A pK Determination Method for Proteins from Titration Curves using Principle Component Analysis

Jaesool Shim and Prashanta Dutta
School of Mechanical and Materials Engineering, Washington State University, Pullman, WA 99164

Cornelius F. Ivory
School of Chemical Engineering and Bioengineering, Washington State University, Pullman, WA 99164

DOI 10.1002/aic.11528
Published online July 9, 2008 in Wiley InterScience (www.interscience.wiley.com).

Principal component analysis (PCA) technique is used for the precise determination of pK values from titration curves of amphoteric molecules, such as proteins and amino acids. A regression model is developed based on the effective charge of amphoteric molecules. Partial domain method, a new algorithm, is introduced to minimize or remove errors in extracted pK values from experimental (titration curve) data with hidden errors, such as sampling errors or experimental uncertainties. For the validation of our algorithm, 3 pK values are determined from the titration curves of glutamic acid and lysine. The extracted pKs are in exact agreement with the original pKs if partial domain method is used to obtain pKs from an exact (uncertainty free) titration curve. In addition, the effects of various uncertainty levels (1%, 2% and 3% noise) in the experimental titration curves are tested using both partial and full domain methods. Analytic results show that partial domain method is very effective in reducing the errors in extracted pKs especially at larger value of redundant number. Next, partial domain based PCA method is applied to extract 5 pKs from the experimental titration curve of hen egg-white lysozyme protein. These pK values are then used to simulate the isoelectric focusing of lysozyme protein in the presence of 25 biprotic ampholytes in a 2-D (two-dimensional) straight microchannel. The transient, as well as focused state behaviors of lysozyme protein are compared between original titration curve (20 pKs) and approximated titration curve (5 pKs) cases. Although there are minor differences at the early stages of focusing, the focusing time, position, and shapes of protein and ampholytes are identical at focused state. © 2008 American Institute of Chemical Engineers *AIChE J,* 54: 2238–2249, 2008

Keywords: pK extraction, principal component analysis, titration curves, proteins, IEF

Introduction

The transport and separation of charged macromolecules, such as protein and DNA in a pH gradient buffer are highly dependent on dissociation/equilibrium constants (aka pK; $pK = -\log_{10} K$).[1–3] Due to the polarizability of charged

Correspondence concerning this article should be addressed to P. Dutta at dutta@mail.wsu.edu molecules at different pH, the dynamics of these molecules are very complicated *in vitro*/*vivo* chemical processes.[4] For this reason, the precise prediction of dissociation constants of charged macromolecules plays an essential role in studying their dynamics in various biochemical and pharmaceutical fields.[5–6] Moreover, precise prediction of pKs helps in explaining many aspects of protein behavior, as well as molecular activities.[7] For instance, the exact knowledge of pK values in ionic states is necessary to fully understand protein separation and/or purification, chemical reagent interaction with charged molecules, and protonation in chemical process.[8]

In the past, the pK calculation has normally been carried out in five ways: The mean field Poisson-Boltzmann method,[9–10] empirical methods,[11] molecular dynamics (MD)-based methods,[12–13] capillary electrophoresis method,[14–15] and titration curve method.[16–19] Neto et al.[9] presented an iterative solution of Poisson-Boltzmann equation for the determination of the pK value of monoprotonic acid, where the pH of dilution experiments was used as an input data. The empirical method which has been first developed by Li et al.[11] uses the relationship between pK values of ionizable residues and their structures. The molecular dynamic method utilizes the free-energy perturbation and the ratio of equilibrium constants for the deprotonation.[12–13] The capillary electrophoresis method is based on the observance of effective mobility of an ionizable compound in a series of electrolyte solutions.[14] This method provides a simple, automated approach for the measurement of pKa values in the range 2–11, but it requires precise ionic mobility measurements under appropriate experimental conditions. In an earlier work, Valentini et al.[15] also described a way to correlate experimental pH with mobility by isoelectric focusing (IEF). Their work revealed how to determine pK values for simple cationic and anionic species, as well as mono-mono-valent amphoteric molecules. Finally, the titration curve method employs a statistical regression on an experimentally obtained titration curve.[16] Mosher and his colleagues used protein titration curve approach for electrophoretic separation of proteins.[17–19] Although the overall theoretical framework to determine pKs is well defined in titration curve method, the experimental uncertainty of titration procedure is a major roadblock in accurate pK determination. In addition to that random extraction of data from titration curves might introduce more error in the system.

In this study, we developed an analytic method to accurately predict pKs from titration curve with aforementioned latent experimental errors or uncertainties. The procedures outlined in this article are as follows. First, we describe the theoretical procedure to predict the pK values from a titration curve. We modified the "effective mobility concept" of the ionized groups used in the capillary electrophoresis method to the "mean charge concept". Hence, we can use titration curves as our regression data, while the capillary electrophoresis method requires experimental mobilities as regression data. Second, we used the PCA (principal component analysis) technique to get rid of the latent errors of titration curves. Finally, we introduced the partial domain method to minimize the pK extraction error further. This robust analytic method is able to determine the multiple pK values from a titration curve and to reduce the pK prediction errors caused by hidden errors from the random samples, or the erroneous titration curves due to experimental uncertainties.

pK Determination by the Mean Charge in Electrophoretic Process

In a soup of partially-ionized solutes, the charge characteristics of macromolecules are determined by a set of dissociation reactions.[8,20] Hence, one can treat each component C_i, as being composed of a set of $N + 1$ species S_i, which are summed over j to recover each of the i components in the system with a total of N dissociable groups, i.e.

$$C_i = \sum_{j=1}^{N+1} S_{ij} \tag{1}$$

This relation allows us to treat all weak acids and bases, ampholytes, proteins and even water, which dissociates into hydronium and hydroxide ions, as being mathematically equivalent. Therefore, the mean charge $\langle z_i \rangle$, of a component C_i, can be defined as follows

$$\langle z_i \rangle = \frac{\sum\limits_{j=1}^{N+1} z_{ij} s_{ij}}{C_i} \tag{2}$$

where z_{ij} is the j^{th} species charge. The basic "mass-action" relationship between charge-adjacent species of the same component is

$$S_{ij+1} + H_2O = S_{ij} + H_3O^+. \tag{3}$$

S_{i1} is the most electronegative species, and S_{iN+1} is the most positive. If the dissociation reactions are fast, we also have N algebraic relations among the component i species

$$K_{ij} = \frac{C_H S_{ij}}{S_{ij+1}} \tag{4}$$

where C_H is the concentration of hydronium, and K_{ij} is the dissociation constant. Therefore, using dissociation constant (K_{ij}), Eq. 2 can be rewritten as

$$\langle z_i \rangle = \frac{\sum\limits_{j=1}^{N+1} z_{ij} S_{ij}}{\sum\limits_{ij}^{N+1} S_{ij}} = \frac{\sum\limits_{j=1}^{N} \left(z_{ij} C_H^{j-1} \prod\limits_{k=j}^{N} K_{ik} \right) + z_{iN+1} C_H^N}{\sum\limits_{j=1}^{N} \left(C_H^{j-1} \prod\limits_{k=j}^{N} K_{ik} \right) + C_H^N} \tag{5}$$

Rearranging dissociation constants, Eq. 5 yields a set of linear equation as

$$\left[\sum_{j=1}^{N} \left(z_{ij} - \langle z_i \rangle \right) \left(C_H^{j-1} \prod_{k=j}^{N} K_{ik} \right) \right] = \left(\langle z_i \rangle - z_{iN+1} \right) C_H^N \tag{6}$$

In order to obtain the solution of linear equations for $\prod_{k=j}^{N} K_{ik}$, mean charge ($\langle z_i \rangle$), and hydronium concentration (C_H) are needed from the titration curve. After extracting (m) numbers of random data ($\langle z_i \rangle$, C_H), from titration curve, the tensor form of Eq. 6 can be changed to a matrix form, as

$$[M][\Psi] = [\Gamma] \tag{7}$$

where $$[M] = \begin{bmatrix} \big(z_{i1} - \langle z_i(1)\rangle\big) & \big(z_{i2} - \langle z_i(1)\rangle\big)C_H(1) & \cdots & \big(z_{iN-1} - \langle z_i(1)\rangle\big)\big(C_H(1)\big)^{N-2} & \big(z_{iN} - \langle z_i(1)\rangle\big)\big(C_H(1)\big)^{N-1} \\ \big(z_{i1} - \langle z_i(2)\rangle\big) & \big(z_{i2} - \langle z_i(2)\rangle\big)C_H(2) & \cdots & \big(z_{iN-1} - \langle z_i(2)\rangle\big)\big(C_H(2)\big)^{N-2} & \big(z_{iN} - \langle z_i(2)\rangle\big)\big(C_H(2)\big)^{N-1} \\ \vdots & \vdots & & \vdots & \vdots \\ \big(z_{i1} - \langle z_i(m)\rangle\big) & \big(z_{i2} - \langle z_i(m)\rangle\big)C_H(m) & \cdots & \big(z_{iN-1} - \langle z_i(m)\rangle\big)\big(C_H(m)\big)^{N-2} & \big(z_{iN} - \langle z_i(m)\rangle\big)\big(C_H(m)\big)^{N-1} \end{bmatrix}$$

$$[\Psi] = \begin{bmatrix} \prod_{k=1}^{N} K_{ik} \\ \prod_{k=2}^{N} K_{ik} \\ \vdots \\ \prod_{k=N-1}^{N} K_{ik} \\ \prod_{k=N}^{N} K_{ik} \end{bmatrix} \quad \text{and} \quad [\Gamma] = \begin{bmatrix} \big(\langle z_i(1)\rangle - z_{iN+1}\big)\big(C_H(1)\big)^N \\ \big(\langle z_i(2)\rangle - z_{iN+1}\big)\big(C_H(2)\big)^N \\ \vdots \\ \big(\langle z_i(m-1)\rangle - z_{iN+1}\big)\big(C_H(m-1)\big)^N \\ \big(\langle z_i(m)\rangle - z_{iN+1}\big)\big(C_H(m)\big)^N \end{bmatrix}.$$

Matrices M and Γ are known from titration curve data, while matrix Ψ contains our desired outputs. From Eq. 7, the matrix Ψ can be obtained as

$$[\Psi]_{N\times 1} = [M]^{+}_{N\times m}[\Gamma]_{m\times 1} \tag{8}$$

where the superscript "+" denotes the pseudo inverse based on least-square method. In general, when the matrices M and Γ have errors due to input and output noises, a system of equations forms an ill conditioned problem, which introduces significant uncertainty in the results primarily due to rank deficiency of the coefficient matrix M. To avoid such uncertainties, PCA has been widely used as a classical statistical method.[21–22] The PCA is a way of identifying eigenvalues in the coefficient matrix and removing the eigenvalues which may cause errors.[23] In the following section we briefly describe the PCA method for our problem.

In any matrix system, the coefficient matrix $M_{m\times N}$ can be represented by using singular value decomposition

$$[M] = [U][S][V]^H \tag{9}$$

where U is the $m \times m$ unitary matrix, V is the $N \times N$ unitary matrix, the superscript H is the Hermitian, and S is a $m \times N$ diagonal matrix, i.e.

$$[S] = \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix} \tag{10}$$

where T is a diagonal matrix having positive diagonal elements in descending order as

$$[T] = diag(s_1, s_2, \ldots, s_r) \quad \text{and} \quad s_1 \geq s_2 \geq \cdots \geq s_r \tag{11}$$

When some eigenvalues ($S_{i,i\leq r}$) of the diagonal matrix are (T) less than a threshold ($\varepsilon = 10^{-6}$), those s_i values are set to zero. Thus

$$[M]^{+}_{N\times m} = [V]_{N\times N}[S]^{+}_{N\times m}[U]^{H}_{m\times m} \tag{12}$$

where

$$[S]^{+}_{N\times m} = \begin{bmatrix} \Lambda^{+} & 0 \\ 0 & 0 \end{bmatrix} \tag{13}$$

and

$$\Lambda^{+} = diag\left(\frac{1}{s_1}, \frac{1}{s_2}, \ldots, \frac{1}{s_{i-1}}, 0, \ldots, 0\right) \tag{14}$$

Using Eqs. 8–14, one can obtain the matrix Ψ. Now, rearranging the matrix Ψ yields the dissociation constants K as

$$P_i = \begin{bmatrix} K_{iN} \\ K_{iN-1} \\ \vdots \\ K_{i2} \\ K_{i1} \end{bmatrix} = \begin{bmatrix} \prod_{k=N}^{N} K_{ki} \\ \left(\prod_{k=N-2}^{N} K_{ik}\right) \Big/ \left(\prod_{k=N-1}^{N} K_{ik}\right) \\ \vdots \\ \left(\prod_{k=2}^{N} K_{ik}\right) \Big/ \left(\prod_{k=3}^{N} K_{ik}\right) \\ \left(\prod_{k=1}^{N} K_{ik}\right) \Big/ \left(\prod_{k=2}^{N} K_{ik}\right) \end{bmatrix} \tag{15}$$

Note that the number of data points (m) extracted from the titration curve needs to be larger than the number of dissociation constants (N), to obtain an average statistical effect from a titration curve. The average effects of titration data is of great importance for cases where the fluctuations of the titration curve occur.

Mathematical Model of Ampholyte Based Isoelectric Focusing

The mathematical model for ampholyte based IEF is governed by a set of mass conservation equations for each component (i) as

$$\frac{\partial C_i}{\partial t} + \nabla \cdot \left[\langle \mu_i \rangle \vec{E} C_i - D_i \nabla C_i\right] = 0. \tag{16}$$

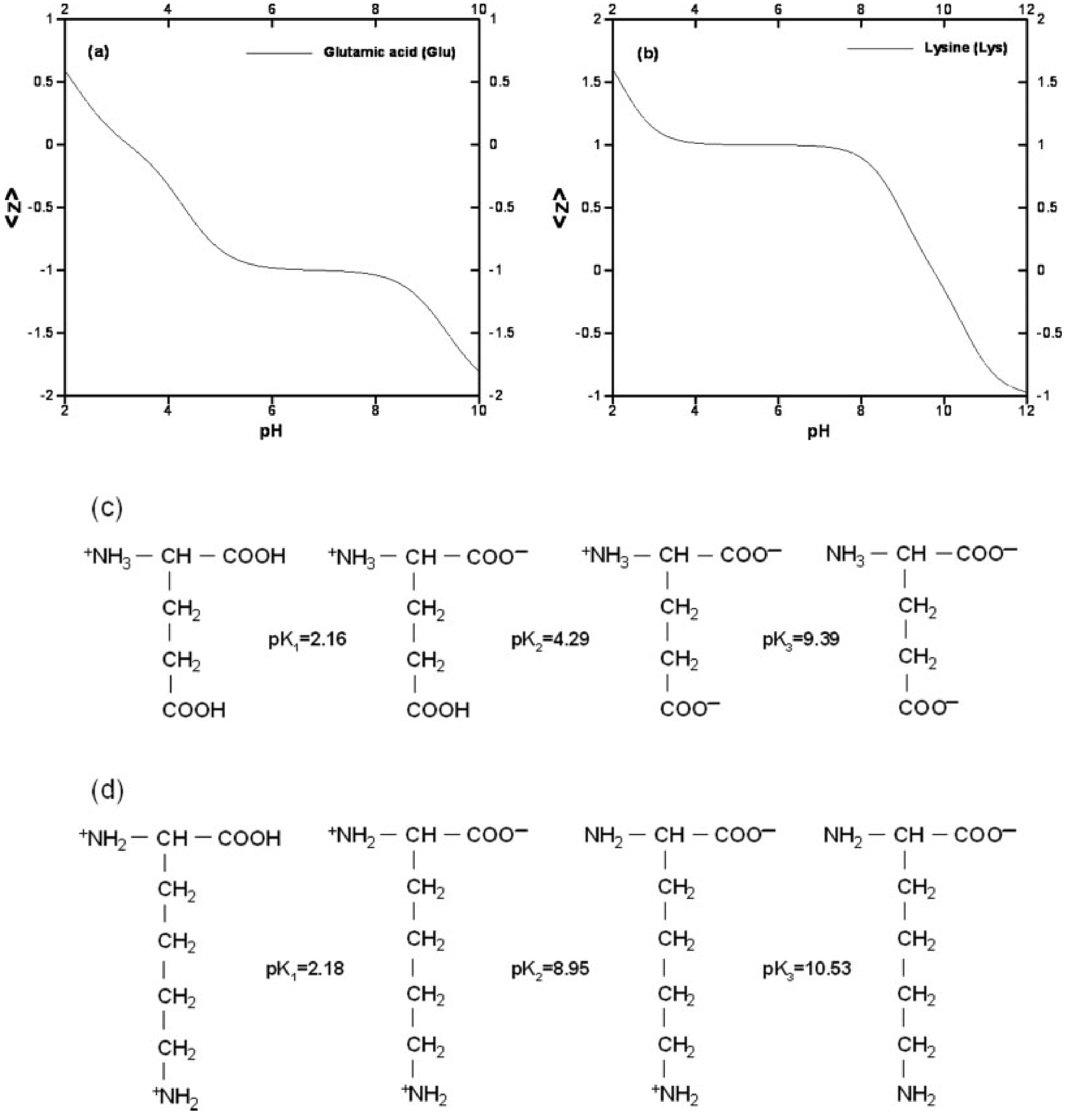

Figure 1. Titration curves (a-b) and the 4 possible structures (c-d) of glutamic acid and lysine.[26–27]

For both glutamic acid and lysine, the structures (left to right) are cation, zwitterions, R-group, and anion. The pK values of glutamic acid are 2.16, 4.29 and 9.39 and those of lysine are 2.18, 8.95 and 10.53.

the charge conservation as

$$\nabla \cdot \left(\sigma \vec{E}\right) = F \sum_{i=1}^{L} \sum_{j=1}^{N+1} z_{ij} D_{ij} \nabla \cdot \nabla S_{ij} \tag{17}$$

and the electroneutrality equation as

$$F \sum_{i=1}^{L} \langle z_i \rangle C_i = 0 \tag{18}$$

where σ is the conductivity of the buffer, which can be defined as $\sigma = F \sum_{i=1}^{L} \sum_{j=1}^{N+1} z_{ij} \mu_{ij} S_{ij}$, F is the Faraday constant, L is the total number of ionic components, $\vec{E}$ is the applied electric field, D_{ij} is the diffusion constants of each species (j), and $\langle \mu_i \rangle$, and D_i are the electrophoretic (effective) mobility and diffusion constant of component (i), respectively. The electrophoretic mobility, $\langle \mu_i \rangle$ can be expressed as $\langle z_i \rangle \omega_i$, where ω_i is the absolute mobility. Equation 17 present one mass conservation equation for each component (C_i) including hydronium and hydroxyl ions. Note that the algebraic

Table 1. Comparison of pK Extraction Errors Between Partial Domain Method and Full Domain Method

Method	Glutamic Acid (Glu)								Lysine (Lys)							
	Full Domain Method				Partial Domain Method				Full Domain Method				Partial Domain Method			
Noise (%)	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
pK_1 (% Error)	2.16	2.16	2.16	2.18	2.16	2.16	2.16	2.16	2.18	2.18	2.21	2.21	2.18	2.18	2.18	2.18
	(0)	(0)	(0)	(0.9)	(0)	(0)	(0)	(0)	(0)	(0)	(1.4)	(1.4)	(0)	(0)	(0)	(0)
pK_2 (% Error)	4.29	4.30	4.16	3.98	4.29	4.29	4.28	4.28	8.95	6.56	5.36	5.17	8.95	8.95	8.95	8.94
	(0)	(0.2)	(3)	(7)	(0)	(0)	(0.2)	(0.2)	(0)	(26)	(40)	(42)	(0)	(0)	(0)	(0.1)
pK_3 (% Error)	9.39	5.84	4.26	4.02	9.39	9.27	8.88	8.74	7.54	5.50	4.92	4.85	10.53	10.53	10.50	10.49
	(0)	(38)	(55)	(55)	(0)	(1.2)	(5.4)	(6.9)	(28)	(47)	(53)	(54)	(0)	(0)	(0.3)	(0.4)

The exact values of pK_1, pK_2 and pK_3 are 2.16, 4.29 and 9.39 for glutamic acid, and 2.18, 8.95, 10.53 for lysine, respectively. 4000 data points (*m*) are used for both glutamic acid and lysine.

electroneutrality condition allows us to eliminate one differential equation from the set which needs to be solved numerically; it is convenient to choose to eliminate hydronium (hydroxyl) equation. Now using equilibrium reaction $K_W = C_{OH}C_H$, one can write the modified electroneutrality equation as

$$C_H - \frac{K_W}{C_H} = -\sum_{i=1}^{L-2} \langle z_i \rangle C_i. \tag{19}$$

where C_H and C_{OH} are the hydronium and hydroxyl concentration, and K_W is the equilibrium constant.

This IEF model neglects heat generation due to Joule heating effects, i.e., temperature is assumed constant throughout the channel. Electrokinetic flow is not considered here since IEF channel is coated with methylcellulose or other chemicals to suppress electroosmosis.[24–25] For the species that make up a particular component, the absolute mobilities and diffusivities of all charge states are assumed to be same.

The mass conservation equations are subjected to zero net flux at the anolyte and catholyte reservoirs, and on the channel surfaces. The boundary conditions for the charge conservation equation consist of imposition of insulating boundary conditions $\left(\vec{E} \cdot \vec{n} = 0\right)$ on the solid channel surfaces and constant electric potentials at the anolyte and catholyte reservoirs.

Results and Discussion

Principal component analysis (PCA) is used to remove the hidden errors of titration curve data, and the partial domain method is introduced to find pK values from titration curve with higher accuracy. This new analytic technique is validated for both simple acidic and basic functional groups and real protein (Hen egg-white lysozyme) to specifically demonstrate the usefulness of this algorithm in handling components with multiple charge states.

Model validation for amino acids

All peptides and polypeptides, including proteins, consist of a sequence of polymers of α-amino acids. There are 20 different α-amino acids, and each α-amino acid can be distinguished by the carboxylic acid (—COOH), the amino (—NH_2) and the ionizable R-groups. Thus, there are four possible structures of the α-amino acid, such as cation, zwitterion, R-group and anion, and they are related by the equilibrium constants.

We chose the glutamic acid (Glu) and lysine (Lys) as the sample acidic and basic functional groups for the pK extraction from their titration curves. Figure 1 shows the titration curves of amino glutamic acid and lysine along with their different charge states. For amino glutamic acid (Figure 1c), the pK values are 2.16, 4.29 and 9.39, respectively, while the pK values of lysine (Figure 1d) are 2.18, 8.95 and 10.53, respectively.[26] Both full domain and partial domain methods are employed for extraction of pKs from their titration curves. The full domain analysis uses the entire titration curve as a computational domain for PCA method. On the other hand, the partial domain analysis divides the titration curves into a number of subdomains, and each subdomain is analyzed individually using the PCA method. For example, in Figure 1a, the full domain analysis uses the entire pH range between 2 and 10 to extract *m* data points in order to find pK_1, pK_2 and pK_3 of glutamic acid. However, the partial domain analysis of glutamic acid is carried out in two divided pH ranges: $2 \leq pH \leq 6$ and $6 \leq pH \leq 10$. In other words, matrices M and Γ are constructed separately for each subdomain using Eqs. 6–7. Table 1 shows the regression results obtained from the PCA technique described in the pK Determination by the Mean Charge in Electrophoretic Process section for both full and partial domain methods. Note that full domain method estimates all three pKs, while partial domain method quantifies pK_1 and pK_2 from the first domain and pK_3 and pK_2 from the second domain. In the partial domain method, pK_2 values come from both subdomains. It is noteworthy to mention that the value of pK_2 obtained from each subdomain is different, but these repeating pK_2 are very close (less than 10% difference). Thus, for this case, the selection of proper pK_2 is guided by the pH of the domain. For instance, in lysine, the pK_2 predicted from domain 1 ($2 \leq pH \leq 7$) is 8.93, while the pK_2 estimated from domain 2 ($7 \leq pH \leq 12$) is 8.95. Hence, we choose 8.95 as the proper value of pK_2 for lysine. A similar technique is used for the pK_2 of glutamic acid, where pK_2 is estimated from the domain 1 ($2 \leq pH \leq 6$).

For glutamic acid, both full domain and partial domain methods estimate the exact pKs if the titration curves data are error free. However, the full domain analysis fails to predict the exact value of pK_3 for lysine despite the error free (0% noise) titration curve used for this analysis. However, the partial domain analysis can provide a very good estima-

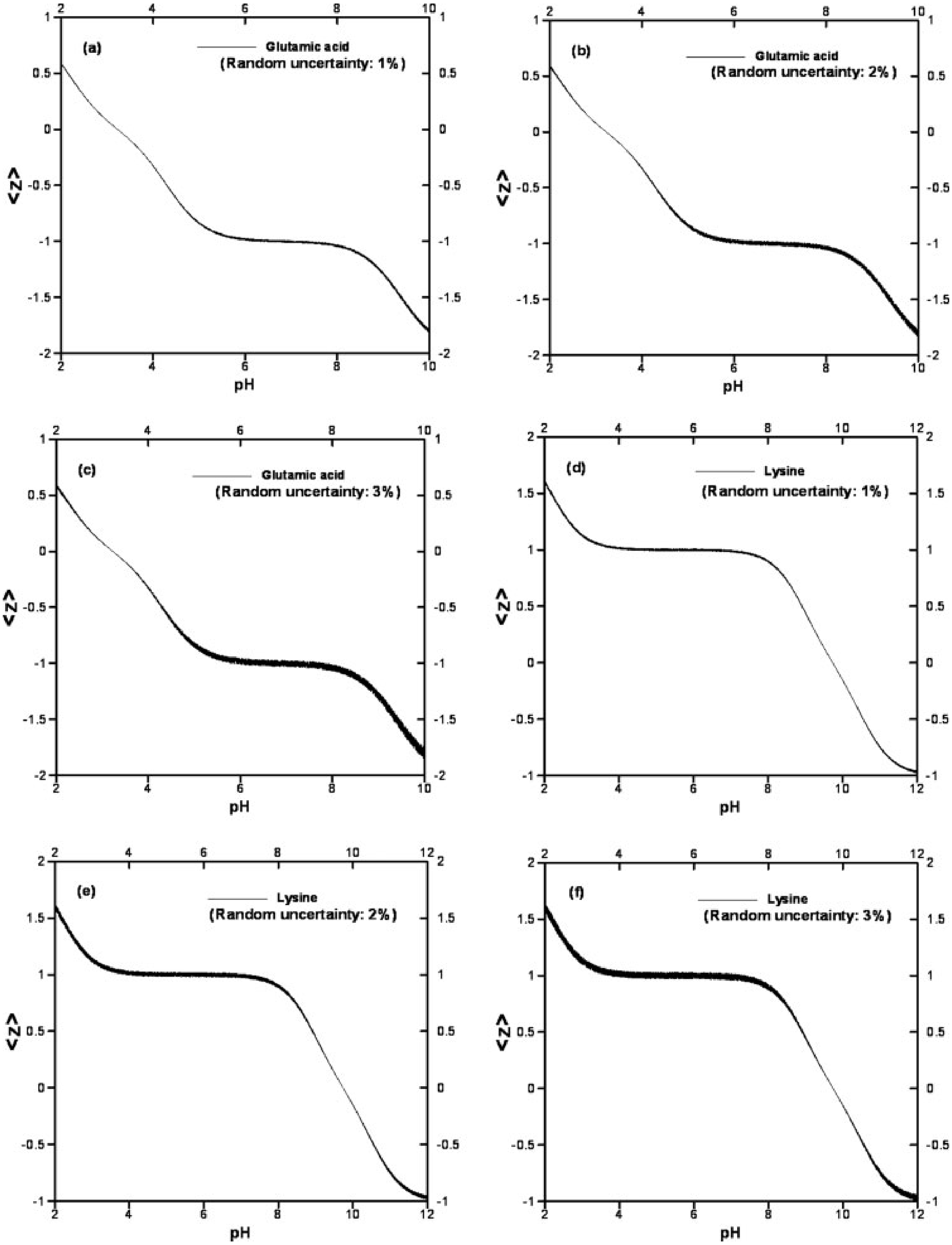

Figure 2. Titration curves with 1%, 2% and 3% random noises for glutamic acid (a, b, and c), and lysine (d, e, and f), respectively.

tion for all pK values. This discrepancy is due to the fact that our method determines the product of dissociation constants ($\prod_{k=1}^{N} K_{ik}$) using pseudo-inverse technique. Hence, the error in the solution increases if the number of dissociation constants (Ks) increases, or if the pK values are more inclined toward the highly basic side.

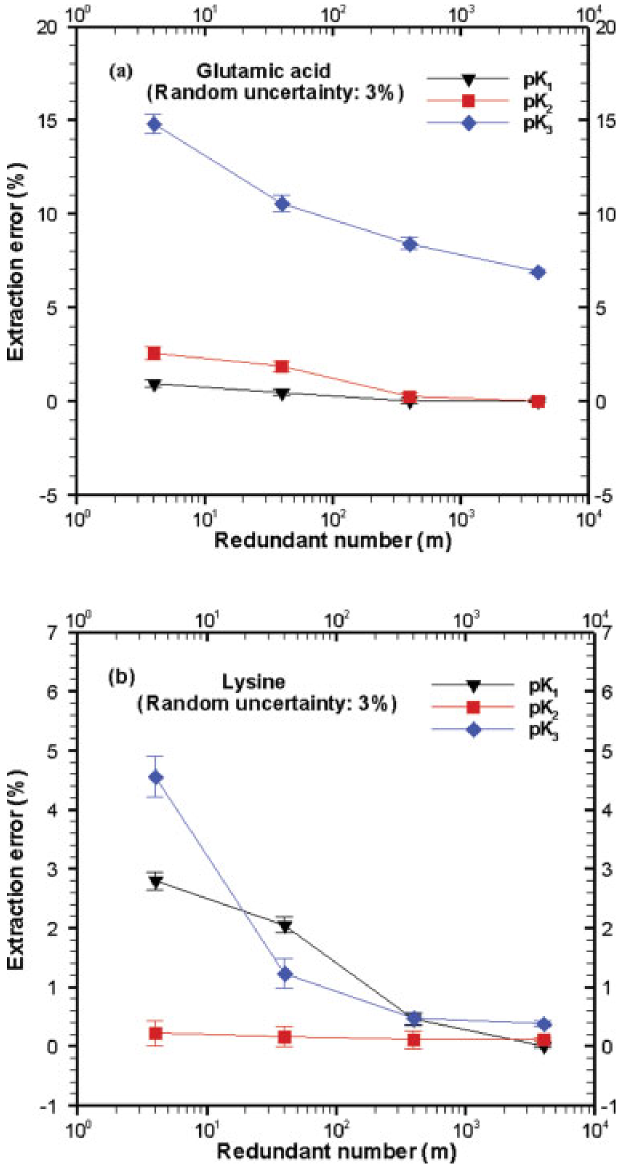

Figure 3. The effect of the number of input data on the prediction error.

All prediction errors are calculated for 3% random uncertainty/noise by the partial domain method for (a) glutamic acid, and (b) lysine. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Effects of experimental uncertainty

In practice, all experiments have some errors due to experimental uncertainty. For this reason, the effect of data error on the prediction of pKs is important to demonstrate the usefulness of the proposed analytic method. To investigate this, we conducted the error analysis by introducing various levels of random noise (1%, 2% and 3%) into the original titration curves (Figure 2). Errors in predicted pK values are shown in Table 1 for glutamic acid and lysine under different experimental uncertainties/noises, using both full and partial domain methods. For both amino acids, the full domain method provides the same pK_1 estimations with those of the partial domain method for various uncertainty levels. However, for pK_2 and pK_3, the error in full domain method grows drastically with the experimental noises. These prediction errors can be reduced significantly by employing partial domain method for both glutamic acid and lysine (Table 1). Thus, it proves that the partial domain method is a very good analytic tool to determine pK values from titration information. For partial domain method, we also studied the effects of the number of input data (m) from the titration curve for various levels of experimental uncertainties. Figure 3 shows the estimation error reduction with numbers of input data or redundant number (m). Note that for prediction of three pKs (pK_1, pK_2 and pK_3), at least three data points are needed to form a matrix 3 × 3; so we conducted the pK estimations for $m = 4$, 40, 400 and 4,000. In the partial domain method, the estimation error decreases with numbers of input data. From this analytic

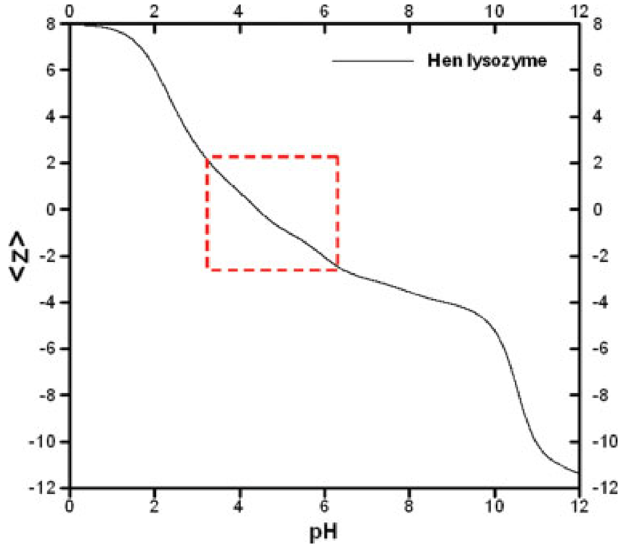

Figure 4. Titration curve of hen egg-white lysozyme protein constructed by amino acid composition.

The dotted block denotes the domain of interest ($3 \leq pH \leq 6$) for extracting pK values by the partial domain method. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 2. Residue and Apparent pK Values of Hen Egg-White Lysozyme (adapted from[2])

Acidic	pK	Basic	pK
Glu7	2.6	Arg	12
Glu35	6.1	Lys1	10.8
Asp48	4.3	Lys13	10.5
Asp52	3.4	Lys33	10.6
Asp66	1.6	Lys96	10.8
Asp101	4.5	Lys97	10.3
Asp18	2	Lys116	10.4
Asp87	2.1	His15	5.8
Asp119	2.5	α-NH2	7.9
Cys	9	–	–
Tyr20	10.3	–	–
Tyr23	9.8	–	–
Tyr53	12.1	–	–
α-COOH	3.1	–	–

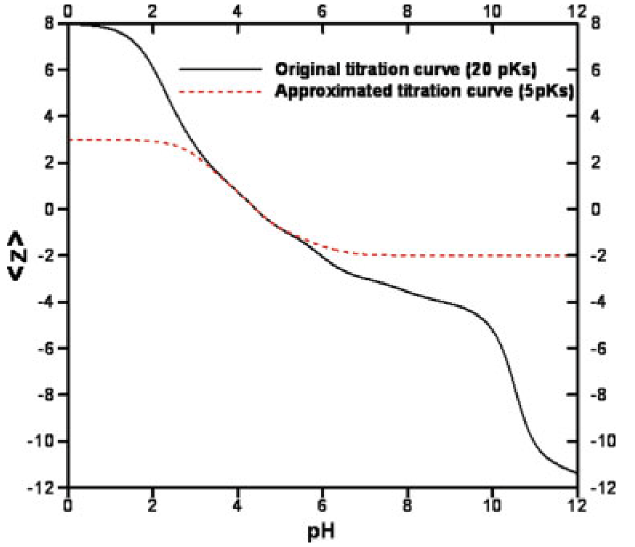

Figure 5. Comparison between the original (20 pKs) and approximated (5 pKs) titration curves.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

study, we can conclude that the partial domain based estimation method has noise stability on the latent noise sources.

Table 3. Electrochemical Properties of 25 Biprotic Ampholytes Used for IEF Simulation

	pK			
Component	pK_1	pK_2	pI	Absolute Mobility
Ampholyte 1	1.50	4.50	3.00	3.00E-08
Ampholyte 2	1.75	4.75	3.25	3.00E-08
Ampholyte 3	2.00	5.00	3.50	3.00E-08
Ampholyte 4	2.25	5.25	3.75	3.00E-08
Ampholyte 5	2.50	5.50	4.00	3.00E-08
Ampholyte 6	2.75	5.75	4.25	3.00E-08
Ampholyte 7	3.00	6.00	4.50	3.00E-08
Ampholyte 8	3.25	6.25	4.75	3.00E-08
Ampholyte 9	3.50	6.50	5.00	3.00E-08
Ampholyte 10	3.75	6.75	5.25	3.00E-08
Ampholyte 11	4.00	7.00	5.50	3.00E-08
Ampholyte 12	4.25	7.25	5.75	3.00E-08
Ampholyte 13	4.50	7.50	6.00	3.00E-08
Ampholyte 14	4.75	7.75	6.25	3.00E-08
Ampholyte 15	5.00	8.00	6.50	3.00E-08
Ampholyte 16	5.25	8.25	6.75	3.00E-08
Ampholyte 17	5.50	8.50	7.00	3.00E-08
Ampholyte 18	5.75	8.75	7.25	3.00E-08
Ampholyte 19	6.00	9.00	7.50	3.00E-08
Ampholyte 20	6.25	9.25	7.75	3.00E-08
Ampholyte 21	6.50	9.50	8.00	3.00E-08
Ampholyte 22	6.75	9.75	8.25	3.00E-08
Ampholyte 23	7.00	10.00	8.50	3.00E-08
Ampholyte 24	7.25	10.25	8.75	3.00E-08
Ampholyte 25	7.50	10.50	9.00	3.00E-08
UNIT				[cm^2/Vs]

The initial (uniform) concentration of each ampholyte is 1.0 mM.

Model validation for proteins

Unlike simple amino acid groups and/or ampholytes, the possible structures of a protein are very complicated. Besides being diverse and varied in ionic solutes, a protein has a variety of pK values that determine its characteristics. Hence, to simulate the transient behavior of proteins in ampholyte based IEF, it is important to know the pK values from the titration curves. So far, the roughly approximated pK values have been employed for the IEF simulation. In this work, we describe how to extract dominant pK values from a titration curve as inputs for computer simulation of IEF using partial domain method.

First, a titration curve is formed for hen lysozyme protein using amino acid composition method[26–27] from experimental data of Kuramitsu and Hamaguchi.[2] Figure 4 shows the titration curves obtained from protein structures/residues and their experimental pK values are presented in Table 2. The p*I* value of the hen egg-white Lysozyme is determined to be 4.45. Next, we employed partial domain method on the constructed titration curve (Figure 4) in order to estimate the dominant pK values to determine the protein behavior in the transient IEF. Although titration curve is constructed for the full pH range ($0 \leq pH \leq 12$), a pH range close to protein p*I* point is selected for the partial domain method. For instance, the pH range considered for hen egg-white lysozyme (p*I* = 4.45) is between 3 and 6. We predicted five pK values by

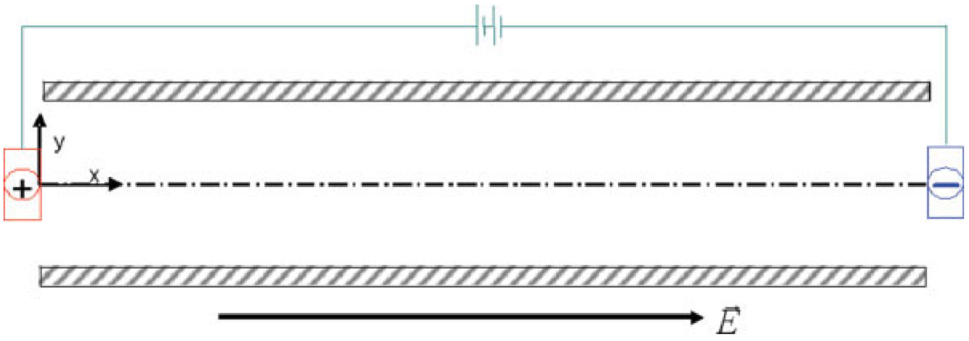

Figure 6. 2-D planar microchannel (1 cm × 100 μm) used to simulate the ampholyte based isoelectric focusing in a pH range of 3 to 9.

The anodic potential is 300 V at the left, while the cathode is set to be ground (0 V) at the right. The initial concentration of ampholytes and protein is uniform thoughout the channel. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 4. Electrochemical Properties of Model Protein Used in the IEF Simulation

pK	Original Titration Curve (20 pKs)	Approximate Titration Curve (5 pKs)
pK_1	1.6	3.1
pK_2	2.0	3.4
pK_3	2.1	4.3
pK_4	2.5	4.5
pK_5	2.6	5.6
pK_6	3.1	–
pK_7	3.4	–
pK_8	4.3	–
pK_9	4.5	–
pK_{10}	5.8	–
pK_{11}	6.1	–
pK_{12}	7.9	–
pK_{13}	9.8	–
pK_{14}	10.3	–
pK_{15}	10.4	–
pK_{16}	10.5	–
pK_{17}	10.6	–
pK_{18}	10.8	–
pK_{19}	10.8	–
pK_{20}	12.1	–

The initial (uniform) concentration of protein is 0.1 mM for the simulation. The absolute mobility (ω) of protein is 3.79E-05 cm^2/Vs for both cases.

the partial domain method to reduce the computational cost and time. The predicted pK values are 3.1, 3.4, 4.3, 4.5 and 5.6. For the validation of these pK values, we recalculated the titration curve using amino acid composition method.[27] Figure 5 illustrates both titration curves: the solid line is the titration curve calculated from 20 pK values, and the dotted line is from 5 pKs. The titration curve calculated from the 5 pKs is in a good agreement with that obtained from 20 pKs within the pH range between 3 and 6. This is because the pH range of 3 to 6 is used to extract five pK values using partial domain method.

IEF simulation

A 2-D finite volume method is used to solve the mass conservation equations given by Eq. 16, together with the charge conservation equation defined in Eq. 17. Details of the numerical scheme are presented in an earlier publication.[8] Briefly, discretized algebraic equations are obtained at each grid point for the mass and charge conservation equations. The power-law scheme is used to form coefficients of algebraic equations.[28] The tridiagonal matrix algorithm (TDMA) is used to solve the discretized algebraic equations along a grid line, and a line by line iteration is employed until

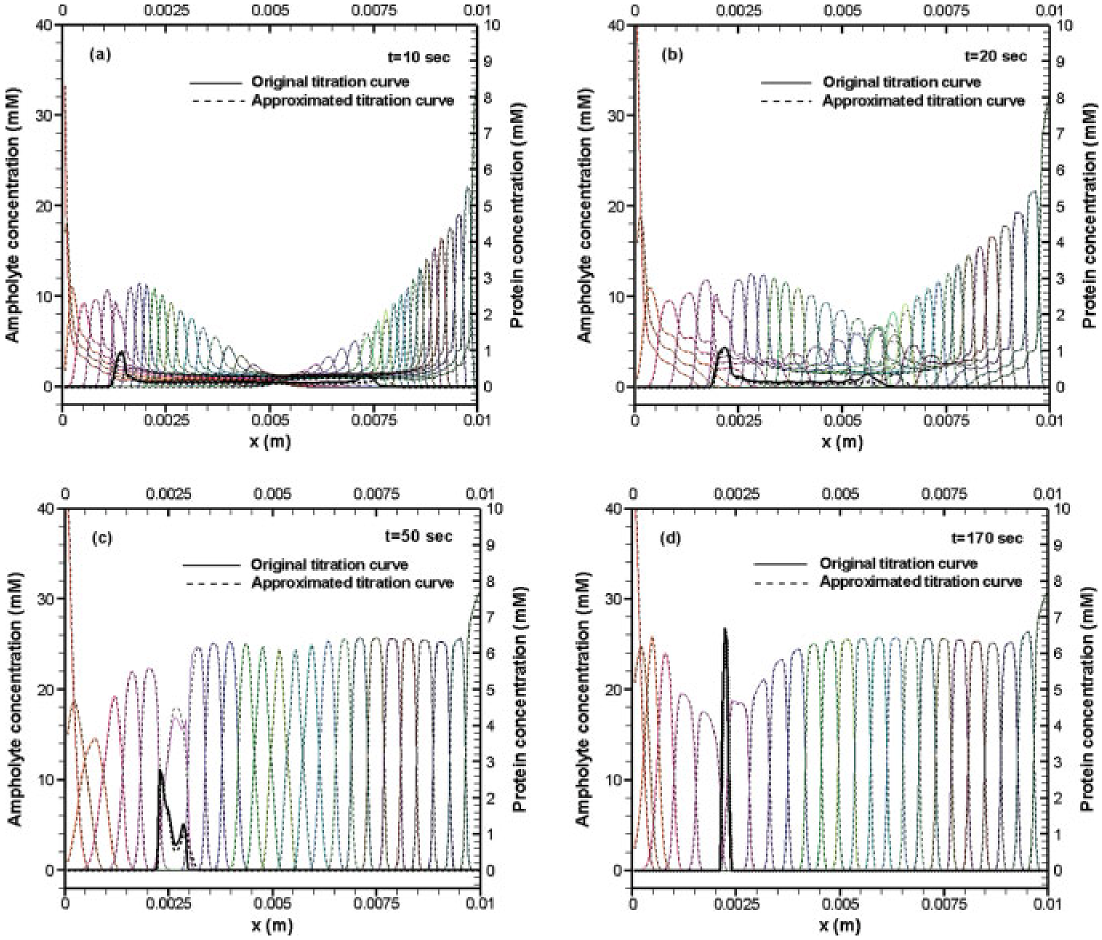

Figure 7. Comparison of transient behaviors of ampholytes and protein for original and approximated titration curves.

The numerical results are extracted at the channel centerline shown in Figure 6. Transient results are obtained at (a) 10 s, (b) 20 s, and (c) 50 s. The simulation reaches focused state at 170 s (d). The nominal electric field used for this simulation is 300 V/cm. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

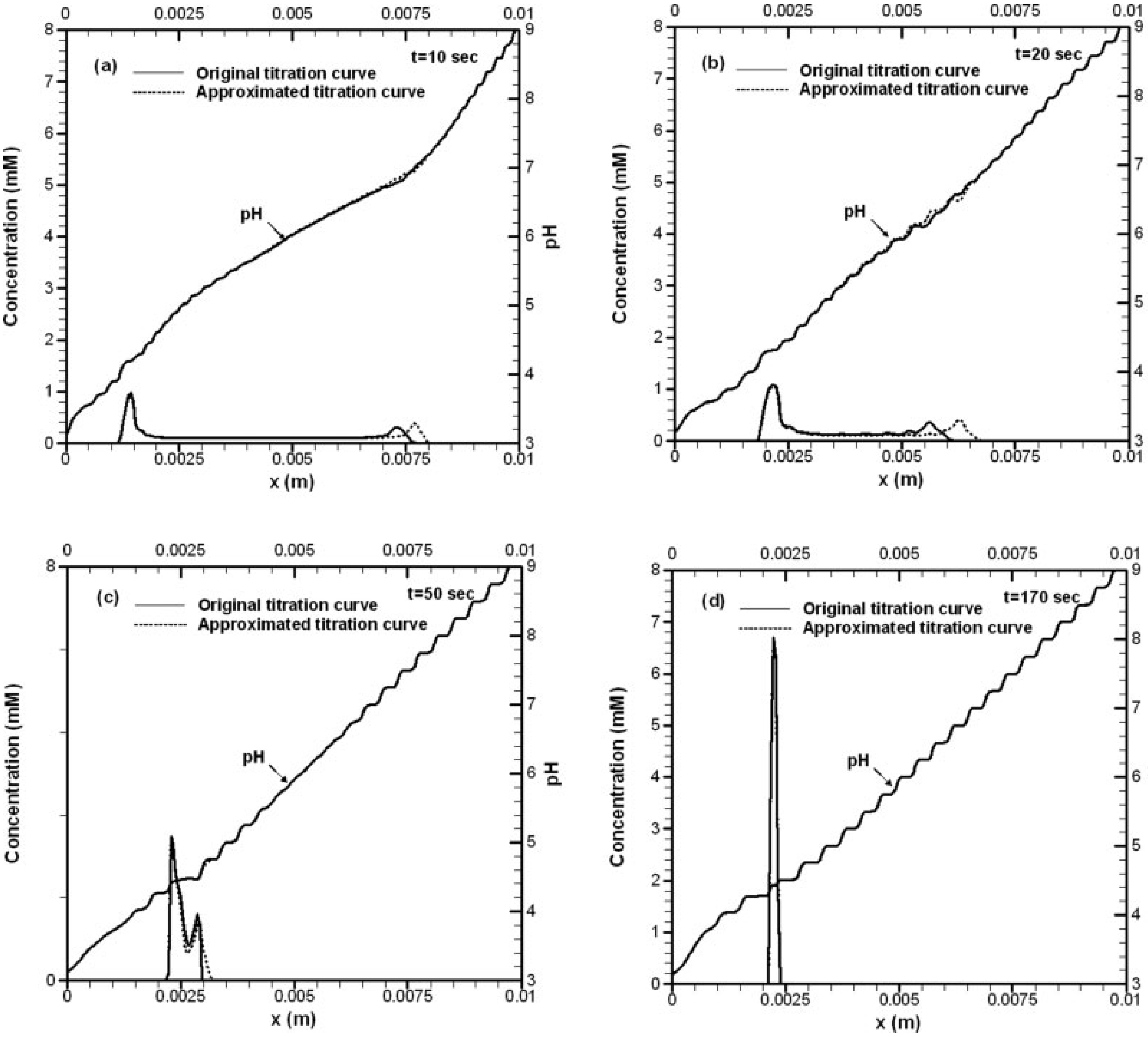

Figure 8. Comparison of pH curves and protein shape for original and approximated titration curves.

All other simulation conditions remain the same as Figure 7.

converged results are obtained throughout the computational domain. In the case of the electroneutrality equation, the Newton-Raphson method is used to obtain the concentration of hydronium (hydroxyl) ions. In our simulation, the convergence criteria are 10^{-4} for mass conservation, and 10^{-5} for charge conservation and chemical electroneutrality.

For the IEF simulation, 25 biprotic carrier ampholytes are selected within a pH range of 3 to 9, and hen lysozyme protein is allowed to focus in a 1 cm long straight microchannel (Figure 6) under a nominal electric field of 300 V/cm. For comparison purpose, two titration curves are considered for this protein: The original titration curve based on 20 pKs and an approximated titration curve constructed based on 5 extracted pKs (Figure 5). The physicochemical properties of ampholytes and proteins are presented in Tables 3 and 4, respectively. Although our numerical model can handle distinct mobility values for each species, as well as component, the same absolute mobilities of ampholytes are considered for all species. Initially protein and each ampholyte are uniformly distributed throughout the channel at a concentration of 0.1 mM and 1 mM, respectively.

The main objective of this IEF simulation is to demonstrate the ability of approximated titration curve obtained from our proposed technique (partial domain method) in replicating the properties of an experimentally obtained (original) titration curve. The transient, as well as focused state behavior of protein and ampholytes are illustrated in Figure 7 for both the original (20 pKs), and approximated (5 pKs) titration curves. It is important to note that both titration curves overlap in the near pI region ($3 \leq pH \leq 6$), but they deviate at other pH. Figure 7 shows that both protein and ampholytes form two peaks initially at the left and right side of the channel, and these peaks gradually move toward their pI points before forming one peak for each component. This is a signature characteristic of ampholyte-based IEF in a straight channel.[8] One important point to note that the migration speeds of protein double peaks are different for the original and approximated titration curves, although the same absolute mobility is used for protein in both numerical simulations. During the transient states, the left (anodic) peaks of protein for both titration curves stay at the same position, but the right peaks of the protein migrate toward the anodic side with different speeds (Figure 7a,b). This is because the approximated titration curve has zero gradients beyond the near p*I* region, while the original titration curve maintains charge dependence throughout the pH region ($0 \leq pH \leq 12$) as shown in Figure 5. However, at 50 s, as two peaks of protein approach to their focal point (p*I*), the speed-lag of right peak is mitigated due to the same shape of titration curve near the p*I* region (Figure 7c). For both cases, the protein and ampholytes reached focused state at 170 s (Figure 7d), and the focused protein has the identical shape for both original and approximated titration curves. This speed-lag of right protein peak during the focusing process can be further explained from the pH profile presented in Figure 8. In Figures 8a,b, the pH curves are slightly different between two simulations. The pH curves are affected by the nonlinear coupling between protein species and ampholytes, through mass and charge interaction. However, the pH curves become almost same as the two peaks start merging, and eventually two pH curves form the same shape when a focused state is reached at 170 s. (Figure 8d). This simulation suggests that one should consider the protein pK values very carefully close to the protein focal point (p*I*), because the titration curve around the protein's focal point is of primary importance for the final shape of the protein distribution.

Conclusions

PCA technique has been successfully applied to determine pK values from titration curves of amino acids and protein. The partial domain method provided more accurate results compared to the full domain method for various acidic and basic amino acids, such as glutamic acid and lysine. The partial domain method is particularly effective if the titration curve contains hidden errors due to experimental uncertainty or sampling error. For the partial domain method, the maximum errors in predicting pKs are less than 7 and 2% for glutamic acid and lysine, respectively, even though the titration curve contains significant experimental noise or uncertainty. In addition, in the partial domain method, the pK extraction error can be reduced significantly by increasing the number of data points from the titration curve. The partial domain method was also very effective in extracting pK values from the titration curve of hen egg-white lysozyme protein. Although experimental titration results show 20 pKs for lysozyme protein, only 5 pK values are extracted in the near p*I* region using partial domain method to minimize the simulation time and expenses. Our finite volume based numerical technique reveals that the electrophoretic characteristics of lysozyme protein differ only slightly if 5 extracted pK values are used as opposed to 20 original pKs. However, the focused state behavior of lysozyme protein and ampholytes are exactly same for original titration curve (20 pKs), and approximated titration curve (5 pKs). Simulation results also show that the focusing time is same for both cases, and a steady state can be reached in a 2-D microchannel in less than three minutes for a nominal electric field of 300 V/cm.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. CBET0626471.

Literature Cited

1. Kaufman JJ, Semo NM, Koski WS. Microelectrometric titration measurement of the p/Ca's and partition and drug distribution coefficients of narcotics and narcotic antagonists and their pH and temperature dependence. *J Med Chem.* 1975;18:647–655.
2. Kuramitsu S, Hamaguchi K Analysis of the acid-base titration curve of hen lysozyme. *J Biochem.* 1980;87:1215–1219.
3. Peacocke A. Historical article: Titration studies and the structure of DNA. *TRENDS Biochem Sci.* 2005;30:160–162.
4. Lucas LH, Ersoy BA, Kueltzo LA, Joshi SB, Brandau DT, Thyagarajapuram N, Peek LJ, Middaugh CR. Probing protein structure and dynamics by second-derivative ultraviolet absorption analysis of cation-pi interactions. *Prot Sci.* 2006;15:2228–2243.
5. van Vlijmen HW, Schaefer M, Karplus M. Improving the accuracy of protein pK(a) calculations: Conformational averaging versus the average structure. *Prot: Struct Funct Genet.* 1998;32:145–158.
6. Szakacs Z, Hagele G. Accurate determination of low pK values by H-1 NMR titration. *Talanta.* 2004;62:819–825.
7. Tynan-Connolly BM, Nielsen JL. pKD: re-designing protein pK(a) values. *Nucleic Acids Res.* 2006;34:48–51.
8. Shim J, Dutta P, Ivory CF. Modeling and simulation of IEF in 2-D microgeometries. *Electrophoresis.* 2007;28:572–586.
9. Neto AA, Filho ED, Fossey MA, Neto JR. **pK** determination. A mean field, Poisson-Boltzmann approach. *Phys Chem B.* 1999;103: 6809–6814.
10. Kahyaoglu A, Jordan, F. Direct proton magnetic resonance determination of the pKa of the active center histidine in thiolsubtilisin. *Prot Sci.* 2002;1:965–973.
11. Li H, Robertson AD, Jensen JH, Very fast empirical prediction and rationalization of protein pK(a) values. *Prots: Struct Funct Bioinform.* 2005;61:704–721.
12. Ivanov I, Chen B, Raugei S, Klein ML. Relative pK(a) values from first-principles molecular dynamics: The case of histidine deprotonation. *J Phys Chem B.* 2006;110:6365–6371.
13. Doltsinis NL, Sprik M. Theoretical pK(a) estimates for solvated P(OH)(5) from coordination constrained Car-Parrinello molecular dynamics. *Phys Chem Chem Phys.* 2003;5:2612–2618.
14. Salwa KP, Sneha P, Karen D, Heather W, Colin FP. Determination of acid dissociation constants by capillary electrophoresis. *J Chromatogr A.* 2004;1037:445–454.
15. Valentini L, Gianazza E, Righetti PG. pK determinatinations via pH-mobility curves obtained by isoelectric focusing electrophoresis. *J Biochem Biophys.* 1980;6:323–338.
16. Meloun M, Syrovy T, Bordovska S, Vrana A. Reliability and uncertainty in the estimation of pK (a) by least squares nonlinear regression analysis of multiwavelength spectrophotometric pH titration data. *Anal Bioanal Chem.* 2007;387:941–955.
17. Mosher RA, Dewey D, Thormann W, Saville DA, Bier M. Computer simulation and experimental validation of the electrophoretic behavior of proteins. *Anal Chem.* 1989;61:362–366.
18. Mosher RA, Gebauer P, Caslavska J, Thormann, W. Computer simulation and experimental validation of the electrophoretic behavior of proteins. *Anal Chem.* 1992;64:2991–2997.
19. Thormann W, Huang TM, Pawliszyn J, Mosher RA. High-resolution computer simulation of the dynamics of isoelectric focusing of proteins. *Electrophoresis* 2004;25;324–337.
20. Shim J, Dutta P, Ivory CF. Effects of ampholyte dissociation constants on protein separation in on-chip isoelectric focusing. *J Nanosci Nanotechnol.* 2008;8:3719–3728.

21. Powell RE, Seering W. Multichannel structural inverse filtering. *ASME J Vibr Acoust Stress Reliab Des*. 1984;106:22–28.
22. Ahn B, Shim J. Force prediction by indirect force measurement and pseudo inverse Technique. *Kor Soc Prec Eng*. 2002;19:43–50.
23. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosoph Mag*. 1901;6(2):559–572.
24. Cui H, Horiuchi K, Dutta P, Ivory CF. Isoelectric focusing in a poly (dimethylsiloxane) microfluidic chip. *Anal Chem*. 2005;77:1303–1309.
25. Cui H, Horiuchi K, Dutta P, Ivory CF. Multistage isoelectric focusing in a polymeric microfluidic chip. *Anal Chem*. 2005;77:7878–7886.
26. Mosher RA, Saville DA, Thormann W. *The Dynamics of Electrophoresis*. VCH: Weinheim; 1992.
27. Mosher RA, Gebauer P, Thormann W. Computer-simulation and experimental validation of the electrophoretic behavior of proteins. *J Chromatogr*. 1993;638:155–164.
28. Shim J, Dutta P, Ivory CF. Finite volume methods for isotachophoretic separation in microchannels: part A. *Num Heat Trans*. 2007; 52:441–461.

Manuscript received Jan. 11, 2008, revision received Feb. 26, 2008, and final revision received Apr. 24, 2008.