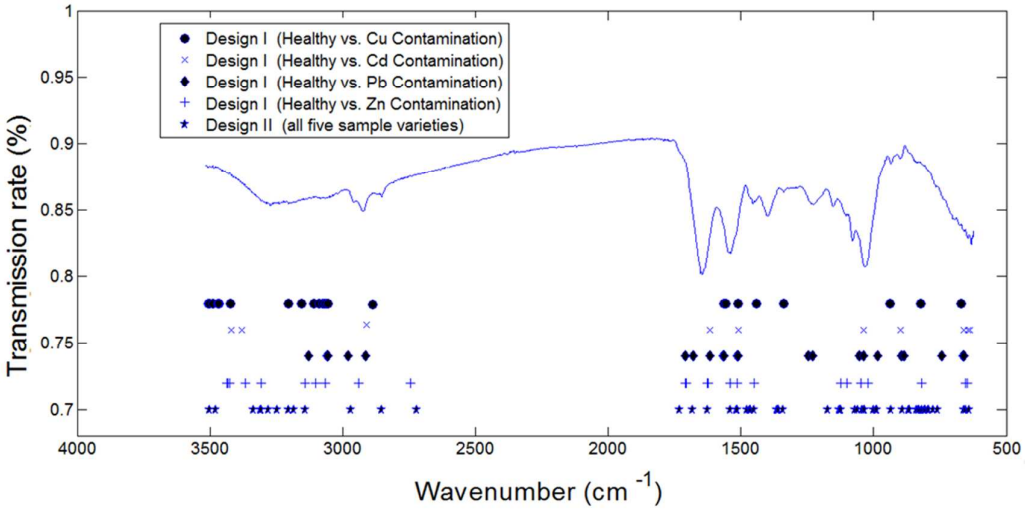


Figure 2 Spectral variables selected using CARS for Designs I and II.

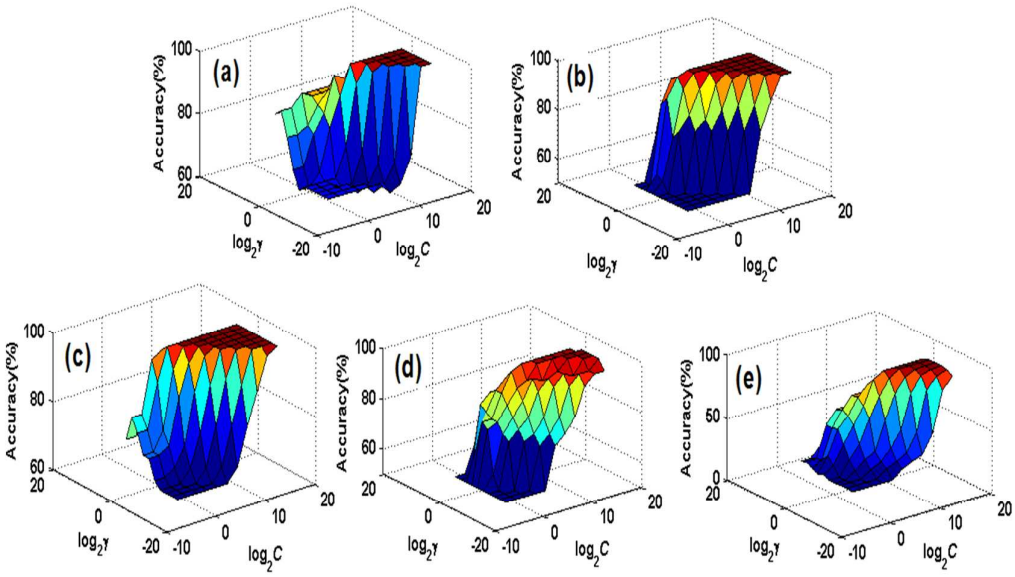


Figure 3 Training accuracy surface with parameters for the CARS-SVM model (a) Design I (Healthy vs. Cd), (b) Design I (Healthy vs. Cu), (c) Design I (Healthy vs. Pb), (d) Design I (Healthy vs. Zn), (e) Design II

Table 1 Heavy metal content statistical values of *Tegillarca granosa* samples

Sample	mean ±SD(mg/kg)			
	Cd content	Cu content	Pb content	Zn content
Healthy	1.832 ± 0.52	7.5± 1.71	0.611± 0.17	32.41± 7.49
Contaminated by Cd	91.83 ± 10.12	6.5± 2.11	0.416± 0.12	22.41± 8.49
Contaminated by Cu	1.132± 0.12	97.08 ± 14.21	0.501± 0.11	27.41± 15.49
Contaminated by Pb	2.332 ± 0.92	5.3± 1.01	241.12 ± 41.17	38.41± 11.49
Contaminated by Zn	1.832± 0.12	9.5± 3.21	0.811 ± 0.27	232.41± 18.59

Table 2 Results of SVM model for Design I

		Training (%)			Prediction(%)		
	Method*	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Healthy vs. Cd	CAR	94.8	97.7	96.3	97.0	95.0	96.3
	SPA	94.0	92.7	93.3	94.7	90.0	92.8
	GA	91.3	92.7	91.5	92.3	94.7	93.2
Healthy vs. Cu	CAR	98.7	95.3	96.0	95.6	94.3	95.3
	SPA	94.4	91.7	93.7	94.0	91.0	92.0
	GA	92.3	90.1	90.5	92.3	90.7	91.2
Healthy vs. Pb	CAR	95.3	98.0	96.7	95.3	97.3	96.1
	SPA	94.0	90.7	92.3	94.7	91.3	93.8
	GA	91.3	94.7	92.5	92.3	95.7	93.2
Healthy vs. Zn	CAR	98.7	95.7	97.7	97.0	94.4	95.9
	SPA	95.0	92.7	94.3	95.7	91.0	93.8
	GA	92.0	93.7	92.5	92.3	95.7	93.1

* Variable selection method

Table 3 Results of the CARs/SPA/GAs-SVM model for Design II

	Method*	Healthy	Cd	Cu	Pb	Zn
Training	CARs	95.1	96.0	93.9	96.6	92.8
	SPA	90.5	92.0	92.1	90.5	93.6
	GAs	94.7	91.6	92.9	91.6	92.1
Prediction	CARs	96.9	96.3	94.3	93.8	91.7
	SPA	92.7	93.3	91.7	92.7	93.3
	GAs	91.7	91.0	88.0	93.0	89.0

*Variable selection method

Table 4 Results of the CAR-SVM model for Design II

		Healthy	Contaminated by Cd	Contaminated by Cu	Contaminated by Pb	Contaminated by Zn
Training	Healthy	255	2	7	1	9
	Contaminated by Cd	2	262	3	2	2
	Contaminated by Cu	4	3	249	2	6
	Contaminated by Pb	3	4	1	259	3
	Contaminated by Zn	4	2	5	4	256
	Accuracy (%)	95.1	96.0	93.9	96.6	92.8
Prediction	Healthy	31	0	1	0	2
	Contaminated by Cd	0	26	1	1	0
	Contaminated by Cu	1	0	33	0	0
	Contaminated by Pb	0	1	0	30	0
	Contaminated by Zn	0	0	0	1	22
	Accuracy (%)	96.9	96.3	94.3	93.8	91.7