*Journal of Chromatography*, 631 (1993) 145–160
Elsevier Science Publishers B.V., Amsterdam

CHROMSYMP. 2606

# Peak tracking with a neural network for spectral recognition

P. M. J. Coenegracht, H. J. Metting, E. M. van Loo, G. J. Snoeijer and D. A. Doornbos

*Chemometrics Research Group, University Centre for Pharmacy, NL-9713 AW Groningen (Netherlands)*

ABSTRACT

A peak tracking method based on a simulated feed-forward neural network with back-propagation is presented. The network uses the normalized UV spectra and peak areas measured in one chromatogram for peak recognition. It suffices to train the network with only one set of spectra recorded in one chromatogram of a sample to recognize the sample peaks in other chromatograms recorded in different mobile phases. The peak recognition method was used in a mixture design-based optimization of the separation of eight sulphonamides, some of which have very similar spectra. Even peaks in a cluster of four to five overlapping peaks could be recognized.

## INTRODUCTION

### *Peak tracking*

For the optimization of ternary and quaternary mobile phases in HPLC by simultaneous methods, mixture designs are often used [1–3]. Peak tracking is essential for the application of these methods because the capacity factor of each analyte in the sample has to be determined in every design point. Peak recognition has to be discerned from solute identification. In solute identification the chemical identity of a peak is determined by comparison with the retention time or spectrum of a reference standard. If the number of analytes is not known then the issue of the peak homogeneity has to be addressed, otherwise peak recognition is sufficient. If the number of solutes is known it suffices to keep track of the band identity when the chromatographic conditions are changed.

Even if two peaks overlap completely, identification by exclusion is possible if the other peaks are recognized, and both can be given the same value of the capacity factor. Severely overlapping peaks pose a different problem as their identity has to established unequivocally. This may be done by separate injection of the reference substances and comparison of retention times. This is time consuming and reference substances have to be available. If reference substances are not available, comparison of peak areas is another possibility. At least a 20% difference in peak areas is needed for the correct assignment of each peak and the number of peaks has to be equal to the number of solutes in all chromatograms [4]. Moreover, peak areas may change considerably when the changing chromatographic conditions shift the UV spectra of the solutes.

Multiple-wavelength detection considerably increases the information content of the chromatograms and spectral matching can be used to recognize peaks in different runs. Deconvolution techniques can also be used to extract peak profiles and even component spectra for overlapping peaks [5]. By factor analysis methods and a set of reference spectra the presence of a solute in peak cluster can be revealed [target transformation factor analysis (TTFA)]. Both the individual profiles and component UV spectra can be derived by an iterative

*Correspondence to:* P. M. J. Coenegracht, Chemometrics Research Group, University Centre for Pharmacy, Ant. Deusinglaan 2, NL-9713 AW Groningen, Netherlands.

0021-9673/93/$06.00 

procedure (ITTFA) [5]. These methods are very powerful, have complementary characteristics and require expert knowledge for their correct implementation.

Peak size plus relative retention have been combined in a peak tracking procedure [6], but the combination of peak size and spectral information has also been used [6,7]. Reviews of different peak tracking methods have been published recently [4,8,9].

Wright *et al.* [10] used a peak recognition strategy based on comparison of peak integrals determined at different wavelengths in an automated optimization procedure. They stated that the procedure requires only complete resolution of each solute in one of the seven chromatograms resulting from the mixture design and complete resolution of all components in a single chromatogram is not required. The tracking procedure was shown to deal with situations where moderate two component overlap occurs.

The tracking procedure proposed in this paper uses UV spectra collected by a multi-channel diode-array spectrophotometer and peak areas measured at one wavelength. The spectra of all components are measured in one chromatogram, normalized and multiplied by the corresponding peak areas and used for training a neural network. To obtain the training set of the pure spectra, the front or tail of a two-component chromatographic peak has to contain the single pure component and complete resolution is not necessary. The trained network is used for peak recognition and again complete resolution is not necessary. In principle, none of the chromatograms resulting from the mixture design has to be separated completely, the spectral similarity of the different solutes can be high and shifts of the UV spectra on changing the mobile phase composition are permitted.

*Neural network*

Neural networks are mathematical systems designed for parallel processing of information consisting of processing elements, neurons or nodes. The following description applies to a typical feedforward neural network with back-propagation and is restricted to the information necessary to understand its functioning. Books on different types of network [11], the back-propagation network [12] and papers reviewing [13] or describing the theory and application of the back-propagation network [14–17] have appeared recently.

The typical feedforward neural network with back-propagation has three layers of nodes: the input, the hidden and the output layer. The nodes of the input layer accept the input vector $\boldsymbol{X}_k$ (absorbance values measured at different wavelengths of the UV spectrum in this instance) and each node is connected to every node of the hidden layer. Each node of the hidden layer is connected to every node of the output layer that produces the output vector, $\boldsymbol{Y}_k$. These connections are weights which are applied to signals passed from one neuron to the next. The nodes in one layer are not interconnected.

There is one input node per variable in a spectrum. The input nodes pass the weighted input signals to the nodes in the hidden layer. The hidden layer nodes transform the sum of the weighted inputs by a non-linear, sigmoidal transfer function and pass weighted signals to the nodes of the ouput layer. The output nodes produce output signals that are obtained by applying again a sigmoidal transfer function (eqn. 1) upon the sum of the weighted values passed to them.

$$f(x) = \frac{1}{1 + e^{-(x+\theta)}} \tag{1}$$

The bias parameter, $\theta$, can be treated as an additional weight that is always added to the weight vector of every node, but is multiplied by a value of 1 instead of $y_j$ or $y_i$ (see eqns. 2 and 4). It can be regarded as a threshold value at which the output of a neuron is released.

The number of output nodes is equal to the number of components of the sample; the number of hidden nodes is an adjustable parameter and was chosen to be equal to the number of input nodes.

The weights or connection strengths between the nodes in the consecutive layers of the network form two matrices, which contain random numbers in the untrained network. In the trained network the matrices contain the modified weights and represent the knowledge of the network. The weights are modified by supervised learning. During the training pairs of input and output vectors, $\boldsymbol{X}_k$ and $\boldsymbol{T}_k$, are presented to the network. The training set consists of spectra of the solutes and the corresponding target vectors. During the training the produced output

vector, $\boldsymbol{Y}_k$, is repeatedly compared with the target vector, $\boldsymbol{T}_k$. Each time the weights are adjusted in the direction of the correct answer. The change of a weight, $\Delta w_{jk}$, between a node $j$ of the hidden layer and a node $k$ of the output layer is proportional to the difference, $\delta_k$, between an element of the target vector and the corresponding element of the output vector and to the value $y_j$ of the signal passed on by the node of the hidden layer to the output node (delta rule):

$$\Delta w_{jk} = \eta\ \delta_k\ y_j \tag{2}$$

where $\eta$ is the learning rare and

$$\delta_k = (t_k - y_k)\ y_k(1 - y_k) \tag{3}$$

For the calculation of the error term, $\delta_j$, of node $j$ of the hidden layer, comparison with a target value is not possible, but the error has to be calculated from the output layer:

$$\Delta w_{ij} = \eta\ \delta_j\ y_i \tag{4}$$

$$\delta_j = \sum_k (\delta_k\ w_{jk})\ y_j(1 - y_j) \tag{5}$$

where $\Delta w_{ij}$ is the weight between node $i$ of the input layer and node $j$ of the hidden layer and $y_i$ is the output of input node $i$.

To summarize the back-propagation algorithm: $\boldsymbol{X}_k$ is propagated through the network to the output layer. The output $\boldsymbol{Y}_k$ is compared with the target, $\boldsymbol{T}_k$, node by node and used to correct the weights. Corrections for the hidden layer are calculated from the errors of the output layer.

To complete the description of the back-propagation network, the momentum term has to be discussed. The momentum factor, $\alpha$, relates the present value of $\Delta w(t)$ to its value in the previous learning cycle, $\Delta w(t-1)$, and furthers faster convergence with fewer oscillations and avoids entrapment in local minima.

$$\Delta w_{ij}(t) = \eta\ [(1 - \alpha)\ \delta_j y_i + \alpha\ \Delta w_{ij}(t-1)] \tag{6}$$

The neural network programme used in this investigation uses eqn. 6 and a value of 0.9 for $\alpha$. When a neural network is used for spectral recognition one attempts to obtain a model that produces the correct set of outputs for a set of spectra. Such a model is obtained by training the network. The trained neural network should be able to generalize from the examples presented during the training, *i.e.*, spectra of several analytes recorded in a given mobile phase composition, to other inputs that it has not yet seen, *i.e.*, spectra of the same analytes recorded in different mobile phase compositions.

An important issue in the application of a neural network is the choice of the training set: how many examples are necessary and how well has the variable space to be covered by the selected examples to obtain a trained network that produces correct outputs?

In this work we obtained two collections of spectra, one of twelve sulphonamide derivatives and another of fifteen mainly benzene derivatives, recorded in six different mobile phase mixtures composed of water, methanol (MeOH) and acetonitrile (MeCN) (Tables I and VI) to train two networks. Each collection of spectra consists of six sets of twelve and fifteen spectra, respectively. First we investigated how many of the six sets were needed to obtain a trained network that could recognize the spectra in the remaining sets correctly. We also investigated whether the number of training sets could be reduced by multiplying the normalized spectra with a number corresponding to the peak area. Second, a subset of eight sulphonamide derivatives was optimized by a mixture design procedure [3] using the same variable space to test the procedure. To make the test more severe, an old column with a low plate number was used.

## EXPERIMENTAL

### *Apparatus and materials*

The spectra were recorded with a Philips PU4120 diode-array detector. This detector was combined with a Waters Assoc. Model 6000 A pump and a Rheodyne Model 7010 injection valve fitted with a 20-$\mu$l loop. Separations were performed at room temperature (*ca.* 20°C) on a 200 × 4 mm I.D. stainless-steel column packed with Nucleosil RP-8, particle size 5 $\mu$m, $N = 2000$. The flow-rate was 1.0 ml/min. The dead time was measured at all design points by injection of uracil and was 2.0 min.

All calculations, *i.e.*, data handling of the diode-array spectra, peak integration, principal component analysis (PCA) and simulation and training of the neural network were performed on an IBM-compatible AT 286 personal computer with mathematical coprocessor.

Mobile phases were prepared from HPLC-grade methanol and acetonitrile (Labscan) and purified deionized water (Milli-Ro/Milli-Q system; Millipore). In the sulphonamide experiments water was acidified with 1% acetic acid. The test solutes were used as purchased. The sulphonamides, diazepam, nicotinic acid and theophylline were of pharmacopoeial grade and the benzene derivatives of "zur Synthese" grade (Merck, Darmstadt, Germany). The test solutions for recording of the spectra were prepared freshly at concentrations varying from 5 μg/ml for the sulphonamides to between 1 and 150 μg/ml for the other solutes, depending on their molar absorptivities. The sulphonamide solutions were acidified with 1% acetic acid. The concentration of the sulphonamides in the test sample for the optimization varied from about 30 to 160 μg/ml.

The spectra of the sulphonamides, A–L (Fig. 1), constituted the first data set. The second data set was formed by the spectra of aniline (A, 0.27), N-butyl *p*-aminobenzoate (B, 0.57), *p*-cresol (C, 0.63), dimethyl phthalate (D, 0.39), ethylbenzene (E, 0.33), *p*-nitrobenzaldehyde (F, 0.69), nitrobenzene (G, 0.45), 2-phenylethanol (H, 0.21), *n*-propyl *p*-hydroxybenzoate (I, 0.15), *p*-toluidine (J, 0.87), toluene (K, 0.93), diazepam (L, 0.51), nicotinic acid (M, 0.75), theophylline (N, 0.81) and uracil (O, 0.99), where the numbers in parentheses refer to peak area-factors mentioned below.

*Software*

The following programmes were used: neural network simulation, Brainmaker vs. 2 (California Scientific Software, Sierra Madre, CA, USA); diode-array data handling, Philips PU 6003 vs. 3.0 software (Analytical Chromatography); data system, vs. 6.0, Philips PU 6000 integration software; principal components analysis, Unscrambler II, vs. 3.5 (Camo, Trondheim, Norway); and optimization software, POEM (predicting optimum eluent mixtures), vs. 3.1 (University Centre of Pharmacy, Groningen, Netherlands [3]).

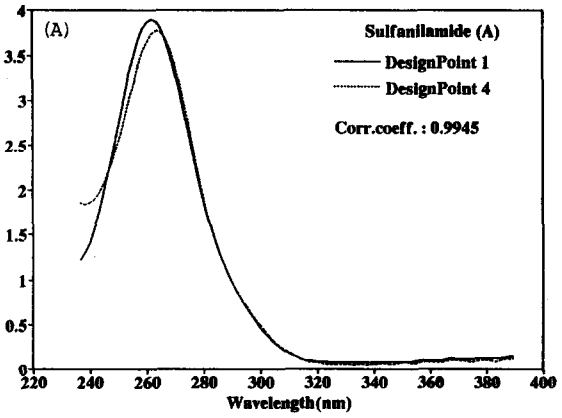
## RESULTS

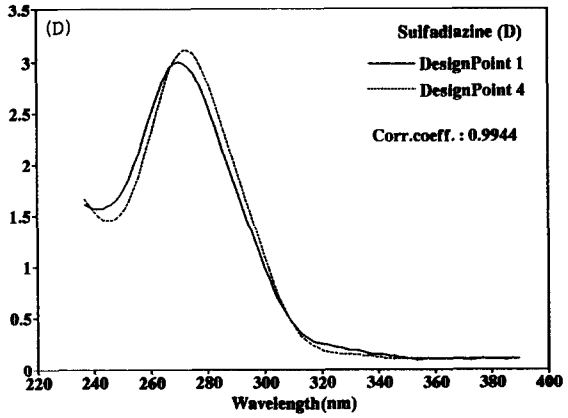
*Neural network training*

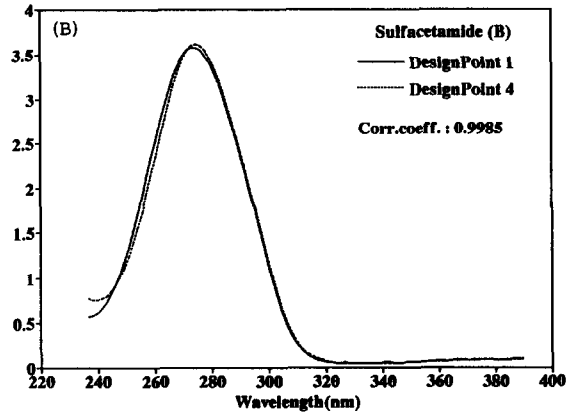
*Sulphonamides.* The spectra of the twelve sulphonamides (Fig. 1A–L) were recorded from 237 to 390 nm in the six eluent mixtures indicated in Table I. One spectrum consisted of 98 absorbance values, which were scaled to the same area under the curve to remove the concentration effect from the spectra. The normalized area is equal to the number of data points in the spectrum. To obtain an impression of the spectral similarity, the correlation coefficients between the twelve spectra recorded in design point 4 of Table I were calculated and all 72 spectra were subjected to a PCA. The results are shown in Table II and Fig. 2. The first two PCs explained 87% of the variation in the data. From the score plot a strong overall similarity is indicated between the spectra of analytes G, J and D, D and E and K and A. This result is corroborated by high values of the corresponding correlation coefficient of the corresponding analytes in one eluent composition (Table II). From the score plot one can conclude that two PCs are not sufficient to identify the analytes.

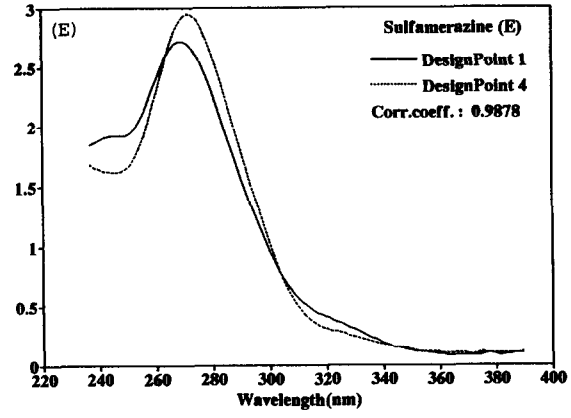
The neural network was designed to have 98 input nodes, 98 nodes in the hidden layer and 12 output nodes, the learning rate $\eta = 1$ and the momentum factor $\alpha = 0.9$. The value of the output nodes can vary between 0 and 1. The network was trained by presenting repeatedly a number of training sets to the network. One set can consist of twelve spectra recorded in one eluent plus twelve output vectors. For example, the spectrum of sulphonamide A is presented together with an output vector consisting of twelve elements, where 1 is the first element and the remaining eleven elements are zeros; the spectrum of sulphonamide B is combined with a vector of which the second element is 1 plus eleven zeros, etc. The training stops if the difference between the output value of each node and the desired value is 0.1. The trained network is presented with a test set of twelve spectra recorded in an eluent that has not been used in the training set and the network is expected to produce the correct output vector. The output vector was assumed to be correct if the value of the correct output node was greater than 0.7 while the remaining output values were smaller than 0.3. These limit values were chosen arbitrarily but are considered to be a severe criterion for making a decision.

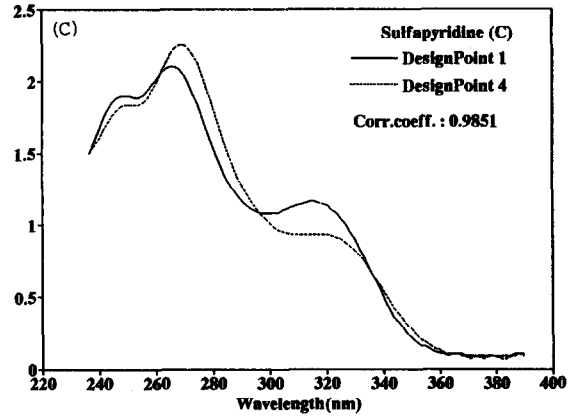
It is to be expected that if a network is trained with a training set of 48 spectra recorded in eluents 1, 2, 3 and 4 of Table I, the network should produce the correct output vectors, because these design points



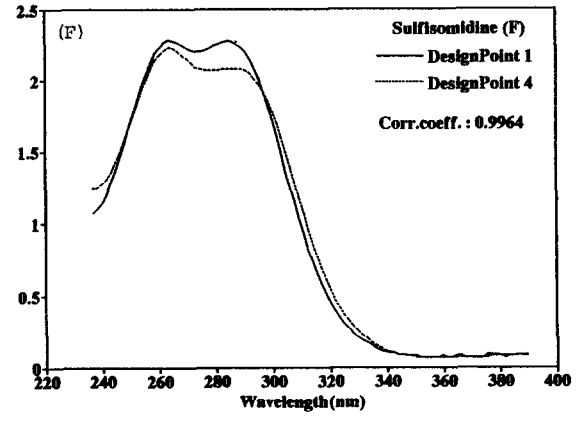




Fig. 1.

(Continued on p. 150)

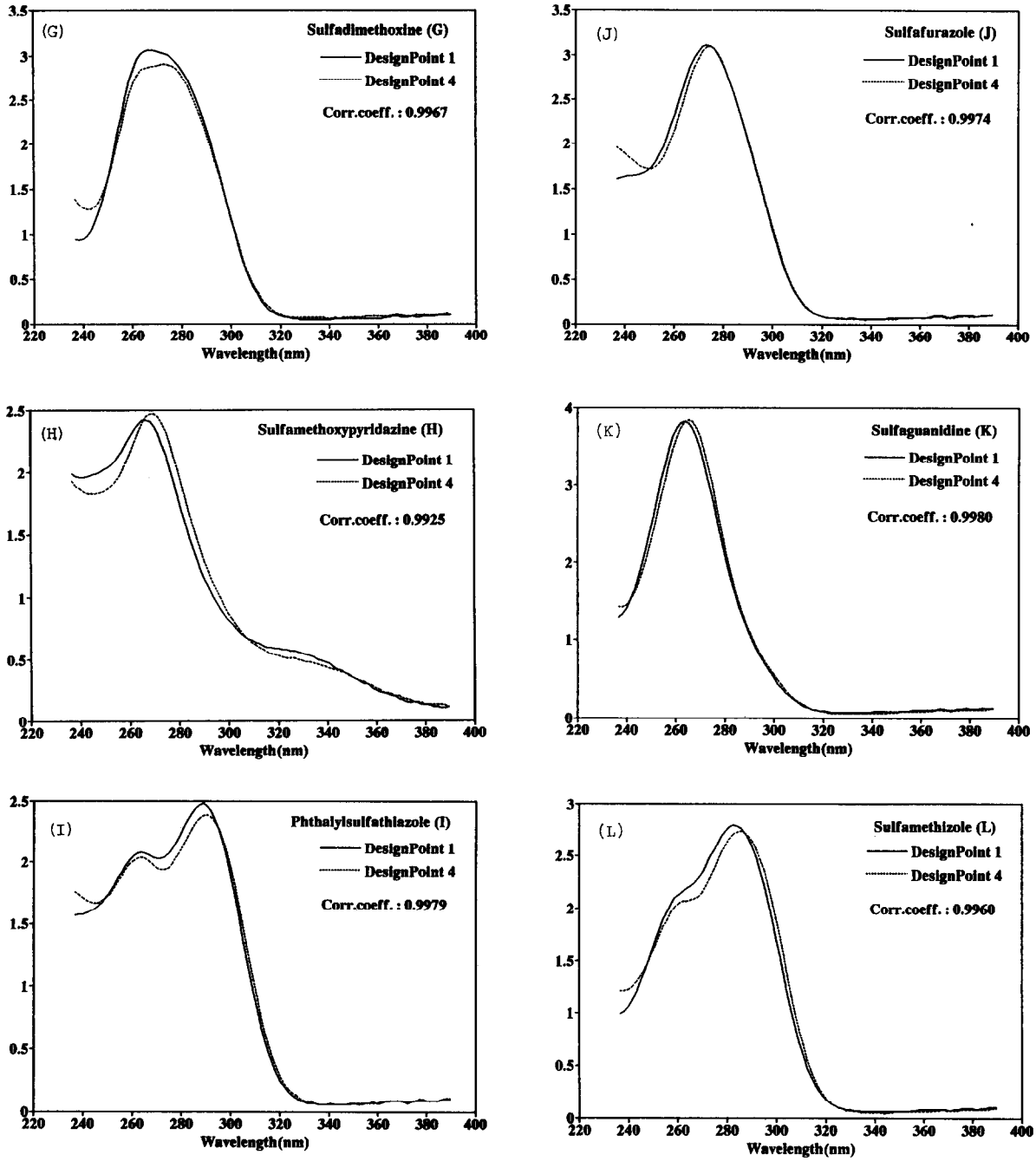



Fig. 1. Spectra of twelve sulphonamides normalized to the same area under the curve. The spectra were recorded in eluent compositions (solid lines) 1 and (dotted lines) 4 in Table I. The correlation coefficient between the two spectra is shown.

TABLE I

ELUENT COMPOSITIONS AT DESIGN POINTS 1–6 (SULPHONAMIDES)

| Design point No. | Fraction | | |
|---|---|---|---|
| | Water | MeOH | MeCN |
| 1 | 0.80 | 0.20 | 0.00 |
| 2 | 0.60 | 0.40 | 0.00 |
| 3 | 0.90 | 0.00 | 0.10 |
| 4 | 0.70 | 0.00 | 0.30 |
| 5 | 0.70 | 0.15 | 0.15 |
| 6 | 0.80 | 0.10 | 0.10 |

span the whole design space. The results of the use of different training and test sets are presented in Table III. The Error column indicates the number of the spectra that were not classified correctly according to the decision criterion, the non-error rate (NER) gives the fraction of correct classifications of the total number of classifications performed; the Cycles column indicates how many times the training set had to be presented to train the network and the Time column gives the time needed for training.

The experiments in Table III were repeated after the addition of random noise with a Gaussian band width of 10% to the training data. The results improved very slightly but the training time increased considerably, up to 3 h for the first experiment in Table III. Therefore, we decided to add no noise to the training data.

Two more transformations of the data were performed to investigate their effect on network training. In both experiments spectra recorded in eluents 1, 2, 3 and 4 were used for training and the spectra for eluents 5 and 6 in Table I for testing.

PCA of the training set resulted in five PCs which explained completely the variation in the data. An object in the training set is now characterized by five scores instead of 98 absorbance values. A new network consisting of five input nodes, twelve nodes in the hidden layer and twelve output nodes needed 15 h for training. After increasing the number of nodes in the hidden layer to 25 the training time was reduced to 2.5 h. Both networks identified correctly the objects of the test set on the basis of their scores on the five PCs.

From the loading plot of the training set, ten wavelengths were selected that contribute most to the variation of spectra in the training set. Now an object is characterized by 10 instead of 98 absorbance values. Another network consisting of ten input nodes, twelve nodes in the hidden layer and twelve nodes in the output layer was configured. It needed 2.2 h to train and produced correct results.

On the basis of these results we decided to use whole spectra and a network of $98 \times 98 \times 12$ nodes for classification because of its good performance in the shortest training time. For peak tracking only one set of twelve spectra recorded in only one eluent

TABLE II

CORRELATION COEFFICIENTS OF SPECTRA OF SULPHONAMIDES RECORDED IN DESIGN POINT 4 OF TABLE I

| Sulphonamide | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1.000 | 0.793 | 0.896 | 0.887 | 0.907 | 0.804 | 0.890 | 0.952 | 0.757 | 0.877 | 0.993 | 0.738 |
| B | | 1.000 | 0.820 | 0.973 | 0.955 | 0.887 | 0.973 | 0.863 | 0.857 | 0.954 | 0.848 | 0.924 |
| C | | | 1.000 | 0.908 | 0.934 | 0.905 | 0.894 | 0.970 | 0.857 | 0.904 | 0.897 | 0.831 |
| D | | | | 1.000 | 0.997 | 0.924 | 0.993 | 0.953 | 0.903 | 0.994 | 0.917 | 0.930 |
| E | | | | | 1.000 | 0.928 | 0.988 | 0.972 | 0.904 | 0.994 | 0.930 | 0.920 |
| F | | | | | | 1.000 | 0.939 | 0.891 | 0.982 | 0.918 | 0.824 | 0.973 |
| G | | | | | | | 1.000 | 0.939 | 0.916 | 0.983 | 0.921 | 0.946 |
| H | | | | | | | | 1.000 | 0.865 | 0.956 | 0.950 | 0.846 |
| I | | | | | | | | | 1.000 | 0.915 | 0.768 | 0.980 |
| J | | | | | | | | | | 1.000 | 0.898 | 0.932 |
| K | | | | | | | | | | | 1.000 | 0.770 |
| L | | | | | | | | | | | | 1.000 |

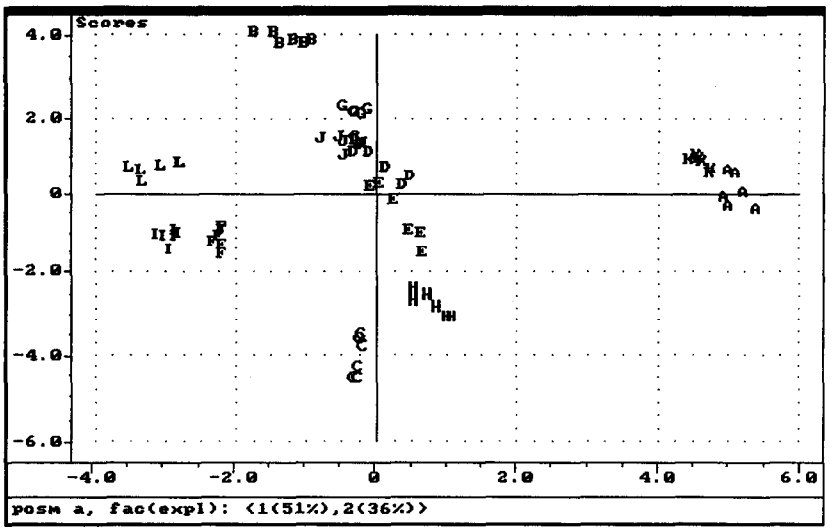



Fig. 2. Score plot of twelve sulphonamide spectra recorded in the six eluent compositions in Table I.

composition should be sufficient to train a network, because then only one reference chromatogram is necessary. Other, more complicated, strategies are possible, but these are not discussed. If we assume that from one chromatogram the spectra and approximate peak areas can be extracted, then this situation can be simulated by multiplying each of the normalized spectra recorded in a given eluent composition with a number that represents the peak area of the analyte in the corresponding chromatogram. The spectra of the sulphonamides were multiplied by the peak area values given in Table IV. The spectra recorded in one design point of Table I were used to train the network and the sets recorded at the five remaining design points were used as test sets. Table V shows the results. Only 21 of the 360 classifications were incorrect according to the decision criterion but only four times was a wrong analyte recognized: the spectrum of sulphadimethoxine (G) was attributed to sulphanilamide (A) three times and to sulphacetamide (B) once. This indicates that the decision criterion is too severe.

*Benzene derivatives.* The second set of spectra consisted of the spectra of eleven benzene derivatives and four other compounds recorded in six different eluent compositions (Table VI). These compounds give acceptable retention times in this factor space, their spectra have more detail and the similarity

TABLE III

RESULTS OF TRAINING WITH DIFFERENT SETS OF NORMALIZED SULPHONAMIDE SPECTRA

| Number of eluent in Table I used | | Error | NER[a] | Cycles | Time (h) |
|---|---|---|---|---|---|
| For training | For testing | | | | |
| 1+2+3+4 | 5+6 | 0 | 1.00 | 129 | 0.50 |
| 1+4 | 5+6 | 1 | 0.96 | 106 | 0.22 |
| 2+3 | 5+6 | 0 | 1.00 | 138 | 0.37 |
| 5 | 1+4 | 6 | 0.75 | 69 | 0.10 |

[a] NER = non-error rate.

TABLE IV

PEAK AREA VALUES FOR SULPHONAMIDES A–L

| Sulphonamide | Peak area | Sulphonamide | Peak area |
|---|---|---|---|
| A | 0.69 | G | 0.55 |
| B | 0.83 | H | 0.48 |
| C | 0.76 | I | 0.34 |
| D | 0.41 | J | 0.20 |
| E | 0.90 | K | 0.27 |
| F | 0.62 | L | 0.97 |

TABLE VI

ELUENT COMPOSITIONS AT DESIGN POINTS 1–6 (BENZENE DERIVATIVES)

| Design point No. | Fraction | | |
|---|---|---|---|
| | Water | MeOH | ACN |
| 1 | 0.60 | 0.40 | 0.00 |
| 2 | 0.40 | 0.60 | 0.00 |
| 3 | 0.65 | 0.00 | 0.35 |
| 4 | 0.45 | 0.00 | 0.55 |
| 5 | 0.55 | 0.25 | 0.20 |
| 6 | 0.45 | 0.25 | 0.30 |

between the different analytes is smaller, but the spectral shifts of the individual analytes in different eluents are larger (Fig. 3). The clusters in Figs. 3 and 2 seem to be similar, but the first two PCs explain 80% of the variance in the spectra of the benzene derivatives against 87% in Fig. 2.

Spectra were recorded from 190 to 346 nm. Per spectrum 100 absorbance values were collected and after normalization the spectra were presented to a network consisting of 100 input nodes, 15 output nodes and a variable number of nodes in the hidden layer. In one experiment the normalized spectra were multiplied by area factors ranging from 0.15 to 0.99 (see Experimental). From the results presented in Table VII one can conclude that the performance of the network decreases as indicated by the NER and that the training time increases as indicated by the number of cycles when the number of hidden nodes is decreased from 120 to 10. The decrease in performance is not linearly related to the number of nodes. There seems to be an optimum number of nodes for the hidden layer at about 120 nodes; with 130 nodes in the hidden layer the performance seems to decrease also. For the existence of an optimum number of nodes in the hidden layer more indications were found from experiments with a network that had two hidden layers.

The number of cycles is an indication of the time necessary to train the network, but the training time also depends on the amount of data in the training set and on the performance of the computer. On a training set of twelve spectra the network is trained in a few minutes on a fast PC. The last two experiments of Table VII show clearly that it is essential to multiply the normalized spectra by an area factor to train a neural network successfully on only one training set.

TABLE V

NER AFTER TRAINING WITH ONE SET OF AREA MULTIPLIED SULPHONAMIDE SPECTRA

| Number of eluent in Table I used | | | | | | |
|---|---|---|---|---|---|---|
| Training | Testing | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | – | 0.83 | 1.00 | 0.83 | 0.83 | 0.92 |
| 2 | 1.00 | – | 1.00 | 1.00 | 1.00 | 1.00 |
| 3 | 1.00 | 1.00 | – | 1.00 | 1.00 | 1.00 |
| 4 | 0.83 | 0.83 | 0.92 | – | 0.92 | 0.92 |
| 5 | 1.00 | 1.00 | 0.92 | 1.00 | – | 0.92 |
| 6 | 0.92 | 0.83 | 1.00 | 1.00 | 0.92 | – |

*Peak tracking and optimization*

It is the aim of this paper to demonstrate the feasibility and some limitations of peak tracking with a neural network in an optimization procedure based on a simplex lattice mixture design. A subsample of eight of the sulphonamides was selected, because the spectra of the sulphonamides are very similar. It was decided to use a ternary mobile phase consisting of water, methanol and acetonitrile, because we prefer this approach to the use of a quaternary eluent consisting of three isoeluotropic binary eluents of water with methanol, acetonitrile and tetrahydrofuran. With the preferred approach it

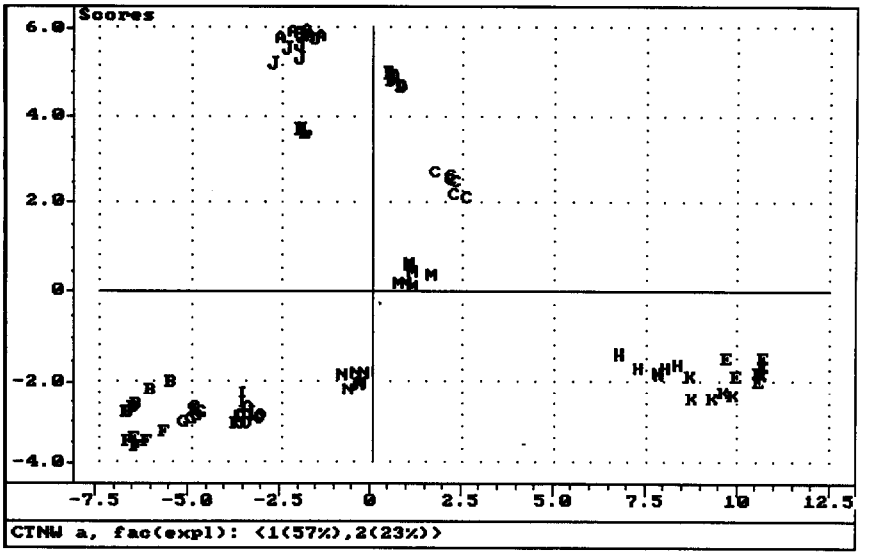



Fig. 3. Score plot of fifteen benzene derivative spectra recorded in the six eluent compositions in Table VI.

is possible to optimize simultaneously the solvent strength and solvent selectivity and no experiments have to be performed to select the appropriate isoeluotropic eluents if one has an approximate knowledge of the required solvent strength [18]. The factor space and the design points are shown in Fig. 4; the corresponding mobile phase compositions and measured peak areas in Table VIII.

Chromatograms were recorded for the nine mobile phases required by the design and peak areas were determined with the integration software at 270 nm (Table VIII). Every 2 s a spectrum was recorded and for each peak of every chromatogram one spectrum was selected at the peak maximum. The selected spectra were normalized and multiplied by the peak areas. In this way a set of eight spectra per chromatogram were obtained to be used for peak recognition by the trained network. The net-

TABLE VII

RESULTS OF TRAINING DIFFERENT NETWORKS WITH DIFFERENT SETS OF NORMALIZED BENZENE DERIVATIVE SPECTRA

| Number of eluent in Table VI used | | Number of hidden nodes | NER | Cycles |
|---|---|---|---|---|
| For training | For testing | | | |
| 1+2+3+4 | 5+6 | 130 | 0.93 | 96 |
| 1+2+3+4 | 5+6 | 120 | 1.00 | 93 |
| 1+2+3+4 | 5+6 | 115 | 0.97 | 104 |
| 1+2+3+4 | 5+6 | 100 | 0.93 | 106 |
| 1+2+3+4 | 5+6 | 58 | 0.90 | 216 |
| 1+2+3+4 | 5+6 | 10 | 0.87 | 235 |
| 5 | 1+2+3+4 | 100 | 0.60 | – |
| 5[a] | 1+2+3+4 | 100 | 1.00 | 75 |

[a] Spectra multiplied by area factor.

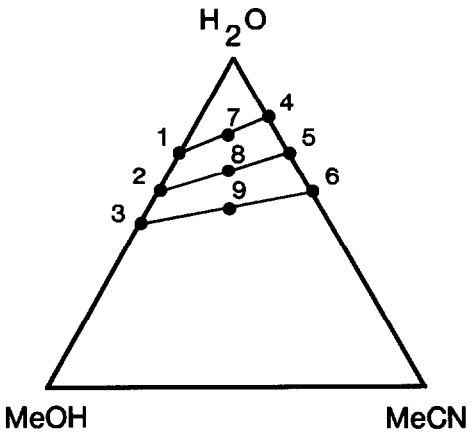



Fig. 4. Mixture triangle which shows the location of the design points in the factor space used for the optimization of the separation of sulphonamides.

work was trained with the spectra obtained from the chromatogram recorded with mobile phase 7 (Fig. 4), which then was taken as the reference chromatogram (Fig. 5G). The reference substances were injected separately in the mobile phase of design point 8 (Fig. 4) to establish the chemical identity of the peaks. The peak areas of the same peak measured in the different chromatograms are not constant (Table VIII) as they are determined for fused peaks by the perpendicular drop method. In a few instances only one integration value was produced by the software for a pair of fused peaks and the individual peak areas were estimated (see Table VIII, values marked [a]) from the produced value using the ratio of the reference chromatogram. The inaccuracy of the area factor probably sometimes affected peak recognition.

In the chromatogram in Fig. 5E (design point 5) peak A was attributed by the network to sulphanilamide and sulphadimetoxine with almost the same output value. In this chromatogram, however, peak G was correctly recognized as sulphadimethoxine and therefore it was safe to assume that peak A belonged to sulphanilamide. The same problem occurred in the chromatogram in Fig. 5H (design point 8) with peaks E and J: two solutes were attributed to each peak. As the remaining peaks of the chromatogram were recognized unequivocally, it was nevertheless possible to attribute the correct solutes to the peaks in question by selecting the two solutes that had not been attributed unequivocally to the other peaks and choosing between them on the basis of the retention time. The recognition of the remaining peaks in chromatograms of design points 1, 2, 4, 5 and 8 (Fig. 5A, B, D, E and H) on the basis of the highest output value of the network was correctly performed without problems.

The chromatograms in Fig. 5C, F and I posed a serious problem owing to very strong overlap of several peaks. In the chromatogram in Fig. 5C the

TABLE VIII

ELUENT COMPOSITIONS AT THE DESIGN POINTS 1–9 OF FIG. 4 AND PEAK AREAS OF THE SULPHONAMIDES

| Design point No. | Fraction | | | Sulphonamide | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Water | MeOH | MeCN | A | F | C | E | H | J | I | G |
| 1 | 0.70 | 0.30 | 0.00 | 2.8 | 7.2 | 4.2 | 7.3 | 2.7 | 1.6 | 0.8 | 5.8 |
| 2 | 0.60 | 0.40 | 0.00 | 3.4 | 6.2[a] | 6.2[a] | 7.5 | 2.8 | 1.8 | 0.9 | 6.0 |
| 3 | 0.50 | 0.50 | 0.00 | 3.0 | 6.4 | 5.0 | 7.5 | 3.0[a] | 1.8[a] | 0.9 | 5.7 |
| 4 | 0.80 | 0.00 | 0.20 | 3.2[a] | 8.2 | 5.5 | 8.7 | 4.0 | 2.0 | 0.9 | 6.5 |
| 5 | 0.70 | 0.00 | 0.30 | 2.4 | 7.4 | 5.5 | 7.8[a] | 3.7[a] | 2.1 | 1.0 | 6.4 |
| 6 | 0.60 | 0.00 | 0.40 | 3.3[a] | 6.8 | 5.5[a] | 7.1 | 3.6[a] | 2.0 | 1.0 | 6.0 |
| 7[b] | 0.76[b] | 0.12[b] | 0.12[b] | 2.1[b] | 6.7[b] | 4.5[b] | 7.3[b] | 2.8[b] | 1.9[b] | 0.9[b] | 5.8[b] |
| 8 | 0.66 | 0.17 | 0.17 | 3.2 | 6.6 | 4.6 | 7.4 | 2.6 | 1.5 | 1.0 | 5.6 |
| 9 | 0.56 | 0.22 | 0.22 | 5.3 | 6.2 | 4.8 | 8.5 | 3.0 | 1.8 | 1.3 | 5.6 |

[a] Estimated values.
[b] Values of reference chromatogram.

merged second and third peaks contain six solutes and peak recognition fails. Normally we would have reduced the design space by moving the lower boundary in the direction of the apex of the triangle after the registration of the chromatogram in point 3 (Fig. 5C), because the resolution is too low. It was

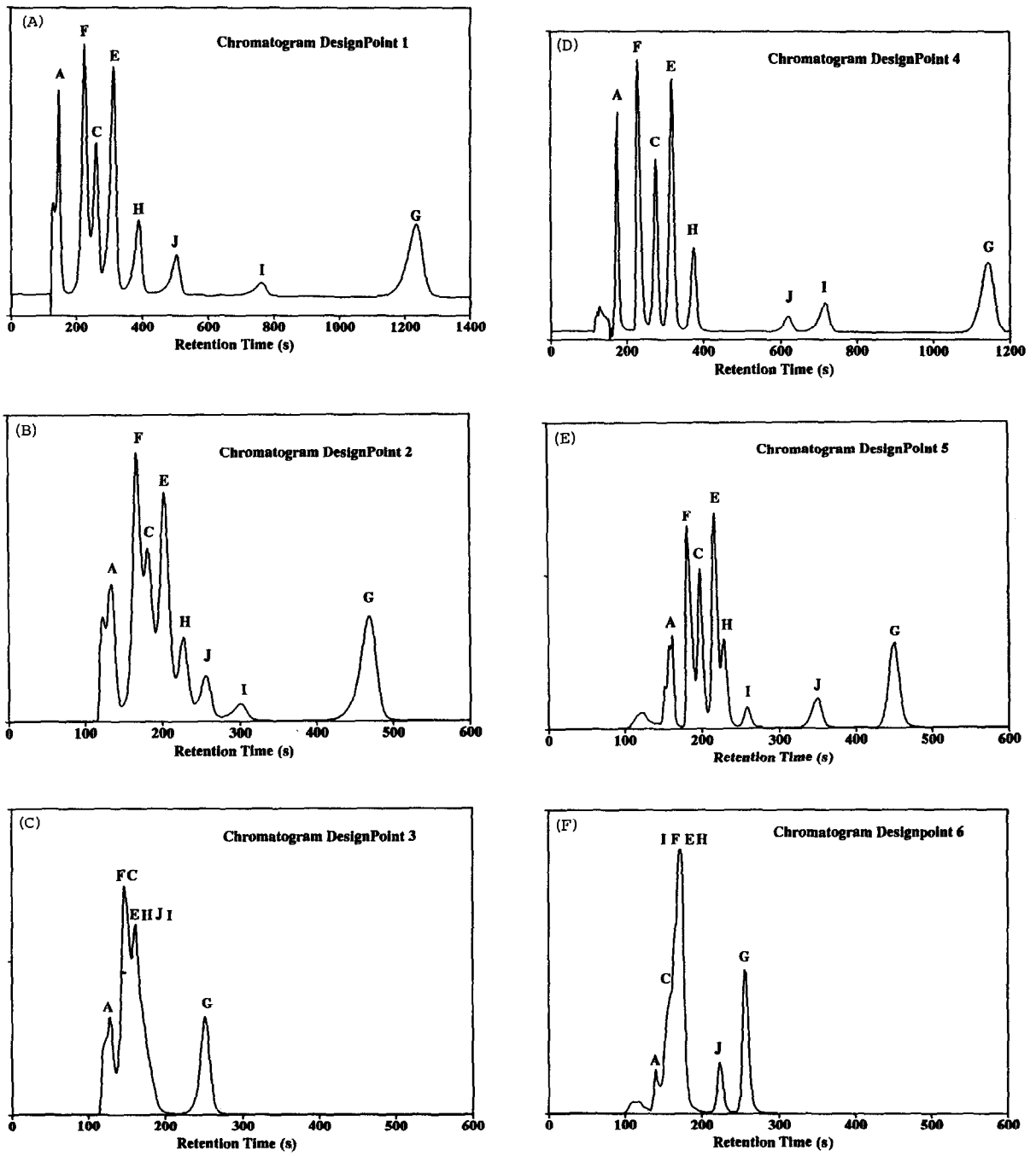



Fig. 5.

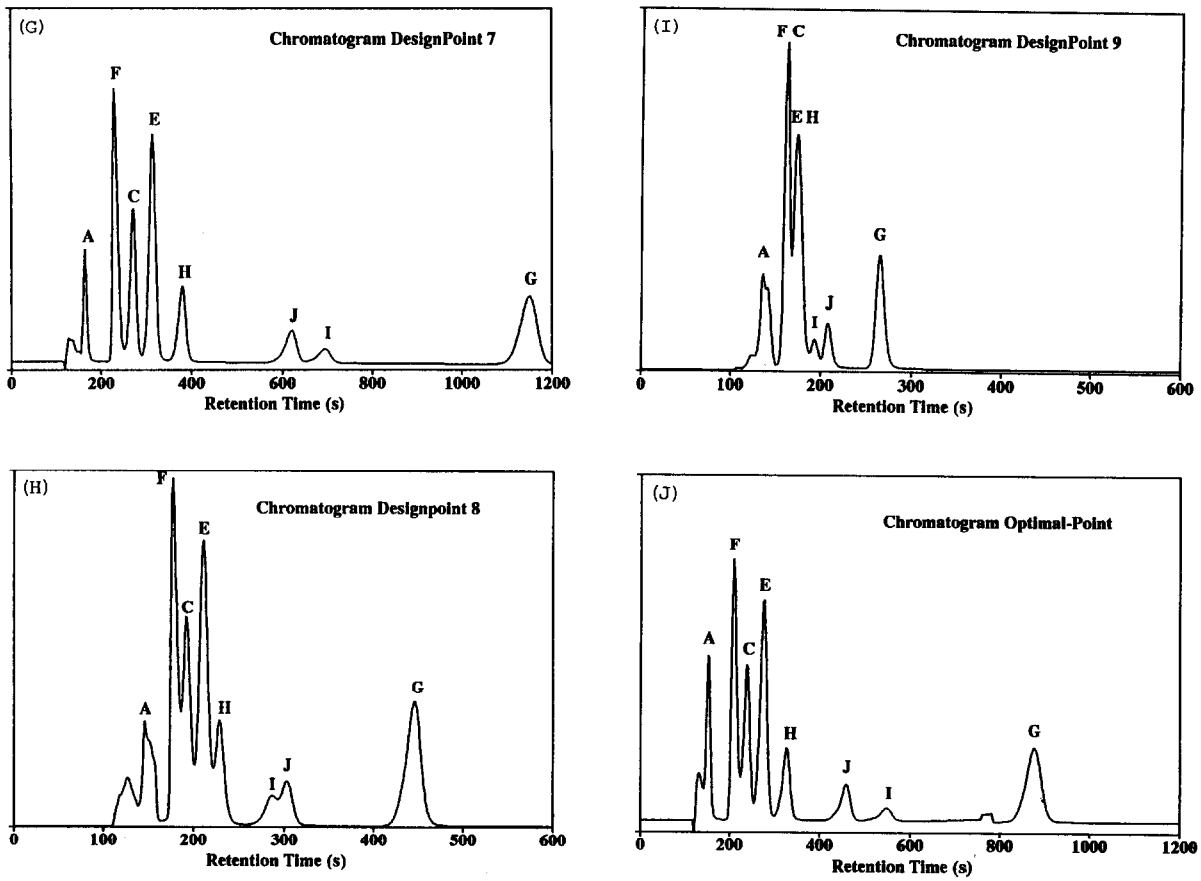



Fig. 5. Chromatograms recorded at the nine design points in Fig. 4 plus one PO chromatogram which offers a good compromise between resolution and analysis time. The first peak in every chromatogram is uracil, which was used to determine the dead time.

decided, however, to continue and divide the original sample of eight solutes into two subsamples. The two subsamples were injected separately. One subsample contained the solutes E, F and J and the other the remaining solutes. Peak recognition was possible in the subsample chromatograms, except in design point 6, where the first subsample still showed strong overlap. Additional single injections of solutes E and F were necessary for the determination of the capacity factor and a total of fourteen injections was needed to complete the determination of the capacity factors of eight solutes in nine mobile phase compositions.

The capacity factors of the eight solutes were used to model the logarithm of the capacity factor of every solute as a function of $x_1$, $x_2$ and $x_3$, which are the fractions of the components of the water–MeOH–MeCN mobile phase. Polynomial models were estimated by multiple linear regression. A grid search of the response surfaces predicted the capacity factor of every solute at all eluent compositions of the design space necessary to construct a grid with a 1% interval in the eluent composition. In every grid point two optimization criteria were calculated: the resolution of the worst separated pair of adjacent peaks, *i.e.*, the minimum resolution or $R_{s\,\mathrm{min}}$, and the capacity factor of the last-eluted peak, $k_{\mathrm{max}}$, which is a measure of the analysis time. A contour map of the $R_{s\,\mathrm{min}}$ (Fig. 6) shows two local maxima between which one may choose. The choice is guided by the second criterion, $k_{\mathrm{max}}$, because it advantageous to obtain a sufficient value of $R_{s\,\mathrm{min}}$ in the shortest

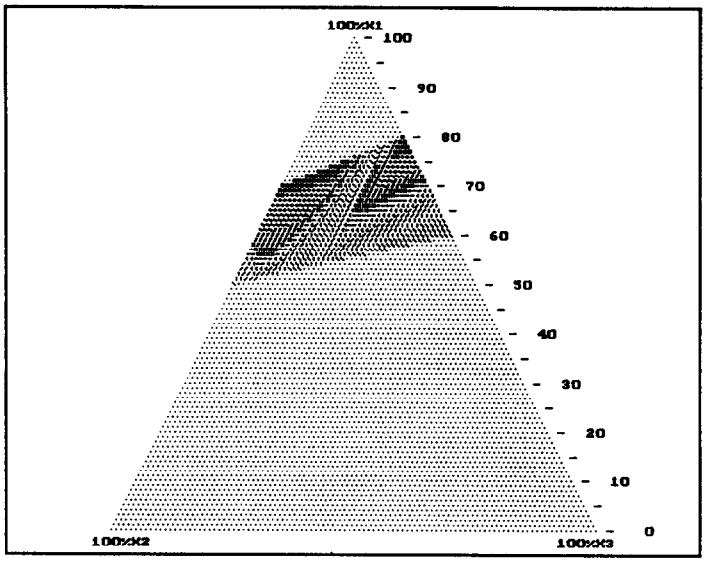



Fig. 6. Response surface of the minimum resolution, $R_{s\,min}$, of the non-isoeluotropic ternary system water ($x_1$)–MeOH ($x_2$)–MeCN ($x_3$). Different symbols correspond to ten different ranges of values of $R_{s\,min}$. The lowest range is indicated by backslashes (0.00–0.16) in the middle and lower right part of the design space and the highest range by black squares (1.44–1.60) at the left and extreme right of the upper boundary of the design space.

analysis time possible, *i.e.*, at the lowest value of $k_{max}$.

The multi-criteria decision making (MCDM) procedure [19,20] is well suited to achieve a quantitative weighting of both criteria against each other. Therefore, an MCDM diagram is constructed, consisting of two perpendicular coordinate axes. On the horizontal or time axis a scale for $k_{max}$ is indicated and on the vertical axis a scale for $R_{s\,min}$ is shown. Every eluent composition of the design space generates one value of $R_{s\,min}$ and one value for $k_{max}$. Hence every eluent composition can be represented by one point in the MCDM diagram. From the resulting cloud of points only the "pareto optimum" (PO) points are

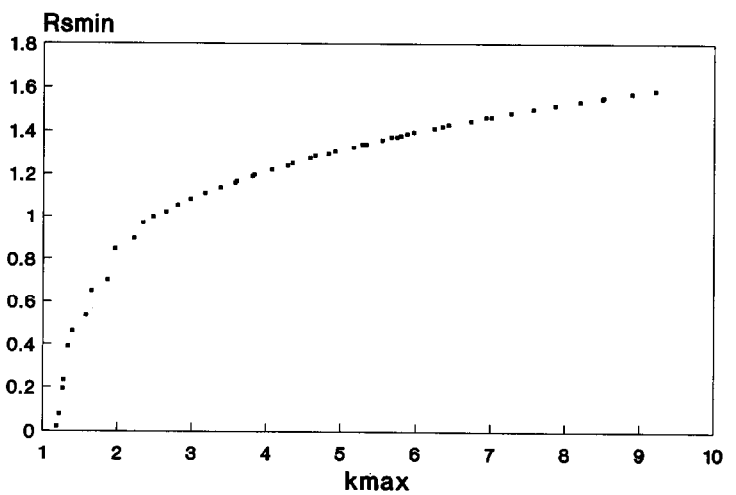



Fig. 7. MCDM diagram. Points represent PO eluent compositions which give the best combinations of $R_{s\,min}$ and $k_{max}$.

TABLE IX

MEASURED PEAK AREAS AND PREDICTED AND MEASURED RETENTION TIMES FOR CHROMATOGRAM IN FIG. 5J

| Peak | Area | Retention time (s) | |
|---|---|---|---|
| | | Measured | Predicted |
| A | 2.9 | 148 | 140 |
| F | 6.8 | 202 | 198 |
| C | 4.7 | 232 | 234 |
| E | 7.4 | 268 | 268 |
| H | 2.9 | 316 | 317 |
| J | 1.9 | 448 | 448 |
| I | 0.8 | 544 | 534 |
| G | 5.7 | 868 | 875 |

shown in the diagram (Fig. 7). A point is called a PO point if no other point in the diagram exists that yields an improvement in one criterion without causing a deterioration of the other criterion. The PO points give the eluent compositions that provide the best possible combinations of the two criteria and show the pay-off between the two criteria. If a value of 1.4 for the $R_{s\,min}$ is acceptable, then the value of $k_{max}$ is 6.2 for an eluent consisting of water–MeOH–MeCN (70:22:8). The corresponding, measured chromatogram is shown in Fig. 5J. Owing to peak asymmetry, the measured $R_{s\,min}$ is lower than the predicted value. The predicted retention times, however, correspond well with the measured retention times (see Table IX, which also shows the measured peak areas). The mean, relative difference between the measured and predicted values is 1.4%. All peaks were correctly recognized by the trained neural network.

## CONCLUSIONS

It has been shown that a simulated feed-forward neural network with back-propagation consisting of an input, hidden and output layer can be used for peak recognition in mixture design-based optimization procedures.

For peak recognition the normalized UV spectrum of the components and the peak areas were used to train the network. Both were required to be able to train the network with only one set of training data, *i.e.*, the spectra and areas measured in only one mobile phase composition. This means that the data of only one reference chromatogram were needed for tracking the peaks in other chromatograms, which were obtained with different mobile phase compositions. If only spectra were used to train the network then the spectra recorded in at least two different mobile phases are necessary to train the network. In that event one reference chromatogram is not sufficient and at least two are needed, which makes the procedure much less attractive.

The spectra and peak areas of eight sulphonamides were used in an illustrative example. Peak recognition of single peaks and of peaks in fused clusters of 2–6 components was possible, although the spectra of some components were very similar and the mean relative standard deviation in the measured peak areas was 13%.

Spectra were recorded at the peak maxima and the peaks in the reference chromatogram were almost completely separated. The results of this investigation, however, suggest that complete separation of the peaks in one reference chromatogram is not a necessary condition for the success of the procedure. It suffices that one can measure for each component one spectrum and one peak area. This may be done in one or more chromatograms, and single but also overlapping peaks can be used, if the spectra can be measured at the peak flanks and the peak areas can be measured with moderate precision. Almost the whole peak tracking procedure was performed with commercially available software, *i.e.*, the measurement of spectra and peak areas and the configuration and running of the network. Only for the spectrum normalization and the multiplication by the peak area did software have to be developed. In our opinion, expert knowledge is not required for the application of this technique. Once the necessary software is available the technique is straightforward and fast: network configuration and training take a few minutes and peak recognition is instantaneous after presentation of the test data to the network. The technique can speed up computer-assisted method development and contribute to the development of automated optimization procedures based on simplex lattice mixture designs.

REFERENCES


1 H. J. G. Debets, *J. Liq. Chromatogr.*, 8 (1985) 2725.
2 J. L. Glajch, J. J. Kirkland and J. M. Minor, *J. Liq. Chromatogr.*, 10 (1987) 1727.
3 P. M. J. Coenegracht, A. K. Smilde, H. J. Metting and D. A. Doornbos, *J. Chromatogr.*, 485 (1989) 195.
4 G. Szepesi, *HPLC in Pharmaceutical Analysis*, Vol. I, CRC Press, Boca Raton, FL, 1990, pp. 151–179.
5 J. K. Strasters, *Thesis*, Technical University of Delft, Delft, 1989.
6 J. L. Glajch, M. A. Quarry, J. F. Vasta and L. R. Snyder, *Anal. Chem.*, 58 (1986) 280.
7 P. J. Schoenmakers, *Optimization of Chromatographic Selectivity*, Elsevier, Amsterdam, 1986.
8 J. W. Dolan, *LC · GC Int.*, 3 (1990) 17.
9 A. F. Fell and B. J. Clark, in R. W. Frei and K. Zech (Editors), *Selective Sample Handling and Detection in HPLC (Journal of Chromatography Library*, Vol. 39A), Elsevier, Amsterdam, 1988, pp. 289–309.
10 A. G. Wright, A. F. Fell and J. C. Berridge, *J. Chromatogr.*, 458 (1988) 335.
11 M. Caudill and C. Butler, *Naturally Intelligent Systems*, MIT Press, London, 1990.
12 J. Stanley and S. Luedeking, *Introduction to Neural Networks*, California Scientific Software, Sierra Madre, CA, 2nd ed., 1989.
13 J. Zupan and J. Gasteiger, *Anal. Chim. Acta*, 248 (1991) 1.
14 T. Wolf, *MC, Magazin für Computerpraxis*, (1990) 92.
15 J. R. Long, V. G. Gregoriou and P. J. Gemperline, *Anal. Chem.*, 62 (1990) 1791.
16 M. Bos, A. Bos and W. E. van der Linden, *Anal. Chim. Acta*, 233 (1990) 31.
17 A. Bos, M. Bos and W. E. van der Linden, *Anal. Chim. Acta*, 256 (1992) 133.
18 J W. Weyland, C. H. P. Bruins and D. A. Doornbos, *J. Chromatogr. Sci.*, 22 (1984) 31.
19 A. K. Smilde, A. Knevelman and P. M. J. Coenegracht, *J. Chromatogr.*, 369 (1986) 1.
20 P. M. J. Coenegracht, A. Knevelman and A. K. Smilde, *J. Liq. Chromatogr.*, 12 (1989) 77.