Published in final edited form as:

Anal Chem. 2010 January 1; 82(1): 307–315. doi:10.1021/ac901982u.

Estimation of Migration-time and Mobility Distributions in Organelle Capillary Electrophoresis with Statistical-Overlap Theory

Joe M. Davis^{*} and

Department of Chemistry and Biochemistry, Southern Illinois University at Carbondale, Carbondale, IL 62901 USA

Edgar A. Arriaga

Department of Chemistry, University of Minnesota, Minneapolis, MN 55455 USA

Abstract

The separation of organelles by capillary electrophoresis (CE) produces large numbers of narrow peaks, which commonly are assumed to originate from single particles. In this paper, we show this is not always true. Here, we use established methods to partition simulated and real organelle CEs into regions of constant peak density and then use statistical-overlap theory to calculate the number of peaks (single particles) in each region. The only required measurements are the number of observed peaks (maxima) and peak standard deviation in the regions, and the durations of the regions. Theory is developed for the precision of the estimated peak number and the threshold saturation above which the calculation is not advisable due to fluctuation of peak numbers. Theory shows that the relative precision is good, when the saturation lies between 0.2 and 1.0, and is optimal when the saturation is slightly greater than 0.5. It also shows the threshold saturation depends on the peak standard deviation, divided by the region's duration. The accuracy and precision of peak numbers estimated in different regions of organelle CEs are verified by computer simulations having both constant and non-uniform peak densities. The estimates are accurate to 6%. The estimated peak numbers in different regions are used to calculate migration-time and electrophoretic-mobility distributions. These distributions are less biased by peak overlap than ones determined by counting maxima and provide more correct measures of the organelle properties. The procedure is applied to a mitochondrial CE, in which over 20% of peaks are hidden by peak overlap.

Introduction

Biological particles including microorganisms, viruses, intact cells, and organelles have negative electrophoretic mobilities that are complex functions of the particles' surface zeta potentials, morphologies, and electrical membrane potentials, as well as the analytical conditions and medium (e.g., electric field and ionic strength)^{1–6}. The mobility of individual particles has been characterized using capillary electrophoresis (CE) and is highly heterogeneous even when particles appear to be identical⁷. When fluorescently labeled particles are analyzed by CE coupled to a laser-induced-fluorescence detector (LIF), the electropherogram consists of a collection of narrow peaks that migrate out in a migration-time window that is dependent on the separation conditions and, most importantly, on the mobility

^{*}corresponding author.

Current address: 733 Schloss Street, Wrightsville Beach, NC 28480, Phone: 910 256 4235, chimicajmd@ec.rr.com

range of the particles in the biological sample. The resultant window can be interpreted as a migration-time distribution, which can be transformed into a mobility distribution.

Mobility distributions are important, because they show quantitative differences among different organelle types⁸ and the influence of organellar content on mobility⁶. For example, they reveal the differences among mitochondria from different muscle types of animals with different ages⁹. The distributions also are indicators of the gentleness or harshness of procedures for sample preparation¹⁰. For distributions to have diagnostic potential, they must be determined without bias. A source of bias addressed by this paper is the overlap of peaks in the CEs, from which the distributions are determined.

A “peak” is defined as a fluorescence intensity profile of a single organelle as it travels through the LIF detector, whereas an “observed peak” is a detected maximum that is comprised of one or more organelle peaks. The numbers of peaks and observed peaks differ because of peak overlap. If we could separate most peaks, we could approximate the migration-time distribution by partitioning a CE into bins and counting the peaks in each. In a recent paper, we used statistical-overlap theory (SOT) and an analogy to the Type-II error of hypothesis testing to predict conditions, for which peak overlap is small enough that the numbers of peaks and observed peaks in such bins are statistically indistinguishable¹¹. Under these conditions, migration-time and mobility distributions can be estimated simply by counting observed peaks in the CE.

For any analysis the organelle concentration has a value, beyond which these conditions are not met because too many peaks overlap. In principal, the analysis can be improved or the sample can be diluted or injected in smaller amounts. However, it may be difficult to improve the analysis. Furthermore, sample dilution and subsequent analysis are not always possible, because organelles have a finite lifetime. Samples also may be limited in size or to a single injection, as are individual cells^{12,13}. In such cases, we may be forced to interpret organelle CEs with many overlapping peaks. In this paper, we develop procedures for using SOT to estimate the actual numbers of peaks in the different bins from the numbers of observed peaks. With these procedures, the total number of peaks in the CE, and the migration-time and mobility distributions, can be estimated.

Because organelle CE entails the injection of only hundreds to thousands of discrete particles, statistical fluctuations among otherwise replicate injections affects the numbers of peaks and observed peaks. Because these numbers differ, the histograms determined from different CEs vary, causing an uncertainty in the migration-time and mobility distributions determined from any one CE. We must know the uncertainty to evaluate the distributions’ reliability.

Theory

Assumptions

The objectives of this paper are to use SOT to estimate accurate and precise numbers of peaks, migration-time distributions, and mobility distributions from organelle CEs. Several approaches exist in SOT^{14–18}, but in all cases a separation is a member of a large ensemble of separations, in which the number, migration times, and heights of peaks are governed by probability distributions. Here, the CE bins are chosen such that the peak density in each is nearly constant. However, the density can vary among different bins, and each bin has its own ensemble. The variation of peak numbers models the variation of the injected number of organelles; the variation of migration times models the variation of mobilities among different organelles. The number of observed peaks also varies among ensemble members.

Review of basic equations

Let m and p equal the numbers of peaks and observed peaks in an ensemble member (i.e., bin), with $\bar{m}$ and $\bar{p}$ equaling the means of these numbers in the ensemble. The means are related by the saturation^{14,17}

$$\alpha = 4\bar{m}\sigma R_s^*/X \quad (1a)$$

which is a metric of peak crowding. Peak numbers and migration times usually are modeled by Poisson statistics for empirical^{15,19,20} and theoretical^{14,16,21,22} reasons. For a Poisson distribution of peaks having constant density and width¹⁴

$$\bar{p} = \bar{m} \exp(-\alpha) \quad (1b)$$

In eq 1a, σ is the peak standard deviation, X is the bin duration, and R_s^* is the average minimum resolution that separates adjacent peaks. The value of R_s^* is not arbitrary but depends on the ensemble's distribution of peak heights²³ and amount of peak overlap²⁴ in a manner that is non-intuitive to most practitioners. The peak-height distribution usually is modeled by an exponential function, which is consistent with the empirically measured or estimated distribution of peak heights in complex mixtures^{19,25–28} and also the prediction of theory^{28,29}. Recently, this R_s^* function was absorbed into the saturation by defining a new variable, the effective saturation α_e ³⁰

$$\alpha_e = \alpha / R_s^* = 4\bar{m}\sigma / X \quad (2a)$$

which allowed eq 1b to be fit by the empirical expression

$$\bar{p} = \bar{m} / (1 + \beta_1 \alpha_e + \beta_2 \alpha_e^2) \quad (2b)$$

over the range, $0 \leq \alpha_e \leq 25$, with $\beta_1 = 0.775 \pm 0.002$ and $\beta_2 = 0.0750 \pm 0.0009$.

Eqs 1 and 2 express the same result. In this paper, we express all theory relative to eq 1, because SOT fundamentally depends on the saturation α . Thus, the effective saturation α_e is not explicitly used, although some findings are reported relative to it. Practitioners may prefer eq 2 because it is independent of R_s^* and avoids the numerical analysis needed with eq 1 (see Procedures section). Conversions between α and α_e are reported in Part One of Supplementary Material to this paper.

For Poisson statistics, the variance of the peak number m , σ_m^2 , among ensemble members is $\bar{m}$. The variance of the observed peak number p , σ_p^2 , is³¹

$$\sigma_p^2 = \bar{m} \exp(-\alpha) [1 - 2\alpha \exp(-\alpha)] \quad (3)$$

Protocol

We partition CEs into bins of known duration X , as shown in Figure 1a, using a standard statistics formula and count the number of observed peaks (maxima) p in each bin. For each bin, we set p equal to $\bar{p}$, eq 1b, and calculate $\bar{m}$ from the known peak standard deviation σ and predicted R_s^* value. Here, σ is assumed constant to model peak widths in CEs having a large post-column sheath flow³². Unlike previous calculations of $\bar{m}$ determined by least-squares fittings to eq 1b of multiple p values from different separations^{19,27,33}, the calculation of $\bar{m}$ is exact, requiring new theory that restricts p to values less than eq 1b, avoids the “double-value” problem¹⁴, and evaluates the calculation’s precision. This theory is discussed here but details are deferred to Supplementary Material for brevity’s sake. These calculations produce an $\bar{m}$ estimate and its standard deviation for each bin. The sum of the $\bar{m}$ ’s is the total number of peaks in the CE, whereas the $\bar{m}$ ’s of different bins determine the discrete SOT-estimated migration-time distribution. The dashed curve, dashed-line histogram, and solid-line histogram in Figure 1b are the actual migration-time distribution in Figure 1a, the SOT-estimated distribution, and the approximate distribution obtained by simply counting the numbers of observed peaks. The graph ordinate is the density, or the number of peaks (or observed peaks) per unit time. The first two distributions agree closely, but the last one is biased because of peak overlap, illustrating the need for SOT calculations. The error bars are the standard deviations of the $\bar{m}$ ’s, which gauge the precision of the SOT-estimated distribution. The inset to Figure 1b is the SOT-estimated mobility distribution, calculated from the SOT-estimated migration-time distribution by mapping the temporal bin boundaries into mobilities. Because the relation between time and mobility is non-linear and because mobilities are signed, the orientation, bin height, and bin width differ from the SOT-estimated migration-time distribution.

Figure 1c is a graph of four model migration-time distributions used to simulate CEs and test the protocol. The distributions are graphed against reduced time ζ (which lies between 0 and 1) and scaled relative to ordinate $f(\zeta)$, such that the area under each is unity. The distributions differ markedly and are assigned names for easy reference (Gaussian, bimodal, asymmetric, and constant). On interpreting the simulations by SOT, we assume the distributions are unknown and subsequently assess their accuracy.

Interpretation of calculated $\bar{m}$

The implications of calculating $\bar{m}$ from a single p value and eq 1b are discussed. Figure 2a is a graph of $\bar{p}$ vs $\bar{m}$ based on eq 1, when peak heights are exponentially random. This curve describes a relation between means. The p value of any bin rarely equals the mean $\bar{p}$, however, because of statistical fluctuation. Figure 2b shows the consequences of equating them. The circle on the $\bar{p}$ vs $\bar{m}$ curve represents the means ($\bar{m}$, $\bar{p}$) of a particular ensemble. The bold curve in the figure’s lower center is the probability distribution of the number of peaks (i.e., the m distribution). The bold curve to the right is the probability distribution of the number of observed peaks (i.e., the p distribution). Their means coincide with the circle. Both distributions are discrete but are shown as continuous functions for simplicity.

Consider the p value identified by the horizontal arrow, a , in Figure 2b. On its identification with eq 1b, it is mapped by the $\bar{p}$ vs $\bar{m}$ curve into the $\bar{m}$ value associated with the vertical arrow, a' . However, this $\bar{m}$ is *not* the ensemble mean. Rather, it is the mean of *another* m distribution, represented by the finely dashed curve. A similar argument can be made for the p and $\bar{m}$ values connected by arrows b and b' ; this $\bar{m}$ is the mean of a third m distribution, represented by the coarsely dashed curve. However, the $\bar{m}$ ’s so determined almost coincide with m values actually belonging to the ensemble, as represented by the squares in Figure 2b. This coincidence is the basis of our analysis: *$\bar{m}$ ’s determined by identifying eq 1b with single random p values form a distribution related to the m distribution.* These determined $\bar{m}$ ’s henceforth are called m_e to distinguish them from the ensemble mean, and they comprise the m_e distribution.

Standard deviation of m_e distribution

The m_e distribution is broadened by our analysis. This is shown in Figure 2c, in which circles represent two different $(\bar{m}, \bar{p})$ coordinates. The arrows have the same meaning as in Figure 2b and show that the m_e distribution broadens as $\partial\bar{p}/\partial\bar{m}$ decreases (the derivative is written as a partial to show that σ/X is constant). This broadening decreases the precision of m_e , reducing its reliability. The standard deviation σ_{m_e} of m_e is derived in Part Two of the Supplementary Material and can be expressed as

$$\sigma_{m_e}(\sigma/X)^{1/2} = \frac{1}{2} \left(\frac{\alpha}{R_s^*} [1 - 2\alpha \exp(-\alpha)] \right)^{1/2} \frac{\exp(\alpha/2) [\alpha^{-1} - d\ln R_s^*/d\alpha]}{\alpha^{-1} - 1 - d\ln R_s^*/d\alpha} \quad (4a)$$

or as the coefficient of variation (CV), equal to $100 \sigma_{m_e}/\bar{m}$

$$\frac{CV}{(\sigma/X)^{1/2}} = 200 \left(\frac{R_s^*}{\alpha} [1 - 2\alpha \exp(-\alpha)] \right)^{1/2} \frac{\exp(\alpha/2) [\alpha^{-1} - d\ln R_s^*/d\alpha]}{\alpha^{-1} - 1 - d\ln R_s^*/d\alpha} \quad (4b)$$

These equations can be evaluated at any α if R_s^* is known. For exponential peak heights, R_s^* can be predicted from theory²⁴; alternatively, the numerical solution so predicted is well fit by the empirical equation

$$R_s^* = R_s^*(\alpha) = 0.725 + \beta_3 \alpha + \beta_4 \alpha^2 + \beta_5 \alpha^3 \quad (4c)$$

over the range, $0 \leq \alpha \leq 3.85$, with $\beta_3 = -0.1910 \pm 0.0001$, $\beta_4 = 0.0039 \pm 0.0001$, and $\beta_5 = 0.00188 \pm 0.00002$. The coefficient 0.725 in eq 4c is the theoretical value of R_s^* at $\alpha = 0$ ²³. The functions σ_{m_e} and CV in eqs 4a and 4b equal or exceed their counterparts for the Poisson distribution, which are $\sqrt{\bar{m}}$ and $100/\sqrt{\bar{m}}$. Monte Carlo simulations of the m_e distribution are reported in Part Three of the Supplementary Material.

For good precision, m_e should be calculated to the left of the maximum of the $\bar{p}$ vs $\bar{m}$ curve in Figure 2a (i.e., the low- $\bar{m}$ side), where $\partial\bar{p}/\partial\bar{m}$ is large. This action also avoids the “double-value problem”¹⁴, i.e., the correspondence of $\bar{p}$ to two $\bar{m}$'s except at the maximum. However, $\bar{p}$ should not be near the maximum for another reason.

Largest interpretable p value and threshold saturation

Figure 2d shows that $\bar{p}$, $\bar{m}$, and α have threshold values on the low- $\bar{m}$ side of the p vs $\bar{m}$ curve, beyond which large outliers of the p distribution can exceed the curve maximum. If this happens, eq 1 has no solution. We can derive a relation among the threshold value of $\bar{p}$, the curve maximum, and the width of the p distribution that assures us that this rarely happens.

The maximum $\bar{p}$ value, $\bar{p}_{\max}$, is found by calculating $\partial\bar{p}/\partial\bar{m}$ from eq 1 and setting it equal to zero

$$\frac{\partial\bar{p}}{\partial\bar{m}} = e^{-\alpha} \left[1 - \bar{m} \frac{\partial\alpha}{\partial\bar{m}} \right] = e^{-\alpha} \left[1 - (\alpha^{-1} - d\ln R_s^*/d\alpha)^{-1} \right] = 0 \quad (5)$$

The expression $\partial\alpha/\partial\bar{m}$ is given by the reciprocal of eq S-9 in Part Two of the Supplementary Material, whereas $d\ln R_s^*/d\alpha$ is evaluated from eq 4c. The solution to eq 5 is given in the Results and Discussion; for the moment, we observe that $\bar{p}_{\max}$ is a multiple of the reciprocal σ/X ratio, i.e., $\bar{p}_{\max} = \delta X/\sigma$, where δ is a constant.

For $\bar{m} \geq 10$, the p distribution is described well by a Gaussian envelope having mean $\bar{p}$ and standard deviation σ_p ^{11,15}. A Gaussian distribution has negligible density three to four standard deviations from its mean. Therefore, the threshold saturation α_t , or the largest saturation at which statistical fluctuations of p cause very few (if any) problems with estimating m_e , is established by the equation

$$\bar{p}_t + \gamma \sigma_p = \bar{p}_{\max} \quad (6a)$$

where $\bar{p}_t$ is the threshold value of $\bar{p}$ and γ is a parameter between 3 and 4. Figure 2d shows the relation among $\bar{p}_t$, the threshold peak number $\bar{m}_t$, γ , σ_p , and $\bar{p}_{\max}$. Using eqs 1a, 1b, and 3, and the expression $\bar{p}_{\max} = \delta X/\sigma$, we can explicitly write eq 6a as

$$\frac{\alpha_t}{4R_s^*(\alpha_t)} \frac{X}{\sigma} \exp(-\alpha_t) + \frac{\gamma}{2} \sqrt{\frac{\alpha_t}{R_s^*(\alpha_t)} \frac{X}{\sigma} \exp(-\alpha_t) [1 - 2\alpha_t \exp(-\alpha_t)]} = \delta \frac{X}{\sigma} \quad (6b)$$

with $R_s^*(\alpha_t)$ equal to eq 4c, as expressed with $\alpha = \alpha_t$. Eq 6b shows the threshold saturation α_t is determined by parameter γ and the reciprocal σ/X ratio.

Procedures

Threshold values

Eq 5 was solved by bisection to determine the scalar δ . For various σ/X , eq 6b was solved for α_t by bisection, with $\gamma = 3$. The threshold $\bar{m}_t$ value then was calculated from α_t , eq 1a, and eq 4c. Values of the threshold effective saturation were calculated from α_t as described in Part One of the Supplementary Material. CEs were simulated to verify α_t .

CE simulations

CEs were simulated as previously described²⁴ along a reduced time coordinate $\zeta = (t - t_o)/^1D$ between 0 and 1, with t , t_o , and 1D equaling time, the time of the first peak's elution, and the CE duration, respectively. The Box Muller transform³⁴ was used to mimic Poisson distributed peak numbers. In any CE, peaks were Gaussians with constant standard deviations and exponentially random heights. The number of observed peaks (maxima) was determined.

Calculation of m_e —The peak number m_e was calculated from the number of observed peaks p in a bin, the σ/X ratio, and eq 1. The calculation was a dual bisection, in which $\bar{m} = m_e$ determined α via eqs 1a and 4c, and then in which $\bar{m}$ was adjusted so that eq 1b equaled p . (Alternatively, practitioners may wish to determine m_e from eq 2b, whose solution is

$$m_e = \frac{1}{8\beta_2(\sigma/X)} \left[\frac{1}{4p(\sigma/X)} - \beta_1 - \sqrt{\beta_1^2 - 4\beta_2 - \frac{\beta_1}{2p(\sigma/X)} + \frac{1}{\{4p(\sigma/X)\}^2}} \right] \quad (7)$$

Evaluating eq 7 is a straightforward computation and simpler than performing the dual bisection, but the m_e so determined is slightly less accurate since it is based on a fit.)

Assessment of eqs 4a and 4b—Five hundred CEs of constant peak density (i.e., single-bin CEs) were simulated for different $\bar{m}$ and for σ/X ratios ranging from 0.00005 to 0.0013. Peak numbers m_e were estimated as described above. To assess accuracy, the percentage error between the average m_e and $\bar{m}$ was calculated. The standard deviation and CV of m_e were compared to eqs 4a and 4b. The minimum of eq 4b was determined with the golden search algorithm³⁴.

Analysis of migration-time distributions—Five hundred CEs were simulated with the four model migration-time distributions in Figure 1c. The ratio of the peak standard deviation to the CE duration, σ^1D , was 8×10^{-5} or 8×10^{-6} (these are realistic values in organelle CE, e.g., $\sigma = 0.008$ s and $^1D = 100 - 1000$ s). Equations for the distributions and their migration times are reported in Part Four of the Supplementary Material.

For each distribution and σ^1D ratio, the mean number of peaks in the entire CE, $\bar{m}_{tot}$, was incremented (e.g., $\bar{m}_{tot} = 100, 200, \dots$) until the saturation of the bin of greatest peak density exceeded the threshold saturation α_t determined as described below. The CE was partitioned into N bins of equal duration X

$$N = \text{trunc} \left[(2p_{tot})^{1/3} \right] + 1 \quad (8)$$

where p_{tot} is the total number of observed peaks in the entire CE and *trunc* means truncation. Eq 8 is one bin larger than the minimum bin number of a common statistics formula³⁵ (with p_{tot} equaling the number of data points for binning) and it determined the σ/X ratio as $N\sigma^1D$ (i.e., $X = ^1D/N$). The peak number m_e in each bin was calculated as described above. The number of peaks in the entire CE, $m_{e,tot}$, was calculated by summing estimated peak numbers in all bins

$$m_{e,tot} = \sum_{i=1}^N m_{e_i} \quad (9)$$

where m is the m_e estimate for the i^{th} bin. The average, standard deviation, and CV of $m_{e,tot}$ values were calculated.

To assess accuracy, the percentage error between the average value of $m_{e,tot}$ and $\bar{m}_{tot}$ was calculated. The standard deviation of $m_{e,tot}$ was compared to its theoretical counterpart calculated by error propagation of eq 9

$$\sigma_{m_{e,tot}} = \left[\sum_{i=1}^N \sigma_{m_{e_i}}^2 \right]^{1/2} \quad (10)$$

with $\sigma_{m_{e_i}}^2$ equaling the variance of the i^{th} bin, as computed from eqs 4a and 4c. The saturation α in these equations was calculated by estimating the mean numbers of peaks and observed peaks in each bin from SOT equations for non-uniform migration-time distributions³⁶,

identifying these estimates with $\bar{p}$ and $\bar{m}_t$, and substituting them in eq 1b. This calculation provides the best estimate of the saturation, with which to associate eq 1b. The theoretical CV was calculated as the product of eq 10 and $100/\bar{m}_{tot}$, and was compared to the CV's determined by simulation.

Average histograms of the four migration-time distributions were calculated for different $\bar{m}_{tot}$ from m_e 's determined from 500 simulations. They were compared to the actual normalized distributions by scaling the histogram areas to equal the average of $m_{e,tot}$, divided by $\bar{m}_{tot}$. The scaling makes it easy to compare distributions for different $\bar{m}_{tot}$ and to assess if the distributions are underestimated or overestimated.

Application to mitochondrial CE

The total number of peaks in, and the migration-time and mobility distributions of, a mitochondrial CE³⁷ were estimated. The CE was acquired at 200 Hz. Maxima produced by baseline noise were removed by clipping as described elsewhere¹¹. In brief, the standard deviation of the baseline noise was estimated from three signal-free regions spanning 50 to 98.6 s, six multiples of it were added to a linear interpolation of the lower noise bound estimated at 145 points, and all signal below the sum was clipped. This action is appropriate for Gaussian noise, which is negligible three or more standard deviations from the mean. The peak standard deviations σ of 10 isolated observed peaks spanning the CE were determined by moments analysis. The migration-time distribution was constructed as described above and transformed into a mobility distribution by mapping the times t of its bin boundaries into mobilities μ

$$\mu = L/(Et) - \mu_{eo} \quad (11)$$

where L is the capillary length (0.4 m), E is the electric field strength (40 kV/m), and μ_{eo} is the electroosmotic mobility ($5.2 \times 10^{-8} \text{ m}^2/\text{V}\cdot\text{s}$).

Results and Discussion

Threshold values of $\bar{p}$ vs $\bar{m}$ curve

By solving eq 5, we find the $\bar{p}$ vs $\bar{m}$ curve maximizes at $\alpha \approx 1.601$. (If R_s^* were constant, it would maximize at $\alpha = 1^{14}$.) On substituting this α and its associated R_s^* value (see eq 4c) into eq 1, we discover that $\bar{p}_{\max}$ is $0.185 X/\sigma$. Thus, the scalar δ in eq 6 is 0.185.

Figure 3a is a graph of the threshold saturation α_t , threshold effective saturation $\alpha_{e,t}$, and logarithm of the threshold peak number $\bar{m}_t$ vs the logarithm of the σ/X ratio, as determined from eq 6b. All thresholds increase with decreasing σ/X because $\sigma_p/\bar{p}$ decreases, allowing $\bar{p}_t$ to move closer to the curve maximum. For a Gaussian p distribution, the probability that p exceeds $\bar{p}_{\max}$ is $[1 - \text{erf}(\sqrt{3}/2)]/2 = 0.00135$. The circles in Figure 3a are α_t values determined by extrapolating p in 50,000 simulations to the expected number of excesses, $67.5 = (0.00135)(50000)$. Least-squares fits to the curves in this figure are reported in Part Five of the Supplementary Material.

Analysis of simulations with constant peak density

Our procedure is based on applying SOT to different bins of a partitioned CE. The peak density in each bin is roughly constant, but it can vary among different bins. We first assess the accuracy and precision of m_e values determined from a single bin of constant peak density.

Accuracy of m_e —Figure 3b is a graph of the percentage error between $m\bar{r}$ and the average m_e of 500 simulations vs saturation α for five σ/X ratios spanning a 26-fold range. (The σ/X ratios in subsequent determinations of migration-time distributions lie within this range.) The results are somewhat scattered but the trend is clear. The accuracy decreases with α , because eq 1 slightly underestimates peak overlap at high saturation. For $\alpha \leq 1$, the accuracy is -6% or better.

Precision of m_e —Figure 3c is a graph of the standard deviation of m_e , σ_{m_e} , multiplied by $(\sigma/X)^{1/2}$ vs α . Figure 3d is a graph of the coefficient of variation, $CV = 100\sigma_{m_e}/m\bar{r}$, divided by $(\sigma/X)^{1/2}$ vs α . The symbols are results of 500 simulations for the same σ/X ratios in Figure 3b; the curves are graphs of eqs 4a and 4b. These equations are based on a propagation of errors (see Part Two of the Supplementary Material) and overestimate σ_{m_e} and CV at large σ_{m_e} but they agree with simulation for $\alpha \leq 0.8$ or so. The bold curves are graphs of

$$\sigma_{m_e}(\sigma/X)^{1/2} = \frac{1}{2} \sqrt{\frac{\alpha}{R_s^*}}; \frac{CV}{(\sigma/X)^{1/2}} = 200 \sqrt{\frac{R_s^*}{\alpha}} \quad (12)$$

which are the limits of eqs 4a and 4b as α approaches zero. On substituting eq 1a into eq 12, we find the limits are $\sigma_{m_e} = \sqrt{m\bar{r}}$ and $CV = 100/\sqrt{m\bar{r}}$, that is, the Poisson limits stated earlier. With increasing α , σ_{m_e} and CV exceed these limits, because the m_e distribution is broadened by the decreasing slope, $\partial\bar{p}/\partial m\bar{r}$, of the $\bar{p}$ vs $m\bar{r}$ curve.

The graph of $CV/(\sigma/X)^{1/2}$ has a shallow minimum at $\alpha \approx 0.524$, with a value of 320.5. The minimum occurs, because Poisson statistics reduces the CV as $m\bar{r}$ (or α) increases, whereas the diminishing slope, $\partial\bar{p}/\partial m\bar{r}$, increases the CV as $m\bar{r}$ (or α) increases. Ultimately the latter effect dominates. It is fortuitous that the minimum is shallow, with roughly the same precision over the range, $0.2 \leq \alpha \leq 1.0$. Regardless of the threshold saturation, we suggest that α not exceed unity if good precision is sought.

Figure 3d shows that the relative precision is best, when σ/X is small. For example, if $m\bar{r}$ is decreased by 10-fold and σ/X is increased by 10-fold, the saturation α doesn't change. Therefore, the right-hand sides of eqs 4a and 4b don't change, but the left-hand sides do. The estimate m_e is 10-fold smaller, but σ_{m_e} is smaller by only $\sqrt{10}$ and the CV is larger by $\sqrt{10}$.

Analysis of simulations with different migration-time distributions

Our estimates from CE simulations having the migration-time distributions in Figure 1c follow the same trends as do simulations of constant peak density. This is not surprising, since the estimates entail repetitive applications of eq 1 to contiguous bins of approximately constant peak density. However, the quantitative results vary with the distribution.

Accuracy of $m_{e,tot}$ —Figure 4 is a graph of the percentage error between the mean total number of peaks $m\bar{r}_{tot}$ and the average $m_{e,tot}$ of 500 simulations vs the logarithm of $m\bar{r}_{tot}$. Each $m_{e,tot}$ was calculated from eq 9 for the four model distributions and the two σ/D ratios, 8×10^{-5} and 8×10^{-6} . The $m\bar{r}_{tot}$ range is greater for the smaller σ/D ratio, because the saturation is proportional to the product of $m\bar{r}$ and σ . The number N of bins ranged from 6 to 45, with σ/X ratios between 1.0×10^{-4} and 1.3×10^{-3} . The symbols at the curve ends have $m\bar{r}_{tot}$ values, at which the saturation in the bin of maximum peak density exceeded the threshold saturation. As in Figure 3b, the results are somewhat scattered but the trend is the same, with the average of $m_{e,tot}$ underestimating $m\bar{r}_{tot}$ at high saturation and an accuracy of roughly -6% or better.

Some $\bar{m}_{tot}$ values greatly exceed typical particle numbers in organelle CE, but they show the potential of theory.

Precision of $m_{e,tot}$ —Figure 5a is a graph of $\sigma_{m_{e,tot}}$ vs $\bar{m}_{tot}$ for $\sigma^1 D = 8 \times 10^{-6}$. The symbols are results of 500 simulations for the four model distributions. The curves were calculated from eq 10 and agree with simulation, indicating the total variance can be modeled by a discrete sum of variances for different bins. As before, $\sigma_{m_{e,tot}}$ exceeds the Poisson limit,

$\sigma_{m_{e,tot}} = \sqrt{\bar{m}_{tot}}$, represented by the bold curve. For a given $\bar{m}_{tot}$, $\sigma_{m_{e,tot}}$ differs for the different distributions because eq 10 and α vary with changing peak density. Figure 5b is the corresponding graph of the coefficient of variation, CV , vs $\bar{m}_{tot}$. As before, minima are found, but the $\bar{m}_{tot}$ at which minimization occurs and the minimum itself differ for different distributions.

The CV is a metric of precision for the entire CE. We may be more interested in good precision in bins of high peak density, since bins of low peak density have little impact on the distribution. By weighting the CV by the fraction of peaks in different bins, we can target $\bar{m}_{tot}$ values for optimal precision where most peaks are found. Unlike most separations, in favorable cases the CV can be adjusted in organelle CE by diluting or concentrating the sample. Accordingly, we define a weighted CV

$$CV_{wt} = \frac{\sum_{i=1}^N \bar{m}_i CV_i}{\bar{m}_{tot}} \approx \frac{\sum_{i=1}^N m_{e,i} CV_i}{m_{e,tot}} \quad (13)$$

where CV_i is the CV of the i^{th} bin. Figure 5c is a graph of CV_{wt} vs $\bar{m}_{tot}$ for the four distributions. The small discontinuities in the curves occur as the bin number changes. As before the $\bar{m}_{tot}$ producing the minimum CV_{wt} differs with the distribution, but now the minimum itself is the same (about 6). The unifying attribute is shown in Figure 5d, which is a graph of the saturation α in different bins vs the reduced time ζ . For nonuniform distributions, the α 's of bins with large peak density are slightly greater than 0.5. This α range produces a small CV for a single bin, as shown in Figure 3d, and it is not surprising that a sum of CV 's weighted by the fraction of peaks follows the same trend. Similar trends to those in Figure 5 are found for the $\sigma^1 D$ ratio, 8×10^{-5} .

Estimation of migration-time distribution—Figure 6 reports graphs of the four normalized distributions $f(\zeta)$ vs reduced time ζ , first shown in Figure 1c, for the two $\sigma^1 D$ ratios. The symbols and error bars are the means and standard deviations of m_e 's determined from 500 simulations for three $\bar{m}_{tot}$ values, scaled such that the distribution area is the average $m_{e,tot}$, divided by $\bar{m}_{tot}$. Each symbol corresponds to a different bin. For each distribution, one value of $\bar{m}_{tot}$ corresponds to the minimum CV_{wt} , whereas the other $\bar{m}_{tot}$ values are smaller and larger. Although all distributions are estimated well, the precision is best at the minimum CV_{wt} if we consider only the bins of large peak density. If possible, separation conditions should be adjusted such that these bins have saturations slightly greater than 0.5.

Application to mitochondrial CE

Figure 7a is a noise-clipped mitochondrial CE containing 1707 observed peaks and first discussed in ref 37. The inset expands the CE and shows overlapping maxima, suggesting that SOT estimates could be useful. The CE first was partitioned among uniform bins. For such bins, however, detail was lost around the maximum peak density near 380 s and slowly varying densities were binned to excess. Accordingly, eq 8 was applied to three regions spanning 0 to

332 s ($p_{tot} = 99$; $N = 6$; $X = 55.3$ s), 332 to 770 s ($p_{tot} = 1355$; $N = 14$; $X = 31.3$ s), and 770 s to 1848.6 s ($p_{tot} = 253$; $N = 8$; $X = 134.8$ s). The average standard deviation σ calculated from 10 observed peaks was 11.1 ms, with a coefficient of variation equaling 8.4. These values are almost constant due to a large post-column sheath flow. It also is likely they are peaks of single mitochondria, because the coefficient of variation is small. The σ/lD ratio is 6.0×10^{-6} ; the σ/X ratios range from 8.2×10^{-5} to 2.0×10^{-4} , for which the threshold saturation α_t is 1.1 or larger (see Figure 3a).

The estimated peak number m_e in each bin was determined from eq 1 and the number of observed peaks p . The saturation α of each bin was approximated from eq 1a as¹⁹

$$\alpha \approx -\ln(p/m_e) \quad (14)$$

and used to evaluate equations dependent on α . Eq 14 was used to estimate α instead of the detailed procedures outlined earlier, because p and m_e are the only metrics we have (i.e., we do not know $\bar{p}$ and $\bar{m}$). The saturation in the bin of greatest peak density is 0.52, which is less than α_t and the upper saturation, $\alpha \approx 1$, for precise determinations of m_e .

Figure 7b reports histograms of migration-time distributions, as determined by estimating m_e (dashed line) and counting observed peaks (solid line). Because the bins have different widths, the ordinate was scaled such that the bin area equals the number of peaks or observed peaks. The error bars are the scaled standard deviations of m_e . The estimated total number of peaks $m_{e,tot}$ is 2163 ± 56 ($CV = 2.6$). Over 20% of the peaks are hidden by peak overlap, mostly in the second region.

The numbers of peaks and observed peaks differ in the three bins of highest density, and they differ somewhat in the two bins to their right (Figure 7b). We evaluated the observed-peak numbers in these bins with our simple procedure based on the Type-II error analogy¹¹. As noted earlier, this procedure allows us to decide if the numbers of peaks and observed peaks are statistically the same. For the three bins of highest density, the procedure is not satisfied, and for the two other bins it shows the numbers of observed peaks are questionable estimates. However, the observed-peak numbers in the other bins satisfy the procedure. Thus, our current work is consistent with our earlier work.

Figure 7c shows the mobility distributions calculated from the migration-time distributions and eq 11. As before, the bin areas equal the numbers of observed peaks or peaks. Because the relation between time and mobility is non-linear, and the mitochondrial and electroosmotic mobilities have opposite signs, the two distribution types appear differently. Most peaks overlap in the mobility range, -2×10^{-8} to -3×10^{-8} m²/V-s.

Conclusions

We have shown that SOT can be used to estimate both accurate and precise migration-time and mobility distributions in partitioned organelle CEs from a series of single observed-peak numbers, as long as the saturation does not exceed unity. Usually, organelle CEs are assumed to be free of peak overlap. Our work shows this is not always true. The mitochondrial CE in Figure 7a has significant peak overlap, requiring SOT calculations to reduce distribution bias. We recommend first screening organelle CEs for peak overlap by our simple procedure based on a Type-II error analogy¹¹, and then making SOT calculations if necessary.

It may be tempting first to transform the migration times of observed peaks into mobilities and then to make the peak-overlap analysis in mobility space, with bins of equal width in mobility.

This is unwise, because peak standard deviations that are constant in time (as in organelle CE with post-column sheath flow) are not constant in mobility. They not only vary from bin to bin but within a bin, complicating the analysis.

We recognize that the number of bins in the SOT-estimated migration-time and mobility distributions is not optimal, because the bin number is calculated from the total number of observed peaks (see eq 8). Ideally, the bin number should be calculated from the total number of peaks, but of course this is unknown prior to the SOT estimation. Fortunately, Figure 6 shows the error in the estimated distributions is small.

Unfortunately, our procedure cannot be extended with high precision to chromatography, except in special cases. Consider gas and liquid chromatography, wherein one might find (for example) 300 observed peaks in a 30-min separation. According to eq 8, these separations would partition among 9 bins of duration $X = 3.33$ min. For peak standard deviations σ of 2 – 4 s, σ/X would range from 0.01 to 0.02. Even the minimum CV of a bin would vary from 32 to 45, as evaluated from eq 4b. Much smaller peak widths would be required for a precise application.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

E.A.A. is supported through an NIH grant AG20866.

Literature Cited

1. Tsoneva IC, Tomov TC. *Bioelectrochemistry and Bioenergetics* 1984;12:253–258.
2. Radko SP, Chrambach A. *J Chromatogr B* 1999;722:1–10.
3. Duffy CF, Gafoor S, Richards DP, Admadzadeh H, O’Kennedy R, Arriaga EA. *Anal Chem* 2001;73:1855–1861. [PubMed: 11338602]
4. Fuller KM, Arriaga EA. *J Chromatogr B* 2004;806:151–159.
5. Pysker MD, Hayes MA. *Langmuir* 2005;21:3572–3577. [PubMed: 15807603]
6. Chen Y, Arriaga EA. *Langmuir* 2007;23:5584–5590. [PubMed: 17402758]
7. Duffy CF, McEathron AA, Arriaga EA. *Electrophoresis* 2002;23:2040–2047. [PubMed: 12210257]
8. Xiong G, Aras O, Shet A, Key NS, Arriaga EA. *Analyst* 2003;128:581–588. [PubMed: 12866871]
9. Ahmadzadeh H, Andreyev D, Arriaga EA, Thompson LV. *J Gerontol Series A: Biol Sci Med Sci* 2006;61A:1211–1218.
10. Fuller KM, Arriaga EA. *Anal Chem* 2003;75:2123–2130. [PubMed: 12720351]
11. Davis JM, Arriaga EA. *J Chromatogr A* 2009;1216:6335–6342. [PubMed: 19632681]
12. Johnson RD, Navratil M, Poe BG, Xiong G, Olson KJ, Ahmadzadeh H, Andreyev D, Duffy CF, Arriaga EA. *Anal Bioanal Chem* 2007;387:107–118. [PubMed: 16937092]
13. Chen Y, Xiong G, Arriaga EA. *Electrophoresis* 2007;28:2406–2415. [PubMed: 17577888]
14. Davis JM, Giddings JC. *Anal Chem* 1983;55:418–424.
15. Martin M, Herman DP, Guiochon G. *Anal Chem* 1986;58:2000–2007.
16. Felinger A, Pasti L, Dondi F. *Anal Chem* 1990;62:1846–1853.
17. Pietrogrande MC, Dondi F, Felinger A, Davis JM. *Chemom Intell Lab Syst* 1995;28:239–258.
18. Dondi F, Bassi A, Cavazzini A, Pietrogrande MC. *Anal Chem* 1998;70:766–773.
19. Herman DP, Gonnord MF, Guiochon G. *Anal Chem* 1984;56:995–1003.
20. Davis JM, Pompe M, Samuel C. *Anal Chem* 2000;72:5700–5713. [PubMed: 11101251]
21. Felinger A. *Anal Chem* 1995;67:2078–2087.

22. Dondi F, Pietrogrande MC, Felinger A. *Chromatographia* 1997;45:435–440.
23. Felinger A. *Anal Chem* 1997;69:2976–2979.
24. Davis JM. *Anal Chem* 1997;69:3796–3805.
25. Nagels LJ, Creten WL, Vanpeperstraete PM. *Anal Chem* 1983;55:216–220.
26. Nagels LJ, Creten WL. *Anal Chem* 1985;57:2706–2711.
27. Dondi F, Kahie YD, Lodi G, Remelli M, Reschiglian P, Bighi C. *Anal Chim Acta* 1986;191:261–273.
28. Pietrogrande MC, Cavazzini A, Dondi F. *Rev Anal Chem* 2000;19:123–156.
29. El Fallah MZ, Martin M. *Chromatographia* 1987;24:115–122.
30. Davis JM, Carr PW. *Anal Chem* 2009;81:1198–1207. [PubMed: 19178343]
31. Rowe K, Davis JM. *Chemom Intell Lab Syst* 1997;38:109–126.
32. Cheng YF, Wu S, Chen DY, Dovichi NJ. *Anal Chem* 1990;62:496–503.
33. Dondi F, Gianferrara T, Reschiglian P, Pietrogrande MC, Ebert C, Linda P. *J Chromatogr* 1990;485:631–645.
34. Press, WH.; Teukolsky, SA.; Vetterling, W.; Flannery, BP. *Numerical Recipes in FORTRAN*. Cambridge University Press; New York: 1992.
35. Terrell GT, Scott DW. *J Am Stat Assoc* 1985;80:209–214.
36. Davis JM. *Anal Chem* 1994;66:735–746.
37. Whiting CE, Arriaga EAJ. *Chromatogr A* 2007;1157:446–453.

Figure 1.

a) Simulated organelle CE. Dashed lines represent bin boundaries spanned by interval X. b) Histograms of SOT-estimated (dashed line) and observed-peak (solid line) migration-time distributions of CE in a). Dashed curve is actual migration-time distribution. Inset is SOT-estimated mobility distribution. c) Graphs of four normalized model migration-time distributions $f(\zeta)$ vs reduced time ζ used to simulate organelle CEs.

Figure 2.

Distribution and precision of m_e . a) Graph of $\bar{p}$ vs $\bar{m}$ for exponentially random peak heights. b) Mapping of p into $\bar{m}$ by the $\bar{p}$ vs $\bar{m}$ curve. Circle represents means of a particular ensemble; bold curves are p and m distributions of the ensemble; dashed curves are m distributions of other ensembles; squares are members of the ensemble's m distribution. c) Graph of $\bar{p}$ vs $\bar{m}$, showing broadening of the m_e distribution from reduction of $\partial \bar{p} / \partial \bar{m}$. Error bars represent standard deviation σ_p of p distributions. d) Relation among upper limit to $\bar{p}$, curve maximum, and width of the p distribution.

Figure 3.

Results for single-bin separations of constant peak density. Symbols in panels b) - d) are simulation results for σ/X equaling 0.00005 ($\times$), 0.0001 ($\circ$), 0.0005 ($\square$), 0.0009 ($\diamond$), and 0.0013 ($\triangle$). a) Graph of threshold saturation α_t , threshold effective saturation $\alpha_{e,t}$, and logarithm of threshold peak number $\bar{m}_t$ vs $\log(\sigma/X)$. Circles are α_t values determined by simulation. b) Graph of percentage error between $\bar{m}$ and average m_e vs saturation α . c) Graph of standard deviation of m_e , σ_{m_e} , multiplied by $(\sigma/X)^{1/2}$ vs α . d) Graph of coefficient of variation, $CV = 100\sigma_{m_e}\bar{m}_t$, divided by $(\sigma/X)^{1/2}$ vs α .

Figure 4.

Accuracy of SOT estimates for migration-time distributions in Figure 1c. Graph of percentage error between $\bar{m}_{tot}$ and average $m_{e,tot}$ of simulations vs $\log(\bar{m}_{tot})$. Line types are the same as in Figure 1c (e.g., solid for Gaussian migration-time distribution). Normal-weight curves are results for $\sigma^1 D = 8 \times 10^{-5}$; bold curves, for $\sigma^1 D = 8 \times 10^{-6}$. Ordinates of symbols are $\bar{m}_{tot}$ values, at which the largest saturation exceeds α_t

Figure 5.

Precision of SOT estimates for migration-time distributions in Figure 1c and $\sigma^1 D = 8 \times 10^{-6}$. a) Graph of standard deviation $\sigma_{m_{e,tot}}$ vs $\bar{m}_{tot}$. Symbols are simulation results (Gaussian ($\circ$), bimodal ($\triangle$), asymmetric ($\diamond$), constant ($\square$)). b) As in a), but for the coefficient of variation CV . c) Graph of weighted coefficient of variation CV_{wt} vs $\bar{m}_{tot}$. d) Graph of saturation α vs reduced time ζ , when $\bar{m}_{tot}$ corresponds to the minimum CV_{wt} . Values of $\bar{m}_{tot}$ are 10750 (Gaussian), 19000 (bimodal), 16000 (asymmetric), and 23000 (constant).

Figure 6.

Scaled SOT-estimated migration-time distributions for $\bar{m}_{tot}$ values less than (circles), equal to (squares), and greater than (diamonds) $\bar{m}_{tot}$ at the minimum CV_{wt} . a) $\sigma^1 D = 8 \times 10^{-6}$, with $\bar{m}_{tot}$ equaling 750, 10750, and 21800 (Gaussian); 1450, 19000, and 34750 (bimodal); 1275, 16000, and 31300 (asymmetric); and 1575, 23000, and 51000 (constant). b) $\sigma^1 D = 8 \times 10^{-5}$, with $\bar{m}_{tot}$ equaling 160, 950, and 1675 (Gaussian); 270, 2000, and 2700 (bimodal); 250, 1650, and 2425 (asymmetric); and 325, 2500, and 4900 (constant).

Figure 7. Experimental organelle CE data. a) Mitochondrial CE. Arrows near time axis are demarcations among three regions. Inset shows peak overlap in region of maximum peak density. b) Histograms of migration-time distributions estimated by SOT (dashed line) and obtained by counting observed peaks (solid line). Main panel shows distributions in second region; inset shows distributions in first and third regions, with break between them. Error bars are SOT-estimated standard deviations. c) Mobility distributions determined from SOT-estimated and observed-peak migration-time distributions.