ELSEVIER

Journal of Chromatography A, 734 (1996) 259–270

JOURNAL OF CHROMATOGRAPHY A

Assessment of chromatographic peak purity by means of artificial neural networks

Yuzhu Hu*, Gewen Zhou, Jihong Kang, Yingxiang Du, Fang Huang, Jianhua Ge

Department of Analytical Chemistry, China Pharmaceutical University, Nanjing 210009, China

Received 24 October 1995; accepted 29 November 1995

Abstract

An improved chemometric approach is proposed for assessing chromatographic peak purity by means of artificial neural networks. A non-linear transformation function with a back-propagation algorithm was used to describe and predict the chromatographic data. The Mann–Whitney *U*-test was used for the concluding the purity of the chromatographic peak. Simulation data and practical analytical data for both pure and mixture samples were analysed with satisfactory results. A prior knowledge of the impurity and the related compound is unnecessary when a slight difference between their chromatogram and spectrum exists. The performance on simulated data sets by this approach was compared with the results from principal component analysis.

Keywords: Peak purity; Artificial neural networks; Chemometrics

1. Introduction

Chromatographic methods are often used for the determination of active components in pharmaceutical analysis. One of the main problems associated with developments in the chromatographic laboratory is assessing the purity of analyte peaks. Many workers are interested in solving this problem for more reasonable applications of chromatographic techniques. Some studies [1,2] have demonstrated that it is possible to detect impurities using liquid chromatography with computer-aided photodiode-array detectors provided that a slight difference between the chromatogram and spectra of components exists. The determination of chemical purity in chromatography can be divided into two categories: instrumental and chemometric methods. The common methods used to assess peak purity with such chromatographic instruments include the following: normalizing and comparing spectra from various peak sections using a match factor [3]; absorbance ratio approach using the absorbance signals at two wavelengths [4]; multiple absorbance ratio [5]; and multiple absorbance ratio correlation [6] with higher sensitivity. Spectral suppression [7], derivative spectroscopy [8] and the spectral derivative null technique [9] were developed for the determination and identification of active analyte peaks with multiple components. Some chemometric approaches, e.g., self-modelling curve resolution [10], princi-

* Corresponding author.

0021-9673/96/$15.00
SSDI 0021-9673(95)01303-2

pal component analysis (PCA) [11,12] and evolving factor analysis [13], have also been proposed for the assessing peak purity. The efficiency of these techniques depends on the level of a priori information available about the system and the degree of analysis desired.

In this paper, we propose a method using artificial neural networks to determine peak homogeneity. The applications of neural networks theory in chemistry have been fully described [14,15]. Artificial neural networks are computational simulations of human cerebral signal processing. This technique has been applied in analytical science as a kind of accurate calibration model [16–20] for multivariate data analysis, e.g., pattern recognition, structure elucidation, process control and many other branches of analytical applications. A cerebellar model arithmetic computer neural network [21] was applied for the deconvolution of overlapping chromatographic peaks. A neural network with normalized UV spectra and peak areas measured in one chromatogram was proposed for peak tracking [22]. From these successful applications, it has been demonstrated that artificial neural networks have the power to derive empirical models from a collection of example cases for systems where the theoretical relationship between the experimental system and the expected model is unclear. The purpose of this paper is to demonstrate the power of neural network techniques using back-propagation in the identification of chromatographic peak purity by both practicable examples and computer simulations.

2. Theory and algorithm

The theoretical basis of the approach in this paper is the back-propagation algorithm used in most artificial neural network applications. The neural networks are usually built from a connecting feed-forward layered structure of neurons. The structure of the networks consists of three layers, namely the input layer, the hidden layer and the output layer. In this study, the first layer of the network, the input layer, consists of 20 absorption values of each chromatographic response in training sets. The second is the hidden layer of neurons receiving the weighted outputs from the input layer and producing output signals inputting to the output layer. The output layer produces the output signals, corresponding to chromatographic responses in training sets.

The spectral signals from the chromatogram can be given by

$$Y_j = \sum_{i=1}^{m} W_{ij} X_i + \delta_j \tag{1}$$

where $(X_1, X_2, X_3, \ldots, X_m)$ represent the absorbance vector inputting to the node, $(W_{1j}, W_{2j}, W_{3j}, \ldots, W_{mj})$ corresponds to the weighting absorptivity vector, Y_j is the output absorbance vector, δ_j is the calculated bias parameter and m is the number of synapses for the neural node.

The non-linear model, which has been shown to be more appropriate as a transformation function for the weighted connection between layers in networks for peak tracking of a spectrum [18–20,22], was used in this study. The expected output vector O_j is from the sum of the weighted inputs which is transformed with a non-linear sigmoidal transfer function:

$$O_j = \frac{1}{1 + e^{-Yj/\theta j}} \tag{2}$$

This function has an output in the range from 0 to 1. From Eq. 1, where Y_j is the weighted sum of the inputs X_j, the bias parameter θ_j is used to modify the shape of the sigmoidal curve.

According to the back-propagation algorithm, X_i is propagated through the network to the output layer. The errors between the output response vector and the expected response vector are used to correct the weights as usually described.

Simulated data were used to establish the performance of this new approach. Both the chromatogram and the spectrum of the components can be simulated by a general Gaussian distribution. The spectral signals for each time point on the chromatogram response curve can be simulated by the following equation:

$$A = \sum_{i=1}^{K} C_i \exp\left\{\frac{1}{2}\left[\frac{(x_j - \mu_{ix})^2}{\sigma_{ix}^2} + \frac{(y_1 - \mu_{iy})^2}{\sigma_{iy}^2}\right]\right\} \quad (3)$$

where C_i is the concentration corresponding the height of the ith chromatographic peak, x_j is the wavelength of the spectrum with the centre wavelength μ_{ix}, y_1 is the retention time of the peak with retention μ_{iy} and σ_{ix} and σ_{iy} are the standard deviation of the peaks of the spectrum and chromatogram corresponding a quarter of the peak width at the peak base. For the purpose of this simulation, x_j, y_1 and C_i are in arbitrary units.

It was noted that the normalized errors should be added to the responses produced by Eq. 3 to match realistic cases. The algorithm for the generation of normally distributed random numbers proposed by Zupan [23] was used. The relative standard deviation for the responses was fixed at 3% to fit the general error level in HPLC and UV determinations.

The data set for a simulated example is listed in Table 1. The response surface of this example calculated by Eq. 3 is shown in Fig. 1. To make the plot clear, the range for the time axis is from $\mu_y - 2\sigma y$ to $\mu_y + 6\sigma y$, and similarly the range of the wavelength axis is from $\mu_x - 2\sigma x$ to $\mu_x + 6\sigma x$.

Since this approach is based on the difference between the spectrum and the chromatogram of a sample, two criteria, the resolution (R_s) and the spectrum similarity (r), are used in this study. The resolution of two adjacent peaks was calculated by the equation

$$R_s = \frac{\mu_{i+1,y} - \mu_{iy}}{2(\sigma_{i+1,y} + \sigma_{iy})} \quad (4)$$

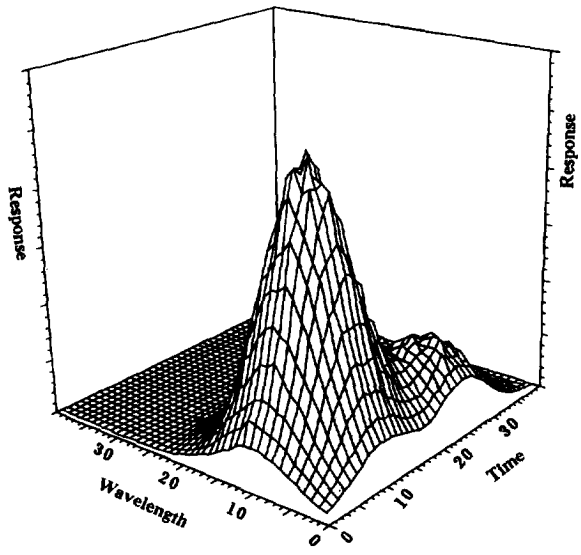

Fig. 1. Response surface for a simulated two-component example. For data sets, see Table 1.

The correlation coefficient [24] was selected to express the spectrum similarity between two spectra for two pure components. For the case in Fig. 1, the resolution $R_s = 0.75$ and the correlation coefficient $r = 0.79$ (see Table 1).

From Fig. 1, it is observed that the first peak produced by the main component is overlapped by the second peak of an impurity. The purpose of this study was to explore the approach when the peak of the impurity is buried by the peak of the main component, which means the resolution of the main component and the impurity should generally be lower than 1.0. On the other hand, the retentions of components should be different and the spectral responses of components should be detectable with a computer-aided photodiode-

Table 1
A case of simulated data sets producing the plot in Fig. 1

Parameter	Main component	Impurity	Resolution (R_s)	Correlation coefficient (r)
C_{max}	100	30		
σ_x	2	2		
σ_y	2	2	0.75	0.79
μ_x	20	18		
μ_y	20	26		

array UV detector. In this case it turns out that a difference in the spectra between one half of the chromatographic peaks and the other half certainly exists. Therefore, we may train the data sets of the front half of each peak and test the responses of the back half using artificial neural network techniques. To abstract more information on the difference of the two parts of spectral data caused by the presence of an impurity, only the range of the peak base is taken into account.

Therefore, the responses which are distributed from time point 1 to 10 on the axis in Fig. 1 were used, i.e., from 16 $(20-2\times2)$ to 24 $(20+2\times2)$ of the retention for this simulated case. Then in this range of the peak base it is divided into nine parts to obtain ten time points. In a similar manner, twenty wavelength points can be obtained in the range of the peak base of the spectrum. One computes twenty spectral responses for each of ten time points by Eq. 3 to obtain a 10×20 matrix of the chromatogram–spectral responses simulated.

Subsequently, ANN is performed based on the spectral data of five time points of the front half as the training sets and five time points of back half as the testing sets. The output of the computation is the ten predictive values of the chromatographic responses. Of course, the predictive values of training sets should coincide well since the ANN model is produced from this set while the results of testing sets depend on the purity of the peak.

If the differences in the experimental and computed values between the training sets and testing sets are statistically significant, the peak purity is very suspicious so that the peak purity can be evaluated. To assess the difference between the training sets and the testing sets, a possible way is to perform a *t*-test on the two groups of differences between experimental values and computed values for both the training and testing sets. However, the *t*-test assumes that the results for each set are normally distributed and it seems that the normal distributions can mostly be obtained with pure substances but not when an impurity is present. For this reason, we propose to adopt a non-parametric test, namely the Mann–Whitney *U*-test [23], for assessing the peak purity as follows.

As shown in Table 2, one first ranks all the data of the absolute differences between the experimental values and computed values in both the training and testing sets, giving rank 1 to the lowest rank, rank 2 to the second, etc. Then we compute the sums of ranks, R_1 and R_2, equals to 17 and 38 in these two groups, respectively. Next, one computes

$$U_1 = 40 - R_1 = 23$$

$$U_2 = 40 - R_2 = 2$$

where 40 is from $n\times n + n(n+1)/2$; n, the number of paired data, equals 5 for this case. The statistical hypothesis is stated as

$$H_0: U_1 = U_2$$

$$H_1: U_1 \neq U_2$$

This test in fact compares the median of two

Table 2
Computation results for data sets in Table 1

	Training sets					Testing sets				
	1	2	3	4	5	6	7	8	9	10
Simulated	0.0135	0.0297	0.0507	0.0795	0.0942	0.0999	0.0870	0.0620	0.0478	0.0412
Predicted	0.0131	0.0290	0.0526	0.0778	0.0953	0.0987	0.0802	0.0581	0.0401	0.0331
Difference	0.0004	0.0007	0.0019	0.0017	0.0011	0.0012	0.0068	0.0039	0.0077	0.0081
Rank	1	2	6	5	3	4	8	7	9	10
			$R_1 = 17$					$R_2 = 38$		
			$U_1 = 23$					$U_2 = 2$		
Conclusion	Impure									

samples. The smaller of the two U values is used to compare with table ($U_{\mathrm{Tab.,\,5.5}} = 4.0$ at $\alpha = 0.10$) for a two-tailed test with ten points. When the computed value is larger than 4.0, the null hypothesis is accepted and one concludes that the peak is pure at the 90% confidence level. Otherwise, the null hypothesis may be rejected and the peak purity is suspect. For this case, because $U = \min(U_1, U_2) = 2.0$, the conclusion is "impure", indicating that the impurity exists.

For samples with a retention time of the impurity longer or shorter than that of the main substance, a difference in ANN models between the training sets and testing sets always exists provided that the chromatogram and spectral data show a difference, and the chromatographic peak purity can therefore be evaluated by the above algorithm.

3. Experimental

3.1. Instrumentation

An HP-1050 liquid chromatograph with computer-aided photodiode-array detectors was used. The column was a stainless-steel column (250 × 4.6 mm I.D.) packed with Spherisorb-C_{18} of particle size 5 μm. Injections of sample of 20 μl were used. All experiments were carried out at room temperature and a flow-rate of 1 ml/min.

3.2. Chemicals

All drugs were of pharmaceutical purity. Stock solutions of the drugs, namely caffeic acid, 3,4-dihydroxybenzoic acid, salicylic acid and 4-hydroxybenzoic acid, were prepared by dissolving accurately weighed 50-mg amounts in 50 ml of methanol. The solutions were mixed and diluted with methanol to the final injected concentrations. Purified water was obtained from unboiled pure water in a quartz glass distillation system. To prepare phosphate buffer with various pH values, 0.05 mol/l sodium dihydrogenphosphate monohydrate was adjusted with 85% phosphoric acid precisely to the final pH value; a PHS-25 digital pH meter was used and the electrodes were calibrated with standard buffer solutions. The buffer solution was filtered through membrane filters (0.5 μm) before use.

3.3. Software

Computer programs for experimental data and simulation data, written in Fortran 77, running on a PC-compatible 80486, 33 MHz computer under MS-DOS 6.2 were developed by the authors. Graphical outputs were produced by Graftool (Release 3.3; 3-D Visions, 1990).

3.4. Procedures

Gaussian elution profiles with the normalized error added at the 3% level were used to simulate chromatograms and UV adsorption spectra with the parameters over a wide range as described later. Experimental data were obtained by using several practical samples. The mobile phase of methanol–phosphate buffer (pH 3.5) were adjusted to a chromatographic resolution of $R_s < 1$ to demonstrate the feasibility of the approach.

For real samples, chromatographic signals were first collected to determine the position of peak base of the main component in an initial test, the peak base was split into nine parts with equal intervals so that ten time points were obtained, then ten chromatogram signals were recorded at each time point after repeated tests. Subsequently, twenty spectral signals were recorded in the range 210–400 nm at equal intervals of 10 nm. In this way, 10 × 20 data sets were obtained for each sample.

The output results of calculations include the resolution of adjacent peaks on the chromatogram, the spectral similarity of components and the predicted responses of the chromatogram for each sample. The conclusion "pure" or "impure" was also shown according to the results of the Mann–Whitney U-test.

4. Results and discussion

To perform artificial neural network techniques, several parameters need to be established. As mentioned above, the magnitude of the chromatogram–spectral signal matrix was fixed at 10×20. In fact, we also tested larger and smaller matrices. It was found that when more points were employed it is probably easy to detect an impurity with a tiny difference but more computation time is needed. The number of nodes in the hidden layer is also a parameter to be adjusted. It was fixed at 5 according to Ref. [11].

The number of iterations is another adjustable parameter. Convergence to a minimum error usually needs many learning iterations, which may range from hundreds to tens of thousands depending on the complexity of the models. We tested various samples, including simulated and real, pure substances and mixtures. Some results are shown in Fig. 2. The relative standard deviation (R.S.D.) is used to evaluate the effect of convergence. Surprisingly, the results show that

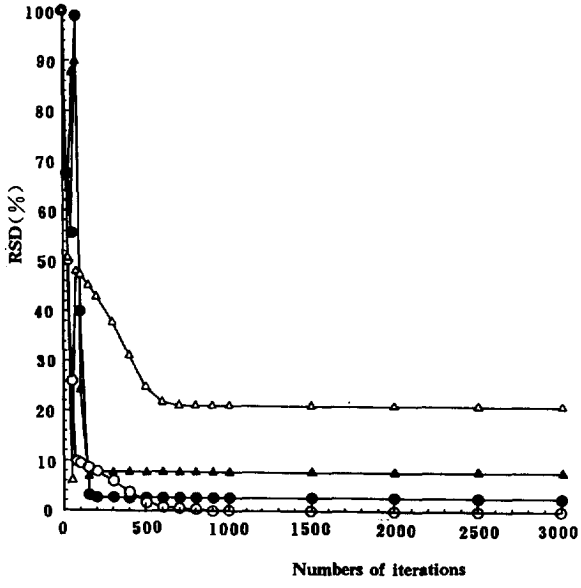

Fig. 2. Results for the number of learning iterations. (1) The example simulated: for data sets, see Table 1. (●) Training sets; (▲) testing sets. (2) A real sample: salicylic acid + 5% 4-hydroxybenzoic acid. (○) Training sets; (△) testing sets.

1000 iterations are sufficient to give a stable estimate value. Fig. 2 also shows that a real sample needs more iterations to give convergence than simulated data sets in which a simple Gaussian peak is used to describe the simulated UV spectra. Owing to the added error at the 3% level in the simulation, the R.S.D. is about 2.7% for the training sets of the simulated data sets whereas the R.S.D. for real samples for training sets is less than 0.1% after reaching convergence. Although training artificial neural networks is generally time consuming, only 16 s are required for 1000 iterations on the 80486 operating at 33 MHz. We also tested the same data on an 80286 computer operating at 25 MHz but more than 15 min were needed.

The gain, the learning rate and the momentum also affect the performance of ANN. We applied a similar method to Fig. 2 to optimize these parameters. The gain and the learning rate are both fixed at 1.0 and the momentum factor at 0.8.

All these parameters were fixed at a certain level in order to perform automatic peak purity control procedures easily in HPLC. The results on simulated data sets and practical examples introduced afterwards show that the procedure behaves well.

The results in Fig. 2 indicated that using the real spectrum of a component is more desirable than using only one simple peak simulated as in our study. However, in the latter approach it is easy to compare more cases with different spectral similarity. Additionally, it will be more precise if the exponentially modified Gaussian (EMG) function [25] is used in modelling the chromatographic peaks. However, it seems too complex to simulate the chromatogram–spectrum for this application. For the above reason, the simple Gaussian peaks with normal noise at the 3% level are still employed in our simulation tests.

Table 3 lists the results for the data from the practical data sets. The first four samples are pure substances and the values of $\min(U_1, U_2)$ are larger than or equal to the criterion value 4 ($\alpha = 0.10$), so that the null hypothesis may be accepted and no significant difference between the training sets and the testing sets exists. For

Table 3
Results for practical examples

Sample[a]		Training sets					Predictive sets					U-test
		1	2	3	4	5	6	7	8	9	10	
1	OP	0.079	0.176	0.312	0.501	0.579	0.512	0.345	0.243	0.129	0.097	$U_1 = 17.0$
	TP	0.075	0.170	0.321	0.484	0.573	0.536	0.327	0.232	0.119	0.089	$U_2 = 8.0$
	Idil	0.004	0.006	0.009	0.018	0.006	0.025	0.018	0.011	0.009	0.008	(Pure)
2	OP	0.066	0.097	0.140	0.190	0.250	0.278	0.339	0.234	0.148	0.048	$U_1 = 17.0$
	TP	0.070	0.100	0.132	0.198	0.256	0.273	0.318	0.248	0.155	0.053	$U_2 = 8.0$
	Idil	0.004	0.003	0.008	0.008	0.006	0.005	0.021	0.014	0.007	0.005	(Pure)
3	OP	0.110	0.202	0.297	0.363	0.471	0.531	0.537	0.542	0.175	0.068	$U_1 = 19.0$
	TP	0.107	0.200	0.290	0.369	0.467	0.556	0.574	0.538	0.182	0.074	$U_2 = 6.0$
	Idil	0.003	0.002	0.007	0.006	0.004	0.025	0.037	0.004	0.007	0.006	(Pure)
4	OP	0.155	0.181	0.206	0.250	0.355	0.454	0.541	0.373	0.164	0.077	$U_1 = 20.0$
	TP	0.144	0.195	0.210	0.253	0.352	0.441	0.560	0.394	0.155	0.083	$U_2 = 5.0$
	Idil	0.011	0.014	0.004	0.003	0.003	0.013	0.019	0.021	0.009	0.006	(Pure)
5	OP	0.118	0.191	0.282	0.430	0.612	0.520	0.315	0.262	0.273	0.182	$U_1 = 25.0$
	TP	0.109	0.198	0.280	0.432	0.610	0.496	0.300	0.230	0.182	0.100	$U_2 = 0.0$
	Idil	0.009	0.007	0.002	0.002	0.002	0.024	0.015	0.032	0.091	0.082	(Impure)
6	OP	0.192	0.321	0.494	0.534	0.544	0.433	0.360	0.294	0.222	0.163	$U_1 = 25.0$
	TP	0.188	0.320	0.492	0.532	0.543	0.455	0.434	0.380	0.321	0.258	$U_2 = 0.0$
	Idil	0.004	0.001	0.002	0.002	0.001	0.022	0.074	0.084	0.099	0.095	(Impure)

OP = simulated values; TP = predicted values; Idil = absolute differences of OP and TP.
[a] Samples: 1 = caffeic acid; 2 = 3,4-dihydroxybenzoic acid; 3 = salicylic acid; 4 = 4-hydroxybenzoic acid; 5 = caffeic acid + 5% 3,4-dihydroxybenzoic acid; 6 = salicylic acid + 5% 4-hydroxybenzoic acid.

the other two samples, the results of the U-test indicate that the null hypotheses is rejected and the conclusion is "impure".

Fig. 3 shows the experimental signals obtained from injection and computation signals from the ANN estimate. For samples 1–4 no impurity can be seen on the chromatogram from the ANN estimate whereas a peak of an impurity in samples 5 and 6 appears as a shoulder on the ANN estimate. The fitted model from computation is of the correct form.

It is certainly insufficient only to perform experiments on the samples in Table 3. To illustrate the characteristics of the method, a simulation test is necessary. Therefore, the experiments on the simulated data sets were carefully designed and carried out.

As discussed in the literature [26,27], spectral similarity, chromatographic resolution and the concentration ratio of overlapping components are main factors that affect the power to detect impurities. Therefore, the data sets produced by Eq. 3 were obtained to observe how these three factors affect the performance of the assessment of peak purity.

To evaluate the sensitivity of detecting an impurity, a criterion value, LD (%), is computed as the detection limit of an impurity which is expressed as the percentage ratio of the concentrations of the least detectable impurity and the main substance.

Considering the interaction of r and R_s, i.e., the effect that one can have on the other, the different parameters were set so as to give r in the range 0–1 at $R_s = 0.30$ and R_s in the range 0–1.2 at $r = 0.52$. The results are shown in Figs. 4–6, where the LD curves express the detection limit of the impurity for a simulated case with one impurity. In the area higher than the line, an impurity can be detected, i.e., the output of calculation for the sample with impurity is "impure".

In order to confirm the feasibility of the proposed method, we compared it with some

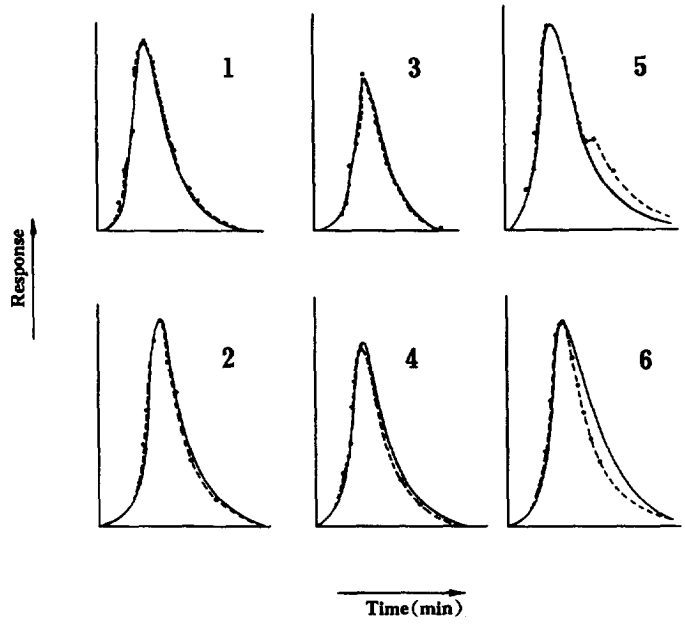

Fig. 3. Comparison of original chromatograms and predictive chromatograms by ANN. Peaks obtained from injection of the sample (solid line) and ANN estimate (dotted line). Samples, see Table 3, Nos. (1–6).

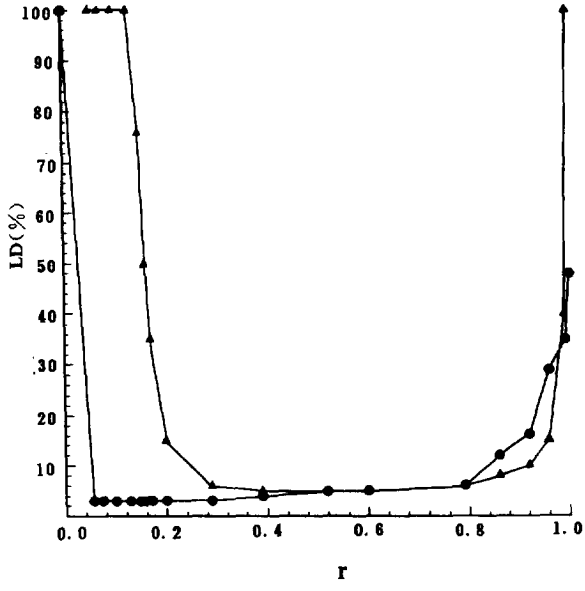

Fig. 4. Plot of LD versus r in a simulated two-component mixture ($R_s = 0.30$) for comparison of ANN and PCA. (▲) ANN; (●) PCA.

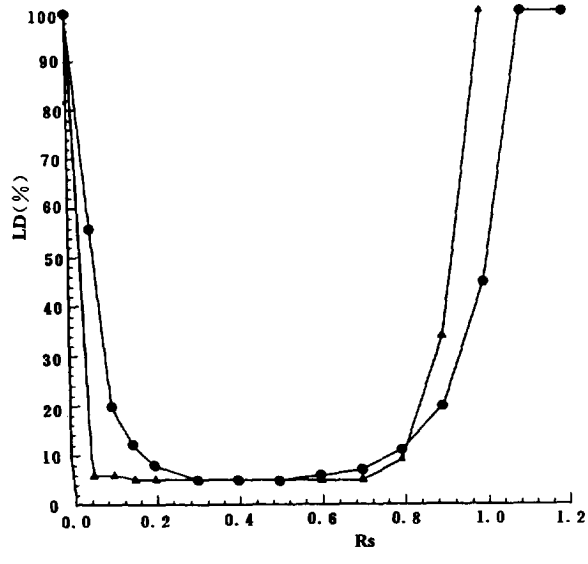

Fig. 5. Plot of LD versus R_s in a simulated two-component mixture ($r = 0.52$) for comparison of ANN and PCA. (▲) ANN; (●) PCA.

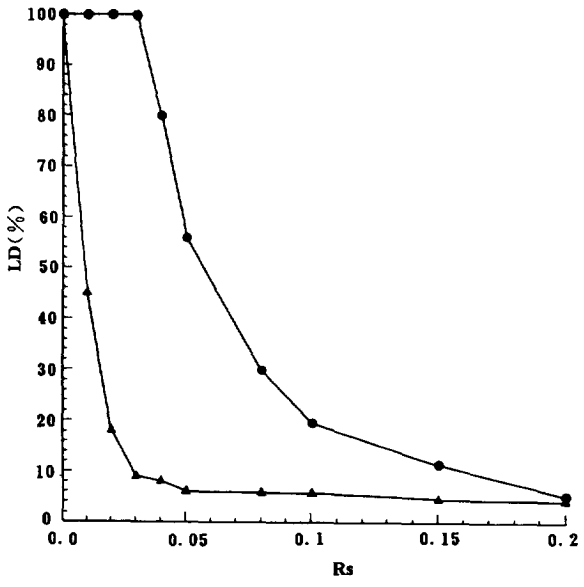

Fig. 6. Plot of LD versus R_s (<0.2) in a simulated two-component mixture ($r = 0.52$) for comparison of ANN and PCA. (▲) ANN; (●) PCA.

other approaches. We used to apply the absorbance ratio approach and principal component analysis for routine analyses in our laboratory [12]. With the first approach, one can observe ratio chromatograms. However, one needs to select carefully the instrumental parameters, e.g., wavelengths and noise threshold level, otherwise erroneous results may be obtained. This means that some preliminary tests need to be carried out using this approach. The PCA algorithm has been well demonstrated for peak purity control in chromatography without assumptions regarding peak shape and location. The potential power of PCA is to abstract the representation of components from the multivariate data of the spectrum and chromatographic signals of the components using eigenvalue decomposition by matrix algebra. According to the algorithm of PCA, the number of spectrally unique components, n, needs to be decided [11,28]. Malinowski [28] proposed the use of a derived probability function to determine the number of factors responsible. However, we found that the correct detection of the number of spectrally unique components (n) using this function depends on the magnitude and characteristics of the error of data sets. It often detected false impurities in a simulation test, i.e., too many impurities were detected. Another approach, i.e., taking the number of maximum eigenvalues with the proportion of the total variance larger than 0.99, seems to be more suitable for the data sets simulated and was therefore adopted in our study, although it cannot take the effect of noise into account.

From Figs. 4 and 5, one first observes that ANN and PCA have similar powers to detect impurities. The lowest value of LD is about 3–5% for the data sets simulated.

In Fig. 4, it is noted that when the spectra of the main substance and the impurity are very similar, i.e., the correlation coefficient r is higher than 0.8, ANN shows a lower LD than PCA, probably because the model of ANN is built based on the data from both chromatogram–spectra and the chromatogram whereas PCA only computes the chromatogram–spectral data. Another reason is that half of the chromatogram–spectral data in ANN are used to build the model, and the difference of the two parts of the data is more evident than the whole data sets evaluated in PCA.

In contrast, when r is lower than 0.4, PCA obviously gives a lower LD value than ANN. It is assumed that some information may be missing in ANN when the spectra of the main substance and the impurity are strongly separated, because PCA computes the whole data sets whereas ANN only uses half.

For the same reason, ANN in Fig. 5 is more sensitive than PCA for detecting an impurity with a similar retention to the main substance. This is shown more clearly in Fig. 6. When $R_s > 0.8$ in Fig. 5, PCA gives a lower LD than ANN.

To conclude, the results show that ANN is more effective than PCA for samples with a lower chromatographic resolution and a greater similarity of spectra between the main component and impurity.

It should be emphasized that the chemical purity cannot always be determined using this method, although it can detect the differences in

the retentions and spectra of the main component and the impurity.

In addition to R_s and r, the sensitivity of determining the main substance is another possible factor affecting the LD value. Both simulated data for a practical sample were tested and the results are shown in Fig. 7. It was found that a concentration of 0.01–0.05 mg/ml of the main substance is appropriate for general applications using this approach.

The simulated data sets were also designed for samples with a main component and an impurity with different retentions and spectra. In the data sets in Table 4, the centre wavelength on the spectra and the retention time on the chromatograms are combined in different ways while all peak widths were fixed at 2. For cases 1–3, the proposed approach can detect the impurity easily. For cases 2 and 3 with different centre wavelengths, the impurity eluted in the front part of the peak while the training was performed. This confirms that this approach can be suitable in cases when the impurity is located in both the front and back halves of the peak. However, for case 4 the neural network approach is incapable of showing the existence of an impurity. It should be noted that the spectral signals of the front and back halves of the chromatographic peak base should be similar owing to the symmetric data sets for μ_x and μ_y of the main component and the impurity for case 4. Therefore, the result that the impurity cannot be detected is reasonable.

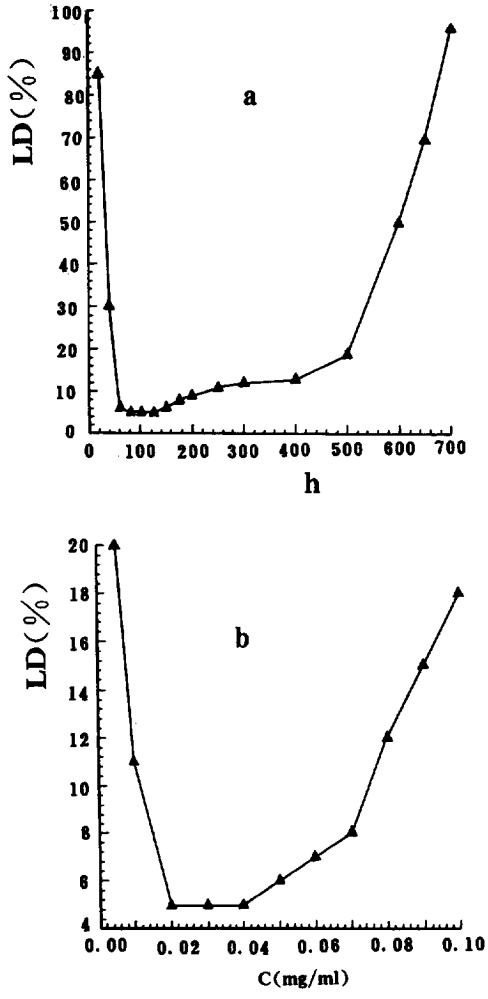

Fig. 7. Plot of LD versus (a) peak height and (b) concentration of injection (mg/ml). (a) Simulated two-component mixture. For data sets, see Table 1, except that the peak height (h) was changed in the test. (b) Real sample: salicylic acid + 5% 4-hydroxybenzoic acid.

Our studies indicate that when a powerful computer is used, the neural networks approach may be capable of assessing chromatographic peaks. Clearly, this approach appears to have the advantage that it is unnecessary to have reference substances for both the sample and the impurity. The sensitivity of this method depends on the differences in the chromatograms and spectra of the component substances. Additionally, the simulation tests demonstrate that it is unnecessary to check whether the retention of the impurity is greater or smaller than that of the main substance, since the approach is based on the difference in the spectra of two parts of the chromatogram.

Compared with PCA, ANN seems more sensitive than PCA to an impurity with a similar retention and spectral response, but it behaves slightly worse than PCA for samples in which the retentions and spectra of the main substance and the impurity are strongly separated, since only half of the data are used. Further, PCA can predict the number of components in a mixture or perform quantitative analysis whereas ANN

Table 4
Computation results for data sets simulated

	Training sets					Testing sets					U-test
	1	2	3	4	5	6	7	8	9	10	
Case 1											
				C_{max}		σ_x	σ_y	μ_x	μ_y		
Main component				100		2	2	20	20		
Impurity				30		2	2	17	23		
OP	0.0383	0.0565	0.0764	0.0942	0.1057	0.1078	0.1004	0.0859	0.0683	0.0509	$U_1 = 25.0$
TP	0.0388	0.0571	0.0767	0.0943	0.1065	0.1109	0.1074	0.0974	0.0834	0.0681	$U_2 = 0.0$
|dif|	0.0005	0.0006	0.0003	0.0001	0.0008	0.0032	0.0070	0.0114	0.0152	0.0172	(Impure)
Case 2											
				C_{max}		σ_x	σ_y	μ_x	μ_y		
Main component				100		2	2	20	23		
Impurity				30		2	2	17	20		
OP	0.0726	0.0853	0.0982	0.1089	0.1145	0.1132	0.1051	0.0924	0.0780	0.0648	$U_1 = 25.0$
TP	0.0681	0.0834	0.0974	0.1074	0.1109	0.1065	0.0943	0.0767	0.0571	0.0388	$U_2 = 0.0$
|dif|	0.0045	0.0018	0.0009	0.0015	0.0036	0.0068	0.0109	0.0157	0.0209	0.0260	(Impure)
Case 3											
				C_{max}		σ_x	σ_y	μ_x	μ_y		
Main component				100		2	2	17	23		
Impurity				30		2	2	20	20		
OP	0.0697	0.0840	0.0977	0.1080	0.1121	0.1084	0.0971	0.0809	0.0632	0.0472	$U_1 = 25.0$
TP	0.0681	0.0834	0.0974	0.1074	0.1109	0.1065	0.0943	0.0767	0.0571	0.0388	$U_2 = 0.0$
|dif|	0.0016	0.0006	0.0003	0.0006	0.0012	0.0019	0.0028	0.0042	0.0061	0.0084	(Impure)
Case 4											
				C_{max}		σ_x	σ_y	μ_x	μ_y		
Main component				100		2	2	17	20		
Impurity				30		2	2	20	23		
OP	0.0401	0.0576	0.0769	0.0948	0.1073	0.1171	0.1075	0.0965	0.0818	0.0663	$U_1 = 17.0$
TP	0.0388	0.0571	0.0767	0.1074	0.0943	0.1065	0.1109	0.0974	0.0834	0.0681	$U_2 = 8.0$
|dif|	0.0022	0.0015	0.0007	0.0006	0.0023	0.0068	0.0137	0.0213	0.0278	0.0235	(Pure)

OP = simulated values; TP = predicted values; |dif| = absolute differences of OP and TP.

cannot so far. Further work is needed on this aspect.

5. Conclusions

The results of this work have shown that the artificial neural network technique has considerable sensitivity for assessing peak purity in liquid chromatography with photodiode-array detection. The accuracy relies upon the magnitude and ratio of the peak heights of the main substance and the impurity, the width of the chromatographic peaks and the characteristics of their UV spectra. The necessary information can be obtained from the sample using a microcomputer without the need for prior knowledge about the impurity. It is possible to use this method to check peak purity on-line for routine use or for quality control in chromatography with photodiode-array detection. Further work will include the investigation of quantitative calibration and

the assessment of peaks consisting of more than two components.

Acknowledgement

This research was supported by the National Science Foundation of China.

References

[1] F. Huang, J.-H. Kang and Y.-Z. Hu, Chin. J. Chromatogr., 11 (1995) 1–5.
[2] L. Huber, Application of Diode-Array Detection in HPLC, Hewlett-Packard, Avondale, PA, 1989.
[3] G.G.R. Seaton, J.G.D. Marr, B.J. Clark and A.F. Fell, Anal. Proc., 23 (1986) 424–426.
[4] P.C. White and T. Catterick, J. Chromatogr., 402 (1987) 135–147.
[5] F.V. Warren, Jr., B.A. Bidlingmeyer and M.F. Delancy, Anal. Chem., 59 (1987) 1897–1907.
[6] J.G.D. Marr, G.G.R. Seaton, B.J. Clark and A.F. Fell, J. Chromatogr., 506 (1990) 289–301.
[7] A.F. Fell, H.P. Scott, R. Gill and A.C. Moffat, J. Chromatogr., 282 (1983) 123–140.
[8] E. Grushka and D. Israeli, Anal. Chem., 62 (1990) 717–721.
[9] A. Grant and P.K. Bhattacharyya, J. Chromatogr., 347 (1985) 219–235.
[10] W.K. Lawton and E.A. Sylvestre, Technimetrics, 13 (1971) 617–623.
[11] P.J. Gemperline and J.C. Hamilton, Anal. Chem., 61 (1989) 2240–2243.
[12] Y.-Z. Hu, J.-H. Kang, J. Yu and F. Huang, J. Chin. Pharm. Univ., 24 (1993) 290–294.
[13] H.R. Keller and D.L. Massart, Anal. Chim. Acta, 246 (1991) 379–390.
[14] J. Zupan and J. Gasteiger, Anal. Chim. Acta, 248 (1991) 1–30.
[15] P.A. Jansson, Anal. Chem., 63 (1991) 357A–362A.
[16] A. Boss, M. Bos and W.E. van der Linden, Anal. Chim. Acta, 256 (1992) 133–144.
[17] M. Hartnett, D. Diamond and P.G. Barker, Analyst, 118 (1993) 347–354.
[18] Ch. Affolter and J.T. Clerc, Chemometr. Intell. Lab. Syst., 21 (1993) 151–157.
[19] Q.C. van Est, P.J. Schoenmakers, J.R.M. Smits and W.P.M. Nijssen, Vibr. Spectrosc., 4 (1993) 263–272.
[20] M. Bos and H.T. Weber, Anal. Chim. Acta, 247 (1991) 97–105.
[21] S.R. Gallant, S.P. Fraleigh and S.M. Cramer, Chemometr. Intell. Lab. Syst., 18 (1993) 41–57.
[22] P.M.J. Coengracht, H.J. Metting, E.M. van Loo, G.J. Snoeijer and D.A. Doornbos, J. Chromatogr., 631 (1993) 145–160.
[23] J. Zupan, Algorithms for Chemists, Wiley, Chichester, 1989, pp. 51–54.
[24] J.B. Castledine, A.F. Fell, R. Mobin and B. Sellberg, J. Chromatogr., 592 (1992) 27–36.
[25] J.P. Foley and J.G. Dorsey, J. Chromatogr. Sci., 22 (1984) 44–46.
[26] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman, Chemometrics: a Textbook, Elsevier, Amsterdam, 1988, pp. 51–54.
[27] S. Ebel and W. Mueck, Chromatographia, 25 (1988) 1075–1086.
[28] E.R. Malinowski, J. Chemometr., 1 (1987) 33–40.