Scholarly Research Exchange

SRX Pharmacology • Volume 2010 • Article ID 481497 • doi:10.3814/2010/481497

Research Article

Recent Progress in the pK_a Estimation of Druglike Molecules by the Nonlinear Regression of Multiwavelength Spectrophotometric pH-Titration Data

Milan Meloun,[1] Tomáš Syrový,[1] and Jahan Ghasemi[2]

[1] *Department of Analytical Chemistry, University of Pardubice, Pardubice 532 10, Czech Republic*
[2] *Chemistry Department, Faculty of Sciences, K. N. Toosi University of Technology (KNTU), 16167 Tehran, Iran*

Correspondence should be addressed to Milan Meloun, milan.meloun@upce.cz

Received 19 October 2009; Revised 3 December 2009; Accepted 21 December 2009

Recent developments in the computational diagnostic tools for the pK_a estimation of druglike molecules carried out by the nonlinear regression of multiwavelength spectrophotometric pH-titration data are demonstrated on the protonation equilibria of silybin. The factor analysis of spectra predict the correct number of components when the *signal-to-error ratio SER* is higher than 10. The mixed dissociation constants of the drug silybin at ionic strength $I = 0.03$ and a temperature of 25°C were determined using two different programs, SPECFIT32 and SQUAD(84). A proposed experimental and computational strategy for the determination of the dissociation constants is presented. The dissociation constant pK_a was estimated by nonlinear regression of the $\{pK_a, I\}$ data at 25°C with SQUAD (and SPECFIT); that is, $pK_{a1} = 6.898(0.022)$ and $6.897(0.002)$; $pK_{a2} = 8.666(0.021)$ and $8.667(0.012)$; $pK_{a3} = 9.611(0.010)$ and $9.611(0.004)$; $pK_{a4} = 11.501(0.008)$ and $11.501(0.007)$. While great progress has been achieved in terms of the reliability of the protonation model estimation, among the most efficient diagnostics of the nonlinear regression of multiwavelength pH-spectra are the goodness-of-fit test, Cattel's scree plot of the factor analysis, spectra deconvolution, the signal-to-error SER ratio analysis, and other tools of efficient spectra analysis.

1. Introduction

Protonation constants, or acid dissociation constants, are very important both in the analysis of drugs and in the interpretation of their mechanisms of action. Spectrophotometry is a convenient method for pK_a determination in very dilute aqueous solutions (about 10^{-5} to 10^{-6} M), provided that the compound possesses pH-dependent light absorption due to the presence of a chromophore in proximity to the ionization centre. Much more information can be extracted if multivariate spectrophotometric data are analyzed by means of an appropriate multivariate data analysis method *cf.* [1–20]. In previous work [21–33], the authors have shown that the spectrophotometric method can be used in combination with suitable chemometric tools for the determination of protonation constants β_{qr} or acid dissociation constants pK_a, even for sparingly soluble drugs.

The Biopharmaceutics Classification system BCS [34, 35] classifies every pharmaceutical active ingredient into one of 4 groups based on two basic characteristics: solubility and permeability. We decided to complete such information and to study the protonation equilibria of a pharmaceutical active ingredient of natural origin, silybin, which possesses low water solubility. While special attention was paid to the methodology, we also studied the protonation equilibria of silybin, in order to verify the validity of the results. Silybin is a flavanolignan belonging to a group of biologically active substances found in the milk (or Saint-Mary) thistle, *Silybum marianum (L.) Gaertner*. These substances are almost certainly produced in the plant by a radical coupling of flavonoid and coniferyl alcohol [36]. Although the plant has been used since ancient times for the treatment of liver and stomach diseases, the first representative of the group, silybin, was described only in the late 1960s. The group of flavanolignans extracted from the seeds of

Saint-Mary thistle was originally called by one common name, silymarin, used interchangeably with the name of its main component, silybin. In 1974 Wagner et al. [37] proposed giving the name silymarin to the whole group of active flavanolignan-like substances and the names silybin, silydianin, and silychristin to its main constituents. The major component of silymarin is silybin, which constitutes 60–70% of the drug [38]. The protonation constants of silybinin at various ionic strengths have been studied at various temperatures [24]. However, in only a few cases has the dependence of the protonation constants on ionic strength been systematically investigated, and the thermodynamic dissociation constant estimated. The reliability of dissociation constants obtained by the regression analysis of potentiometric or pH-spectrophotometric data is dependent upon (i) the calibration of the glass electrode cell, (ii) the algorithm used, (iii) the instrumental method used and the parameters selected for refinement, and (iv) a strategy of efficient experimentation. Much work has been put into developing methods for the resolution of multicomponent spectra but less work has been carried out to reveal the limitations of these methods and in the estimation of the minor components of the resolved spectra. Approaches to determining the rank of the absorbance matrix A are based on pure principal component analysis (PCA). Generally, PCA will extract some of the noise, that is, the experimental and/or random error which will usually be represented by the principal components with smallest size or variance. When no noise is present in the spectra, the number of eigenvalues of the covariance matrix A^TA larger than zero is equivalent to the number of components r, assuming that the spectra of components in the mixture are linearly independent. As all real data always contain experimental noise, the number of eigenvalues different from zero is usually larger than the number of components, r. Experimental and/or random error can mask the identification of the true dimensionality of a given data set. In any study of this type, the level of "experimental noise" will be a critical factor. It is therefore necessary to have a consistent definition of the *signal-to-noise ratio (SNR)*, so that the impact of this parameter can be critically assessed. Traditional approaches to SNR are typically based on the ratio of the maximum signal to the maximum noise value. As an alternative, the concept of instrumental error has again been employed, and the *signal-to-error ratio (SER)* is defined where, as an error level, the instrumental standard deviation of absorbance, $s_{\text{inst}}(A)$, is used. Attention should be paid to the methods' ability to detect a minor component in the presence of major ones. The detection limit or the smallest relative concentration of the minor component present depends on several factors, such as (i) the spectral similarity of the minor component to the others, (ii) instrumental resolution, (iii) noise level and noise type, and (iv) the signal-to-noise ratio *SNR* with respect to the minor component.

The regression methods include traditional least-squares curve fitting approaches, based on the previous postulation of a chemical model, that is, the postulation of a set of species defined by their stoichiometric coefficients and formation constants, which are then refined by least-squares minimization. These mathematical procedures require the fulfillment of the mass-balance equations and the mass-action law. The most relevant algorithms are SQUAD(84) [6–11] and SPECFIT32 [14–17, 39].

This paper describes the current status of computational diagnostic tools in the pK_a estimation of a drug carried out by the nonlinear regression of the multiwavelength spectrophotometric pH-titration data. The dissociation constants of the drug, silybin, at an ionic strength $I = 0.03$ and at 25°C, are estimated to prove the reliability of the whole regression procedure.

2. Theoretical

2.1. Procedure for the Determination of the Protonation Constants. The protonation equilibria between the anion L (the charges are omitted for the sake of simplicity) of a drug and the proton H are considered to form a set of the variously protonated species $\text{L}, \text{LH}, \text{LH}_2, \text{LH}_3, \ldots$, and so forth, which have the general formula L_qH_r in a particular chemical model and which are represented by n_c the number of species, $(q,r)_i$, $i = 1, \ldots, n_c$, where index i labels their particular stoichiometry; the overall protonation (stability) constant of the protonated species, β_{qr}, may then be expressed as

$$\beta_{qr} = \frac{\left[\text{L}_q\text{H}_r\right]}{\left([\text{L}]^q[\text{H}]^r\right)} = \frac{c}{l^q h^r}, \tag{1}$$

where the free concentration $[\text{L}] = l$, $[\text{H}] = h$, and $[\text{L}_q\text{H}_r] = c$. As each aqueous species is characterized by its own spectrum, for UV/VIS experiments and the ith solution measured at the jth wavelength, the Lambert-Beer law yields the absorbance, $A_{i,j}$, being defined as

$$A_{i,j} = \sum_{n=1}^{n_c} \varepsilon_{j,n} c_n = \sum_{n=1}^{n_c} \left(\varepsilon_{qr,j} \beta_{qr} l^q h^r\right)_n, \tag{2}$$

where $\varepsilon_{qr,j}$ is the molar absorptivity of the L_qH_r species with the stoichiometric coefficients q, r measured at the jth wavelength for n_s solutions with known total concentrations of $n_z = 2$ basic components, c_{L} and c_{H}, at n_w wavelengths. The multicomponent spectra analysing program SQUAD(84) [8] may adjust β_{qr} and ε_{qr} for a given absorption spectra set by minimising the residual-square sum function, U,

$$\begin{aligned} U &= \sum_{i=1}^{n} \sum_{j=1}^{m} \left(A_{\text{exp},i,j} - A_{\text{calc},i,j}\right)^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \left(A_{\text{exp},i,j} - \sum_{k=1}^{p} \varepsilon_{j,k} c_k\right)^2 \\ &= \text{minimum}, \end{aligned} \tag{3}$$

where $A_{i,j}$ represents the element of the experimental absorbance response-surface of size $n_s \times n_w$ (Figures 1(a), 1(b)) and the independent variables c_k are the total concentrations of the basic components c_{L} and c_{H} being

adjusted in n_s solutions. This means that the predicted absorbance-response surface is fitted to given spectral data, with one dimension representing the dependent variable (absorbance), and the other two dimensions representing the independent variables, namely. the total component concentrations (or pH) of n_s solutions, at n_w wavelengths. Another popular program is SPECFIT/32 [39], based on singular value decomposition and nonlinear regression modelling using the Levenberg-Marquardt method for the determination of stability constants from spectrophotometric titration data. The experimental and computational schemes for the determination of the protonation constants of the multicomponent system are taken from Meloun et al. [2] and the details for the computer data treatment are collected in the *Supporting Information.*

All spectra evaluation may be performed with the factor analysis INDICES algorithm [30] in the *S*-Plus programming environment [40]. Most index methods are functions of the number of principal components $PC(k)$'s, with the use of which the spectral data are usually plotted against an integer index k, $PC(k) = f(k)$. When the $PC(k)$ reaches, the value of the instrumental error of the spectrophotometer used, $s_{\text{inst}}(A)$, the corresponding index k^* represents the number of light-absorbing components in a mixture, $p = k^*$. In a scree plot the value of *PC(k)* decreases steeply with increasing *PC*s as long as the *PC*s are significant. When k is exhausted, the indices fall off, some even displaying a minimum. At this point, $p = k^*$ for all indices. The index values at this point can be predicted from the properties of the noise, which may be used as a criterion to determine p, [30].

2.2. Computational Schema for Protonation Model Building with SPECFIT/32. An experimental and computational scheme for the protonation model building of a multicomponent and multiwavelength system was proposed by Meloun et al., compare [2, page 226] or [8, 31], and is here revised with regard to SPECFIT/32 and INDICES application.

(1) *Instrumental error of absorbance measurements,* $s_{\text{inst}}(A)$: the INDICES algorithm, compare. [30], should be used to evaluate $s_{\text{inst}}(A)$. The Cattel's scree plot of $s_k(A) = f(k)$ of the Wernimont-Kankare procedure consists of two straight lines intersecting at $\{s_k^*(A);\ k^*\}$, where k^* is the matrix rank for the system and the instrumental error of the spectrophotometer used, $s_{\text{inst}}(A) = s_5^*(A)$ reaching a value of 0.25 mAU in range of 225–360 nm for the Cintra 40 (GBC, Australia) spectrophotometer employed. This value can be used for prediction of the *signal-to-error ratio*, (*SER),* for experimental data. It was proven that the indices are able to accurately predict the correct number of components that contribute to a set of absorption spectra for data sets with *SER* of equal to or higher than 10.

(2) *Experimental design:* Simultaneous monitoring of absorbance and pH during titrations is used in a titration, when the total concentration of one of the components changes incrementally over a relatively wide range, but the total concentrations of the other components change only by dilution.

(3) *Number of light-absorbing species:* a qualitative interpretation of the spectra aims to evaluate the quality of the dataset and remove spurious data, and to estimate the minimum number of *factors*, that is, contributing aqueous species, which are necessary to describe the experimental data. The INDICES [30] determine the number of dominant species present in the equilibrium mixture. The method can detect minor components and predicts the correct number of components for data sets with the signal-to-error ratio SER of at least equal to or higher than 10. For the signal value S in a numerator of the ratio *S/E*, the absorbance difference for the jth-wavelength at the ith-spectrum $\Delta_{ij} = A_{ij} - A_{i,\text{acid}}$ can be used, where $A_{i,\text{acid}}$ is the limiting spectrum of acid form of drug measured. This absorbance change Δ_{ij} is then divided with the instrumental standard deviation $s_{\text{inst}}(A)$ and resulting ratio $\Delta/s_{\text{inst}}(A)$ represents here the signal-to-error ratio SER of the spectra studied. This SER ratio is examined for all absorbance matrix elements in the whole range of wavelength λ and is compared with the limiting SER value. It was proven that when the ratio $\Delta/s_{\text{inst}}(A)$ is equal to or higher than 10, the factor analysis is able to predict the correct number of components in equilibrium mixture.

(4) *Choice of computational strategy of regression process:* the input data should specify whether β_{qr} or $\log \beta_{qr}$ values are to be refined with an application of two procedures of nonlinear regression.

(5) *The initial estimates of predicted parameter* β_{qr} *from molecular structure*: it is wise before starting a regression to analyze actual experimental data, to search for scientific library sources to obtain a good default for the number of ionizing groups, and numerical values for the initial guess as to relevant protonation constants and the probable spectral traces of all the expected components.

(6) *Diagnostics indicating a chemical model:* when the minimization process of a regression spectra analysis terminates, some efficient diagnostic criteria are examined to determine whether the results should be accepted [32, 33].

1st diagnostic—the physical meaning of the parametric estimates: the physical meaning of the protonation constants, associated molar absorptivities, and stoichiometric indices is examined: β_{qr} and ε_{qr} should be neither too high nor too low, and ε_{qr} should not be negative. The empirical rule that is often used is that a parameter is considered to be significant when the relation $s(\beta_j) \times F_\sigma < \beta_j$ is met and where F_σ is equal to 3 at a 99.9% statistical probability level.

2nd diagnostic—the physical meaning of the species concentrations: there are some physical constraints which are generally applied to concentrations of species and their molar absorptivities: concentrations and molar absorptivities must be positive numbers. Moreover, the calculated distribution of the free concentration of the basic components and the variously protonated species of the chemical model should show realistic molarities, that is, down to about 10^{-8} M.

3rd diagnostic—parametric correlation coefficients: partial correlation coefficients, r_{ij}, indicate the interdependence of two parameters, that is, the stability constants β_i and β_j, when others are fixed in value.

4th diagnostic—goodness-of-fit test: to identify the "best" or true chemical model when several are possible or

proposed, and to establish whether the chemical model represents the data adequately, the residuals e should be carefully analyzed. The goodness-of-fit achieved is easily seen by examination of the differences between the experimental and calculated values of absorbance, $e_i = A_{\text{exp},i,j} - A_{\text{calc},i,j}$. One of the most important statistics calculated is *the standard deviation of the absorbance*, $s(A)$, calculated at the termination of the minimization process as $s(A) = \sqrt{U\min/df}$, where $U_{\min}$ stands for the residuals-square-sum function in minimum and df is the degree of freedom. This is usually compared with the standard deviation of absorbance calculated by the INDICES program [30] $s_k(A)$ and the instrumental error of the spectrophotometer used $s_{\text{inst}}(A)$ and if it is valid that $s(A) \leq s_k(A)$, or $s(A) \leq s_{\text{inst}}(A)$, then the fit is considered to be statistically acceptable. Some realistic empirical limits are employed: for example, when $s_{\text{inst}}(A) \leq s(A) \leq 0.002$, the goodness-of-fit is still taken as acceptable, while $s(A) > 0.005$ indicates that a good fit has not been obtained. Alternatively, the statistical measures of residuals e can be calculated to examine the following criteria: the *residual bias* $\bar{e}$ should be a value close to zero; the *residual standard deviation* $s(e)$ being equal to the absorbance standard deviation $s(A)$ should be close to the *instrumental standard deviation* $s_{\text{inst}}(A)$*;* the *residual skewness* $g_1(e)$ should be close to zero for a symmetric distribution of residuals; the *residual kurtosis* $g_2(e)$ should be close to 3 for a Gaussian distribution of residuals.

The details of the computer data treatment are given in the *Supporting Information.*

3. Experimental

3.1. Chemicals and Solutions. Silybin was generously donated by IVAX-CR, Czech Republic. A silymarin extract of pharmacopoeial quality (DAB IX) was prepared from *Silybum marianum,* var. Silyb (L.) Gaertn. (Asteraceae). Individual component was isolated and purified by ethylacetate extraction, crystallization, and chromatography. The final purity achieved was Silybin: IVAX-CR company standard AB023, Batch no. 190194, 97.5% (HPLC).

Perchloric acid, 1 M, was prepared from conc. $HClO_4$ (p.a., Lachema Brno) using redistilled water and standardized against HgO and NaI with a reproducibility of less than 0.20%. *Sodium hydroxide*, 1 M, was prepared from pellets (p. a., Aldrich Chemical Company) with carbon dioxide-free redistilled water and standardized against a solution of potassium hydrogenphthalate using the Gran Method with a reproducibility of 0.1%. The preparation of other solutions from analytical reagent-grade chemicals has been described previously [31].

3.2. Apparatus and pH-Spectrophotometric Titration Procedure. The apparatus used and the pH-spectrophotometric titration procedure have been described previously [31].

3.3. Software Used. Computations relating to the determination of the dissociation constants were performed by regression analysis of the UV/VIS spectra using the SQUAD(84) [8] and SPECFIT/32 [39] programs. Most of graphs were plotted using ORIGIN 7.5 [41] and S-Plus [40]. Qualitative interpretation of the spectra with the use of the INDICES program [30] aims to evaluate the quality of the dataset and remove spurious data, and to estimate the minimum number of *factors*, that is, contributing aqueous species, which are necessary to describe the experimental data, and determine the number of dominant species present in the equilibrium mixture.

3.4. Supporting Information Available. Complete experimental and computational procedures, input data specimen, and corresponding output in numerical and graphical forms for the programs INDICES, SQUAD(84), and SPECFIT/32 are available free of charge online at http://meloun.upce.cz/ and in the block *DATA*.

4. Results and Discussion

Recently, silybin was studied in our laboratory [24] and this drug was therefore taken as an example of a drug acid for the demonstration of the reliability of a protonation model and of protonation constants estimation, because of two problems: the first problem in the evaluation of the protonation equilibria of the drug silybin is the extensively overlapping equilibria, as the difference of two consecutive dissociation constants is less than 3 (here about 1.2). Such close equilibria are always difficult to evaluate and therefore the user should carefully prove the reliability of each dissociation constant estimation. A distribution diagram of the relative concentrations of all of the variously protonated species demonstrates the overlapping protonation equilibria for three close consecutive dissociation constants. The second problem concerns the small differences between the molar absorptivities in the variously protonated species within a spectrum. It may happen that nonlinear regression fails when small differences of absorbance are of the same magnitude as instrumental noise, $s_{\text{inst}}(A)$.

The deprotonated silybin LH_4 form exhibits several isosbestic points in spectra. Each isosbestic point usually indicates a simple two-state equilibrium. pH-spectrophotometric titration enables absorbance-response data (Figures 1(a) and 2(a)) to be obtained for analysis by nonlinear regression, and the reliability of parameter estimates (pK's and ε's) can be evaluated on the basis of the goodness-of-fit test of residuals (Figures 1(b) and 2(d)). The A-pH curves at 243, 250, 288, and 327 nm (Figure 2(c)) show that a dissociation constant can be indicated. However, as the changes in spectra are small with deprotonation, the variously protonated species exhibit similar and overlapping absorption bands. The small shift of a band maximum to lower wavelengths in the spectra set is indicated, Figure 2(a). The adjustment of pH value from 6 to 12 causes the absorbance to change, so that the monitoring of the variously protonated components of the protonation equilibrium is rather uncertain. As the changes in spectra are small, very precise measurement of absorbance is required for a reliable

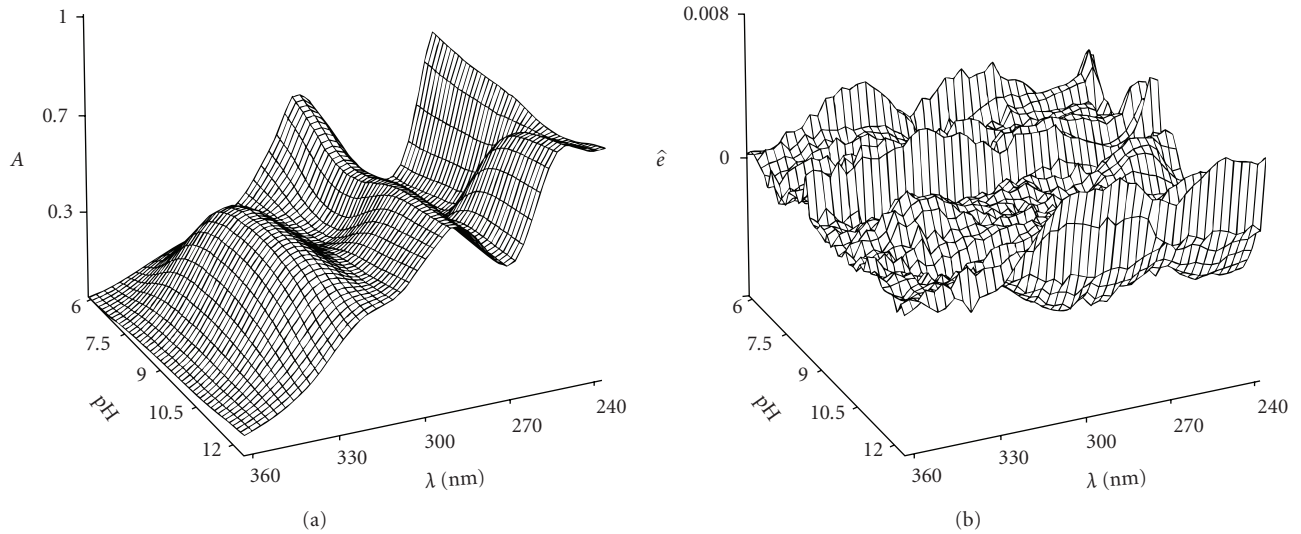

FIGURE 1: The 3D-absorbance-response-surface representing (a) the measured multiwavelength absorption spectral dependence on pH at 25°C and $I = 0.03$ for silybin, and (b) the 3D-residuals map after nonlinear regression performed by SQUAD (and by SPECFIT in brackets) for silybin, which exhibits 780 residuals of the *residual bias* $\bar{e} = 3.50E - 17$ (2.48*E*-08) being close to zero and the *residual standard deviation* $s(e) = 1.01$ mAU (0.88 mAU) being close to the *instrumental standard deviation* $s_{\text{inst}}(A) = 0.30$ mAU (0.30 mAU)*;* the *residual skewness* $g_1(e) = 0.29$ (-0.35) is close to zero and indicates a symmetric distribution of residuals; the *residual kurtosis* $g_2(e) = 2.43$ (3.21) is close to 3 and indicates a Gaussian distribution of residuals. The accuracy test of the bias proves that the bias is not significantly different from zero, (S-Plus).

estimation of the overlapping deprotonation equilibrium studied.

In the first step of the regression spectra analysis, the number of light-absorbing species is estimated by the factor analysis INDICES algorithm (Figure 2(b)). The position of the break point on the $s_k(A) = f(k)$ curve in the factor analysis scree plot is calculated and gives $k^* = 5$ with corresponding coordinate $\log s_k^*(A) = -3.5$, that is, $s_k^*(A) = 0.30$ mAU, which also represents the actual instrumental error $s_{\text{inst}}(A)$ of the spectrophotometer used. Due to the large variations in the indicator values, these latter are plotted on a logarithmic scale. All other selected methods of the modified factor analysis in the INDICES algorithm estimate the five light-absorbing components L, LH, LH_2, LH_3, and LH_4 of the protonation equilibrium. The number of light-absorbing species p can be predicted from the index function values by finding the point $p = k$, where the slope of the index function $PC(k) = f(k)$ changes, or by comparing $PC(k)$ values to the instrumental error $s_{\text{inst}}(A) \approx 0.30$ mAU. This is the common criterion for determining p. Very low values of $s_{\text{inst}}(A)$ prove that reliable spectrophotometer and experimental techniques were used.

The dissociation constant and two molar absorptivities of silybin calculated for 39 wavelengths of 20 spectra constitute $(5 \times 39) + 4 = 199$ unknown regression parameters which are estimated and refined by SQUAD(84) or SPECFIT32 in the first run. The reliability of the parameter estimates may be tested with the use of following diagnostics.

The first diagnostic value indicates whether all of the parametric estimates β_{qr} and ε_{qr} have physical meaning and reach realistic values, for silybin at 25°C and ionic strength $I = 0.03$ with SQUAD (and SPECFIT); that is, $pK_{a1} = 6.898(0.022)$ and 6.897(0.002), $pK_{a2} = 8.666(0.021)$, and 8.667(0.012); $pK_{a3} = 9.611(0.010)$ and 9.611(0.004); $pK_{a4} = 11.501(0.008)$ and 11.501(0.007). As the standard deviations $s(\log \beta_{qr})$ of parameters $\log \beta_{qr}$ and $s(\varepsilon_{qr})$ of parameters ε_{qr} are significantly smaller than their corresponding parameter estimates, all of the variously protonated species are statistically significant at a significance level of $\alpha = 0.05$. The physical meaning of the dissociation constant, molar absorptivities, and stoichiometric indices is examined. The absolute values of $s(\beta_j)$, $s(\varepsilon_j)$ give information about the last U-contour of the hyperparaboloid in the neighbourhood of the pit, $U_{\min}$. For well-conditioned parameters, the last U-contour is a regular ellipsoid, and the standard deviations are reasonably low. High s values are found with ill-conditioned parameters and a "saucer"-shaped pit. The relation $s(\beta_j) \times F_\sigma < \beta_j$ should be met, where F_σ is equal to 3 at a 99.9% statistical probability level. The set of standard deviations of ε_{pqr} for the various wavelengths, $s(\varepsilon_{qr}) = f(\lambda)$, should have a Gaussian distribution; otherwise, erroneous estimates of ε_{qr} are obtained. Figure 2(e) shows the estimated molar absorptivities of all of the variously protonated species ε_{L}, ε_{LH}, of silybin in dependence on wavelength.

The second diagnostic tests whether all of the calculated free concentrations of the five variously protonated species on the distribution diagram of the relative concentration expressed as a percentage have physical meaning, which proved to be the case (Figure 2(f)). Numerical values of protonation constants and molar absorption coefficients may not seem very interesting, and a graphical presentation is more illustrative; graphs are also efficient diagnostic tools in the search for the most probable chemical model. A distribution diagram makes it easier to judge the contributions of

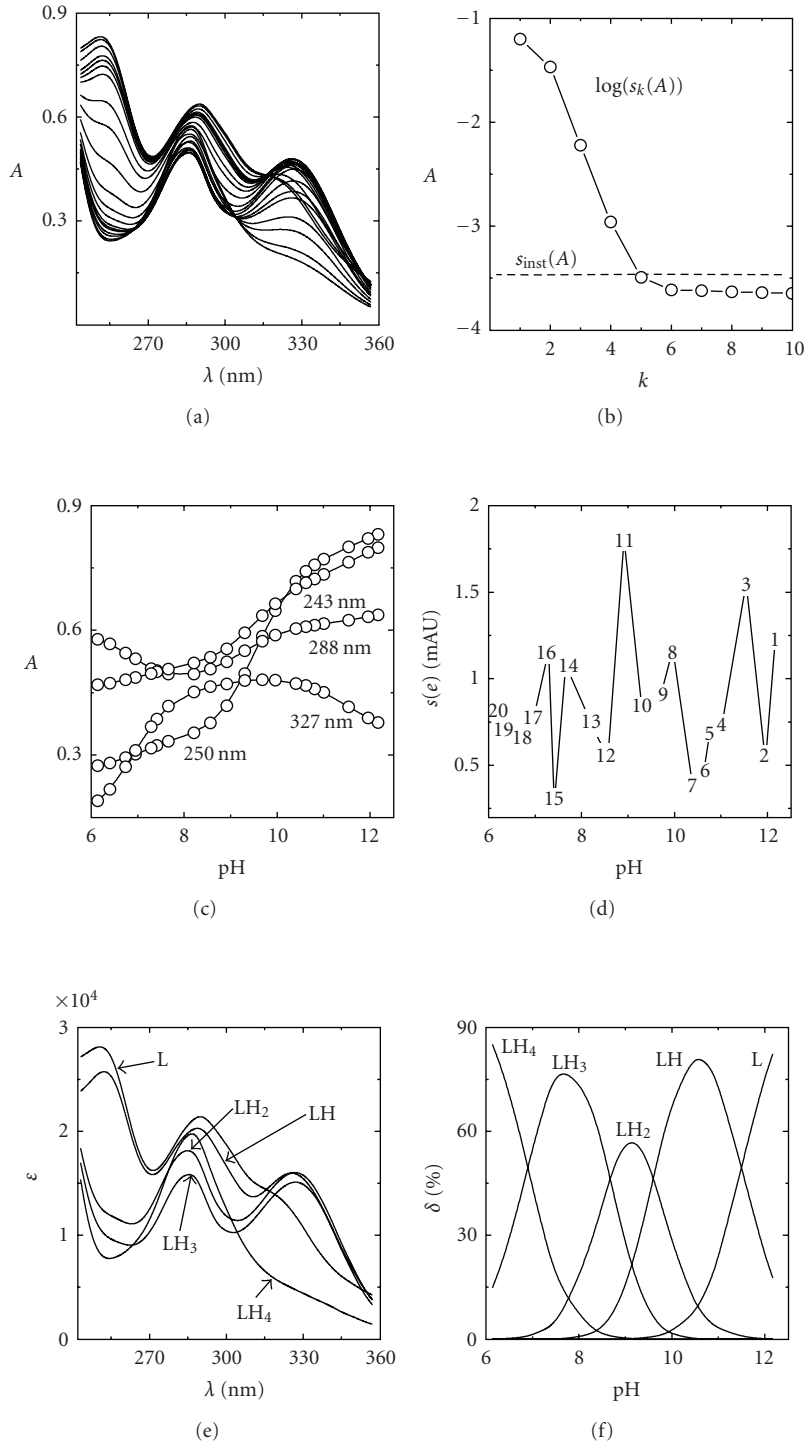

Figure 2: The nonlinear regression analysis of the protonation equilibria model and factor analysis of silybin: (a) Absorption spectral dependence on pH at 25°C, (b) Cattel's scree plot of the Wernimont-Kankare procedure for the determination of the number of light-absorbing species in the mixture $k^* = 5$ leads to the actual instrumental error of the spectrophotometer used $s_{\text{inst}}(A) = 0.30$ mAU (INDICES in S-Plus), (c) The absorbance versus pH curves for 243 nm, 250 nm, 288, and 327 nm in dependence on pH at 25°C, (d) Detecting influential outlying spectra with the use of the goodness-of-fit test and the plot of the residual standard deviation $s(e)$ versus pH for 20 spectra in dependence on pH at 25°C and $I = 0.03$, (e) Pure spectra profiles of molar absorptivities versus wavelengths for variously protonated species L, LH, LH_2, LH_3, and LH_4, (f) Distribution diagram of the relative concentrations of both variously protonated species L, LH, LH_2, LH_3, and LH_4, of silybin as a function of pH at 25°C is calculated from estimates by SQUAD and SPECFIT: $pK_{a1} = 6.898(0.022)$ and $6.897(0.002)$; $pK_{a2} = 8.666(0.021)$ and $8.667(0.012)$; $pK_{a3} = 9.611(0.010)$ and $9.611(0.004)$; $pK_{a4} = 11.501(0.008)$ and $11.501(0.007)$, (SQUAD and SPECFIT, ORIGIN).

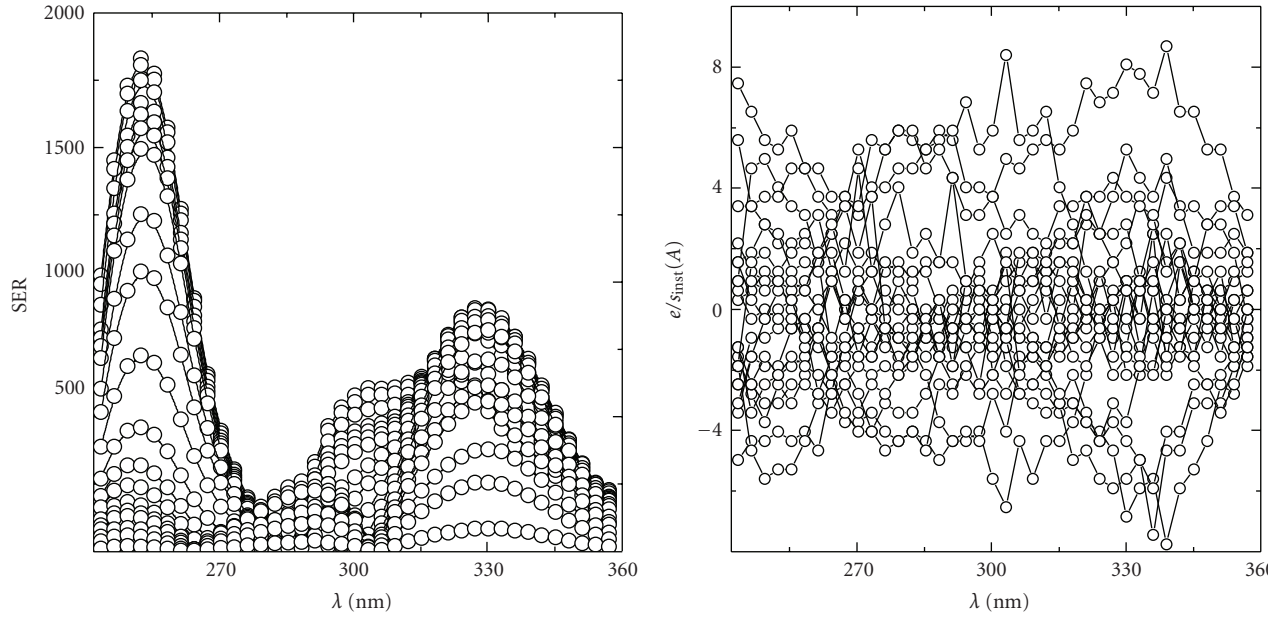

FIGURE 3: The plot of absorbance changes in the spectrum (a) means that the value of the absorbance difference for the jth-wavelength of the ith-spectrum $\Delta_{ij} = A_{ij} - A_{i,\text{acid}}$ is divided by the instrumental standard deviation $s_{\text{inst}}(A)$ and the resulting ratios $SER = \Delta/s_{\text{inst}}(A)$ are plotted as a function of wavelength λ for all absorbance matrix elements where $A_{i,\text{acid}}$ means the limiting spectrum of the acid form of the drug measured. This ratio is compared to the limiting SER value for silybin to test if the absorbance changes Δ_{ij} are significantly larger than the instrumental noise $s_{\text{inst}}(A)$. The plot of the ratio $e/s_{\text{inst}}(A)$, that is, the ratio of residuals divided by the instrumental standard deviation $s_{\text{inst}}(A)$ as a function of wavelength λ for all residual matrix elements (b) for silybin tests if the residuals **e** are of the same magnitude as the instrumental noise $s_{\text{inst}}(A)$.

individual species to the total concentration quickly. Since the molar absorptivities will generally be in the range of 10^3–$10^5\,\text{l}\cdot\text{mol}^{-1}\cdot\text{cm}^{-1}$, species present at less than *ca.* 0.1% relative concentration will affect the absorbance significantly only if their ε is extremely high. The diagram shows that the protonation equilibria of five species L, LH, LH_2, LH_3, and LH_4 are strongly overlapping and shows the estimated molar absorptivities of all the variously protonated species ε_{L}, ε_{LH}, $\varepsilon_{\text{LH}_2}$, $\varepsilon_{\text{LH}_3}$ and $\varepsilon_{\text{LH}_4}$ of silybin in dependence on wavelength. Some spectra overlap considerably here, and such cases may cause some resolution difficulties in a nonlinear regression approach.

The third diagnostic concerning the matrix of correlation coefficients around 0.9 proves that there is an absence of an interdependence of any pair of protonation constants of silybin except for species LH_1 versus LH_2, and LH_3 versus LH_4. The significant correlation of these two pairs may be explained by the too closely overlapping protonation constants, which concern overlapping equilibria.

The fourth diagnostic concerns the goodness-of-fit (Figure 2(d)). The goodness-of-fit achieved is easily seen by examination of the differences between the experimental and calculated values of absorbance, $e_i = A_{\text{exp},i,j} - A_{\text{calc},i,j}$. Examination of the spectra and of the graph of the predicted absorbance response-surface through all the experimental points should reveal whether the results calculated are consistent and whether any gross experimental errors have been made in the measurement of the spectra. One of the most important statistics calculated is the standard deviation of absorbance, $s(A)$, calculated from a set of refined parameters at the termination of the minimization process. It is usually compared with the standard deviation of absorbance calculated by the INDICES program [30], $s_k(A)$, and if $s(A) \leq s_k(A)$, or $s(A) \leq s_{\text{inst}}(A)$, the instrumental error of the spectrophotometer used, the fit is considered to be statistically acceptable. This proves that the $s_5(A)$ value is equal to 0.30 mAU. Numerical values of statistical measures of the residuals now indicate very good fitness, and also prove that the minimum of the elliptic hyperparaboloid U was reached: the residual mean $\bar{e} = 3.50 \times 10^{-17}$ (SPECFIT gives -2.48×10^{-8}) proves that there is no bias or systematic error in the spectra fitting. The mean residual $|\bar{e}| = 0.67$ (SPECFIT gives 0.68) mAU and the residual standard deviation $s(e) = 1.01$ (SPECFIT gives 0.88) mAU have sufficiently low values. The standard deviation of absorbance $s(A)$ after termination of the minimization process is always equal or lower than 1 mAU, and the proposal of a good chemical model and reliable parameter estimates is thus proven. The skewness $g_1(e) = 0.29$ (SPECFIT gives 0.35) is quite close to zero and proves a symmetric distribution of the residuals set while the kurtosis $g_2(e) = 2.43$ (SPECFIT gives 3.21) is reasonably close to 3 supporting a Gaussian distribution.

Although this statistical analysis of residuals gives the most rigorous test of the degree-of-fit, realistic empirical limits must be used. The statistical measures of all residuals e prove that the minimum of the elliptic hyperparaboloid U

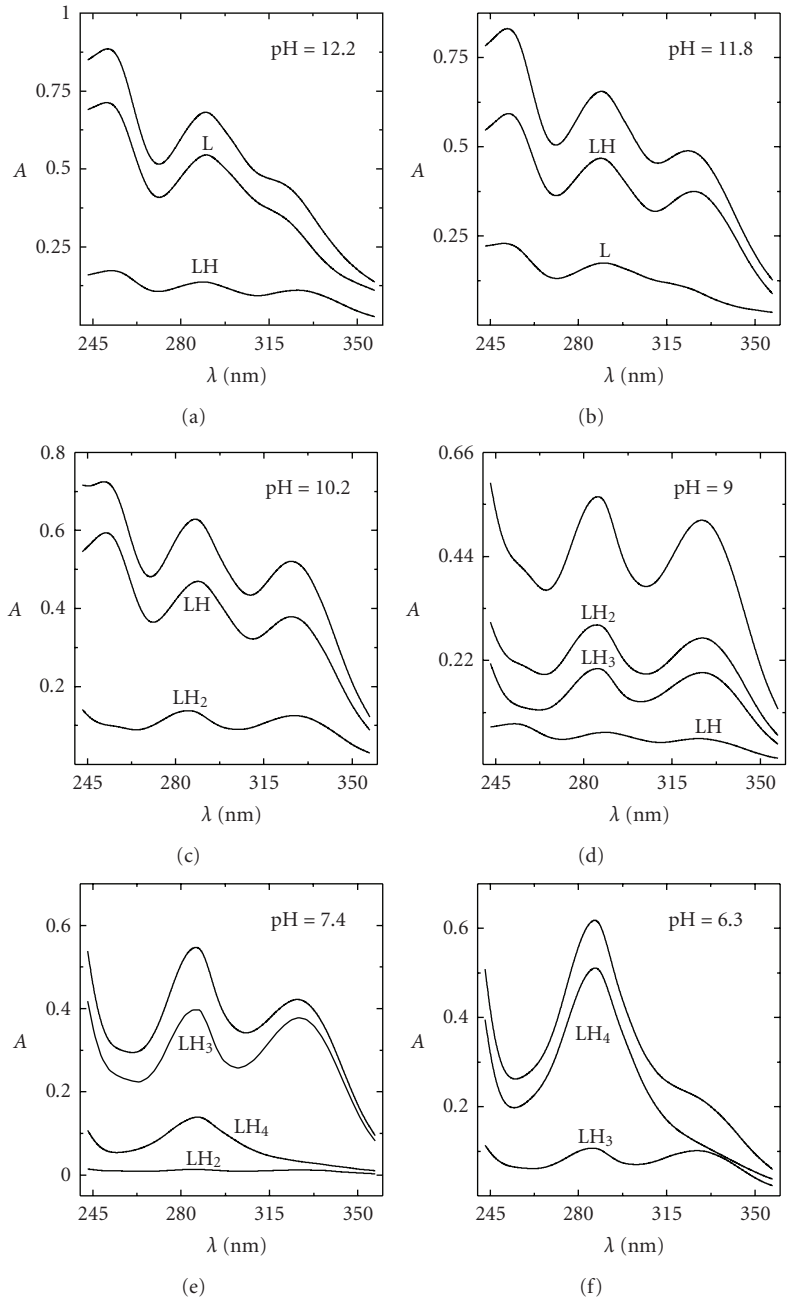

Figure 4: Deconvolution of the experimental absorption spectrum of silybin for 39 wavelengths into spectra of the individual variously protonated species L, LH, LH_2, LH_3, LH_4 in solution (above) and the statistical analysis of the residuals (below) of each particular absorption spectrum for a selected value of pH equal to: (a) 12.2, (b) 11.8, (c) 10.2, (d) 9.0, (e) 7.4, and (f) 6.3. The charges of the species are omitted for the sake of simplicity (SQUAD, ORIGIN).

is reached: the residual standard deviation $s(e)$ always has sufficiently low values of lower than 1 mAU. The criteria of resolution used for the hypotheses were (1) a failure of the minimization process in a divergency or a cyclisation, (2) an examination of the physical meaning of the estimated parameters to see if they were both realistic and positive, and (3) the residuals should be randomly distributed about the predicted regression spectrum, and systematic departures from randomness were taken to indicate that either the chemical model or the parameter estimates were unsatisfactory.

To express changes of absorbance in the spectral set, the absorbance differences for the jth wavelength of the ith spectrum $\Delta_i = A_{ij} - A_{i,\text{acid}}$ were calculated so that from the absorbance value of the spectrum measured at the actual pH, the absorbance value of the acidic form was subtracted. The absorbance difference Δ_i was then divided with an actual instrumental standard deviation $s_{\text{inst}}(A)$ of

spectrophotometer used, the resulting value representing the *signal-to-error* value (*SER*). The left part of Figure 3 brings the graph of the SER in dependence on wavelength in the measured range for silybin. When SER is larger than 10, a factor analysis is able to predict the correct number of light-absorbing components in equilibrium mixture. To prove that nonlinear regression is able to analyze such a data set, the residuals set was compared with the instrumental noise $s_{\text{inst}}(A)$. If the ratio $e/s_{\text{inst}}(A)$ is of similar magnitude, that is, nearly equal to one, it means that sufficient curve fitting by nonlinear regression of spectra set was achieved, and that the minimization process found the minimum of residual-square-sum function $U_{\min}$. The right part of Figure 3 shows a comparison of the ratio $e/s_{\text{inst}}(A)$ in dependence on wavelength for all three drugs measured. From the figure, it is obvious that most of the residuals are of the same magnitude as the instrumental noise and therefore the regression found reliable estimations of the unknown parameters.

As additivity of absorbance should be proven, the sum of all the absorbances of the variously protonated species at a given wavelength should be equal to the experimental absorbance. Resolution of each experimental spectrum into the spectra for the individual variously protonated species shows whether the experimental design, that is, the pH range was efficient. If for a particular pH value the spectrum consists of just a single component, further spectra for that pH or similar value would be redundant even though they should improve only the precision. In concentrations or pH ranges where more components contribute significantly to the spectrum, several spectra should be measured. Figure 4 proves that deconvolution of the spectra into absorbance increments for the individual species helps in planning future efficient experimentation, and that the selected pH values have been chosen efficiently. Spectrum deconvolution seems quite a useful tool in the proposal of a strategy for efficient experimentation. Such a spectrum provides sufficient information for a regression analysis which monitors at least two species in equilibrium, where none of them is a minor species. A minor species has a relative concentration in a distribution diagram of less than 5% of the total concentration of the basic component c_{L}. When, on the other hand, only one species is prevalent in solution, the spectrum yields quite poor information into the regression analysis, and the parameter estimate is rather unsure and definitely not reliable enough.

5. Conclusions

When drugs are very poorly soluble, then pH-spectrophotometric titration may be used with the non-linear regression of the absorbance-response-surface data instead of a potentiometric determination of dissociation constants. Regression diagnostics represent procedures for examination of the *regression triplet* (*data, model, method*) for the identification of (a) the data quality for a proposed model (b) the model quality for a given set of data, and (c) the fulfillment of all least-squares assumptions. The reliability of the dissociation constants of the drug silybin may be proven with goodness-of-fit tests of the absorption spectra measured at various pH. The dissociation constant pK_a was estimated by nonlinear regression of $\{pK_a, I\}$ data at 25°C with SQUAD (and SPECFIT), that is, pK_{a1} = 6.898(0.022) and 6.897(0.002); pK_{a2} = 8.666(0.021) and 8.667(0.012); pK_{a3} = 9.611(0.010) and 9.611(0.004); pK_{a4} = 11.501(0.008) and 11.501(0.007) at $I = 0.03$. Goodness-of-fit tests for various regression diagnostics enabled the reliability of the parameter estimates to be determined. Most indices always predict the correct number of components, and even the presence of a minor one when the *signal-to-error ratio (SER)* is higher than 10. The Wernimont-Kankare procedure in the factor analysis program INDICES performs reliable determination of the instrumental standard deviation of spectrophotometer used $s_{\text{inst}}(A)$ and correctly predicts the number of light-absorbing components present. This procedure also solves ill-defined problems with severe collinearity in the spectra, very small changes in spectra, and with overlapping equilibria.

Acknowledgments

The financial support of the IGA Grant Agency (Grant no. NR9055-4/2006 and NR9831-4/2008) and of the Czech Ministry of Education (Grants no. MSM0021627502 and MSM0021627501) is gratefully acknowledged.

References

[1] F. R. Hartley, C. Burgess, and R. M. Alcock, *Solution Equilibria*, Ellis Horwood, Chichester, UK, 1980.

[2] M. Meloun, J. Havel, and J. Högfeldt, *Computation of Solution Equilibria*, Ellis Horwood, Chichester, UK, 1988.

[3] M. Meloun and J. Havel, *Computation of Solution Equilibria, Part 1. Spectrophotometry*, vol. 25, Folia Facultatis Scientarum Naturalium Universitatis Prkynianae Brunensis (Chemia), Brno, Czech Republic, 1984.

[4] M. Meloun and J. Havel, *Computation of Solution Equilibria, Part 2. Potentiometry*, vol. 26, Folia Facultatis Scientarum Naturalium Universitatis Prkynianae Brunensis (Chemia), Brno, Czech Republic, 1985.

[5] L. G. Sillén and B. Warnqvist, "Equilibrium constants and model testing from spectrophotometric data, using LETAGROP," *Acta Chemica Scandinavica*, vol. 22, p. 3032, 1968.

[6] D. J. Leggett, Ed., *Computational Methods for the Determination of Formation Constants*, Plenum Press, New York, NY, USA, 1985.

[7] J. Havel and M. Meloun, "General computer programs for the determination of formation constants from various types of data," in *Computational Methods for the Determination of Formation Constants*, D. J. Leggett, Ed., pp. 221–289, Plenum Press, New York, NY, USA, 1985.

[8] M. Meloun, M. Javůrek, and J. Havel, "Multiparametric curve fitting-X. A structural classification of programs for analysing multicomponent spectra and their use in equilibrium-model determination," *Talanta*, vol. 33, no. 6, pp. 513–524, 1986.

[9] D. J. Leggett and W. A. E. McBryde, "General computer program for the computation of stability constants from absorbance data," *Analytical Chemistry*, vol. 47, no. 7, pp. 1065–1070, 1975.

[10] D. J. Leggett, "Numerical analysis of multicomponent spectra," *Analytical Chemistry*, vol. 49, no. 2, pp. 276–281, 1977.

[11] D. J. Leggett, S. L. Kelly, L. R. Shiue, Y. T. Wu, D. Chang, and K. M. Kadish, "A computational approach to the spectrophotometric determination of stability constants-II. Application to metalloporphyrin-axial ligand interactions in non-aqueous solvents," *Talanta*, vol. 30, no. 8, pp. 579–586, 1983.

[12] J. J. Kankare, "Computation of equilibrium constants for multicomponent systems from spectrophotometric data," *Analytical Chemistry*, vol. 42, no. 12, pp. 1322–1326, 1970.

[13] P. Gans, A. Sabatini, and A. Vacca, "Investigation of equilibria in solution. Determination of equilibrium constants with the HYPERQUAD suite of programs," *Talanta*, vol. 43, no. 10, pp. 1739–1753, 1996.

[14] H. Gampp, M. Maeder, C. J. Meyer, and A. D. Zuberbühler, "Calculation of equilibrium constants from multiwavelength spectroscopic data—I: mathematical considerations," *Talanta*, vol. 32, no. 2, pp. 95–101, 1985.

[15] H. Gampp, M. Maeder, C. J. Meyer, and A. D. Zuberbühler, "Calculation of equilibrium constants from multiwavelength spectroscopic data—II: specfit: two user-friendly programs in basic and standard fortran 77," *Talanta*, vol. 32, no. 4, pp. 251–264, 1985.

[16] H. Gampp, M. Maeder, C. J. Meyer, and A. D. Zuberbühler, "Calculation of equilibrium constants from multiwavelength spectroscopic data—III: model-free analysis of spectrophotometric and ESR titrations," *Talanta*, vol. 32, no. 12, pp. 1133–1139, 1985.

[17] H. Gampp, M. Maeder, C. J. Meyer, and A. D. Zuberbühler, "Calculation of equilibrium constants from multiwavelength spectroscopic data—IV: model-free least-squares refinement by use of evolving factor analysis," *Talanta*, vol. 33, no. 12, pp. 943–951, 1986.

[18] K. Y. Tam and K. Takács-Novák, "Multi-wavelength spectrophotometric determination of acid dissociation constants: a validation study," *Analytica Chimica Acta*, vol. 434, no. 1, pp. 157–167, 2001.

[19] Z.-P. Chen, Y.-Z. Liang, J.-H. Jiang, Y. Li, J.-Y. Qian, and R.-Q. Yu, "Determination of the number of components in mixtures using a new approach incorporating chemical information," *Journal of Chemometrics*, vol. 13, no. 1, pp. 15–30, 1999.

[20] Z.-P. Chen, J.-H. Jiang, Y. Li, H.-L. Shen, Y.-Z. Liag, and R.-Q. Yu, "Smoothed window factor analysis," *Analytica Chimica Acta*, vol. 381, no. 2-3, pp. 233–246, 1999.

[21] M. Meloun, J. Čapek, P. Mikšík, and R. G. Brereton, "Critical comparison of methods predicting the number of components in spectroscopic data," *Analytica Chimica Acta*, vol. 423, no. 1, pp. 51–68, 2000.

[22] M. Meloun and M. Pluhařová, "Thermodynamic dissociation constants of codeine, ethylmorphine and homatropine by regression analysis of potentiometric titration data," *Analytica Chimica Acta*, vol. 416, no. 1, pp. 55–68, 2000.

[23] M. Meloun and P. Černohorský, "Thermodynamic dissociation constants of isocaine, physostigmine and pilocarpine by regression analysis of potentiometric data," *Talanta*, vol. 52, no. 5, pp. 931–945, 2000.

[24] M. Meloun, D. Burkoňová, T. Syrový, and A. Vrána, "Thermodynamic dissociation constants of silychristin, silybin, silydianin and mycophenolate by the regression analysis of spectrophotometric data," *Analytica Chimica Acta*, vol. 486, no. 1, pp. 125–141, 2003.

[25] M. Meloun, T. Syrový, and A. Vrána, "Determination of the number of light-absorbing species in the protonation equilibra of selected drugs," *Analytica Chimica Acta*, vol. 489, no. 2, pp. 137–151, 2003.

[26] M. Meloun, T. Syrový, and A. Vrána, "The thermodynamic dissociation constants of ambroxol, antazoline, naphazoline, oxymetazoline and ranitidine by the regression analysis of spectrophotometric data," *Talanta*, vol. 62, no. 3, pp. 511–522, 2004.

[27] M. Meloun, T. Syrový, and A. Vrána, "The thermodynamic dissociation constants of losartan, paracetamol, phenylephrine and quinine by the regression analysis of spectrophotometric data," *Analytica Chimica Acta*, vol. 533, no. 1, pp. 97–110, 2005.

[28] M. Meloun, T. Syrový, and A. Vrána, "The thermodynamic dissociation constants of haemanthamine, lisuride, metergoline and nicergoline by the regression analysis of spectrophotometric data," *Analytica Chimica Acta*, vol. 543, no. 1-2, pp. 254–266, 2005.

[29] M. Meloun, M. Javůrek, and J. Militký, "Computer estimation of dissociation constants. Part V. Regression analysis of extended Debye-Hückel law," *Mikrochimica Acta*, vol. 109, no. 5-6, pp. 221–231, 1992.

[30] INDICES, http://meloun.upce.cz/, and the block Algorithms.

[31] M. Meloun, S. Bordovská, T. Syrový, and A. Vrána, "Tutorial on chemical model building and testing to spectroscopic data with least-squares regression," *Analytica Chimica Acta*, vol. 580, no. 1, pp. 107–121, 2006.

[32] M. Meloun, J. Militký, and M. Forina, *Chemometrics for Analytical Chemistry, Vol. 1. PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester, UK, 1992.

[33] M. Meloun, J. Militký, and M. Forina, *Chemometrics for Analytical Chemistry, Vol. 2. PC-Aided Regression and Related Methods*, Ellis Horwood, Chichester, UK, 1994.

[34] G. L. Amidon, H. Lennernas, and V. P. Shah, "A theoretical basis for a biopharmaceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability," *Pharmaceutical Research*, vol. 12, no. 3, pp. 413–420, 1995.

[35] US Department of Health and Human Service, FDA Center for Drug Evaluation and Research (CDER) Draft Guidance: Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate Release Solid Dosage Forms Containing Certain Active Moieties/Active Ingredients Based on a Biopharmaceutic Drug Classification System. CDERGUID-2062 DFT.WPD, January 1999.

[36] *The Merck Index, An Encyclopedia of Chemicals, Drugs and Biologicals*, Merck & Co., Whitehouse Station, NJ, USA, 13th edition, 2001.

[37] H. Wagner, P. Diesel, and M. Seitz, "Chemistry and analysis of silymarin from Silybum marianum Gaertn," *Arzneimittel-Forschung/Drug Research*, vol. 24, no. 4, pp. 466–471, 1974.

[38] V. Šimánek, V. Křen, J. Ulrichová, J. Vičar, and L. Cvak, "Silymarin: what is in the name ...? An appeal for a change of editorial policy," *Journal of Hepatology*, vol. 32, no. 2, pp. 442–444, 2000.

[39] SPECFIT/32: Spectrum Software Associates, 2004, 197M Boston Post Road West, Marlborough, Mass, USA, http://www.bio-logic.info/rapid-kinetics/index.html.

[40] S-PLUS, http://www.insightful.com/products/splus/.

[41] ORIGIN: OriginLab Corporation, One Roundhouse Plaza, Suite 303, Northampton, Mass, USA.