Pharmaceutical Fingerprinting in Phase Space. 2. Pattern Recognition

ARTICLE *in* ANALYTICAL CHEMISTRY · AUGUST 1999

Impact Factor: 5.64 · DOI: 10.1021/ac981346j · Source: PubMed

Anal. Chem. **1999,** *71,* 2431−2439

Pharmaceutical Fingerprinting in Phase Space. 2. Pattern Recognition

Igor V. Tetko*

Department of Biomedical Applications, Institute of Bioorganic and Petroleum Chemistry, Murmanskaya 1, Kyiv-660 253660, Ukraine

Tatjana I. Aksenova and Alla A. Patiokha

Institute of Applied System Analysis, Prospekt Peremogy 37, 252056 Kyiv, Ukraine

Alessandro E. P. Villa

Laboratoire de Neuro-heuristique, Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, Lausanne CH-1005, Switzerland

William J. Welsh

Department of Chemistry and Center for Molecular Electronics, University of Missouri−St. Louis, St. Louis, Missouri 63121

Walter L. Zielinski

Division of Drug Analysis, U.S. Food and Drug Administration, St. Louis, Missouri 63101

David J. Livingstone

ChemQuest, Cheyney House, 19-21 Cheyney Street, Steeple Morden, Herts SG8 0LP, U.K., and Centre for Molecular Design, University of Portsmouth, Portsmouth, Hants PO1 2EG, U.K.

The current study introduces an approach for pattern recognition of drug manufacturers according to their HPLC trace impurity data. This method considers signals in phase space and accounts for two different types of noise: additive and perturbative. The pharmaceutical fingerprints are estimated as mean trajectories of HPLC trace impurity data and are used as reference models for recognition of new data by the minimal length classifier. The chromatographic trace organic impurity patterns collected from six different manufacturers of L-tryptophan are analyzed as an example. The prediction ability of the new method tested using three different cross-validation procedures remains about 95% even if the number of available data in the training sets decreases by 5 times. The accuracy of prediction in phase space is superior compared to results calculated using a Window Preprocessing method and artificial neural networks. The difference in performance between new and previous methods becomes more significant under particular conditions that are more adequate for practical application of the method. In addition, the current approach enables simple and comprehensive interpretation of the calculated results.

Selection of the feature space and its metric are crucial problems for pattern recognition. For example, if signals $x^j(t)$, $t \in [0, T]$ are analyzed, the pattern recognition of different classes of signals, $p = 1, ..., m$, can be done in the k-dimensional feature space $\mathbf{R}^k$ of $x(t_i)$, $i = 1, ..., k$, using the ordinary Euclidean norm. The rationale for this approach assumes that signals in each class j are described as the sum of the mean signal and an additive noise. Simple methods of pattern recognition such as a minimal length classifier (MLC), etc., can be successfully applied to classify data of such kind.

We have shown in the accompanying article[1] that real signals, i.e., HPLC chromatograms, are also characterized by the presence of perturbative noise. This noise contributes to variations in peak retention times (i.e., to nonstationarity of the signals along the time axis) and distances between chromatograms are large in the ordinary Euclidean space. This fact significantly complicates the structure of the analyzed classes, making their differentiation difficult.

A possible solution to this problem consists of data preprocessing and/or application of more complex pattern recognition methods, for example, artificial neural networks. In our previous studies,[2−4] we have developed an integrated approach for clas-

* Present address: Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, CH-1005 Lausanne, Switzerland. Fax: ++41-21-692-5505. E-mail (Switzerland): itetko@eliot.unil.ch. E-mail (Ukraine): tetko@bioorganic.kiev.ua.

(1) Aksenova, T. I.; Tetko, I. V.; Ivakhnenko, A. G.; Villa, A. E. P.; Welsh, W. J.; Zielinski, W. L. *Anal. Chem.* **1999**, *71*, 2423−2430.

(2) Welsh, W. J.; Lin, W.; Tersigni, S. H.; Collantes, E.; Duta, R.; Carey, M.; Zielinski, W. L.; Brower, J.; Spencer, J. A.; Layloff, T. P. *Anal. Chem.* **1996**, *68*, 3473−3482.

10.1021/ac981346j CCC: $18.00
Published on Web 05/29/1999

sification of chromatograms based on window preprocessing (WP)[2,4] and the use of wavelet packets.[3] These approaches were used successfully to analyze samples of L-tryptophan (LT) drug substances from commercial production lots of different manufacturers.

There are, however, potential drawbacks associated with WP. This procedure converts the original HPLC data (i.e., peak height vs time) within the trace impurity pattern region into compact sets of integers suitable for ANN analysis. By doing so, it effectively reduces the dimension of the problem by a factor of 10−100. However, it was shown that the predictive ability of the WP method can be seriously impaired if the peak manifold for the active ingredient in the formulation is located within the window boundaries.[2] This problem is mainly due to the presence of perturbative noise along the time axis, and this type of noise is not removed by WP. Another problem is the arbitrary nature by which the number of windows and boundaries in WP are chosen. Inasmuch as the validity and reliability of results from this approach will depend critically on the level of expertise of the user, this method may not be applicable to real-world problems.

The accompanying article[1] considered a more sophisticated model of time series signals by its analysis in phase space. It was shown that this model accounts for two components of noise, i.e., additive noise and perturbative noise. The suitability of this model for description of HPLC signals recorded using the same column was confirmed by the Kolmogorov−Smirnov (K−S) test of normality. The current study extends the application of the proposed model for pattern recognition of LT manufacturers. We show that the proposed approach reduces the problem to recognition of a mixture of normal distributions. This provides reliable identification of HPLC manufacturers, especially for classification across different columns.

INPUT DATA

HPLC Data. The present study was conducted on the same HPLC data as previously investigated, i.e., chromatographic profiles obtained on L-tryptophan (LT) drug substance from two production lots of six different commercial LT manufacturers. Two markers, M1 and M2, were added to each sample to bracket the retention times of the peaks associated with the LT samples and to normalize data as indicated elsewhere. Each chromatogram was represented by 899 points located between the LT peak manifold and the M2 peak marker (see Figure 1 in the accompanying article[1]). These data were used as the initial source for analysis in phase space as well as for the preprocessing scheme briefly described below.

Windows Preprocessing Scheme. In this scheme the fingerprinting region was divided into 22 time windows of equal length. Each window was analyzed to locate the highest peak h_{max} within it. The resulting series of 22 h_{max} values was then converted to a corresponding series of integer values (designated *H1*−*H22*) according to a procedure described elsewhere.[2] An additional series of 22 input entries (*N1*−*N22*) was obtained, representing the number of nonnoise peaks in each of the 22 time windows taken in sequential order. Finally, two parameters were included to provide a cumulative statistic for the entire fingerprint region (i.e., the 22 time windows), viz., *AllPeaks* (corresponding to the total number of nonnoisy peaks) and *HPeaks* (corresponding to the number of peaks having an h_{max} value greater than the value of the M2 marker). The complete set of these 46 parameters for each chromatogram served as the initial input for an artificial neural network. This method gave the best prediction ability for WP.[2−4]

Optimization of Parameters for WP. Our previous study with an application of ANN pruning methods found that the pattern recognition of these data could be mainly restricted to windows 10−16. This region of data, designated region R2, corresponded to absolute time t = 400−650 ms of chromatograms. The parameters optimized by ANN were *H12*, *H14*−*H16*, *N10*, *N11*, *N15*, *Hpeaks* and *HR2Peaks*. The last parameter counted the total number of very high peaks, i.e., peaks with height greater than that of marker M2, in the region R2. The use of the optimized set of parameters increased the prediction ability of ANNs. More details on the data handling and processing of the chromatograms are given elsewhere.[2−4]

The baseline drift of chromatograms was corrected by extraction of a minimum value of HPLC signal detected in a window of ±25 ms around the analyzed point as described elsewhere.[1]

Training and Test Sets. The 253 chromatograms in this classification study included 3−5 replicates for every combination of 6 LT manufacturer, 2 lots, and 3 HPLC columns. Three different procedures to test performance of the classification methods were used.

The 6-fold cross-validation procedure was employed as previously indicated.[2] The chromatograms were partitioned into six separate combinations of training and test sets (i.e., runs 1−6) in such a way that (1) no chromatogram in the test set would encounter any of its replicates in the training set and (2) each unique combination of LT manufacturer, lot, and HPLC column was included in a test set just once (Table 1). The sample size of the resulting data sets was 209−215 chromatograms for training and 38−44 chromatograms for testing. This cross-validation procedure used approximately five-sixths of the data for training of classifiers, while only one-sixth of data was used for testing of the prediction ability of the methods. Thus, the training data set contained comprehensive information about all parameters of data, including lot-to-lot and column-to-column variation. However, it is possible that in real world applications such information is only partially available to the investigator. Thus, two additional cross-validation methods were implemented to better test the relative performance of the analyzed pattern recognition methods in conditions that are more similar to practical application of the pharmaceutical fingerprinting and to better evaluate the sensitivity of the developed methods lot-to-lot and column-to-column variations.

The 3-fold cross-validation procedure was developed to test performance of different pattern recognition methods for column-to-column variations. In this procedure the training set was formed by data from both commercial lots of all manufacturers recorded with the same column. The remaining chromatograms formed the test sets (Table 1). This procedure used only one-third of all

(3) Collantes, E. R.; Duta, R.; Welsh, W. J.; Zielinski, W. L.; Brower, J. *Anal. Chem.* **1997**, *69*, 1392−1397.
(4) Tetko, I. V.; Villa, A. E. P.; Aksenova, T. I.; Zielinski, W. L.; Brower, J.; Collantes, E. R.; Welsh, W. J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 660−668.

Table 1. Number of Chromatograms in the Test Sets

run	LT manufacturer A	B	C	D	E	F	total
			Sixfold Cross-Validation				
1	10 (1-X)[a]	9 (2-X)	6 (1-Y)	6 (2-Y)	6 (1-Z)	6 (2-Z)	43
2	6 (2-Z)	6 (1-X)	8 (2-X)	6 (1-Y)	6 (2-Y)	6 (1-Z)	38
3	6 (1-Z)	6 (2-Z)	10 (1-X)	9 (2-X)	6 (1-Y)	6 (2-Y)	43
4	6 (2-Y)	5 (1-Z)	6 (2-Z)	9 (1-X)	10 (2-X)	6 (1-Y)	42
5	6 (1-Y)	6 (2-Y)	6 (1-Z)	6 (2-Z)	10 (1-X)	10 (2-X)	44
6	10 (2-X)	6 (1-Y)	6 (2-Y)	5 (1-Z)	6 (2-Z)	10 (1-X)	43
total	44	38	42	41	44	44	253
			Threefold Cross-Validation				
			{1,2}-{Y,Z}[b]				
1	24[d]	23	24	23	24	24	142
			{1,2}-{X,Z}				
2	32	26	30	29	32	32	181
			{1,2}-{X,Y}				
3	32	27	30	30	32	32	183
total	88	76	84	82	88	88	506
			Sixfold-Bis Cross-Validation				
			{1,2}-{Y,Z}, 2-X				
1	34	32	32	32	34	34	198
			{1,2}-{Y,Z}, 1-X				
2	34	29	34	32	34	34	197
			{1,2}-{X,Z}, 2-Y				
3	38	32	36	35	38	38	217
			{1,2}-{X,Z}, 1-Y				
4	38	32	36	35	38	38	217
			{1,2}-{Y,X}, 2-Z				
5	38	33	36	36	38	38	219
			{1,2}-{Y,X}, 1-Z				
6	38	32	36	35	38	38	217
total	220	190	210	205	220	220	1265

[a] The composition of each test set chromatogram by lot (1 or 2) and HPLC column (X = Vydac 1, Y = Vydac 2, Z = Waters) is indicated in parentheses. [b] The test sets for all manufacturers are composed of the same combinations of lots and HPLC columns indicated in parentheses {}.

available data in the training set, while two-thirds of the data were used in the test set.

The 6-fold-bis cross-validation procedure used the test and training sets that were, to some extent, reversed in comparison to the 6-fold cross-validation mentioned above (Table 1). In this procedure only data recorded using the same column for one lot of all manufacturers (one-sixth of all available data) constituted the training set, while the rest of data (five-sixths of all cases) were used in the test set. This cross-validation procedure perfectly fit the demands for practical monitoring of the commercial firms, when HPLC profiles of manufacturers are collected in a database using a reference column and single commercial production lot of manufacturers. These data, afterward, are retried to monitor the production process and to analyze HPLC profiles recorded with similar, but not the same column.

METHODS

Neural networks employed in this study were fully connected feed-forward back-propagation networks with one hidden layer and bias neurons.[5] ANN training was accomplished using the SuperSAB algorithm.[6] The logistic $f(x) = 1/(1+e^{-x})$ activation function was used for both hidden and output nodes. The number of input nodes was conditioned by the number of included parameters while the number of neurons in the hidden layer was selected to be 5, as proposed in ref 4. Six output nodes (one for each LT manufacturer) were used for coding and prediction of the analyzed manufacturers. The output node with the highest numerical value was taken as the predicted LT manufacturer for a single network. The early stopping over ensemble technique was used to avoid overfitting/overtraining of neural networks and to improve their ability to generalize.[7,8] Each analyzed artificial neural network ensemble was composed of $M = 200$ networks. The LT manufacturer associated with a given chromatogram corresponded to that manufacturer predicted by the majority of the ANNs forming the ensemble. A prediction was considered "incorrect" if it was impossible to classify correctly the chromatogram according to LT manufacturer at the 95% level of confidence. More details on the ANN method can be found elsewhere.[4] In summary, we selected the parameters of the ANN method that calculated the best prediction ability[2,4] in order to provide a fair comparison of the ANNs performance with that of the new approach.

The computer codes for the ANN and pruning algorithms were programmed in ANSI C++. The calculations were performed on the HP Workstation Cluster at the Swiss Center for Scientific Computing (CSCS).

Phase Fingerprints. It was shown in the accompanying article[1] that chromatograms expressed by $x^i(t)$, $t = 0, 1, ..., T$ fit to the model $x^i(t) = x(t) + \xi(t)$, where $x(t)$ is a solution of an ordinary differential equation

$$\frac{\mathrm{d}^q x}{\mathrm{d}t^q} = f\left(x,...,\frac{\mathrm{d}^{q-1}x}{\mathrm{d}t^{q-1}}\right) + F(x,...,t) \qquad (1)$$

where $\xi(t)$ is the additive noise, i.e., the sequence of independent identically distributed random variables with zero mean and limited variance, q is order of the equation, and $F(\ )$ is a perturbative function of a random process with zero mean. The equation

$$\frac{\mathrm{d}^q x}{\mathrm{d}t^q} = f\left(x,...,\frac{\mathrm{d}^{q-1}x}{\mathrm{d}t^{q-1}}\right) \qquad (2)$$

describes a self-oscillating system with a stable limit trajectory given by

$$\mathbf{x}^0(t) = (x_1^0(t), ..., x_q^0(t))^{\mathrm{T}} \qquad (3)$$

in phase space with coordinates $x_1 = x$, $x_2 = \mathrm{d}x/\mathrm{d}t$, ..., $x_n = \mathrm{d}^{q-1}x/\mathrm{d}t^{q-1}$.

Each chromatogram i was considered in phase space $\mathbf{x}^i(t) = (x_1^i(t), ..., x_q^i(t))^{\mathrm{T}}$ and was interpreted as one trajectory or one

(5) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Publishers: New York, 1993.
(6) Tollenaere, T. *Neural Networks* **1990**, *3*, 561−573.
(7) Tetko, I. V.; Villa, A. E. P. *Neural Networks* **1997**, *10*, 1361−1374.
(8) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.

solution of eq 1. These trajectories lay in the neighborhood of the stable limit trajectory $\mathbf{x}^0(t)$. The limit trajectory describes a chromatogram recorded without noise. A useful property of the limit trajectory is that it corresponds to an asymptote of the mean trajectory of analyzed signals if the number of averaged trajectories in phase space increases infinitely. Thus, the mean trajectory of signals can be used to characterize HPLC manufacturers. We refer to the mean trajectory in phase space as the *phase fingerprint* of a manufacturer. Because of differences in physical properties of HPLC columns, each manufacturer should be characterized not by one, but by several phase fingerprints separately calculated for each column.[1] Use of phase fingerprints instead of trajectories of the analyzed signals in the time domain extends the feature space, since it takes into consideration not only the signal itself but also its derivatives. Because of the arbitrary character of the time scale, the signal x and its derivatives $\mathrm{d}^j x/\mathrm{d}t^j$ are essentially on different scales. To solve the problem of relative scaling of components of signal in phase space, the maximum absolute value of signal and derivatives of all chromatograms recorded from the same manufacturer were normalized to intervals of unit length as described in the accompanying article.[1] The phase fingerprints calculated in such a way were used for pattern recognition and classification of new chromatograms, as described in the next sections.

Estimation of signal derivatives was done using integral operators.[9] The kernel function and the regularization parameter were selected as described in the accompanying article.[1]

Measure of Distance in Phase Space. Although the theory of the model for the description of signals in phase space according to eqs 1 and 2 was elaborated more than 30 years ago,[10,11] no applications of this approach for pattern recognition have been reported so far. We describe how simple considerations can be used to derive reliable pattern recognition methods in this space.

Pattern recognition in phase space requires the introduction of some function that will be used as a distance in this space. The Euclidean norm in the time domain is not suitable because, due to perturbative noise, chromatograms produced by the same manufacturer and measured with the same column can be far from one another in the time domain. That is why another function should be used. New variables designated θ and $\mathbf{n}(\theta)$ were previously introduced (see Figure 3 at ref 1) to describe the trajectories of analyzed signals. The first variable, the phase θ, is a movement time along the limit trajectory from a starting point P_0 to the orthogonal projection N of the analyzed point M at the limit trajectory. The values of the phase are in the range $0 \le \theta < T$. The vector $\mathbf{n}(\theta)$ connects analyzed point M on the signal trajectory and its orthogonal projection N on the limit trajectory (the distance MN, equal to $|\mathbf{n}(\theta)|$, is the minimal distance between M and the limit trajectory). We propose to estimate the distance between a chromatogram $\mathbf{x}^i$ and a phase fingerprint $\mathbf{x}_p^0$, $p = 1, ..., m$ as

$$\begin{cases} \rho(\mathbf{x}_p^0, \mathbf{x}^i) = \left(\dfrac{\mathrm{d}(\mathbf{x}_p^0, \mathbf{x}^i)}{\mathrm{d}(\mathbf{x}_p^0, \mathbf{0})}\right)^{1/2}, \quad \text{where} \\ d(\mathbf{x}_p^0,\mathbf{x}^i) = \int\limits_{[0,T]} ||\mathbf{n}(\theta)||\, \mathrm{d}\theta = \int\limits_{[0,T]} ||\mathbf{x}_p^0(\theta) - \mathbf{x}^i(t(\theta))||\, \mathrm{d}\theta, \text{ and} \\ \mathrm{d}(\mathbf{x}_p^0, \mathbf{0}) = \int_{[0,T]} ||\mathbf{x}_p^0(\theta)||\, \mathrm{d}\theta \end{cases} \tag{4}$$

Here, $||\cdot||$ is an ordinary Euclidean norm calculated in phase space and vector $\mathbf{0}$ corresponds to the zero chromatogram $x(\theta) \equiv 0$, $\theta = 1, ..., t$, i.e., the chromatogram without any peaks. The term $\mathrm{d}(\mathbf{x}^0,\mathbf{0})$ normalizes (to the same scale) distances that are calculated between analyzed signal $\mathbf{x}^i$ and phase fingerprints $\mathbf{x}_p^0$. The zero chromatogram $\mathbf{x}^i = \mathbf{0}$ is used as the reference point. The distance to this chromatogram is equal to $\rho(\mathbf{x}_p^0,\mathbf{0}) = 1$ for any analyzed phase fingerprint $\mathbf{x}_p^0$, $p = 1, ..., m$.

The distances $\rho(\mathbf{x}^0,\mathbf{x}_p^i)$ calculated by eq 4 make it possible to consider the analyzed problem as a pattern recognition of normal distributions with the ordinary Euclidean norm in a transformed space of features. Vector $\mathbf{n}(\theta)$ and components of this vector $\mathrm{d}^j(x^i)(t(\theta))/\mathrm{d}t^j - \mathrm{d}^j(x^0)(t(\theta))/\mathrm{d}t^j$ ($j = 0, ..., q - 1$ is order of differentiation) are given by asymptotically normal distributions with zero mean. That is why vectors $\mathbf{x}^i(t(\theta))$ are also distributed according to the normal distribution and $\mathbf{x}^0(t(\theta))$ corresponds to the asymptotically unbiased consistent estimation of the mathematical expectation of $\mathbf{x}^i(t(\theta))$ for any θ. Thus in the space of features $\mathbf{x}^i(t(\theta))$, $\theta = 0, 1, ..., T$ the analyzed problem is reduced to the recognition of a mixture of normal distributions. Then, it is possible to apply simple classification methods, such as MLC, for the pattern recognition of new data. According to this method, the minimum distance $\min_p \rho(\mathbf{x}_p^0,\mathbf{x}^i)$ calculated for analyzed phase fingerprints $\mathbf{x}_p^0$, $p = 1, ..., m$ predicts the firm manufacturer of the analyzed sample. The calculation of the distances $\rho(\mathbf{x}_p^0,\mathbf{x}^i)$ in the transformed space is reduced to

$$\begin{cases} \mathrm{d}(\mathbf{x}^0,\mathbf{x}^i) = \displaystyle\sum_{\theta\in[0,T]} \sum_{0\le j\le q-1} \left(\frac{\mathrm{d}^j(x^0)}{\mathrm{d}t^j}(t(\theta)) - \frac{\mathrm{d}^j(x^i)}{\mathrm{d}t^j}(t(\theta))\right)^2 \\ \mathrm{d}(\mathbf{x}^0,\mathbf{0}) = \displaystyle\sum_{\theta\in[0,T]} \sum_{0\le j\le q-1} \left(\frac{\mathrm{d}^j(x^0)}{\mathrm{d}t^t}(t(\theta))\right)^2 \end{cases} \tag{5}$$

This corresponds to an approximation of the integrals in eq 4 calculated for discrete phases $\theta = 0, 1, ..., T$, $\Delta\theta = 1$. The partition of intervals in the time domain is uniform for the analyzed signals, i.e., $t = 0, 1, ..., T$, $\Delta t = 1$. However, the similar partition for the phase θ of chromatograms, i.e., $\theta = 0, 1, ..., T$, $\Delta\theta = 1$, is in general irregular in time, i.e., $t(\theta_{i+1}) - t(\theta_i) \neq 1$.

Calculation of the Function of Time. An estimation of the distance $\rho(\mathbf{x}_p^0,\mathbf{x}^i)$ between phase fingerprint $\mathbf{x}_p^0$ and analyzed chromatogram $\mathbf{x}^i$ requires estimation of the function $t(\theta)$. To make this estimation, we raise the hypothesis that the trajectory of the analyzed signal is characterized by the limit trajectory $\mathbf{x}_p^0$. The time $t(\theta)$ is distributed according to the Gaussian distribution $N(\theta,\sigma)$. Therefore $t(\theta) \in (\theta - \delta, \theta + \delta)$, where δ depends on the variance $\sigma(t)$, i.e.

$$t(\theta_i) \in I, \qquad I = (\theta_i - \delta, \theta_i + \delta) \cap (\theta_{i-1}, T) \cap (0, T) \tag{6}$$

(9) Aksenova, T. I.; Shelekhova, V. Yu. *SAMS* **1995**, *18*, 159−163.
(10) Bogoljubov, N. N.; Mitropolsky, Y. A. *Asymptotic Methods in the Theory of Non-Linear Oscillations*, 2nd ed.; Gordon and Breach: New York, 1961.
(11) Gudzenko, L. I. *Izv. Vuzov Radiophys.* **1962**, *5*, 573−587.

Table 2. Performance of the Method in Phase Space x, x', x'' for the Sixfold Cross-Validation Scheme

data set	A	B	C	D	E	F	total errors	% correct classif
Training Sets[b]								
$\delta = 1$	0	4	2	6	17	0	29 (23)	98 (98)
$\delta = 2$	0	2	3	7	11	0	23 (16)	98 (98)
$\delta = 3$	0	2	10	11	9	0	32 (22)	97 (98)
$\delta = 4$	0	2	17	11	6	0	36 (25)	97 (98)
$\delta = 5$	0	5	27	8	0	2	42 (34)	97 (97)
$\delta = 6$	0	12	33	8	1	4	58 (50)	95 (95)
$\delta = 7$	1	16	52	3	4	4	80 (77)	94 (93)
Test Sets								
$\delta = 1$	1	1	6	41	0	0	49 (8)	81 (96)
$\delta = 2$	1	0	6	41	0	0	48 (7)	81 (97)
$\delta = 3$	1	0	10	41	0	0	52 (11)	79 (95)
$\delta = 4$	1	1	18	41	0	0	61 (20)	76 (91)
$\delta = 5$	3	1	16	37	0	0	66 (29)	74 (86)
$\delta = 6$	13	3	24	34	2	0	76 (42)	70 (80)
$\delta = 7$	15	3	24	30	3	0	81 (51)	68 (76)

[a] The results calculated without manufacturer D are indicated in parentheses. [b] For this cross-validation scheme there are 1265 chromatograms in the training and 253 chromatograms in the test sets.

We estimate the time $t(\theta)$ as

$$t(\theta) = \arg\min_{t \in I} ||\mathbf{x}^0(\theta) - \mathbf{x}^i(t)|| \quad (7)$$

This formula calculates a function that can be used as time in eq 5. Let us note that this function should be calculated using phase fingerprints with which the analyzed chromatogram is compared.

The magnitude of the variance $\sigma(t)$ is a function of time. Thus it can be different for various parts of the analyzed chromatogram. For example, the variance is equal to zero for the left most point of the chromatograms, i.e., for the position of marker M2. This is because chromatograms were normalized at the same duration (t(M2) = 899 for all chromatograms), as is indicated in the Input Data section. Other examples of values of $\sigma(t)$ calculated for the highest peak of the chromatograms are indicated in Table 2 of the accompanying article.[1] Thus, the selection of δ presents significant difficulties. We decided to use this value as a parameter of the algorithm and to select it as an integer that leads to minimal error in recognition of samples during training. For simplicity, the same value was used for all manufacturers.

Selection of the δ value by optimization of error for training sets is valid if data in the training and test sets come from the same Gaussian distributions, i.e., have the same phase fingerprints. It was shown in the accompanying article[1] that this is the case for data from both lots of manufacturers A, B, C, E, and F recorded using the same column. However, data of manufacturer D are characterized by different phase fingerprints for each lot. Thus the prediction ability of the current method in lot-to-lot prediction of this manufacturer can be low. In addition, because phase fingerprints are different for data recorded using different columns, the current method (in its present form) should not be applied for such analysis.

Use of Absolute Magnitudes of Signals. Chromatograms recorded for different manufacturers can be characterized by different orders of amplitude of the peaks. For example, an amplitude of the maximal peak for manufacturer F is about 3−7 units, while that of manufacturer E is on the order of 30−70 units. This is valuable information that can be used in pattern recognition of LT manufacturers. Unfortunately, it is lost following normalization of signal and its derivatives, as is proposed in section 2.1. To use this information for classification of manufacturers, a mean value $\overline{P_m}$ and dispersions of magnitudes of the highest peak $\sigma(\overline{P_m})$ were estimated for each manufacturer m according to the distribution of data in the training set. An amplitude of the highest peak P_i was estimated for each chromatogram to be analyzed (see Table 1 of accompanying article[1]). The analysis of the chromatogram was done if and only if $P_i \in [\overline{P_m} - 3\sigma, \overline{P_m} + 3\sigma]$. Otherwise, we considered that this chromatogram cannot be generated by the manufacturer m, i.e., assumed that the pairwise distance between this chromatogram and phase fingerprints of this manufacturer is very large. Such prescreening of possible manufacturers also decreased the time required for data analysis.

General Scheme of Pattern Recognition. The present application of pattern recognition in phase space consisted of the following steps.

(1) Analysis in the training set

(1.1) analysis of data collected for the same manufacturer p

(1.1.1) calculation of derivatives of signals;

(1.1.2) normalization of signal and derivatives to unit interval

(1.1.3) calculation of phase θ, $t_p(\theta)$, and $\mathbf{n}_p(\theta)$ values

(1.1.4) application of K−S goodness-of-fit test to verify normality of $t_p(\theta)$ and components of the vector $\mathbf{n}_p(\theta)$ values; separation of data for different distributions, if required

(1.1.5) calculation of the phase fingerprint

(1.2) estimation of a mean value $\overline{P_m}$ and dispersion of magnitudes of the highest peak $\sigma(\overline{P_m})$ for each manufacturer $m = 1, ..., 6$

(1.3) selection of parameter δ to be used in the function of time by minimization of the prediction error for the training set

(2) pattern recognition of a new chromatogram

(2.1) calculation of derivatives of the chromatogram

(2.2) normalization of its signal and derivatives to unit interval

(2.3) comparison of the amplitude of the highest peak with corresponding intervals for analyzed manufacturers

(2.4) calculation of functions of time $t(\theta)$ and distances between manufacturer and the trajectory of the signal in phase space

(2.5) The manufacturer of the chromatogram is the one that has the minimum distance between its phase fingerprint and the analyzed chromatogram (i.e., the MLC was used).

RESULTS

A previous analysis has shown that variability of components of vector $\mathbf{n}(\theta)$ satisfied the K−S test for all manufacturers except manufacturer D.[1] Data calculated for this manufacturer were described as a mixture of two normal distributions each formed by data recorded for one of two analyzed lots. A similar analysis of $t(\theta)$ indicated a mixture of three normal distributions for all manufacturers, and also for each lot of manufacturer D. Each of these distributions was formed by chromatograms recorded using the same HPLC column. Thus, three different phase fingerprints could be calculated for manufacturer A, B, C, E, and F and 6 (3 columns times 2 lots) for manufacturer D.

Sixfold Cross-Validation Scheme. In this scheme only approximately one-sixth of the data, that is one combination of lot and a column, were used as the test sets in one run. The other available data were used as the training set (Table 1).

Preliminary Analysis of Data. The model proposed in the current study has several parameters that should be fixed for an application of the algorithm, i.e., order of equation (eq 1) and value of parameter δ (eq 6). As a rule of thumb, the third-order equation (phase space x, x', and x'') was selected for preliminary analysis, as discussed in the accompanying article.[1] The selection of parameter δ will be discussed further in the text.

The K−S goodness-of-fit test was applied to analyze data in each training set. This test detected three normal distributions for manufacturers A, B, C, E, and F and five such distributions for manufacturer D. The composition of data in the distributions was the same as that calculated using previous analysis of all data (i.e., the data were separated according to three HPLC columns), and for manufacturer D they were also separated according to two commercial lots. Let us note that one of three distributions for manufacturers A, B, C, E, and F was composed of data from one lot only, because the complementary data composed the test set. The phase fingerprints were calculated for each distribution.

Several values of parameter $\delta = 1, ..., 7$, $\Delta\delta = 1$ were analyzed (Table 2). The minimum error for training set was found using $\delta = 2$ and 3. These values of parameter provided the minimum error for the test set too. The performance of the method was high for data from the training set. However, its prediction ability was low for the test set, mostly due to an extremely poor recognition of manufacturer D. For example, 100% of prediction errors were calculated for this manufacturer for parameter $\delta = 1-4$. The prediction performance of the method was very high for both the training and test sets (about 98%) if manufacturer D was excluded from the analysis.

Analysis of phase space combined with a simple decision-making algorithm provides a clear interpretation of the calculated results. We found that recognition of chromatograms of all manufacturers except D was done using the phase fingerprint calculated for the second lot of the same column that was in the test set, i.e., when both test and training data were from the same normal distribution. Thus, it was possible to significantly decrease the amount of data in the training set, i.e., to use only the data of the corresponding lot of the same manufacturer, without affecting the prediction accuracy for these manufacturers. Recognition of test sets is done according to phase fingerprints calculated for the same normal distribution, and this explains the high prediction ability of our method for these manufacturers. However, the data recorded for manufacturer D in test and training sets were always from different normal distributions. This violated the basic assumptions of the method, and recognition of manufacturer D cannot be performed correctly.

Optimization of Dimension of Phase Space. An analysis was done to verify if phase space x, x', and x'' was optimal for the current task. The parameter δ was selected to be $\delta = 3$ for all manufacturers. For this analysis we excluded manufacturer D since its presence could produce spurious results in some cases. All possible combinations of signal and its derivatives up to third order (x''') were analyzed (Table 3). The best prediction ability was calculated using phase space x, x', and x'', i.e., phase space that was used in our preliminary analysis. Thus, this phase space was selected for all other analyses reported in this study.

Table 3. Performance of the Method in Different Phase Spaces for the Sixfold Cross-Validation Scheme

	manufacturer						
	A	B	C	D	E	F	total errors
x	13	2	6	X	2	2	25
x′	9	4	18	X	0	0	31
x″	9	2	27	X	0	0	38
x‴	6	2	28	X	0	3	39
x,x′	1	1	10	X	0	0	12
x, x″	3	1	10	X	0	1	15
x, x‴	1	0	16	X	0	0	17
x′, x″	6	0	22	X	0	0	28
x′, x‴	3	0	25	X	0	0	28
x″, x‴	6	0	24	X	2	0	32
x, x′, x″	1	0	10	X	0	0	11
x, x′, x‴	1	0	19	X	0	0	20
x, x″, x‴	1	0	18	X	0	0	19
x′, x″, x‴	6	0	24	X	0	0	30
x, x′, x″, x‴	0	0	18	X	0	0	18

Compensation for a Shift of HPLC Chromatograms Recorded from Different Columns. The accompanying article[1] showed that major differences between chromatograms recorded using different columns for the same data (e.g., data collected from the same lot of the same manufacturer) consist of large shifts due to different properties of columns and were essentially nonlinear along the phase fingerprint region. Since all the data were normalized for the same duration using markers, all shifts between chromatograms were equal to zero at times corresponding to the markers. Thus, the maximal values of shifts appeared above the central area between markers. Let us note that since the markers did not coincide with the start and the end of the fingerprint region, the largest shift was observed in the left ($t = 0$) and the central part ($t = 450$) of the fingerprint region. Approximate values of these shifts were estimated in the accompanying article.[1] The shifts were on the order of 80−110 and 20−40 units for peaks recorded using Vydac 2 and Vydac 1, respectively, compared to that peaks of Waters for phases $\theta = \{0-700\}$. The shifts between columns decreased to 0 (position of marker M2, $t = \theta = 899$) for phases $\theta = \{700-899\}$ time units.[1] However, this region did not contain significant peaks and thus did not influence the pattern recognition of the manufacturers.

The information about the presence of significant nonlinear shift between columns can be used for the pattern recognition. It was shown in the accompanying article that phase fingerprints calculated for different columns of the same lot were similar in phase space despite a presence of significant shifts between them in the time domain. Thus, in principle, a correct recognition of manufacturers (and, especially, that of manufacturer D) could be done using their phase fingerprints recorded with other columns. Unfortunately, this could not be done in the previous analysis. The problem was in eqs 7 and 8 that were used for calculating the function of time. These equations restricted the maximal delay in time between the phase fingerprint and the analyzed chromatograms to the parameter δ and thus prevented recognition of chromatograms by phase fingerprints recorded with other columns. This was not a problem for analysis of manufacturers A, B, C, E, and F, because they had a small lot-to-lot variance, and

Table 4. Prediction Ability of Pattern Recognition Methods[a]

method	A	B	C	D	E	F	total errors[b]	% correct classif
Sixfold Cross-Validation								
ANN	2	1	2	18	5	0	28	89
ANN (optim.)	0	0	7	6	0	0	13	95
MLC	1	0	10	41	0	0	52	79
MLC (shift)	0	1	8	3	0	0	12	95
Threefold Cross-Validation								
ANN	67	32	59	21	42	85	306	40
ANN (optim)	50	35	51	11	25	70	242	52
ANN (shift)	73	27	25	13	25	32	195	61
ANN (optim, shift)	0	39	6	16	13	47	121	76
MLC (shift)	8	2	0	8	0	0	18	96
Sixfold-Bis Cross-Validation								
ANN	70	54	122	135	60	176	617	51
ANN (shift)	87	64	75	123	44	79	472	63
ANN (optim)	89	96	170	59	69	161	644	49
ANN (optim, shift)	37	105	93	49	46	125	455	64
MLC (shift)	9	5	44	125	0	0	183	86
MLC (shift)	9	5	44	X	0	0	58	95

[a] ANN = artificial neural network applied to parameters generated according to WP. (optim) = set of parameters in WP optimized as indicated in ref 4. (shift) = the chromatograms and windows in WP were aligned to compensate shifts due to use of different columns as indicated in the article. MLC = minimal length classifier applied to analysis of chromatograms in phase space. [b] The total number of chromatograms for different cross-validation schemes is different (see Table 1).

Table 5. Prediction Ability of Analysis in Phase Space for the Sixfold-Bis Cross-Validation Scheme

compositions of test and training subsets	A	B	C	D	E	F	total errors	% correct classif
Column-to-Column Variation								
same lot & different columns	0	0	16	10	0	0	26 (16)[a]	95 (96)
Lot-to-Lot Variation								
different columns & lots	9	3	20	77	0	0	109 (32)	78 (92)
same col & different lots	0	2	8	38	0	0	48 (10)	81 (95)
total for different lots	9	5	28	115	0	0	157 (42)	79 (93)

[a] The results calculated without manufacturer D are indicated in parentheses.

their test chromatograms could be successfully recognized by phase fingerprints calculated for a corresponding lot from the training set. However, manufacturer D could not be recognized in such a way due to significant lot-to-lot variations for this manufacturer. Thus, correct recognition of data from this manufacturer could be only done using phase fingerprints recorded with other HPLC columns.

The shifts between chromatograms were corrected by translating the chromatograms and increasing the value of parameter δ. The chromatograms recorded for the Waters column were used as reference ones while chromatograms recorded with Vydac 1 and Vydac 2 columns were shifted on the left side for 30 and 90 time units, respectively. The parameter $\delta = 15$ was selected to compensate for the largest variance of the shifts in the same peaks recorded using different chromatograms (i.e., 30 ± 15 and 90 ± 15 covered all ranges of shifts between corresponding columns). These values of parameters were fixed for all analyzed manufacturers.

A compensation of the shift significantly improved the prediction ability of the pattern recognition (Table 4). This prediction ability was superior to the best results calculated for this training/test set protocol using WP and artificial neural networks. Even manufacturer D was correctly recognized by the pattern recognition method in all except 3 cases.

Threefold Cross-Validation Scheme. An application of ANN using parameters calculated with WP provided a low prediction ability (Table 4). The prediction ability of parameters optimized with pruning methods was slightly better compared to the use of the total set of parameters. The high prediction ability of ANN (and, probably, of other pattern recognition methods applied to this task) was mainly due to the low lot-to-lot variations observed for these data. Unfortunately, the previous methods were unable to compensate for nonlinear shifts between chromatograms for different HPLC columns.

We tried to compensate for shifts in the HPLC data for WP. The size of one window in WP was $899/22 \approx 40$ units. We tried to align the windows for different HPLC columns by using data recorded with column Waters as a reference point, while chromatograms recorded with Vydac 1 and Vydac 2 columns were shifted to the left side for 40 (1 time window) and 80 time units (2 time windows), respectively. The windows that could not match each other after the shift were not considered in the analysis. An analysis of the data aligned by this method provided a significant improvement of prediction ability of the ANNs. The best results were calculated using the optimized sets of parameters.

An analysis in phase space was done using the same parameters of the correction for a shift as indicated in the previous sections. The prediction ability of MLC was significantly better in comparison to results calculated with ANNs.

Sixfold-Bis Cross-Validation Scheme. The same correction for the shifts of chromatograms as proposed in the previous section was used. We found that the pattern recognition of manufacturer D was extremely poor in this scheme. This manufacturer was incorrectly classified in 125 out of 205 cases (89% of errors). The prediction ability of the pattern for all other manufacturers except this one was about 94%.

The current cross-validation scheme enables the possibility of directly estimating the lot-to-lot and column-to-column sensitivity of the method (Table 5). Similar to the previous results, analysis of lot-to-lot variance showed a significant change in the production process of manufacturer D. The chromatograms recorded for two lots of this manufacturer were completely different and could not be correctly recognized by the classifier. The prediction ability of the method was on average higher (by about 3%) when data from different lots were analyzed using the same column compared to cases when data from different lots were predicted using different columns.

The sensitivity of column-to-column variations was already tested in the 3-fold cross-validation scheme, where about 96% of correct recognition were found. Comparable results (95% of correct prediction) were calculated in this cross-validation scheme too. For this analysis we considered only cases when data from the same lot were predicted using different columns. The prediction ability of the method for manufacturer D was very similar to that calculated for other manufacturers. Thus, it is possible to conclude

that the quality of data within one lot of this manufacturer was approximately the same and analogous to that of other manufacturers.

The prediction ability of an ANN applied for this cross-validation scheme was lower than that of MLC. Compensation for the shift between chromatograms increased the generalization ability of neural networks; however, it still remained significantly less than that of analysis in phase space.

DISCUSSION

The current study analyzed the chromatographic trace organic impurity patterns collected from six different manufacturers of L-tryptophan. This data set was selected for the current study because it was extensively analyzed by several pattern recognition methods, which permits comparison of performance of the new and previous results. We have shown that analysis of HPLC data in phase space gives a reliable identification of drug manufacturers, Furthermore, the generalization ability of this method is superior to ANNs and WP.

Explicit information about the physical characteristics of analyzed HPLC columns, i.e., existence of large shifts between HPLC profiles recorded for the same data with different columns, was used. This information can be easily evaluated during routine analysis. A possible approach consists of using some reference data for which its spectra are recorded using both the reference column and the column applied for new data analysis. An estimation of shifts of essential peaks in both columns provides the information that is necessary for correct application of the proposed method. This method is recommended if it is impossible to use the same column for measurements of original data under investigation. For example, HPLC profiles of original drugs could be recorded using the reference column, stored in some computer database and later retrieved for further analysis. However, if HPLC profiles for data in training and test sets could be recorded using the same column, no correction factors are required. The simplest version of analysis with selection of parameter δ according to the performance of MLC on the training set can be applied to analyze these data. Let us note that such analysis could provide more accurate prediction (about 98% ca. 92−96% if correction for shifts between columns is required) of data. Such analysis can be recommended to monitor product consistency. The presence of significant variance in HPLC profiles will indicate changes in production processes of pharmaceutical products.

It should be noted that inversions of peak retention can sometimes occur (although the relative amplitudes of their respective signals should remain relatively constant). Fortunately, it was observed that this was more likely to occur with very small peaks that make nonsignificant contributions to the sample composition. The problem of inversion can be the case if the impurities will interact in different ways with the used columns and the order of their retention times will be altered. Peak reversal is more common if the same sample is run on HPLC columns with different chemical compositions, e.g., different polarities or/and column packings, offering different solute−stationary phase interactions. The three HPLC columns used in the current study were chosen to have the same column dimensions and presumably "equivalent" column packings (5 μm, 300 A, C_{18} reversed phase). Thus the inversion problem made very small contributions to the overall sample composition in our sample.[2] It is also very important to note that the algorithm proposed in the current study does not use any assumption that peaks should be shifted in one direction only. Thus, the inversion problem will not affect the algorithm so long as the inversion is within parameter *d* that is used in matching of peaks. In general, however, we advise using columns containing a (reversed phase) packing from the same lot produced by the same vendor in order to minimize the problem of inversion of retention times.

ANNs applied to parameters optimized with WP in a 6-fold cross-validation scheme provided prediction ability superior to MLC applied without compensation of shift between different HPLC columns. However, after compensation for such shifts the performance of MLC was similar for the 6-fold cross-validation scheme and significantly higher for the 3-fold and 6-fold-bis cross-validation schemes in comparison to ANNs.

These results can be explained by considering the fact that in the case of a 6-fold cross-validation scheme more heterogeneous data were available for ANN training. These training data contained in an implicit way some information about the presence of significant shifts between different columns. Thus, the neural networks were able to learn these dependencies in an implicit way and account for the problem of shifts. On the contrary, the analysis in phase space was based on the simple pattern recognition method of MLC and was unable to incorporate this information in implicit form. It is remarkable that if information about shifts was included in MLC in explicit form, the prediction abilities of both approaches were on average very similar.

The 3-fold and 6-fold-bis cross-validation schemes employed to train neural networks did not incorporate any implicit information about the presence of significant shifts between columns. Thus these methods were unable to account for the shifts and their prediction ability was very poor. An alignment of windows for neural network training increased their prediction ability, but it was not optimal. Thus, the prediction ability of ANN remained significantly lower compared to analysis in phase space. Let us note that in this study a very simple method, i.e., translation along time axis, was used to compensate for shifts between chromatograms. Probably, more sophisticated schemes, like nonlinear mapping techniques[12] could be more relevant for the data aligning for the neural network and could significantly improve the prediction ability of this method. Another approach would be to incorporate in the neural network learning some prior information about the presence of significant shifts between data in training and test sets, for example to use "hints",[13] i.e., the method that is a popular technique in financial market prediction.

The generalization of the optimized set of parameters for the neural network was better than that of a full set of parameters using all training/set protocols with and without alignment of windows. These results confirm our previous finding about the importance of optimization of a set of parameters for neural network training. Only parameters calculated for windows 10−16 (i.e., 7 out of 22 windows) participated to the neural network training using the optimized set of parameters. Thus, the shifts between columns for this localized region of data were approximately constant and the alignment was more efficient for the optimized set of parameters compared to that of the total set.

(12) Livingstone, D. J.; Hesketh, G.; Clayworth, D. *J. Mol. Graphics* **1991**, *9*, 115−188.
(13) Abu-Mostafa, Y. S. *Neural Comput.* **1995**, *7*, 639−671.

The prediction ability of the proposed method for manufacturer D was low if different lots from this manufacturer were used in test and training sets. This result as well as comparison of phase fingerprints calculated for both lots of this manufacturer unambiguously indicate the presence of significant changes in the production process of this manufacturer. This suggests that such an analysis in phase space could be a suitable tool to monitor the product consistency of drug manufacturers.

Manufacturer C had the second largest number of misclassifications for two of three analyzed cross-validation schemes. We can assume that there is some instability in the production process of this manufacturer that influences its pattern recognition.

It is very important to note the insensitivity of the proposed approach to the number of data used in the training set. A 5-fold decrease of the number of chromatograms in the training set does not significantly affect the predictive performance of this method, which remains about 95% for all tested cross-validation schemes. That is why the proposed approach represents a promising tool for pattern recognition and monitoring of production processes of drug manufacturers.

ACKNOWLEDGMENT

This study was partially supported by NATO HTECH.LG 972304, INTAS-Ukraine 95-0060, INTAS-OPEN 97-168, and the Swiss National Science Foundation FNRS 2150-045689.95 grants. The overall project is supported in part by equipment grants from the Center for Molecular Electronics of the University of Missouri—St. Louis and by a contract with the FDA Division of Drug Analysis, St. Louis, administered by Thomas P. Layloff. The authors express their appreciation to Samuel W. Page of the FDA Center for Food Safety and Nutrition, Washington, DC, and Robert Hill of the Centers for Disease Control (CDC), Atlanta, GA, for providing the samples of the L-tryptophan bulk substance used in these studies. We also thank Alexey G. Ivakhnenko and Tamara N. Kasheva for their helpful suggestions.

Received for review December 4, 1998. Accepted March 22, 1999.

AC981346J