Available online at www.sciencedirect.com

ScienceDirect

Analytica Chimica Acta 580 (2006) 107–121

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Tutorial on a chemical model building by least-squares non-linear regression of multiwavelength spectrophotometric pH-titration data

Milan Meloun[a,*], Sylva Bordovská[a], Tomáš Syrový[a], Aleš Vrána[b]

[a] *Department of Analytical Chemistry, University of Pardubice, 532 10 Pardubice, Czech Republic*
[b] *IVAX Pharmaceuticals, s.r.o., 747 70 Opava, Czech Republic*

Received 27 April 2006; received in revised form 4 July 2006; accepted 20 July 2006
Available online 26 July 2006

Abstract

Although the modern instrumentation enables for the increased amount of data to be delivered in shorter time, computer-assisted spectra analysis is limited by the intelligence and by the programmed logic tool applications. Proposed tutorial covers all the main steps of the data processing which involve the chemical model building, from calculating the concentration profiles and, using spectra regression, fitting the protonation constants of the chemical model to multiwavelength and multivariate data measured. Suggested diagnostics are examined to see whether the chemical model hypothesis can be accepted, as an incorrect model with false stoichiometric indices may lead to slow convergence, cyclization or divergence of the regression process minimization. Diagnostics concern the physical meaning of unknown parameters $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$, physical sense of associated species concentrations, parametric correlation coefficients, goodness-of-fit tests, error analyses and spectra deconvolution, and the correct number of light-absorbing species determination. All of the benefits of spectrophotometric data analysis are demonstrated on the protonation constants of the ionizable anticancer drug 7-ethyl-10-hydroxycamptothecine, using data double checked with the SQUAD(84) and SPECFIT/32 regression programs and with factor analysis of the INDICES program. The experimental determination of protonation constants with their computational prediction based on a knowledge of chemical structures of the drug was through the combined MARVIN and PALLAS programs. If the proposed model adequately represents the data, the residuals should form a random pattern with a normal distribution $N(0, s^2)$, with the residual mean equal to zero, and the standard deviation of residuals being near to experimental noise. Examination of residual plots may be assisted by a graphical analysis of residuals, and systematic departures from randomness indicate that the model and parameter estimates are not satisfactory.

Keywords: Spectrophotometric titration; Dissociation constant; Protonation; 7-Ethyl-10-hydroxycamptothecine; Anticancer drug; SPECFIT; SQUAD; INDICES; PALLAS; MARVIN

1. Introduction

The accurate determination of protonation constants is often required in various chemical, biochemical and pharmaceutical fields as the protonation constants of organic reagents and drugs play a fundamental role in many analytical and medical procedures. If a drug is poorly soluble then, instead of a potentiometric determination of dissociation constants, pH-spectrophotometric titration may be used with the non-linear regression of the absorbance-response-surface data. Spectroscopic methods are, in general, highly sensitive and are as such suitable for studying protonation equilibria solutions [1–26]. If the components involved can be obtained in pure form, or if their spectral responses do not overlap, such analysis is trivial. For many systems, particularly those with similar components, this is not the case, and these have been difficult to analyze. There are several advantages when using multiwavelength data as compared to selecting a single wavelength: (a) Determination of the pure spectra for all species and intermediates of the equilibria mixture. (b) Application of a wide range of model-free analyses, from simple factor analysis to indicate the number of species (e.g., INDICES [12]) to sophisticated analysis based on evolving

[*] Corresponding author. Tel.: +420 466037026; fax: +420 466037068.
E-mail address: milan.meloun@upce.cz (M. Meloun).

0003-2670/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2006.07.043

factor analysis. (c) The need to determine a “good” wavelength to follow the actual equilibrium or reaction is eliminated. (d) The analysis of multiwavelength data is often significantly more robust.

Since the mid-1960s, computers have acquired an ever-greater importance in the evaluation of equilibrium measurement data using the full spectrum in order to determine the stability (protonation) constants $\boldsymbol{\beta}_{qr}$ and molar absorptivities $\boldsymbol{\varepsilon}_{qr}$. The most widespread programs and algorithms for determining the stability constants from absorbance data are LETAGROP-SPEFO [4], SQUAD [5–10], PSEQUAD [5], HYPERQUAD [23], SPECFIT [24–26,34] and more recently DATAN [27–32] and BeerOz [33]. All these computational approaches are based on the initial proposal of stoichiometries of species which define the chemical equilibrium model, are based on mass-action law and mass balance equations, and also involve least-squares curve-fitting procedures. Such programs, for example, SQUAD(84) [7], contain functional blocks for (i) determination of the number of light-absorbing species, (ii) regression estimation of $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$, (iii) a rigorous goodness-of-fit test, (iv) an error analysis, which includes calculation of the confidence interval of the parameters, correlation coefficients and residual-squares-sum function contours and other statistics, and (v) individual spectrum deconvolution. Splitting a program structure into such logical units helps to elucidate its anatomy, and to understand the *modus operandi* of a sophisticated program [7–10,33,34].

In the context of this tutorial, a solution equilibria study is represented by the investigation of protonation of ionizable drug acids and encompasses the identification of the correct number of the various species which absorb light and the determination of the associated protonation constants. As the protonation equilibria of some certain drugs have been studied systematically in our laboratory [13–18,21,22], the authors have tried to complete the tutorial procedure from chemical model building and testing to double checked spectra least-squares regression with two programs, SQUAD(84) and SPECFIT/32, and to determine protonation constants of the poorly soluble anticancer drug 7-ethyl-10-hydroxycamptothecin. This compound (CAS No. 86639-52-3, molecular formula $C_{22}H_{20}N_2O_5$, molecular weight 392.40 and dissociation constants were not yet estimated), used here as an example only, is the pharmacologically active metabolite of the anticancer drug irinotecan, used globally in the first line treatment of advanced metastatic colorectal cancer. Concurrently, the experimental determination of protonation constants was combined with their computer prediction based on a knowledge of chemical structures [50] using the MARVIN [51] and PALLAS [52] programs.

2. Theoretical

2.1. Protonation constants by regression spectra analysis

An acid–base equilibrium of the drug studied is described in terms of the protonation of the Brönstedt base L^{z-1} according to the equation $L^{z-1} + H^+ \rightarrow HL^z$ characterized by the protonation constant

$$K_H = \frac{a_{HL^z}}{a_L^{z-1} a_{H^+}} = \frac{[HL^z]}{[L^{z-1}][H^+]} \frac{y_{HL^z}}{y_{L^{z-1}} y_{H^+}}$$

Dissociation reactions realized at constant ionic strength termed “mixed dissociation constants”, are defined as

$$K_{a,j} = \frac{[H_{j-1}L] a_{H+}}{[H_j L]}$$

These constants are found in experiments where pH values are measured with glass and reference electrodes, standardized with the practical $pH(S) = pa_{H^+}$ activity scale recommended internationally [1,2]; $pH(S) = p(a_{H^+})_c + \log \rho_s$ where index c means molar (and, if relevant, molal m concentrations) and ρ_s is the density of the solvent. For aqueous solutions and temperatures up to 35 °C, this correction is less than 0.003 pH units. The value of $[H_{j-1}L]/[H_jL]$ may be determined by spectrophotometric-pH titration when a determination of the mixed dissociation constant pK_a is performed, cf. ref. [2,3]. If the protonation equilibria between the anion L (the charges are omitted for the sake of simplicity) of a drug and a proton H are considered to form a set of variously protonated species L, LH, LH_2, LH_3, etc., which have the general formula L_qH_r in a particular chemical model and are represented by n_c the number of species $(q, r)_i$, $i = 1, \ldots, n_c$ where index i labels their particular stoicheiometry, then the overall protonation (stability) constant of the protonated species, β_{qr}, may be expressed as

$$\boldsymbol{\beta}_{qr} = \frac{[L_qH_r]}{[L]^q[H]^r} = \frac{c}{l^q h^r}$$

where the free concentration $[L] = l$, $[H] = h$ and $[L_qH_r] = c$. 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 relates the absorbance, $A_{i,j}$, being defined as

$$A_{i,j} = \sum_{n=1}^{n_c} \varepsilon_{j,n} c_n = \sum_{n=1}^{n_c} (\varepsilon_{qr,j} \boldsymbol{\beta}_{qr} l^q h^r)_n$$

where $\boldsymbol{\varepsilon}_{qr,j}$ is the molar absorptivity of the L_qH_r species with the stoichiometric coefficients q, r measured at the jth wavelength. The absorbance $A_{i,j}$ is an element of the absorbance matrix $\boldsymbol{A}$ of size $(n_s \times n_w)$ being measured for n_s solutions with known total concentrations of $n_z = 2$ basic components, c_L and c_H, at n_w wavelengths. The rank of the matrix $\boldsymbol{A}$ is obtained from the equation $\text{rank}(\boldsymbol{A}) = \min[\text{rank}(\boldsymbol{E}), \text{rank}(\boldsymbol{C})] \leq \min(n_w, n_c, n_s)$. Since the rank of $\boldsymbol{A}$ is equal to the rank of $\boldsymbol{E}$ or $\boldsymbol{C}$, whichever is the smaller, and since $\text{rank}(\boldsymbol{E}) \leq n_c$ and $\text{rank}(\boldsymbol{C}) \leq n_c$, then provided n_w and n_s are equal to or greater than n_c, it will only be necessary to determine the rank of matrix $\boldsymbol{A}$, which is equivalent to the number of dominant light-absorbing components [2,3,11,12].

Two families of algorithms for data interpretation can be distinguished, based on the types of constraints applied in the spectra interpretation. The first family, originally implemented in the program SQUAD(75) [5], uses the constraint of a non-linear thermodynamic speciation model. A non-linear least-squares method is used to optimise the absorptivity coefficients and equilibrium constants of formation of the absorbing species. The multicomponent spectra analysing program SQUAD(84) [7] can adjust $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$ for absorption spectra by minimising the residual-square sum function (RSS),

$$\mathrm{RSS} = \sum_{i=1}^{n_\mathrm{s}}\sum_{j=1}^{n_\mathrm{w}}(A_{\mathrm{exp},i,j} - A_{\mathrm{calc},i,j})^2$$

$$= \sum_{i=1}^{n_\mathrm{s}}\sum_{j=1}^{\mathrm{n_w}}\left(A_{\mathrm{exp},i,j} - \sum_{k=1}^{n_\mathrm{c}}{}_{-j,k}c_k\right)^2 = \mathrm{minimum}$$

where $A_{i,j}$ represents the element of the experimental absorbance response-surface of size $n_\mathrm{s} \times n_\mathrm{w}$ 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. Unknown parameters are the best estimates of the protonation constants, $\boldsymbol{\beta}_{qr,i}$, $i = 1, \ldots, n_\mathrm{c}$, which are adjusted by the SQUAD(84) regression algorithm. At the same time, a matrix of molar absorptivities ($\boldsymbol{\varepsilon}_{qr,j}$, $j = 1, \ldots, n_\mathrm{w})_k$, $k = 1, \ldots, n_\mathrm{c}$, as non-negative reals is estimated, based on the current values of protonation constants. For a set of current values of $\boldsymbol{\beta}_{qr,i}$, the free concentrations of ligand l for each solution are calculated, as h is known from pH measurement. Then, the concentrations of all the species in the equilibrium mixture $[\mathrm{L}_q\mathrm{H}_r]_j$, $j = 1, \ldots, n_\mathrm{c}$ are obtained; they represent n_s solutions of the matrix $\boldsymbol{C}$. The calculated standard deviation of absorbance $s(A)$ and the Hamilton R-factor are used as the most important criteria for a fitness test. If, after termination of the minimization process, the condition $s(A) \approx s_\mathrm{inst}(A)$ is met and the R-factor is less than 1%, the hypothesis of the chemical model is taken as the most probable one and is accepted. SQUAD(75) [5] and its successors (e.g., SQUAD(84) [7]), have been used successfully in many complexation or protonation equilibria [35–38] studies.

Another popular program is the commercial SPECFIT/32 [34], based on the algorithm developed by Gampp and coworkers [24–26], and the similar modular program BeerOz (Matlab) [33] for the determination of stability constants from spectrophotometric titration data. The method referred to as "model-free" does not require any assumption as to the chemistry of the system other than the number of active complexes present, not any assumptions as to the nature of absorbing complexes, their stoichiometry or a thermodynamic model. The solution is retrieved using constraints such as non-negativity for concentrations and absorptivities, closure (the sum of the concentrations of some species should be equal to a known quantity) and unimodality (only one maximum in the concentration profiles). The latest version of SPECFIT/32 [34] makes use of a multiwavelength and multivariate spectra treatment and enables a global analysis for equilibrium and kinetic systems with singular value decomposition and non-linear least-squares regression modeling using the Levenberg–Marquardt method [39], and has been used in many papers [23–26,34,40–44]. Factor analysis is used as a powerful tool for the determination of independent components in a given data matrix is used.

2.2. *Procedure for protonation model building and testing*

An experimental and computational scheme for protonation model building and testing, and for the determination of protonation constants of a multicomponent system was proposed by Meloun et al., cf. page 226 in ref. [2] or ref. [7] and is here extended and revised with regard to SPECFIT/32:

(1) *Instrumental error of absorbance measurements*, $s_\mathrm{inst}(A)$: The INDICES algorithm cf. ref. [12] should be used with solutions of potassium dichromate to evaluate $s_\mathrm{inst}(A)$. The Cattel's scree plot of $s_k(A) = f(k)$ consists of two straight lines intersecting at $\{s_k^*(A); k^*\}$ where k^* is the matrix rank for the system. Since $k^* = 1$ for one component $\mathrm{K_2Cr_2O_7}$ in solution, the value of $s_k(A)$ for $k^* = 1$ is a good estimate of the instrumental error of the spectrophotometer used, $s_\mathrm{inst}(A) = s_1^*(A)$ reaching a value of 0.25 mA U for the Cintra 40 (GBC, Australia) spectrophotometer employed.

(2) *Experimental design*: Since preparation of a large number of separate solutions is tedious, simultaneous monitoring of absorbance and pH during titrations is valuable [7]. In a titration, the total concentration of one of the components changes incrementaly over a relatively wide range, but the total concentrations of the other components change only by dilution, or not at all if they are present at the same concentration in the titrant and titrand. However, the absorbance cannot be varied over a large range without decreasing the precision of its measurement, and is effectively confined to a range of about one order of magnitude, e.g., $0.1 < A < 1.2$, though the range of concentrations measured can be increased by use of different path-lengths, e.g., 5, 1 and 0.1 cm. The protonation equilibria of drugs are usually studied in the ultraviolet and visible region, 190–760 nm. The wavelength range selected is such that every species makes a significant contribution to the absorbance; little information is obtained in regions of great spectral overlap or where the molar absorptivities of two or more species are linearly interdependent, as the change of absorbance following changes in c_L and c_H becomes rather small. If only a small number of wavelengths is used those of maxima or shoulders should be chosen, because small errors in setting the wavelength are then less important. It is best to use wavelengths at which the molar absorptivities of the species differ greatly, or a large number of wavelengths spaced at equal intervals.

(3) *Number of light-absorbing species*: A qualitative interpretation of the spectra aims to evaluate of the quality of the dataset and remove spurious data, and to estimate the minimum number of *factors*, i.e. contributing aqueous species, which are necessary to describe the experimental data. The INDICES [12] determine the number of dominant species present in the equilibrium mixture. In this algorithm the various indicator function PC(k) techniques developed to deduce the exact size of the true component space can be classified into two general categories: (a) precise methods based upon the knowledge of the experimental error of the absorbance data, $s_\mathrm{inst}(A)$, and (b) approximate methods

requiring no knowledge of the experimental error, $s_{\mathrm{inst}}(A)$. In general, more precise and most inclining methods are based on *the first criterion* concerning the procedure of finding the point where the slope of the indicator function $\mathrm{PC}(k) = f(k)$ changes. Each "real" factor corresponding to an actual absorbing species in solution will cause a dramatic decrease in $\mathrm{PC}(k)$ value, whereas superfluous factors cause only very small decreases. In reality, though, noise also contains systematic contributions, either from instrumental or from physical factors, and the break in the slope may not be very clear on graphs. Elbergali et al. [45] therefore proposed derivatives to improve the identification of the number of components. The *derivative criteria*, $\mathrm{S.D.}(k)$ are based on the points where the slope changes and reaches a maximum. The $\mathrm{S.D.}(k)$ is defined as $\mathrm{S.D.}(k) = \log[\mathrm{PC}(k+1)] - 2 \times \log[\mathrm{PC}(k)] + \log[\mathrm{PC}(k-1)]$ and $p - k$ should be at the first maximum of the $\mathrm{S.D.}(k)$ function. The *third derivative* $\mathrm{TD}(k)$ value crosses zero and reaches a negative minimum which can be used as a criterion. The $\mathrm{TD}(k)$ is defined as $\mathrm{TD}(k) = \log[\mathrm{PC}(k+2)] - 3 \times \log[\mathrm{PC}(k+1)] + 3 \times \log[\mathrm{PC}(k)] - \log[\mathrm{PC}(k-1)]$ and p should be equal to k where $\mathrm{TD}(k)$ has its first minimum. The change in slope can also be found by calculating the *derivatives ratio*, $\mathrm{ROD}(k)$ by $\mathrm{ROP}(k) = \{\mathrm{PC}(k-1) - \mathrm{PC}(k)\}/\{\mathrm{PC}(k) - \mathrm{PC}(k+1)\}$. Ideally $\mathrm{ROD}(k)$ should have a maximum at the point where $k = p$.

(a) *Precise indices*: Besides the first criterion applied, indicator function $\mathrm{PC}(k)$ methods are also based on a comparison of an actual index $\mathrm{PC}(k)$ of the method used with the experimental error of the instrument used, $s_{\mathrm{inst}}(A)$. These have been described elsewhere [12]:

1. *Kankare's residual standard deviation*, $s_k(A)$. The $s_k(A)$ values for different numbers of components k are plotted against an index k, $s_k(A) = f(k)$, and the number of significant components is an integer $p = k$ for which $s_k(A)$ is close to the instrumental error of absorbance $s_{\mathrm{inst}}(A)$, [11,12]. When no outliers (grossly erroneous points) are present in the spectra examined, $s_k^*(A) \leq s_{\mathrm{inst}}(A)$ is valid. Outliers are detected, and corrected and the $s_k^*(A) = f(k)$ plot is then recalculated; the spectra are then free from gross errors and ready to be analyzed by the regression program.
2. *Residual standard deviation*, $\mathrm{R.S.D.}(k)$, is used analogously to the previous method $s_k(A)$.
3. *Average error criterion*, $\mathrm{AE}(k)$, is used analogously to the preceding method $s_k(A)$.
4. *Bartlett χ^2 criterion*, $\chi^2(k)$ is used when the true number of significant components corresponds to the first k value for which $\chi^2(k)$ is less than critical $\chi^2(k)_{\mathrm{expected}} = (n-k)(m-k)$.

(b) *Approximate methods*: A more difficult problem is to deduce the number of components without relying on an estimation of the instrumental error of absorbance, $s_{\mathrm{inst}}(A)$: only the first criterion remains. Most of the techniques presented are empirical functions [12]. Eigenvalues g_k are conventionally used as a measure of the size of a principal component [46]. The first p eigenvalues, called a set of primary eigenvalues, contain a contribution from the real components and should be considerably larger than those containing only noise. The second set, called the secondary eigenvalues contains $(o - p)$ eigenvalues and these are referred to as non-significant eigenvalues.

1. *Exner function, $\psi(k)$*: The Exner $\psi(k)$ function may be used for the identification of the true dimensionality of the data. Exner proposed that $\psi = 0.3$ can be considered a fair correlation, $\psi = 0.2$ can be considered a good correlation and $\psi = 0.1$ an excellent correlation. This means that for $\psi < 0.1$ the corresponding k can be taken as the number of light-absorbing species in solution; the first criterion is, however, often preferred as the more reliable one.
2. *Scree test, RPV(k)*: The scree test for the identification of the true dimensionality of a data set is based on the observation that the residual variance should level off before those dimensions containing random error are included in the data reproduction. When the residual percentage variance is plotted against the number of k PC dimensions used in data reproduction, $\mathrm{RPV}(k) = f(k)$, the curve should drop rapidly and level off at some point. According to the first criterion, the point where the curve begins to level off, or where a discontinuity appears, is taken to be the dimensionality of the data space [47,48].
3. *Imbedded error function, IE(k)*: The imbedded error function IE(k) is an empirical function [48] developed to identify those k latent variables which contain error without relying upon an estimate of the error associated with the absorbance data matrix. The imbedded error is a function of the error eigenvalues. The behavior of the $\mathrm{IE}(k)$ function, as long as k varies from 1 to o, can be used to deduce the true dimensionality of the data. The $\mathrm{IE}(k)$ function should decrease as the true dimensions are used in the data reproduction. When the true dimensions are exhausted, however, and the error dimensions are included in the reproduction, the $\mathrm{IE}(k)$ should increase.
4. *Factor indicator function, IND(k)*: The factor indicator function $\mathrm{IND}(k)$ is an empirical function which appears more sensitive than the $\mathrm{IE}(k)$ function in identifzing the true dimensionality of an absorbance data matrix [47]. This function, like the $\mathrm{IE}(k)$ function, reaches a minimum when the correct number of latent variables or k PC dimensions is employed in the data reproduction. It has however been observed that the minimum is more pronounced and/or can often occur even in situations where the $\mathrm{IE}(k)$ function exhibits no minimum.
5. *Ratio of eigenvalues calculated by smoothed PCA and those by ordinary PCA, RESO(k)*: The recommended procedure for determining the number

of components in mixtures using RESO(k) contains principal components analysis for the measured spectra set using the SVD algorithm to find the eigenvalues g_i^0 which correspond to ordinary PCA. Details may be found in the original paper describing RESO [49]. The testing criterion calculates the index RESO_i^a or the ratios between $g_{a,i}^s$ and g_i^0 for different a and plot $\log(\mathrm{RESO}_i^a)$ versus component number. It estimates the number of components by examining the $\log(\mathrm{RESO}_i^a)$ versus component number plots. RESO then locates the number of $\log(\mathrm{RESO}_i^a)s$ which are very close to each other and do not change substantially with the variation of k in comparison to the remaining $\log(\mathrm{RESO}_i^a)s$. This is the number of components existing in the mixture examined.

(4) *Choice of computational strategy*: The input data should specify whether $\boldsymbol{\beta}_{qr}$ or $\log \boldsymbol{\beta}_{qr}$ values are to be refined whether multiple regression (MR) or non-negative linear least-squares (NNLS) are desired [5,7], whether baseline correction has to be performed, etc. In description of the model, it should be indicated whether the protonation constants are to be refined or held constant, and whether molar absorptivities are to be refined.

(5) *Previously reported or theoretically predicted parameter* $\boldsymbol{\beta}_{qr}$ *estimates*: 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 stability (protonation) constants and the probable spectral traces of all the expected components. This information assists in enabling the use of very good values close to final results as the necessary initial guesses in the minimization process. This is critical when the numbers of unknowns are high and the risk of local minima destroys the output of non-linear regression analysis of the spectroscopic data.

Two programs, PALLAS [51] and MARVIN [52] provide a collection of powerful tools for making *a prediction* of the pK_a values of any organic compound on the basis of base on the structural formulae of the compounds, using approximately 300 Hammett and Taft equations. Depending on the nature of the chemical structure and based on the hypothesis that the ionization state of a particular group is dependent upon its subenvironments constituted by its neighboring atoms and bonds, a hierarchical tree is constructed from the ionizing atom outward. This contains the atoms directly connected to the root atom at the first level, those bonded to the first level at the second level, and so on. Ab initio quantum mechanics calculations have been used extensively, as have semiempirical quantum mechanics [50].

(6) *Diagnostic criteria indicating a correct chemical model*: When the minimization process of a regression spectra analysis terminates, some diagnostic criteria are examined to determine whether the results should be accepted. An incorrect hypothesis on the chemical model leads to divergency, cyclization, or the failure of the minimization. To attain a good chemical model, the following diagnostics should be considered:

First diagnostic—the physical meaning of the parametric estimates: The physical meaning of the stability (protonation) constants, associated molar absorptivities, and stoichiometric indices is examined: $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$ should be neither too high nor too low, and $\boldsymbol{\varepsilon}_{qr}$ should not be negative. The absolute values of $s(\beta_j)$, $s(\varepsilon_j)$ give information about the last RSS-contour of the hyperparaboloid in neighborhood of the pit, $\mathrm{RSS}_\mathrm{min}$. For well-conditioned parameters, the last RSS-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 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. The set of standard deviations of $\boldsymbol{\varepsilon}_{pqr}$ for various wavelengths, $s(\boldsymbol{\varepsilon}_{qr}) = f(\lambda)$, should have a Gaussian distribution; otherwise erroneous estimates of $\boldsymbol{\varepsilon}_{qr}$ are obtained. High parameter standard deviations are often caused either by termination of the minimization process before a minimum is reached or high non-linearity in the regression model.

Second 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, i.e. down to about 10^{-8} M. Since a species present at about 1% relative concentration or less in an equilibrium behaves as numerical noise in a regression analysis, a distribution diagram makes it easier to judge the contributions of the individual species to the total concentration quickly. Since the molar absorptivities will be generally be in the range $10^3 - 10^5$ L mole^{-1} cm^{-1}, species present at low concentration, e.g. less than ca. 0.1% relative concentration, will affect the absorbance significantly only if their ε is extremely high. They may represent an "enough to interfere but not enough to determine" specimen.

Third diagnostic—parametric correlation coefficients: Partial correlation coefficients, r_{ij}, indicate the interdependence of two parameters, i.e. stability constants β_i and β_j, when others are fixed in value. Fundamentally, all of these correlation coefficients may have values between −1 and +1, where zero indicates complete independence, and +1 or −1 indicates complete correlation. Two completely correlated species cannot be included in a chemical model, because the relevant protonation constants are strongly correlated and an increase or decrease of one component may compensated for the other.

Fourth diagnostic—goodness-of-fit test: This diagnostic contains the most important criteria for testing the correctness of the hypothetical chemical model proposed. To identify the "best" or true chemical model when several are possible or proposed, and to establish whether or not 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, $e_i = A_{\exp,i,j} - A_{\text{calc},i,j}$. One of the most important statistics calculated is *the standard deviation of the absorbance*, $s(A)$, calculated from a set of refined parameters at the termination of the minimization process. This is usually compared with the standard deviation of absorbance calculated by the INDICES program [12] $s_k(A)$ 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. Although this statistical analysis of residuals [53] gives the most rigorous test of the degree-of-fit, 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 $s(A) > 0.005$ indicated 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 mean* (known as the *residual bias*) $\bar{e}$ should be a value close to zero; the *mean residual* $|\bar{e}|$ and 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; a Hamilton R-factor of relative fit, expressed as a percentage ($R \times 100\%$), of <0.5% is taken as an excellent fit, but a value of >2% is taken to be a poor one. The R-factor gives a rigorous test of the null hypothesis H_0 (giving R_0) against the alternative H_1 (giving R_1). H_1 could be rejected at the α significance level if $R_1/R_0 > R_{(m,n-m,\alpha)}$, where n is the number of experimental points, m the number of unknown parameters, and $(n-m)$ is the number of degrees of freedom. The value of $R_{(m,n-m,\alpha)}$ can be found in statistical tables.

Fifth diagnostic—deconvolution of spectra: Resolution of each experimental spectrum into spectra of the individual species proves whether the experimental design is efficient enough. If for a particular concentration range the spectrum consists of just a single component, further spectra for that range would be redundant. In ranges where many components contribute significantly to the spectrum, several spectra should be measured. If the model represents the data adequately, the residuals should possess characteristics that agree with, or at least do not refute, the basic assumptions: the residuals should be randomly distributed about the A_{calc} values predicted by the regression equation. Systematic departures from randomness indicate that the model is not satisfactory. Examination of plots of the residuals versus λ may assist numerical and/or graphical aids in the analysis of residuals. A study of the signs of the residuals (+ or −) and the sums of the signs can be used. Graphical presentation of the residuals is of great help in the diagnosis: for detection of an outlier, detection of a trend in the residuals, detection of a sign change, detection of an abrupt shift of level in the spectrum, and examination of symmetry and normality in the residuals distribution.

(7) *Search for the best computation a strategy*: Analysis of simulated spectra is usually recommended as it serves to—(a) establish the best computational strategy for an efficient regression analysis, (b) investigate of the sensitivity of each parameter in the chemical model assumed, and (c) examinate of the influence of the instrumental noise of the spectrophotometer used $s_{\text{inst}}(A)$ on the accuracy and precision of the parameters estimated $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$. The details for the computer data treatment are collected in the Supporting Information.

2.3. Reliability of the estimated protonation constants

The adequacy of a proposed regression model with experimental spectra and the reliability of parameter estimates $\boldsymbol{\beta}_{qr,j}$ found (being denoted for the sake of simplicity as b_j, $j = 1, \ldots, m$) and ε_{ij}, $j = 1, \ldots, n_{\text{w}}$, may be examined by a goodness-of-fit test, cf. page 101 in ref. [2]:

(1) *The quality of parameter estimates b_j, $j = 1, \ldots, m$, found* is reviewed according to the variances $D(b_j)$. Often an empirical rule is used: parameter b_j differs significantly from zero when its estimate is greater than 3 standard deviations, $3\sqrt{D(b_j)} < |b_j|$, $j = 1, \ldots, m$.

(2) *The quality of the experimental data* is examined by identification of the influential points (namely outliers) with the use of regression diagnostics, cf. page 62 in ref. [53].

(3) *The quality of curve fit achieved*: the adequacy of the proposed model and m parameter estimates found with n values of experimental data is examined by a goodness-of-fit test based on the statistical analysis of classical residuals. If the proposed model adequately represents the data, the residuals should form a random pattern with a normal distribution $N(0, s^2)$, with the residual mean equal to zero, $\bar{e} = 0$, and the standard deviation of residuals $s(e)$ being near to noise, i.e. the experimental error ε of the absorbance measured. Systematic departures from randomness indicate that the model and parameter estimates are not satisfactory. Examination of residual plots may be assisted by graphical analysis of the residuals, cf. pages 289 and 290 in ref. [53].

3. Experimental

3.1. Chemicals and solutions

7-Ethyl-10-hydroxycamptothecin were purchased from Molcan Corporation, Canada, with a purity of 98.5% (HPLC). Potassium hydroxide, 1 M, was prepared from an exact weight of pellets (p.a., Aldrich Chemical Company) with carbon-dioxide free redistilled water. The solution was stored for several days in a polyethylene bottle. This solution was standardized against a solution of potassium hydrogen-phthalate using the Gran method with a reproducibility of 0.1%. Potassium chloride (p.a. Lachema Brno) was not purified further. Buffers and other solutions were prepared from analytical-reagent grade chemicals. Twice-redistilled water was used in the preparation of solutions.

3.2. Apparatus and pH-spectrophotometric titration procedure

The free hydrogen-ion concentration h was measured via emf on an OP-208/1 digital voltmeter (Radelkis, Budapest) with a precision of ±0.1 mV using a G202B glass electrode (Radiometer, Copenhagen) and an OP-8303P commercial SCE reference electrode (Radelkis, Budapest). The spectrophotometric multiple-wavelength pH-titration was carried out as follows: an aqueous solution 20.00 cm^3 containing 10^{-5} mol/L drug, 0.100 mol/L hydrochloric acid and 10 mL indifferent solution KCl for adjustment of constant value of an ionic strength was titrated with standard 1.0 mol/L KOH at 298 K and 20 absorption spectra were recorded. Titrations were performed in a water-jacketed double-walled glass vessel of 100 mL, closed with a Teflon bung containing the electrodes, an argon inlet, a thermometer, a propeller stirrer and a capillary tip from a micro-burette. All pH measurements were carried out at 25.0 °C ± 0.1° and 37.0 °C ± 0.1°. When the drug was titrated, a stream of argon gas was bubbled through the solution both to stir and to maintain an inert atmosphere. The argon was passed through an aqueous ionic medium by prior passage through one or two vessels also containing the titrand medium before entering the corresponding titrand solution. The burettes used were syringe micro-burettes of 1250 μL capacity (META, Brno) with a 2.50 cm micrometer screw, [54]. The polyethylene capillary tip of the micro-burette was immersed into the solution when adding reagent, but withdrawn after each addition in order to avoid leakage of the reagent during the pH read out. The micro-burette was calibrated by 10 replicate determinations of the total volume of delivered water by weighing on a Sartorius 1712 MP8 balance with results evaluated statistically, leading to a precision of ±0.015% in added volume over the whole volume range. The solution was pumped into the cuvette and spectrophotometric measurement was performed with the use of a Cintra 40 spectrophotometer (GBC, Australia).

3.3. Software used

Computation related to the determination of dissociation constants was performed by regression analysis of UV/VIS spectra using the SQUAD(84) [7] and SPECFIT/32 [34] programs. Most of graphs were plotted using ORIGIN 7.5 [55]. For prediction of pK_a on base of the molecule structure the programs PALLAS [51] and MARVIN [52] were used. The factor analysis was performed with program INDICES [12].

3.4. Supporting information available

Complete experimental and computational procedures, input data specimen and corresponding output in numerical and graphical form for both programs, SQUAD(84) and SPECFIT/32 are available free of charge via the Internet at http://meloun.upce.cz and in the block *DATA* of a menu.

4. Results and discussion

The SQUAD(84) spectra analysis starts with data smoothing followed by a factor analysis using the INDICES program. The experimental spectra are obtained for the titration of an alkaline 1.02 × 10^{-4} M 7-ethyl-10-hydroxycamptothecine solution by a standard solution of 1 M HCl (or HClO$_4$) to adjust pH value. Comparison of both SQUAD and SPECFIT regression program treatments, with the proposed strategy for an efficient experimentation in protonation constants determination is presented. pH-spectrophotometric titration enables the absorbance-response-surface data on Fig. 1 to be obtained for analysis with non-linear regression. As the actual SQUAD version used has a limited dimension and input can contain 20 spectra only, an efficient spectra sample 20 × 39 ($n_s \times n_w$) was used (Fig. 2) for regression analysis.

The number of light-absorbing species p can be predicted from the indices function values by finding the point $p = k$ where the slope of Cattel's indices function PC(k) = f(k) changes, or by comparing PC(k) values with the instrumental error $s_{\mathrm{inst}}(A)$. This is the common criterion for determining p on Fig. 3. Very low values of $s_{\mathrm{inst}}(A)$ prove that relatively reliable spectrophotometer and experimental technique were used. Due to the large variations in the indices values, their logarithms in all nine selected

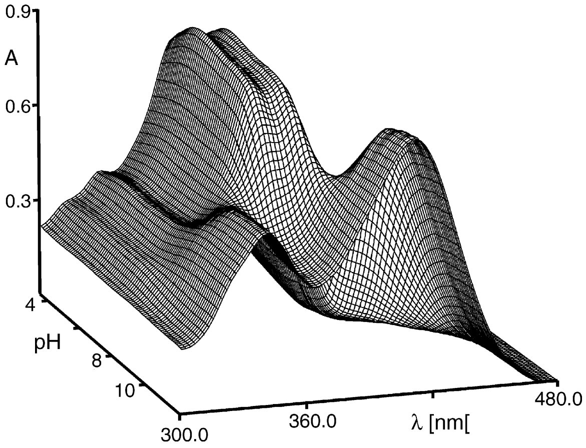

Fig. 1. The 3D-absorbance-response-surface representing 26 absorption spectra of protonation equilibria of 7-ethyl-10-hydroxycamptothecine in dependence on pH at 25 °C (S-Plus).

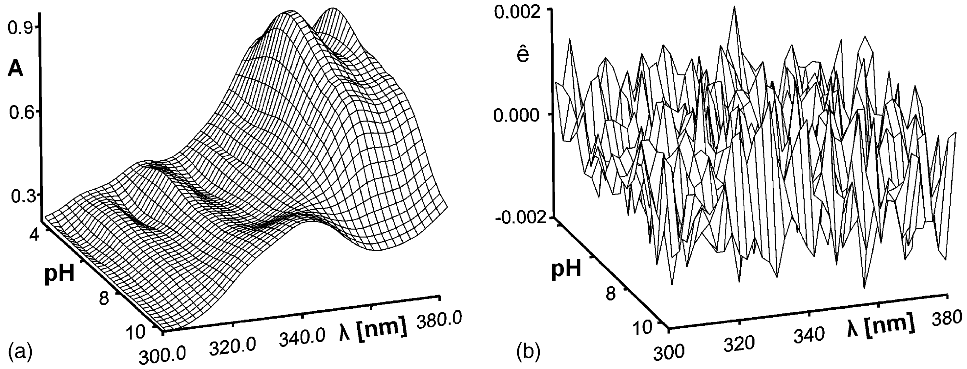

Fig. 2. (a) 3D-absorbance-response-surface representing a sample of 17 absorption spectra taken from the set on Fig. 1; (b) 3D-overall diagram of residuals representing the response surface indicating the quality of goodness-of-fit after removal of influential outlying spectra (S-Plus).

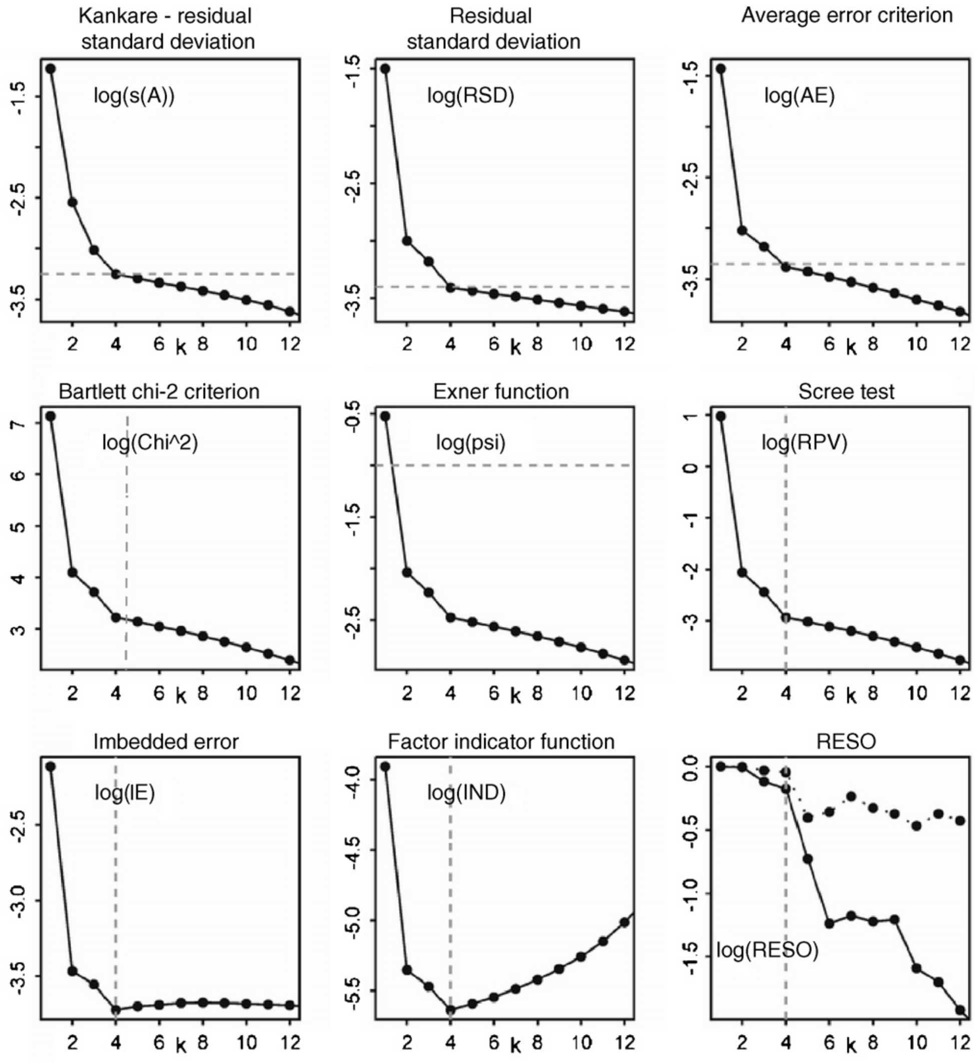

Fig. 3. Cattel's scree plot for the determination of the number of light-absorbing species in mixture $k^* = 4$ and the actual instrumental error of the spectrophotometer used $s_4^*(A) = 0.56\,\mathrm{mA\,U}$ (Kankare). The logarithm dependence of 9 indices methods as a function of the number of principal components k for the pH-absorbance matrix: *first row*—Kankare's residual standard deviation, $s_k(A)$; residual standard deviation, R.S.D; average error criterion, AE; *second row*—Bartlett χ^2 criterion; Exner ψ function; scree test RPV; *third row*—imbedded error function IE; factor indicator function IND; RESO function. All methods lead to the same conclusion $k^* = 4$ (INDICES in S-Plus).

methods as a function of the number of principal components k for the drug analyzed were used.

For the indices methods in Fig. 3 (Kankare's residual standard deviation $s_k(A)$, the residual standard deviation R.S.D. and the average error criterion AE) the horizontal line denotes the value of the instrumental error, $s_{\text{inst}}(A)$. The best approximation of $s_{\text{inst}}(A)$ for 7-ethyl-10-hydroxycamptothecin was found for $k=4$, while higher values of k do not lead to any significant decrease of $s_k(A)$. The position of a break-point on the $s_k(\text{A})=f(k)$ curve in the scree plot is calculated and gives $k^*=4$ with the corresponding co-ordinate $s_4^*(A) = 0.56\,\text{mA U}$ which also represents the instrumental error $s_{\text{inst}}(A)$ of the spectrophotometer used. For the Bartlett χ^2 criterion, the horizontal line denotes a magnitude of χ^2_{krit} and the vertical line separates values of k for which H_0 was accepted. In the case of the approximate indices methods for the Exner ψ function, the value $\psi \leq 0.1$ is achieved for $k=4$ while higher values of k do not bring a significant decrease, in the value ψ. For the scree test RPV, the curve of dependence RPV(k) = f(k) begins to level off at some point of k. This $k=4$ value is considered to be the dimensionality of the absorbance data space. For the imbedded error function IE there is a minimum of $k=4$ on the curve of the function IE = f(k). Similarly, for the factor indicator function, a minimum of $k=4$ on the curve of the function IND = f(k) is reached. The RESO method also leads to $k=4$ species in a mixture. It may concluded that (a) generally, the most reliable indices methods seem to be those based on a knowledge of the instrumental error of absorbance, $s_{\text{inst}}(A)$, (b) indices methods are all based on finding the point where the slope of the indices function changes, and (c) precise methods based on a knowledge of the instrumental error of absorbance $s_{\text{inst}}(A)$ should be preferred.

When there are more than three components, derivative methods can be used: when the curve PC(k) = f(k) does not exhibit a clear break-point, the second derivative localizes this break more reliably. The *derivative criteria* are based on the point where the slope changes and reaches a maximum in Fig. 4. The second

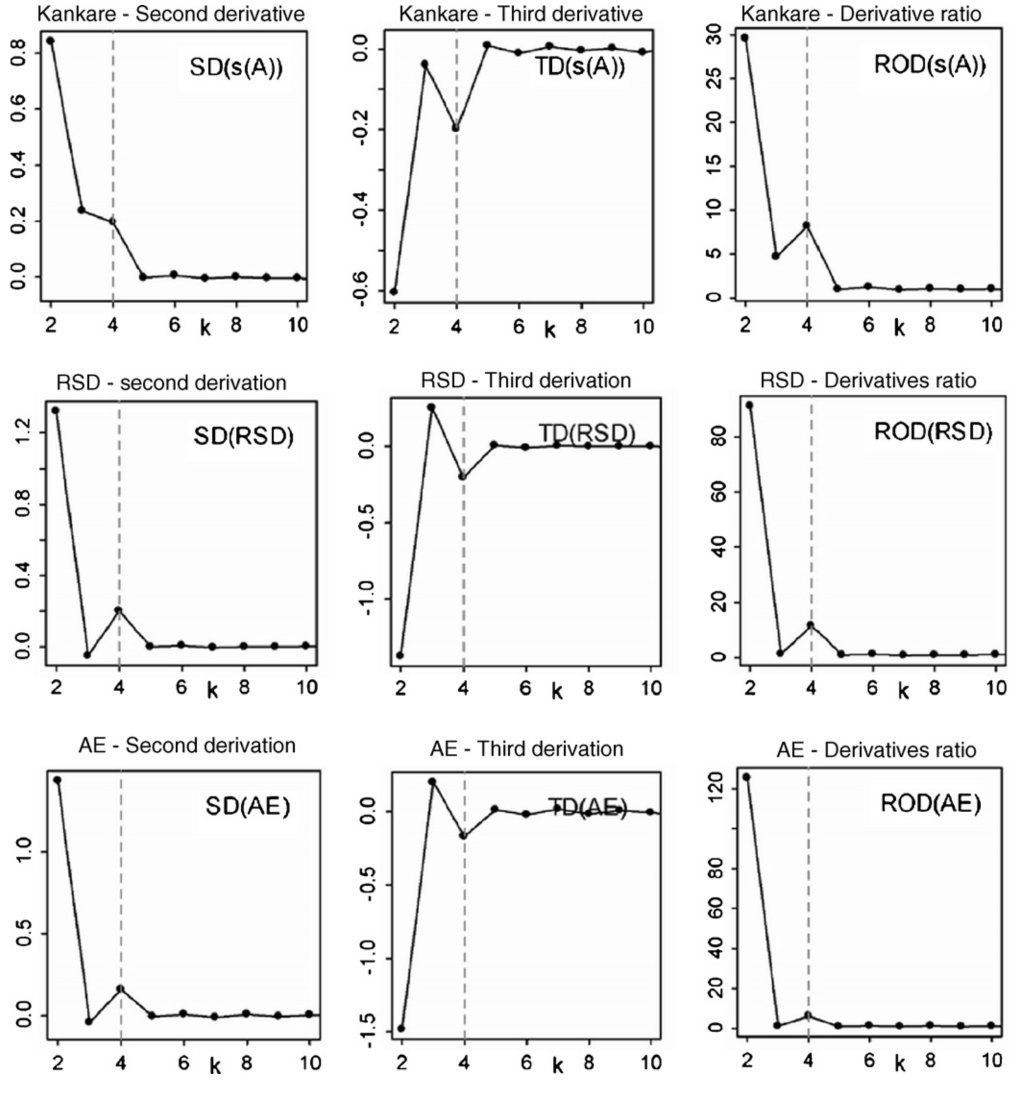

Fig. 4. The derivatives detection criteria of some indices functions applied to the absorbance data from Fig. 3: *first row*—the second derivative of the Kankare residual standard deviation S.D.($s_k(A)$) (left); the third derivative TD($s_k(A)$) (middle); and the derivatives ratio ROD($s_k(A)$) (right); *second row*—the second derivatives of the residual standard deviation S.D.(R.S.D.); the third derivative TD(R.S.D.) (middle); and the derivatives ratio ROD(R.S.D.) (right); *third row*—the second derivatives of the average error function S.D.(AE); the third derivative TD(AE) (middle);and the derivatives ratio ROD(AE) (right); (INDICES in S-Plus).

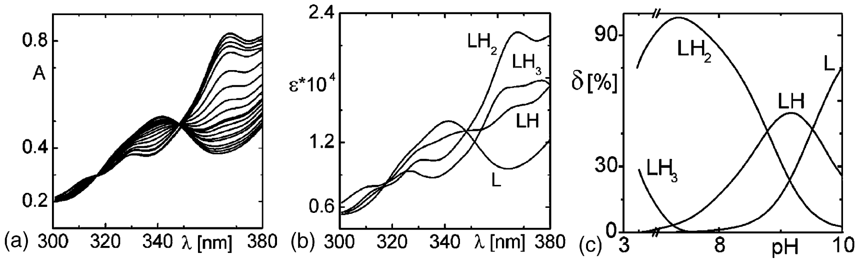

Fig. 5. (a) The absorption spectra of 7-ethyl-10-hydroxycamptothecine for various pH values, (b) pure spectra profiles of molar absorptivities vs. wavelengths for variously protonated species L, LH, LH_2, LH_3, (c) distribution diagram of the relative concentrations of all of the variously protonated species L, LH, LH_2, LH_3 of 7-ethyl-10-hydroxycamptothecine in dependence on pH (SQUAD, ORIGIN).

derivative S.D.(k) and $p-k$ should be at the first maximum of the S.D.(k) function. The *third derivative* TD(k) value crosses zero and reaches a negative minimum which can be used as a criterion. The change in slope can also be found by calculating the *derivatives ratio* ROD(k). Ideally ROD(k) should have a maximum at the point where $k=p$. A more difficult problem is to deduce the numer of components without relying on an estimation of the instrumental error of absorbance, $s_{\mathrm{inst}}(A)$; then only the first criterion remains. All three index methods predict the four variously protonated light-absorbing species of drug 7-ethyl-10-hydroxycamptothecine in protonation equilibrium, $k=4$.

Two sets of simulated and experimental absorption spectra were used to examine the applicability of both algorithms to the determination of protonation constants. Three protonation constants and four molar absorptivities of 7-ethyl-10-hydroxycamptothecine for 39 wavelengths and 20 spectra (Figs. 2 and 5) constitute the unknown parameters which are refined by the MR algorithm in the first run of the SQUAD program. In the second run, the NNLS algorithm makes the final refinement of all the previously found parameter estimates with all the molar absorptivities kept non-negative. The reliability of the parameter estimates may be tested with the use of SQUAD(84) diagnostics in Table 1:

The first diagnostic indicates whether all parametric estimates $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$ have physical meaning and attain realistic values. As the standard deviations $s(\log \boldsymbol{\beta}_{qr})$ of parameters $\log \boldsymbol{\beta}_{qr}$ and $s(\boldsymbol{\varepsilon}_{qr})$ of parameters $\boldsymbol{\varepsilon}_{qr}$ are significantly smaller than their corresponding parameter estimates, all the variously protonated species are statisticaly significant. Fig. 5 shows the estimated molar absorptivities of four of the variously protonated species ε_{L}, $\varepsilon_{\mathrm{LH}}$, $\varepsilon_{\mathrm{LH_2}}$, and $\varepsilon_{\mathrm{LH_3}}$ of the anticancer drug 7-ethyl-

Table 1
The best chemical model found for a protonation equilibrium of 7-ethyl-10-hydroxycamptothecine using double checked non-linear least squares regression analysis of multiwavelengths and multivariate pH-spectra with SQUAD(84) and SPECFIT/32 (bold) for $n_s=17$ spectra measured at $n_w=39$ wavelengths for $n_z=2$ basic components L and H, forming $n_c=4$ variously protonated species

L_qH_r	Estimated protonation constants		Partial correlation coefficients		
	$\log \boldsymbol{\beta}_{qr}$	$s(\log \boldsymbol{\beta}_{qr})$	L_1H_1	L_1H_2	L_1H_3
L_1H_1	9.516, **9.519**	0.022, **0.035**	1	–	–
L_1H_2	18.299, **18.306**	0.041, **0.015**	0.997	1	–
L_1H_3	21.346, **21.395**	0.062, **0.018**	0.6232	0.6198	1

	SQUAD	SPECFIT
Determination of the number of light-absorbing species by factor analysis		
Number of light-absorbing species, k^*	4	4
Residual standard deviation, $s_k^*(A)$	0.56	Not estimated
Goodness-of-fit test by the statistical analysis of the residuals		
Residual square sum, RSS	3.35×10^{-4}	2.38×10^{-4}
Residual mean $\bar{e}$ bar [mA U]	-2.05×10^{-8}	Not estimated
Mean residual $\lvert\bar{e}\rvert$ [mA U]	0.58	Not estimated
Standard deviation of residuals, $s(e)$ [mA U]	0.81	0.6
Residual skewness $\hat{g}_1(e)$	−0.12	Not estimated
Residual kurtosis $\hat{g}_2(e)$	2.12	Not estimated
Hamilton R-factor [%]	0.15	Not estimated
ε (all species) vs. λ	Realistic	Realistic

The charges of the ions are omitted for the sake of simplicity and the standard deviations of the parameter estimates are in the last valid digits in brackets. The resolution criterion and reliability of parameter estimates found are proven with goodness-of-fit statistics such as the residual square sum RSS, the standard deviation of absorbance after termination of the regression process, $s(A)$ [mA U], the residual standard deviation by factor analysis $s_k(A)$ [mA U], the mean residual $\bar{e}$, the residual standard deviation $s(e)$, the residual skewness $g_1(e)$ and the residual kurtosis $g_2(e)$ proving a Gaussian distribution; Hamilton R-factor [%] and non-negative and realistic estimates of calculated molar absorptivities of all variously protonated species ε vs. λ.

10-hydroxycamptothecine in dependence on wavelength. Some spectra overlap, and such cases may cause some resolution difficulties in a non-linear regression approach.

The second diagnostic tests whether all of the calculated free concentrations of variously protonated species on the distribution diagram have physical meaning, which proved to be the case (Fig. 5). The diagram shows that one overlapping protonation equilibrium exists.

The third diagnostic concerning the matrix of correlation coefficients in Table 1 proves that there is an interdependence of one pair of protonation constants of 7-ethyl-10-hydroxycamptothecine r (β_{11} versus β_{12}). The significant correlation of this pair, pK_{a2} = 8.79 and pK_{a3} = 9.51, may be explained by proximate dissociation constants, which associated with the overlapping equilibria.

The fourth diagnostic concerning the goodness-of-fit (Fig. 6 left) indicates three outlying spectra, nos. 1, 4 and 18. After removing the outliers, the plot of $s(e)$ and $|\bar{e}|$ for each spectrum proves that the $s_4(A)$ value is equal to 0.56 mA U and is close to the standard deviation of absorbance when the minimization process terminates, $s(A) = 0.81$ mA U (Table 1). The statistical measures of all residuals from Fig. 6 prove that the minimum of the eliptic hyperparaboloid RSS is reached: the residual mean $\bar{e} = -2.05 \times 10^{-8}$ proves that there is no bias or systematic error in the spectra fitting. The mean residual $|\bar{e}| =$ 0.58 mA U and the residual standard deviation $s(e) = 0.81$ mA U (and 0.60 SPECFIT) have sufficiently low values. The skewness $g_1(e) = -0.12$ is close to zero and proves a symmetric distribution of the residuals set, while the kurtosis $g_2(e) = 2.12$ is close to 3 proving a Gaussian distribution. The Hamilton R-factor of relative fitness is 0.15%, proving an excellent achieved fitness, and therefore the parameter estimates may be considered as reliable.

The fifth diagnostic, the spectra deconvolution on Fig. 7, shows the deconvolution of the experimental spectrum into

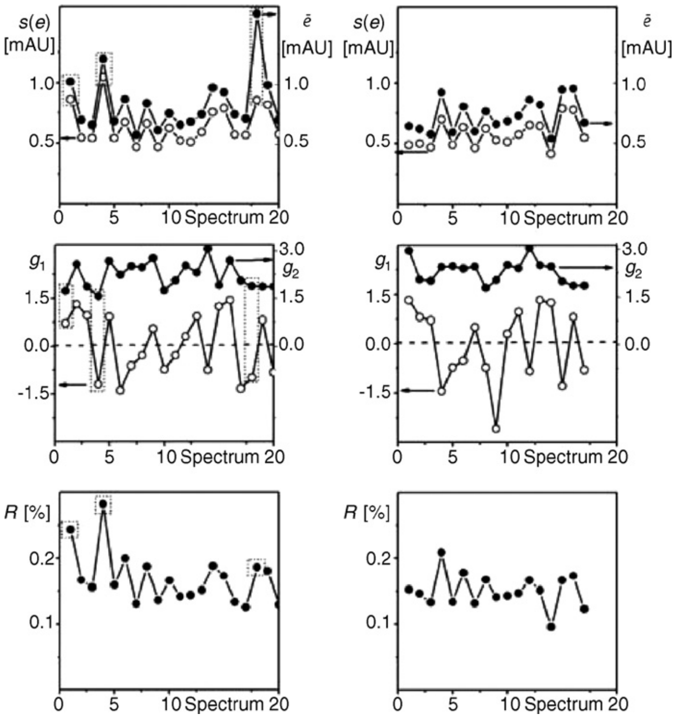

Fig. 6. Detecting and removing influential outlying spectra with the use of the goodness-of-fit test. Achieved spectra fitness before (left) and after (right) removing outliers. Rectangles indicate outliers: *first row*—the plot of the residual standard deviation $s(e)$ and the mean residual $|\bar{e}|$ indicates spectra nos. 1, 4 and 18 as the outliers; *second row*—test of residual distribution symmetry using skewness g_1 and kurtosis g_2; *third row*—a Hamilton R-factor of relative fit expressed as a percentage of an excellent curve-fitting can be used for the detection of outliers (SQUAD, ORIGIN).

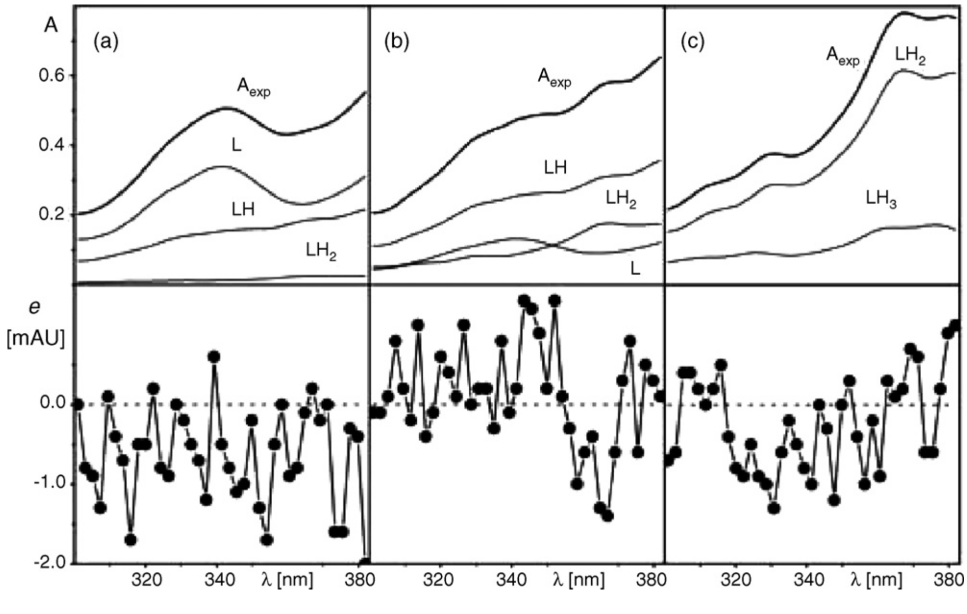

Fig. 7. Deconvolution of the experimental absorption spectrum of 7-ethyl-10-hydroxycamptothecine for 39 wavelengths into spectra of the individual variously protonated species L, LH, LH_2, LH_3 in solution (above) and the statistical analysis of the residuals (below) of each particular absorption spectrum for a selected value of pH equal to: (a) 10.070, (b) 9.478 and (c) 7.231 (SQUAD, ORIGIN).

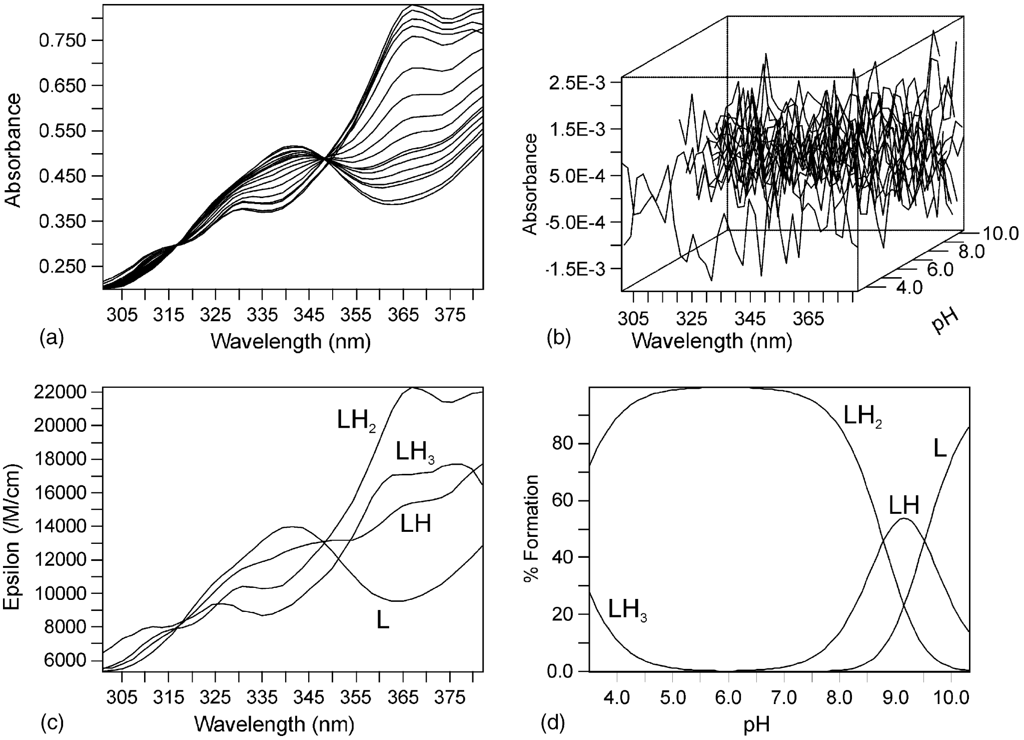

Fig. 8. Typical SPECFIT working environment testing a chemical model hypothesis of four variously protonated species L, LH, LH_2, LH_3 of 7-ethyl-10-hydroxycamptothecine in dependence on pH: (a) the measured absorption spectra for various pH values; (b) the 3D-presentation map of residuals; (c) pure spectra profiles of molar absorptivities vs. wavelengths for all of the variously protonated species; (d) distribution diagram of the relative concentrations of all of the variously protonated species (SPECFIT).

spectra of the individual variously protonated species to examine whether the experimental design is efficient. Spectrum deconvolution seems to be a quite useful tool in the proposal of an efficient experimentation strategy. 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 prevails in solution, the spectrum yields quite poor information into the regression analysis, and the parameter estimate is somewhat uncertain, and definitely not reliable enough. To test the reliability of protonation constants at different ionic strengths, a goodness-of-fit test is applied with the use of a statistical analysis of the residuals, and the results are given in Tables 1 and 2. For the drug studied, the most efficient tools, such as the Hamilton R-factor, the mean residual and the standard deviation of residuals, are applied: as the R-factor in all cases reaches a value of less than 0.2%, an excellent fitness and reliable parameter estimates are indicated. The standard deviation of absorbance $s(A)$ after termination of the minimization process is always better than 1.0 mA U, and the proposal of a good chemical model and of reliable parameter estimates are proven.

The SPECFIT/32 program found the same estimates of parameters $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$ and of associated species concentrations, parametric correlation coefficients, goodness-of-fit test, error analysis and spectra deconvolution, and a typical SPECFIT working environment testing a chemical model hypothesis of four variously protonated species L, LH, LH_2, and LH_3 of 7-ethyl-10-hydroxycamptothecine in dependence on pH is given in Fig. 8, and of four another species L, L_2H, L_2H_2, and L_2H_3 are in Fig. 9.

The first problem in the evaluation of the protonation equilibria of the first drug concerns the strongly overlapping equilibria because the difference between the two consecutive dissociation constants is $\log \beta_{12} - \log \beta_{11} = 0.82$, which is less than three pH units (the rule of overlapping equilibria). Such close equilibria are always difficult to evaluate and therefore the user should carefully prove the reliability of each protonation constant estimation. A distribution diagram of the relative concentrations of all of the variously protonated species demonstrates the overlapping protonation equilibria for two

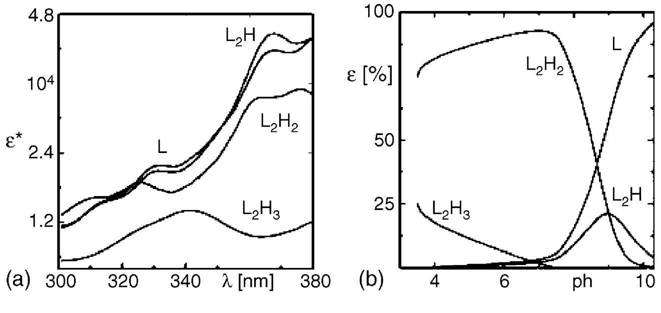

Fig. 9. Testing a chemical model hypothesis of 4 variously protonated species L, L_2H, L_2H_2 and L_2H_3, of 7-ethyl-10-hydroxycamptothecine in dependence on pH: (a) pure spectra profiles of molar absorptivities vs. wavelengths for all of the variously protonated species and (b) distribution diagram of the relative concentrations of all of the variously protonated species (SQUAD, ORIGIN).

close consecutive protonation constants. To investigate the reliability of the protonation constants estimation, a simulated data set should also be employed, using the block of the acid dissociation simulate function of SPECFIT/32 program. The quantity of added noise in the generated absorption spectra is $s_{\mathrm{inst}}(A) = 0.5$ mA U. A spectra set was generated for protonation constants $\log \beta_{11} = 9.51$, $\log \beta_{12} = 18.30$ and $\log \beta_{13} = 21.39$ (it means $\mathrm{p}K_{\mathrm{a}1} = 3.09$, $\mathrm{p}K_{\mathrm{a}2} = 8.79$ and $\mathrm{p}K_{\mathrm{a}3} = 9.51$). The wavelength and pH range of the spectra are used agree with the experimental spectra set 301–382 nm, with step 2.13 nm and pH range from 3.50 to 10.30, respectively.

Seeking the best chemical model of protonation equilibria, four various hypotheses of the stoichiometric indices q and r of L_qH_r acid were tested in order to find the model which best represents the simulated and experimental data (Table 2). The factor analysis of the INDICES program leads to 4 light-absorbing components and the instrumental standard deviation $s_k(A) = 0.21$ mA U for the simulated data and $s_k(A) = 0.56$ mA U for the experimental data. Therefore, not more than four variously protonated species should be tested here. Both data sets lead to the same conclusion: that two hypotheses of the chemical model cannot be distinguished with the use of the degree-of-fit test as the resolution criterion, i.e. the second hypothesis of species L, LH, LH_2, LH_3, and the fourth hypothesis of species L, L_2H, L_2H_2, L_2H_3 in Table 2. True chemical model could be determined with the use of a new experimental strategy, for example, applying measurement for higher concentration of drug. After the degree-of-fit test, the quality of the plot of molar absorptivities $\boldsymbol{\varepsilon}_{pq,j}$, $j = 1, \ldots, n_{\mathrm{w}}$ of all of the variously protonated species in dependence on wavelength λ on Figs. 8 and 9 is examined to ascertain whether the curves are realistic enough.

Table 2
The search for a protonation equilibria model of 7-ethyl-10-hydroxycamptothecine using non-linear least-squares regression analysis of multiwavelengths and multivariate pH-spectra of Table 1 when (a) simulated data, and (b) experimental data were used

q, r	Given $\log \boldsymbol{\beta}_{qr}$	Estimated $\log \boldsymbol{\beta}_{qr}$ using a hypothesis of chemical model no.			
		1	2	3	4
(a)					
1, 1	9.51	9.10(1)	9.51(1)	–	–
1, 2	18.3	12.11(5)	18.30(2)	–	–
1, 3	21.39	–	21.39(3)	–	–
2, 1	–	–	–	–	13.36(1)
2, 2	–	–	–	22.97(1)	22.31(3)
2, 3	–	–	–	–	25.39(3)
2, 4	–	–	–	39.56(3)	–
2, 6	–	–	–	47.12(5)	–
Degree-of-fit test by the statistical analysis of residuals as the resolution criterion					
$s(A)$ or $s(e)$ [mA U]		1.21	0.39	2.6	0.38
$s_k(A)$ [mA U], p		0.21, 4	0.21, 4	0.21, 4	0.21, 4
$\bar{e}$		0.71	0.24	1.39	0.23
$g_1(e)$		0.36	−0.3	0.36	−0.25
$g_2(e)$		6.8	5.2	5.72	6.34
R-factor [%]		0.21	0.07	0.44	0.06
ε (all species) vs. λ are		Realistic	Realistic	Realistic	Realistic
Model hypothesis		Rejected	Accepted	Rejected	Accepted
(b)					
1, 1		9.12(1)	9.52(2)	–	–
1, 2		12.07(13)	18.30(4)	–	–
1, 3		–	21.35(6)	–	–
2, 1		–	–	–	13.39(3)
2, 2		–	–	23.10(1)	22.37(5)
2, 3		–	–	–	25.40(7)
2, 4		–	–	40.06(3)	–
2, 6		–	–	47.30(8)	–
Degree-of-fit test by the statistical analysis of residuals as the resolution criterion					
$s(A)$ or $s(e)$ [mA U]		1.86	0.81	2.78	0.85
$s_k(A)$ [mA U], p		0.56, 4	0.56, 4	0.56, 4	0.56, 4
$\bar{e}$		1.22	0.58	1.65	0.6
$g_1(e)$		0.57	−0.12	−0.34	−0.03
$g_2(e)$		4.75	2.12	3.74	2.12
R-factor [%]		0.35	0.15	0.5	0.15
ε (all species) vs. λ		Realistic	Realistic	Realistic	Realistic
Model hypothesis		Rejected	Accepted	Rejected	Accepted

5. Conclusions

When a drug is poorly soluble then instead of a potentiometric determination of dissociation constants, multiwavelength spectrophotometric pH-titration may be analyzed with the least-squares non-linear regression. The reliability of the dissociation constants of ionizable drug may be proven with goodness-of-fit tests of the absorption spectra measured at various pH. The criteria of resolution used for the hypotheses in question form the main part of the diagnostic tutorial proposed: (1) the number of light-absorbing species is estimated by factor analysis of the spectra set, (2) failure of the minimization process in a divergency or a cyclization; (3) examination of the physical meaning of the estimated parameters $\boldsymbol{\beta}_{qr}$ and $\boldsymbol{\varepsilon}_{qr}$ and of associated species concentrations if both were realistic and positive; and (4) residuals randomly distributed about the predicted regression spectrum, systematic departures from randomness being taken to indicate that either the chemical model or parameter estimates were unsatisfactory. However, they are cases when the fitness test may not always lead to the straightforward answer about the chemical model namely when several mathematical solutions (models) are valid and all these models tested fit points well.

Acknowledgments

The financial support of the Grant Agency IGA (Grant No. NR9055-4/2006) and of the Ministry of Education (Grant No. MSM253100002) is gratefully acknowledged.

References

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

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

[3] M. Meloun, J. Havel, Computation of Solution Equilibria, 1. Spectrophotometry, Folia Fac. Sci. Nat. Univ. Purkyn. Brunensis (Chemia), vol. XXV, Brno 1984. 2. Potentiometry, vol. XXVI, Brno 1985.

[4] L.G. Sillén, B. Warnqvist, Equilibrium constants and model testing from spectrophotometric data, using LETAGROP, Acta Chem. Scand. 22 (1968) 3032.

[5] SQUAD:, in: D.J. Leggett (Ed.), Computational Methods for the Determination of Formation Constants, Plenum Press, New York, 1985 (a) pp. 99–157, (b) pp. 291–353.

[6] J. Havel, M. Meloun, in: D.J. Leggett (Ed.), Computational Methods for the Determination of Formation Constants, Plenum Press, New York, 1985 (a) p. 19 and (b) p. 221.

[7] SQUAD(84): M. Meloun, M. Javůrek, J. Havel, Multiparametric curve fitting. X. A structural classification of program for analysing multicomponent spectra and their use in equilibrium-model determination, Talanta 33 (1986) 513–524.

[8] D.J. Leggett, W.A.E. McBryde, General computer program for the computation of stability constants from absorbance data, Anal. Chem. 47 (1975) 1065–1070.

[9] D.J. Leggett, Numerical analysis of multicomponent spectra, Anal. Chem. 49 (1977) 276–281.

[10] D.J. Leggett, S.L. Kelly, L.R. Shine, Y.T. Wu, D. Chang, 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 30 (1983) 579–586.

[11] J.J. Kankare, Computation of equilibrium constants for multicomponent systems from spectrophoto-metric data, Anal. Chem. 42 (1970) 1322–1326.

[12] M. Meloun, J. Čapek, P. Mikšík, R.G. Brereton, Critical comparison of methods predicting the number of components in spectroscopic data, Anal. Chim. Acta 423 (2000) 51–68.

[13] M. Meloun, M. Pluhařová, Thermodynamic dissociation constants of codeine, ethylmorphine and homatropine by regression analysis of potentiometric titration data, Anal. Chim. Acta 416 (2000) 55–68.

[14] M. Meloun, P. Černohorský, Thermodynamic dissociation constants of isocaine, physostigmine and pilocarpine by regression analysis of potentiometric data, Talanta 52 (2000) 931–945.

[15] M. Meloun, D. Burkoňová, T. Syrový, A. Vrána, The thermodynamic dissociation constants of silychristin, silybin, silydianin and mycophenolate by the regression analysis of spectrophotometric data, Anal. Chim. Acta 486 (2003) 125–141.

[16] M. Meloun, T. Syrový, A. Vrána, Determination of the number of light-absorbing species in the protonation equilibria of selected drugs, Anal. Chim. Acta 489 (2003) 137–151.

[17] M. Meloun, T. Syrový, A. Vrána, The thermodynamic dissociation constants of ambroxol, antazoline, naphazoline, oxymetazoline and ranitidine by the regression analysis of spectrophotometric data, Talanta 62 (2004) 511–522.

[18] M. Meloun, T. Syrový, A. Vrána, The thermodynamic dissociation constants of losartan, paracetamol, phenylephrine and quinine by the regression analysis of spectrophotometric data, Anal. Chim. Acta 533 (2005) 97–110.

[19] M. Meloun, J. Čapek, T. Syrový, Number of species in complexation equilibria os SNAZOXS or Naphtylazoxine 6S an Cd, Co, Cu, Ni, Pb and Zn ions by PCA of UV-VIS spectra, PDF, Talanta 66 (2005) 547–561.

[20] M. Meloun, T. Syrový, Number of species in complexation equilibriua of *o*-, *m*- and *p*-CAPAZOXS with Cd^{2+}, Co^{2+}, Ni^{2+}, Pb^{2+} and Zn^{2+} ions by PCA of UV-VIS spectra, Talanta, in press.

[21] M. Meloun, T. Syrový, A. Vrána, The thermodynamic dissociation constants of haemanthamine, lisuride, metergoline and nicergoline by the regression analysis of spectrophotometric data, Anal. Chim. Acta 543 (2005) 254–266.

[22] M. Meloun, M. Javůrek, J. Militký, Computer estimation of dissociation constants. Part V. Regression analysis of extended Debye-Hückel law, Microchim. Acta 109 (1992) 221–231.

[23] HYPERQUAD: P. Gans, A. Sabatini, A. Vacca, Investigation of equilibria in solution. Determination of equilibrium constants with the HYPERQUAD suite of programs, Talanta 43 (1996) 1739–1753.

[24] SPECFIT: H. Gampp, M. Maeder, Ch.J. Mayer, A.D. Zuberbühler, Calculation of equilibrium constants from multiwavelength spectroscopic data—I: Mathematical considerations, Talanta 32 (1985) 95–101.

[25] SPECFIT: H. Gampp, M. Maeder, Ch.J. Meyer, A.D. Zuberbühler, Calculation of equilibrium constants from multiwavelength spectroscopic Data–II. SPECFIT: two user-friendly programs in basic and standard fortran 77, Talanta 32 (1985) 257–264.

[26] SPECFIT: H. Gampp, M. Maeder, Ch.J. Meyer, A.D. Zuberbühler, Calculation of equilibrium constants from multiwavelength spectroscopic data—III. Model-free analysis of spectrophotometric and ESR titrations, Talanta 32 (1985) 1113–1133;
SPECFIT: H. Gampp, M. Maeder, Ch.J. Meyer, A.D. Zuberbühler, Calculation of equilibrium constants from multiwavelength spectroscopic data—IV. Model-free least-squares refinement by use of evolving factor analysis, Talanta 33 (1986) 943–951.

[27] J. Ghasemi, A. Niazi, M. Kubista, A. Elbergali, Spectrophotometric determination of acidity constants of 4-(2-pyridylazo)resorcinol in binary methanol-water mixtures, Anal. Chim. Acta 455 (2002) 335–342.

[28] I. Scarminio, M. Kubista, Analysis of correlated spectra data, Anal. Chem. 65 (1993) 409–416.

[29] M. Kubista, R. Sjöback, J. Nygren, Quantitative spectral analysis of multicomponent equilibria, Anal. Chim. Acta 302 (1995) 121–125.

[30] J. Nygren, A. Elbergali, M. Kubista, Unambiguous characterization of a single test sample by fluorescence spectroscopy and solvent extraction without use of standards, Anal. Chem. 70 (1998) 4841–4846.

[31] L. Antonov, G. Gergov, V. Petrov, M. Kubista, J. Nygren, UV–vis spectroscopic and chemometric study on the aggregation of ionic dyes in water, Talanta 49 (1999) 99–106.

[32] J. Nygren, J.M. Andrade, M. Kubista, Characterization of a single sample by combinong thermodynamic and spectroscopic information in spectral analysis, Anal. Chem. 68 (1996) 1706–1710.

[33] BEEROZ: J. Brugger, BeerOz, a set of Matlab routines for the quantitative interpretation of spectrophotometric measurements of metal speciation in solution, Comput. Geosci. in press.

[34] SPECFIT/32: Spectrum Software Associates, 197M Boston Post Road West, Marlborough, MA, 01752 U.S.A., 2004 (http://www.bio-logic.info/rapid-kinetics/specfit.html).

[35] R. Gargallo, R. Tauler, A. Izquierdo-Ridorsa, Influence of selectivity and polyelectrolyte effects on the performance of solft modelling and hard-modelling approaches applied to the study of acid-base equilibria of polyelectrolytes by spectrometric titrations, Anal. Chim. Acta 331 (1996) 195–205.

[36] S.I. Sinkov, E.I. Bozhenko, Complexation behavior of Pu(IV) and Pu(VI) with urea in nitric acid solution, J. Alloys Compd. 271–273 (1998) 809–812.

[37] M. Bernabé-Pineda, M.T. Ramírez-Silva, M.A. Romero-Romo, E. Gonzáles-Vergara, A. Rojas-Hernández, Specrtophotometric and electrochemical determination of the formation constants of the complexes Curcumin-Fe(III)-water and Curcumin-Fe(II)-water, Spectrochim. Acta Part A 60 (2004) 1105–1113.

[38] R. Gargallo, M. Vives, R. Tauler, R. Eritja, Protonation studies and multivariate curve resolution on oligodeoxynucleotides carrying the mutagenic base 2-aminopurine, Biophys. J. 81 (2001) 2886–2896.

[39] D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, J. Soc. Ind. Appl. Math. 11 (1963) 431–441.

[40] J. Ghasemi, Sh. Nayebi, M. Kubista, B. Sjogreen, A new algorithm for the determination of protolytic constants from spectrophotometric data in multiwavelength mode: Calculations of acidity constants of 4-(2-pyridylazo)resorcinol (PAR) in mixed nonaqueous-water solvents, Talanta 68 (2006) 1201–1214.

[41] T. Khayamian, Z. Kardanpour, J. Ghasemi, A New Application of PC-ANN in Spectrophotometric Determination of Acidity Constants of PAR, J. Braz. Chem. Soc. 16 (2005) 1118–1123.

[42] G. Puxty, M. Maeder, K. Hungerbühler, Tutorial on the fitting of kinetics models to multivariate spectroscopic measurements with non-linear least-squares regression, Chemometrics Int. Lab. Syst. 81 (2006) 149–164.

[43] M.C. Aragoni, M. Arca, G. Crisponi, V.M. Nurchi, R. Silvagni, Characterization of the ionization and spectral properties of sulphonephtalein indicators. Correlation with substituent effects and structural features. Part II, Talanta 42 (1995) 1157–1163.

[44] B. Nigović, N. Kujundžić, K. Sanković, D. Vikić-Topić, Complex formation between transition metals and m2-pyrrolidone-5-hydroxamic acid, Acta Chim. Slov. 49 (2002) 525–535.

[45] A.K. Elbergali, J. Nygren, M. Kubista, An automated procedure to predict the number of components in spectroscopic data, Anal. Chim. Acta 379 (1999) 143–158.

[46] T.M. Rossi, I.M. Warner, Rank estimation of excitation–emission matrices using frequency analysis of eigenvectors, Anal. Chem. 58 (1986) 810–815.

[47] E.R. Malinowski, Factor Analysis in Chemistry, second ed., Wiley, New York, 1991.

[48] R.D. Catell, Multivariate Behav. Res. 1 (1966) 245–276.

[49] Z.-P. Chen, J.-H. Jiang, Y. Li, H.-L. Shen, Y.-Z. Liag, R.-Q Yu, Smoothed window factor analysis, Anal. Chim. Acta 381 (1999) 233–246.

[50] Li Xing, R.C. Glen, Novel methods for the prediction of log P, pK and log D, J. Chem. Inf. Comput. Sci. 42 (2002) 796–805.

[51] Pallas: http://compudrug.com/show.php?id=90, http://compudrug.com/show.php?id=36.

[52] Marvin: http://www.chemaxon.com/conf/Prediction_of_dissociation_constant_using_microconstants.pdf and http://www.chemaxon.com/conf/New_method_for_pKa_estimation.pdf.

[53] M. Meloun, J. Militký, M. Forina, Chemometrics for Analytical Chemistry, Vol. 2. PC-Aided Regression and Related Methods, Ellis Horwood, Chichester, 1994;
M. Meloun, J. Militký, M. Forina, Chemometrics for Analytical Chemistry, Vol. 1. PC-Aided Statistical Data Analysis, Ellis Horwood, Chichester, 1992.

[54] M. Meloun, V. Říha, J. Žáćek, Piston microburette for dosing aggressive liquids (in Czech), Chem. Listy 82 (1988) 765.

[55] ORIGIN, OriginLab Corporation, One Roundhouse Plaza, Suite 303, Northampton, MA 01060, USA, 2005.