Pharmaceutical Fingerprinting in Phase Space. 1. Construction of Phase Fingerprints

6 AUTHORS, INCLUDING:

Tetiana Aksenova
Atomic Energy and Alternative Energies Com…
60 PUBLICATIONS   **195** CITATIONS

Igor Tetko
Helmholtz Zentrum München
206 PUBLICATIONS   **6,844** CITATIONS

Alessandro E. P. Villa
220 PUBLICATIONS   **3,217** CITATIONS

Pharmaceutical Fingerprinting in Phase Space. 1. Construction of Phase Fingerprints

Tatjana I. Aksenova

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

Igor V. Tetko*

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

Alexey G. Ivakhnenko

Glushkov Institute of Cybernetics, Academic Glushkov Avenue 20, Kyiv 252207 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

The present study proposes a general method for constructing pharmaceutical fingerprints in the analysis of HPLC trace organic impurity patterns. The approach considers signals in phase space and accounts for two different types of noise: additive and perturbative. The first type, additive noise, contributes to distortion of the absolute values of signal peaks. The second type, perturbative noise, contributes to variations of the retention times of signal peaks and distorts the time scale of the trace organic impurity patterns. The ability of the proposed approach to consider both types of noise significantly distinguishes it from existing methods of data analysis that are usually designed to treat only the additive noise. Analysis of the HPLC signals in phase space eliminates the problem of perturbation noise and enables detection and comparison of similar signal segments recorded at different retention times. The current study analyzes the chromatographic trace organic impurity patterns collected from six different manufacturers of L-tryptophan using three HPLC columns. For five manufacturers the variability of data recorded with the same column are in perfect agreement with the proposed model. A significant variance of parameters is detected for one manufacturer, thus indicating a possible change in its product consistency. The analysis in phase space is also used to explain the previously detected variability of HPLC signals across columns. The accompanying paper reports an application of the proposed approach for the pattern recognition of HPLC data.

The discovery of fraudulent practices at a generic pharmaceutical firm in 1989 led the U.S. Food and Drug Administration (FDA) to conduct large-scale investigations into the pharmaceutical industry. It was determined that instances of fraud might be directly detectable from analytical "fingerprints" that could demonstrate sameness or differences between substances sampled from manufacturers.[1−3] Therefore, rapid and reliable methods were sought that could be used to monitor within- and between-batch product consistency, to examine the effects of process changes in the production of pharmaceutical products, and to determine whether a product marketed today is nominally the same as that which was originally approved.

It is well established that information on the microscopic chemical composition of products provided by chromatographic trace organic impurity patterns represents an important component of the product fingerprint. A comparison of such patterns can often be used to reliably judge whether samples are nominally the same or different and to determine precursor and degradation profiles in the bulk drug.[4] In addition, comparisons of HPLC trace impurity fingerprints have been used to establish a correlation between changes in the synthesis of given lots of a bulk pharmaceutical and occurrences of human pharmacogenic disease (isoxicam with Lyell's syndrome, L-tryptophan with eosinophilia-myalgia syndrome[5,6]).

* Present address: Institut de Physiologie, Université de Lausanne, Rue du Bugnon 7, CH-1005 Lausanne, Switzerland. Fax: ++41-21-692-5505. Tel: ++44-21-692-5534. E-mail (Switzerland): itetko@eliot.unil.ch. E-mail (Ukraine): tetko@bioorganic.kiev.ua.

(1)) Layloff, T. P. *Pharm. Technol.* **1991**, *15*, 146−148.
(2) Kirchhoefer, R. D. *J. AOAC Int.* **1992**, *75*, 577−580.
(3) Haddad, W. FDC Reports, Inc.: Chevy Chase, MD; August 14, 1989.
(4) Inman, E. L.; Tenbarge, H. J. *J. Chromatogr. Sci.* **1988**, *26*, 89−94.
(5) Slutsker, L.; Hoesly, F. C.; Miller, L.; Williams, L. P.; Watson, J. C.; Fleming, D. W. *J. Am. Med. Assoc.* **1990**, *264*, 213-217.

Our recent studies,[7-9] evaluated several computer-based classifiers as potential tools for pharmaceutical fingerprinting. The data used in these studies were chromatographic data files that were generated by HPLC analyses of samples of L-tryptophan (LT) produced by six different manufacturers of LT, as reported earlier.[7] The samples were analyzed by several pattern recognition methods such as artificial neural networks (ANN), K-nearest neighbors, and soft independent modeling of class analogy (SIMCA), as well as by a panel of human experts.

The application of pattern recognition methods for pharmaceutical fingerprinting encounters several problems.[7] One problem is that HPLC trace impurity data are subject to concerns about repeatability and imprecision inasmuch as even HPLC columns that are nominally identical can exhibit variations in peak height (additive noise) and retention time (nonstationarity) for a given sample run under the nominally similar conditions (vide infra).[7,10] The problem of nonstationarity of HPLC signals cannot be eliminated with simple methods, for example, normalization of data to the same duration, that is usually used for HPLC data analysis.[7] This is because nonstationarity in time causes nonlinear deformations of the HPLC signals in time that cannot be corrected using only linear transformations of the time axis. On the contrary, the general methods of clustering, regression analyses and classification, in particular those used in refs 7-9, are based on the assumptions of the presence of additive noise only. Owing to the variations in peak retention times, the distances between chromatograms recorded for the same substance by the nominally similar columns could be large in ordinary Euclidean space (Figure 1). Therefore, the separate classes (e.g., manufacturer) do not represent compact sets in feature space, as is necessary for successful recognition. This explains the low prediction ability of the classifiers (about 80%) calculated for these data using the original feature space.[7]

The prediction accuracy of the classifiers was increased using a Window Preprocessing (WP) scheme.[7] Application of WP compensated to some extent for the nonstability of the chromatograms and reduced the negative effects of lot-to-lot and column-to-column variations on the performance of the classifiers. The prediction accuracy of all three classifiers improved compared to analysis of the nonpreprocessed data. The highest prediction rate (about 94% correct) was obtained with an artificial neural network classifier.[7] Additional improvement of the neural network prediction ability (95%) was calculated following an optimization of the fingerprint region.[9]

Although useful as a preprocessor, the WP method does exhibit some drawbacks. This approach is designed to decrease the dimension of the original feature space and to convert the classes under study into compact sets in a new feature space. This transformation allows the preprocessed data to be treated as models with additive noise. If, however, the window boundaries are located in the peak region, this transformation does not take place and the predictive ability of classifiers can be impaired. It should be also noted that no objective criteria exist for determining the precise number of windows and the location of the boundaries between them.

The present study investigates more complex models for signal preprocessing. An analysis of HPLC signals in phase space is proposed. This technique provides an appropriate description of the nature and peculiarities of the observed signals. The results calculated by this new method are easily interpreted and can be used to monitor consistency of the pharmaceutical product. The accompanying article[11] extends the proposed approach for pattern recognition of HPLC data.

EXPERIMENTAL SECTION

Data Description. The present study was conducted on the same HPLC data as previously investigated,[7-9] i.e., 253 chromatographic profiles obtained on L-tryptophan (LT) drug substance from production lots of six different commercial LT manufacturers designated as A, B, ..., F (Figure 2). The trace impurity patterns were recorded using three different HPLC columns—Waters, Vydac 1, and Vydac 2—from two commercial production lots of each manufacturer. The HPLC columns were of similar type; i.e., they contained the same 5-μm C_{18} reverse-phase packing.[7] For each combination of LT manufacturer, lot, column, and run day, 3-5 replicate chromatograms were recorded. Two markers, M1 and M2, were added to each sample to bracket the retention times of the peaks (Figure 1A). The duration and the amplitude of each data sample were normalized on a linear scale with respect to the amplitude and detector responses of the markers.[7] Each data sample was represented by 899 points located between the LT-peak manifold and the M2 peak marker (Figure 1). The cardinal number of the point was used to measure its absolute time in the chromatograms. More details on the data preprocessing of these chromatograms can be found elsewhere.[7]

Baseline Drift Correction. The analysis in phase space is sensitive to the baseline drift that was observed for these trace impurity patterns (Figure 2). A simple method for correcting the drift was implemented. We found that the half-width of single peaks in the HPLC data was about 10-15 time units. For each point of the chromatogram, the minimum value of the HPLC signal within a time window of ±25 ms was calculated. This minimum value was then subtracted from the signal value at the point under analysis. The data preprocessed in such a way served as the initial source for analysis in phase space.

METHODS

Mathematical Statement of the Problem and General Description of the Model. The mathematical model investigated in the current study is based on the assumption that a recorded HPLC signal can be described as a solution of a nonlinear differential equation with noise. This allows us to describe the noise components of the signal by two sources: additive and perturbative noise. The additive noise contributes to the distortion of the amplitude of the signal $x(t)$, while the perturbative noise contributes primarily to deformation of the signal along the time

(6) Trucksess, M. W.; Thomas, F. S.; Page, S. W. *J. Pharm. Sci.* **1994**, *83*, 720-722.

(7) 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.

(8) Collantes, E. R.; Duta, R.; Welsh, W. J.; Zielinski, W. L.; Brower, J. *Anal. Chem.* **1997**, *69*, 1392-1397.

(9) 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.

(10) Otto, M. *Anal. Chem.* **1990**, *62*, 797A-802A.

(11) Tetko, I. V.; Aksenova, T. I.; Patiokha, A. A.; Villa, A. E. P.; Welsh, W. J.; Zielinski, W. L.; Livingstone, D. J. *Anal. Chem.* **1999**, *71*, 2431-2439.
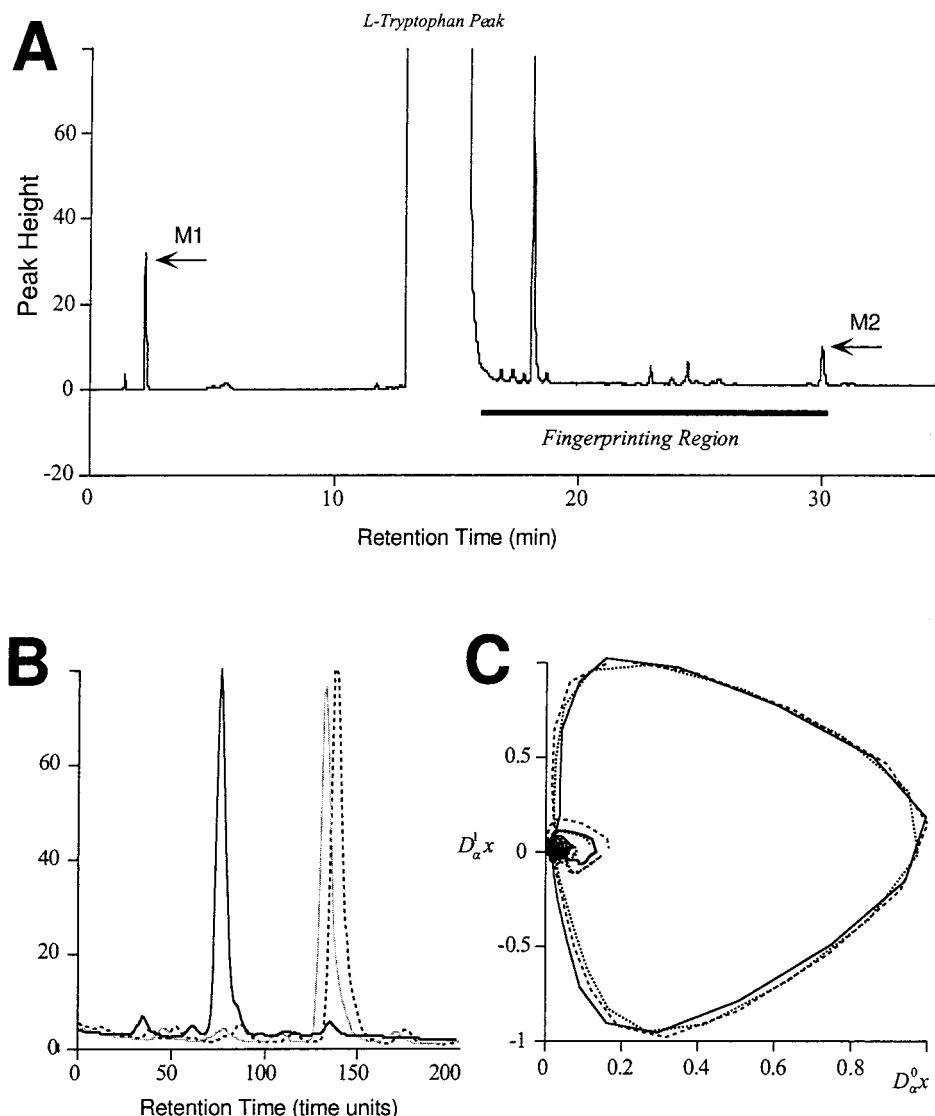
Figure 1. Example of HPLC data recorded from manufacturer E used in the present study following data normalization procedure. (A) The location of the early (M1) and late (M2) markers are indicated by arrows. The thick line indicates the fingerprinting region. (B) Fragments of three chromatograms of manufacturer E recorded using the nominally similar condition (column Vydac, lot 1) in the time domain. The solid line indicates a chromatogram recorded with Vydac 1, and the two remaining lines correspond to chromatograms recorded with the Vydac 2 column. Note that a shift in time of appearance of the peak can provoke difficulties with pattern recognition methods when applied to classification of these signals in the time domain. (C) Trajectories of the same chromatograms in phase space x, x'. An analysis of the data in this space compensates for distortion of the HPLC signals along the time axes.

axis. As a result of the perturbative noise, corresponding identical segments of chromatograms could appear at different moments of time (Figure 1A). Similar signal distortions in both the amplitude and time axes have been observed in the analysis of sun spot activity and of seismogram and cardio- and myogram data,[12−16] and the approach presented in this study has already been demonstrated as successful for modeling these problems.

Let us formulate the mathematical statement of the problem. We suppose that signal $\bar{x}(t) = x(t) + \xi(t)$ is observed at discrete

(12) Chertoprud, A. G.; Gudzenko, L. I. In *Kinetics of Simple Models of Oscillating Theory*, Basov, N. G., Ed.; Nauka: Moscow, 1976; Chapter 8.
(13) Gudzenko, L. I. *Izv. Vuzov Radiophys.* **1962**, 5, 573−587.
(14) Aksenova, T. I.; Mostovoy, S. V.; Mostovoy, V. S.; Osadchuk, A. E.; Shelekhova, V. Yu. *Dopov. Acad. Nauk Ukr.* (*Proc. Ukrainian Acad. Sci.*) **1997**, No. 1, 121−124.
(15) Aksenova, T. I.; Chibirova O. K. *Proceedings of the 18th Annual International Conference of IEEE*, IEEE Press: Amsterdam, 1996.
(16) Gudzenko, L. I.; Orlov, V. N. In *Questions in Clinical Biophysics*, Shultzev G. P., Ed.; Nauka: Moscow,1966.

times $t = 0, 1, ..., T$. The term $\xi(t)$, which corresponds to the additive noise of the signal $x(t)$, is a sequence of independent identically distributed random variables with zero mean and finite variance ($\sigma^2 < \infty$). The space of $\bar{x}(t)$ and t is known as the time domain of the signal. Examples of such signals are the HPLC traces in Figures 1 and 2.

We assume that the observed signal $x(t)$ is a solution of an ordinary differential equation

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

where q is the order of the equation and $F()$ describes a perturbation function of a random process with zero mean and amplitude that is bounded by a small value. The perturbation function is characterized by a small correlation time τ^* determined
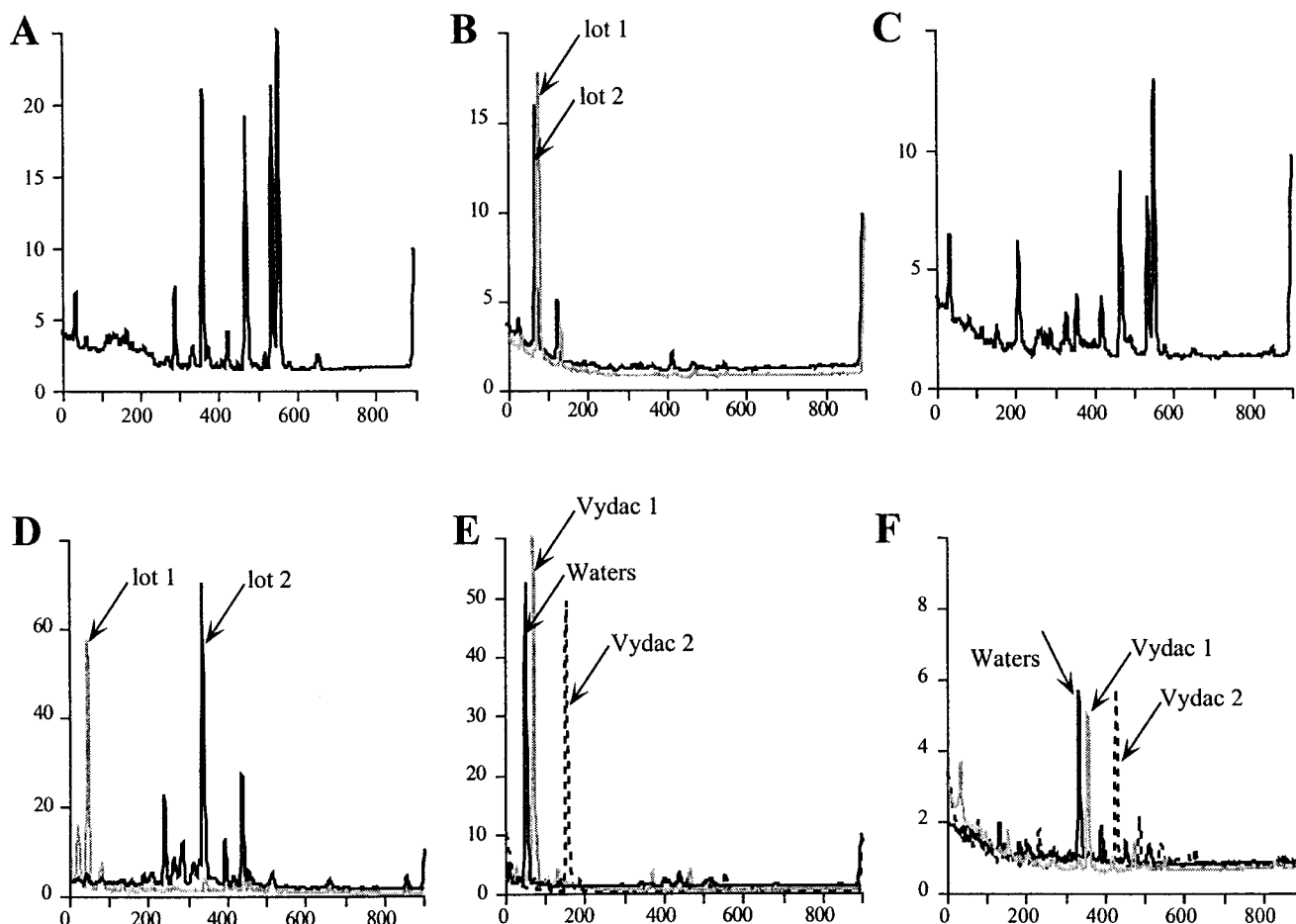
Figure 2. Representative chromatograms from each of the six LT manufacturers (A–F) following the data normalization procedure. The duration of the fingerprint region is 899 time units and the height of marker M2 (see peak at $t = 899$ time units) is 10 for all data. Unless otherwise noted, the data for all manufacturers were recorded using the Vydac 1 column from the first lot. Variations in HPLC profiles due to use of different lots are shown for manufacturers B (there is no essential variation) and D (the profiles for each lot are completely different) with gray lines corresponding to the first lot. The presence of significant shifts due to use of different columns is noted for manufacturers E and F.

as the value for which correlation function $E(F(t)\ F(t - \tau)) \approx 0$ if $\tau > \tau^*$ for any t; i.e., $F(t)$ and $F(t - \tau)$ are uncorrelated random variables if $\tau > \tau^*$.

The mathematical theory of processes described by eq 1 was developed by several authors.[12,17] This theory makes the assumption that the undisturbed equation

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

describes a self-oscillating system with stable limit trajectory[18] $\mathbf{x}^0(t) = (x_1^0(t), ..., x_q^0(t))^T$ in phase space with $x_1 = x$, $x_2 = dx/dt$, ..., $x_q = d^{q-1}x/dt^{q-1}$.

Thus, the mathematical model introduced in eqs 1 and 2 presumes that the observed signal is described by a solution of differential equations with noise presented in the dynamic equation. This model has several useful properties, as described below. For the sake of simplicity, we restrict our analysis to the equation

of second order $q = 2$, but the conclusion can be generalized for an equation of any order. Examples of HPLC signal trajectories in the time domain and in phase space are shown in Figure 1. The phase space has two coordinates x, $y = dx/dt$ for $q = 2$. The undisturbed eq 2 has a stable limit trajectory $x^0(t)$, $y^0(t)$ in this phase space.

It is well-known that the trajectory of the signal tends continuously to the limit trajectory whenever it is found in the limit trajectory neighborhood, independently of initial conditions.[12,18] The perturbation function $F()$ in eq 1 tends to displace the trajectories of the signal out from the limit trajectory. However, if the perturbation is small enough, the trajectories stay in the neighborhood of the limit trajectory $x^0(t)$, $y^0(t)$; i.e., the solutions of eq 1 are similar to one another but they never coincide. Indeed, the chromatograms recorded from the same manufacturer using nominally similar columns (Figure 1C) have very similar trajectories in phase space, but these trajectories do not coincide.

It is convenient for the next analysis to introduce new variables for describing the limit trajectory and signals in phase space. Let us fix an arbitrary point on the trajectory P_0 as a starting point (Figure 3). The position of any other point on the limit trajectory can be described by its phase θ, which is a time movement along the trajectory from the starting point P_0 to the point being

(17) Bogoljubov, N. N.; Mitropolsky, Y. A. *Asymptotic methods in the theory of nonlinear oscillations*, 2nd ed.; Gordon and Breach: New York, 1961.
(18) The mathematical term "stable limit cycle" is also used to designate the limit trajectory $\tilde{x}^0(t)$. See, e.g.: Kreyszig, E. *Advanced Engineering Mathematics*; John Wiley & Sons: New York, Singapore, 1993.
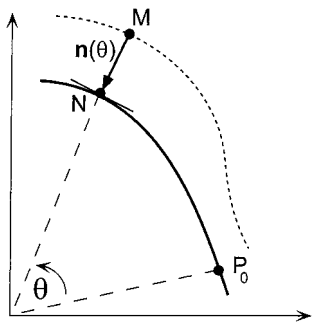
Figure 3. New variables, i.e., θ and $\mathbf{n}(\theta)$, are introduced to describe trajectories of the analyzed signal in phase space. The thick line is the limit trajectory. The length of the vector $|\mathbf{n}(\theta)|$ corresponds to the minimal distance between the signal and the limit trajectory.

analyzed. The phase θ unambiguously characterizes all points of the limit trajectory. The trajectories of the signal with noise can be described by variables $\mathbf{n}(\theta)$ and $t(\theta)$ (Figure 3). Here $\mathbf{n}(\theta)$ is a vector between 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). The second variable $t(\theta)$ is a time movement along the analyzed curve. The time movement is easy to interpret. For example, it corresponds to the retention times of the identified chromatogram peaks. Thus, each signal trajectory is described by variables $\mathbf{n}_i(\theta)$ and $t_i(\theta)$ where i is the number of the trajectory. These variables characterize the deviation of an individual signal from the limit trajectory. The limit trajectory is defined by $\mathbf{n}(\theta) \equiv \mathbf{0}$ and $t(\theta) \equiv \theta$, where $\mathbf{0}$ denotes a vector with all components equal to 0.

The mean trajectory of the signals in phase space converges to the limit trajectory if the number of averaged trajectories increases infinitely:[12,17]

$$\begin{cases} \mathbf{n}^*(\theta) = \lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}\mathbf{n}_i(\theta), & t^*(\theta) = \lim_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}t_i(\theta) \\ \mathbf{n}^*(\theta) \approx \mathbf{0}, & t^*(\theta) \approx \theta \end{cases} \quad (3)$$

The values $t_i(\theta)$ and components of the vector $\mathbf{n}_i(\theta)$ are characterized by an asymptotically Gaussian distribution for any θ; i.e., they can be described as the sum of their means and asymptotically Gaussian noise.[12] This property allows us to estimate the mean trajectory of the signal in phase space. Let us select some reference trajectory of HPLC in phase space x, y and calculate a normal line to this trajectory at some point. This normal line will cross trajectories of all other signals at points $y(t_i(\theta))$, $x(t_i(\theta))$. The mean trajectory is estimated by

$$\begin{cases} \hat{x}^0 = \frac{1}{k}\sum_{i=1}^{k}x_i(t_i(\theta)) \\ \hat{y}^0 = \frac{1}{k}\sum_{i=1}^{k}y_i(t_i(\theta)) \end{cases} \quad (4)$$

The average phase is estimated as $\hat{\theta} = (1/k)\sum_{i=1}^{k}t_i(\theta)$ according to eq 3. Then the function $\hat{x}^0(\theta)$ approximates the mean trajectory in the time domain.

Let us assume that the chromatograms recorded from the same LT manufacturer are characterized by the same limit

trajectory using the above model. In this case all samples recorded from the same manufacturer are neighbors of this limit trajectory. The mean trajectory estimated according to eq 4 corresponds to an asymptotic unbiased estimation of a chromatogram of the manufacturer in both phase space and time domain. We refer to the mean trajectory in phase space as the *phase fingerprint* of the data.

The relevance of the proposed model to the problem under analysis is verified by the Kolmogorov−Smirnov (K−S) test of normality[19] for $t(\theta)$ and for the components of vector $\mathbf{n}(\theta)$.

Estimation of Signal Derivatives. The analysis in phase space requires estimation of higher-order derivatives of the signal, and some computational problems must be addressed. The presence of both types of noise in the signal seriously affects the calculations.

In a recent study,[20] we proposed that the signal trajectory in phase space can be described by coordinates $D_\alpha^0 x, D_\alpha^1 x,...,D_\alpha^{q-1} x$, where D_α^k are integral operators

$$D_\alpha^k f(t) = \int_R \omega_\alpha^k(\tau - t) f(\tau) \, \mathrm{d}\tau \quad (5)$$

and the kernel function ω_α satisfies the following conditions:

(a) $\omega_\alpha(t) = 0, \quad$ if $|t| > \alpha$

(b) $\int_R \omega_\alpha(t) \, \mathrm{d}t = 1 \quad (6)$

(c) ω_α has q continuous derivatives

It was shown[20,21] that D_α^k estimates a smoothed derivative of the signal of the order k with parameter of regularization α. For example, if function $f(t)$ has the derivative $f(t)^k$ then the $D_\alpha^k f(t)$ tends to $f(t)^k$ if $\alpha \to 0$; i.e., application of the integral operators replaces the complex problem of derivative estimation with a simpler calculation of integrals.

Any function satisfying conditions 6 can be used as the kernel. Selection of both the kernel function and the value of the parameter α depends on the order of the derivative to be calculated, the level of additive noise, and the required smoothness of the data. In the present study, a fast algorithm of derivative estimation[20,21] was used to calculate integral operators D_α^k. This algorithm uses the piecewise polynomial kernel and incurs smaller errors of derivative estimation compared with traditional methods based on Gaussian or Sobolev operators.[21] In addition, application of this algorithm is computationally efficient.

Choice of Model Parameters. Previous applications of the proposed model have shown that optimal results for analysis of electrocardiogram and myograms[14,16,22] were provided by the third-order equation, while studies of the sun spot activity used the second-order equation.[12] Applying a rule of thumb, we decided to use the third-order equation for analysis of the HPLC data. The

(19) Press: W. H.; Teukolsky, S. A.; Vetterling, W. T. Flannery, B. P. *Numerical Recipies in C. The Art of Scientific Computing*; Cambridge University Press: Cambridge, 1994; Chapter 14.
(20) Aksenova, T. I.; Shelekhova, V. Yu. *SAMS* **1995**, *18*, 159−163.
(21) Shelekhova, V. Yu. Efficient Algorithms of Derivative Estimation for Noisy Observations. Ph.D. Thesis, Kyiv, Institute of Cybernetics, 1995.
(22) Gudzenko, L. I.; Sorokina, A. E. In *Kinetics of Simple Models of Oscillating Theory*; Basov, N. G., Ed.; Nauka: Moscow, 1976; Chapter 7.
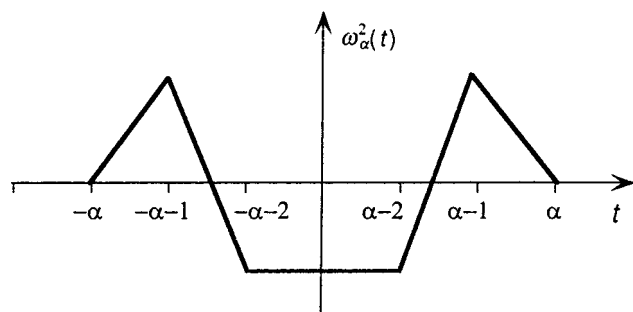
Figure 4. Kernel function $\omega_\alpha^2(t)$ used to estimate the second derivative of the signal.

use of an equation of such order requires estimation of the first and the second derivatives of the signal in phase space. This was done with the piecewise polynomial kernel of third order. The second derivative of this kernel is a piecewise linear function of the form

$$
\omega_\alpha^2 = \begin{cases}
3(\alpha + x)/(1 - 3\alpha + 2\alpha^2), & x \in [-\alpha, -\alpha + 1] \\[4pt]
3(4 - 5\alpha + 2\alpha^2 + x(2\alpha - 1))/ \\
\quad (3 - 11\alpha + 12\alpha^2 - 4\alpha^3), & x \in [-\alpha + 1, -\alpha + 2] \\[4pt]
6/(3 - 11\alpha + 12\alpha^2 - 4\alpha^3), & x \in [-\alpha + 2, 0]
\end{cases} \quad (7)
$$

which is symmetrical on the interval $[0, \alpha]$ (see Figure 4). The kernel itself and its derivatives are determined by integration of ω_α^2 under condition (6) of normalization. This kernel was found by minimization of the error for estimation of derivatives with a fixed value of the regularization parameter.[21]

The use of integral operators for calculation of derivatives requires selection of the parameter α. This parameter corresponds to the time interval $[t - \alpha, t + \alpha]$ that is used for signal smoothing and for calculating its derivatives. The value $\alpha = 3$ was chosen. It is the smallest value that can be used for integer-valued partitioning of the time interval.[20] Selection of a small value for parameter α assumes an absence of significant additive noise in the data being analyzed.

Normalization of Signal and Its Derivatives. Signal x and its derivatives $d^j x/dt^j$ are essentially on different scales. This dependency is due to the scale of time t that is used for differentiation of the signal. Use of different time scales influences the relative magnitude of derivatives compared with the original signal as well as the shape of the phase fingerprints. To solve the problem of relative scaling in phase space, the maximum absolute value of signals and derivatives of all chromatograms recorded from the same manufacturer were normalized on the interval of unit length. This procedure attributes equal relevance to the signal and its derivatives in phase space. The signal x was simply normalized on the interval $[0,1]$. The derivatives $d^j x/dt^j$ were normalized to the unit-length interval in such a way that the zero value of the derivatives was not displaced. Such normalization calculates the derivatives that are all equal to 0 for any flat signal $x \equiv$ constant, $t = 0, ..., T$. The normalized derivatives had maximum and minimum values that were, in general, asymmetric along the time axis. In the samples studied here, we found that

Table 1. Parameters of the Highest Peaks P_{max} and the Results of the Kolmogorov−Smirnov Test Applied for Analysis of Data in Phase Space

LT manufacturer	parameters of the peak P_{max}			$\mathbf{n}(\theta_{pmax})$		$t(\theta_{pmax})$	
	θ_{Pmax}[a]	$E(P_{max})$	$\sigma(P_{max})$	$\sigma(|\mathbf{n}|)$[b]	K−S[c]	$\sigma(t)$	K−S
A	560	28.6	6.06	0.14	true	43	false
B	83	15.6	3.58	0.15	true	36	false
C	429	6.52	5.23	0.26	true	36	false
D	215	134	90.3	0.29	false	163	false
D[d] (lot 1)	53	202	87.2	0.28	true	29	false
D (lot 2)	369	70	10.5	0.13	true	37	false
E	82	55.3	14.9	0.17	true	34	false
F	369	5.2	1.22	0.18	true	37	false

[a] The phase θ_{pmax} corresponds to the time of appearance of the highest peak P_{max} at the limit trajectory, $t_i(\theta)$ corresponds to the time of appearance of this peak for the ith HPLC chromatogram and $n_i(\theta)$ measures the deviation of the ith trajectory from the limit trajectory (see Methods Section and Figure 3 for more details). The average magnitude of peaks $E(P_{max})$ and their absolute standard deviation $\sigma(P_{max})$ are also shown. [b] Dispersion of the length of vector $\mathbf{n}(\theta)$. [c] K−S is the Kolmogorov−Smirnov goodness-of-fit test applied to projection of $\mathbf{n}(\theta_{pmax})$ vector at the signal axis; "false" indicates that analyzed distribution is different from the normal at significance level $p = 0.01$. [d] separate analysis of data collected from two lots of manufacturer D.
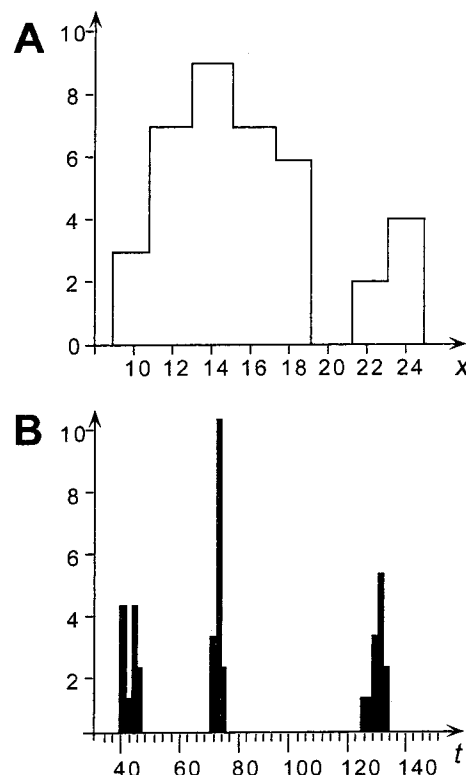

Figure 5. Histogram of a projection $n_x(\theta)$ of vector $\mathbf{n}(\theta)$ on a signal axis and histogram of movement time $t(\theta)$ recorded for the highest peak of manufacturer B. (A) The distribution of $n_x(\theta)$ fits the Kolmogorov−Smirnov test of normality. (B) The distribution of $t(\theta)$ does not fit this test, and three separate subdistributions can be clearly identified for it. Each subdistribution is formed by data recorded by the same HPLC column.

the signal derivatives were almost symmetrical and, thus, their normalized values were approximately in the range $(-0.5, 0.5)$. The data normalized in this way were used to calculate the phase fingerprints. It is important to note that this normalization (and any other similar linear transformation of axes of signals in phase
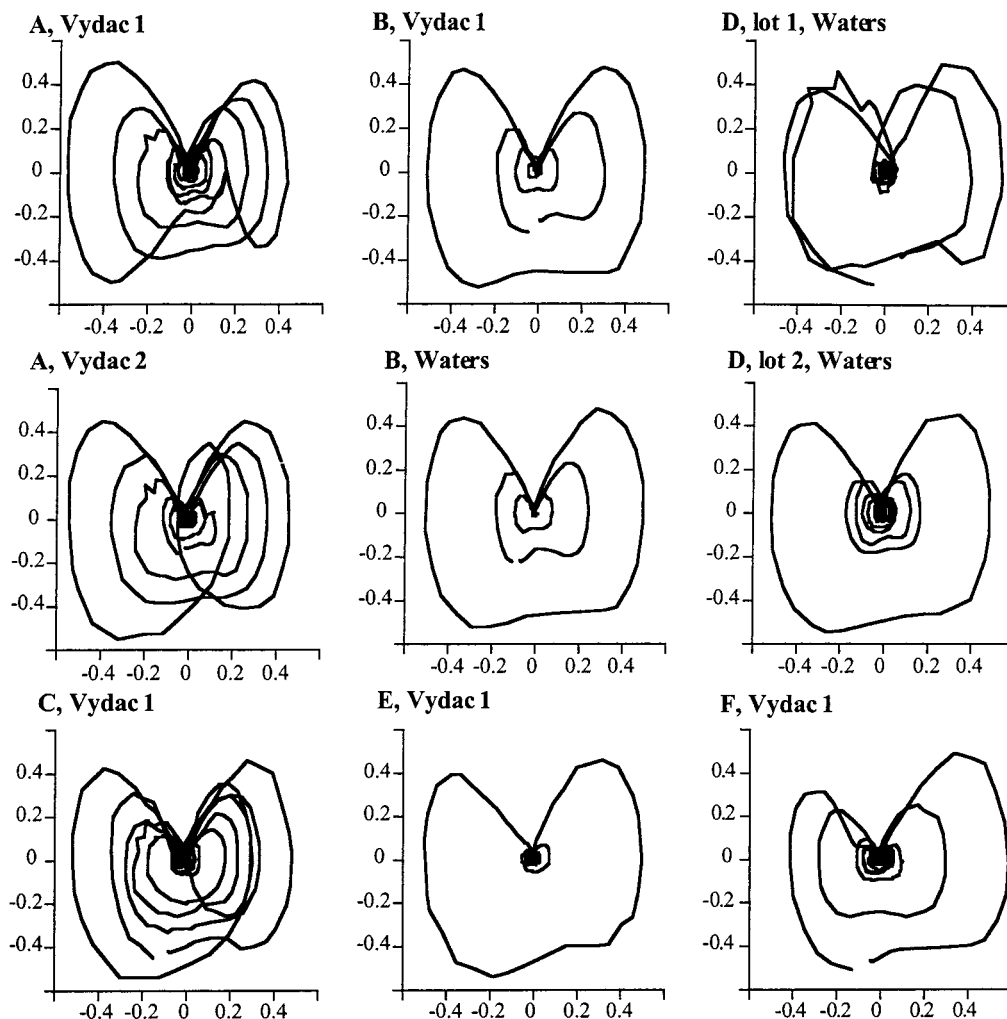
Figure 6. Phase fingerprints of manufacturers A–F in the projection to phase space x', x''. The type of HPLC column used for data recording is indicated. The phase fingerprints for HPLC columns recorded (for example, A, Vydac 1 and A, Vydac A; B, Vydac 1 and B, Waters) are remarkably similar for the same manufacturer but are unlike in shape for different manufacturers. The phase fingerprints calculated for two lots of manufacturer D are completely different even for the same column.

space) does not affect results of the K–S goodness-of-fit test that are reported in the Results.

Selection of Initial Phase for Data Analysis. One particular requirement for applying the K–S test of normality (and before averaging in eq 4) is to identify data having the same phase θ. We assumed that the retention time of the highest peak of chromatograms recorded from the same manufacturer was characterized by the same phase. The data calculated for this peak were used to investigate the distributions of $\mathbf{n}(\theta)$ and $t(\theta)$. It should be noted that the same analysis can be done for any particular chromatographic peak taken from the fingerprint region.

RESULTS

For the first step of this analysis, all data recorded from the same LT manufacturer were pooled together. Application of the algorithm for analysis of these data revealed that the variability of components of $\mathbf{n}(\theta)$ satisfied the K–S goodness-of-fit test (Table 1). For manufacturer B the distribution of the projection of $\mathbf{n}(\theta)$ on the signal axis, i.e., $n_x(\theta)$, was unimodal (Figure 5A). Conversely, the distribution of $t(\theta)$ for the same data was multimodal (Figure 5B) and clearly indicated a mixture of three distributions. The distributions were separated using the criterion of maximum

likelihood. Each of them satisfied the K–S test (Table 1) and represented HPLC data recorded using the same column. A similar analysis revealed the same structure of data for manufacturers A, C, E, and F. The parameters of the K–S test of normality for $n_x(\theta)$ and $t(\theta)$ data are summarized in Table 1.

The distribution of $n_x(\theta)$ for data of manufacturer D was interpreted as a mixture of two normal distributions. After separation using the criterion of maximum likelihood, the distributions satisfied the K–S test. The two distributions corresponded to the two production lots of manufacturer D. We found that the variability of $t(\theta)$ associated with HPLC column was similar to that observed for other manufacturers. Thus, six normal distributions corresponding to combinations of three different HPLC columns and two commercial lots were separated for manufacturer D.

The analysis of data in phase space detected a significant difference in HPLC profiles according to column type for all manufacturers. However, data recorded using the same column and both lots of all manufacturers, except manufacturer D, satisfied the K–S test of normality and the requirements of the mathematical model that we introduced. In the case of manufacturer D, the

Table 2. the Phase $\theta_{P\max}$ and the Dispersion of the Movement Times $\Sigma(t)$ of the Highest Peak $P_{\max}$ Calculated for HPLC Columns[a]

	Waters		Vydac 1			Vydac 2		
LT manufacturer	$\theta_{P\max}$	$\sigma(t)$	$\theta_{P\max}$	$\sigma(t)$	$\Delta(\theta)^a$	$\theta_{P\max}$	$\sigma(t)$	$\Delta(\theta)$
A	513	3.24	550	1.58	37	625	1.60	112
B	43	2.41	74	1.19	31	131	2.34	88
C	514	3.01	552	1.61	38	627	2.17	113
D (lot 1)[b]	1	1.79	30	2.6	29	95	2.73	94
D (lot 2)	335	2.76	354	0.88	19	424	1.97	89
E	44	3.99	75	3.19	31	133	8.76	89
F	335	1.56	355	1.52	20	427	1.60	92

[a] The difference in phases of the highest peaks detected using Vydac 1, 2 and Water HPLC columns. [b] A separate analysis of HPLC data recorded for two lots of manufacturer D was performed. The distributions of movement times $t(\theta)$ fits the Kolmogorov−Smirnov test of normality at the level of significance $p = 0.01$ for all data.

proposed model can be successfully applied only after additional separation of data according to production lots. A significant difference in HPLC profiles detected for two lots of manufacturer D indicated a crucial change in the production process by this manufacturer. This result was confirmed by visual inspection of HPLC profiles of this manufacturer recorded from both lots in time domain (Figure 2) and in phase space (Figure 6).

A separation of the chromatograms according to the type of HPLC columns significantly decreased the dispersion of times $t(\theta)$ calculated for the same phase (cf. Tables 1 and 2). On average, the dispersions of $\sigma(t)$ calculated for various columns were very similar (Table 2). This indicates a similar performance of the columns employed.

The times of appearance of the highest peaks detected by different columns for the same data were always in the order Waters < Vydac 1 < Vydac 2 (see Figure 2 and Table 2). The time delays between columns depended on the absolute position of the peak in the HPLC data. For example, the highest peak of the data from manufacturer E was detected by the Waters column 89 counts in advance of that by Vydac 2 (absolute times for peak with Waters and Vydac 2 were 44 and 133 counts), while this delay was 112 for the highest peaks of data from manufacturer A. The time delay between the highest peaks did not depend on the analyzed manufacturer but only on the absolute retention times of the peak. For example, similar shifts in retention times were calculated for manufacturers A and C, D (lot 2) and F, and B and E. This indicates that HPLC profiles of the same manufacturer detected by different columns are nonlinearly transformed compared one to another. The transformation function does not depend on chromatographic data but only on the absolute retention times of the peaks. Let us note that the shifts between chromatograms decreased to "0" at the position of marker "M2" (since the same marker was used for all columns). We found that this decrease was not evident for phases $\theta = [700-899]$. The shifts between Waters and Vydac 1 and between Waters and Vydac 2 decreased from approximately 40 and 110, respectively, to 0 times units for these phases. On the contrary, for phases $\theta = [0-650]$, the shifts between Waters and Vydac 1, and Waters and Vydac 2 were in the range 20−40 and 80−110, respectively. The region $\theta = [0-650]$ was characterized by the presence of many essential peaks in the trace impurity patterns of analyzed manufacturers

and, thus, was presumably more important for analysis and identification of drug manufacturers and calculation of phase fingerprints compared to the region with phases $\theta = (700-899)$ that did not contain such peaks (Figure 2).

DISCUSSION

The results of the present study show that the general assumption about the possibility to describe HPLC data as a solution of an ordinary differential equation is valid for HPLC data. However, the differences in properties of HPLC columns generate different phase fingerprints for each column. Thus each manufacturer can be described with several phase fingerprints calculated according to the HPLC columns employed.

Our data confirm previous results that found significant column-to-column variations in HPLC data using the SIMCA method.[7] The current analysis suggests that this variation is mainly due to the presence of significant nonlinear transformations and shifts of HPLC signals in the time domain. A comparison of chromatograms recorded for the same data but using different HPLC columns may provide some difficulties in the identification of the original drug manufacturers, especially if analysis is performed in the time domain. The detected shifts influenced times of appearance but did not change the absolute amplitudes of the HPLC peaks, as indicated by the absence of significant variation of amplitudes of the highest peaks detected by different columns for data recorded from the same manufacturer (Table 1). However, the phase fingerprints calculated using different columns are remarkably similar for the same manufacturer despite the nonlinearity over time and, certainly can discriminate among the different manufacturers (Figure 6). This suggests that analysis in phase space could provide a reliable pattern recognition of HPLC data and, perhaps, other applications.

These qualitative results of preliminary HPLC data analysis are used to develop a successful pattern recognition method, as demonstrated in the accompanying article.[11]