awkt: A Physicochemical Parameter Estimation Tool for Capillary Zone Electrophoresis

Christian Löffeld*

June 13, 2017

Abstract

We introduce the open source parameter estimation tool awkt. It is primarily aimed at assisting in the characterization and identification of globular proteins using capillary zone electrophoresis. The program untangles a proteins’ charge-to-size ratio by applying nonlinear least-squares optimization to a particular closed-form solution of the Nernst-Planck equation. Numerical estimates for the diffusion coefficient, the hydrodynamic radius and the net electric charge are computed for all molecular entities that are being detected during the electropherogram analysis process. The program offers control parameters for various operational features, including hypothesis testing and model selection, peak detection, baseline correction and concurrent file processing.

Introduction

Capillary zone electrophoresis (CZE) is a nondestructive electrophoretic separation technique that is utilized in a large variety of chemical and biochemical analysis settings [1–5]. More recently, electrophoretic methods have become important tools in proteomic and genomic endeavors, however most often coupled to a downstream mass spectrometry setup [6, 7]. Given appropriate experimental conditions, capillary zone electrophoresis has the potential to enable deep physicochemical analyses of complex analytes such as proteins on its own. The hydrodynamic radius and the net electric charge of a molecular species are of considerable interest with regards to its characterization. Capillary zone electrophoresis has an intimate handle on these two quantities [8, 9]. Reliable determination of these parameters would transform CZE into a truly multidimensional analytical technique, simultaneously facilitating separation and partial molecule characterization. Furthermore, numerical estimates for any associated physical quantity, such as the molecular weight, may also be obtained given that a mathematical relation between these quantities is available or can be established [10]. In order to harness the intrinsic analytical potential of the capillary electrophoretic process, its mechanistic underpinnings have been investigated widely, [11,12] and references therein. The simulation and modeling of many electrophoretic processes is complex, and often only a numerical solution to the underlying partial differential equation is accessible [13]. Despite being extremely useful in many cases, they however also generally inhibit the possibility for parameter estimation due to the absence of an explicit physicochemical model.

In this letter, we introduce a fast, versatile and reliable C++ tool for parameter estimation using nonlinear least-squares optimization [14]. We model the electrophoretic process of globular proteins using a particular closed-form solution to the Nernst-Planck equation. This approach facilitates physicochemical parameter estimation of all detected molecular species in a potentially very complex mixture.

*Email: christian.loeffeld@gmail.com
Background

The electrophoretic process of a molecular species \( i \) in a capillary filled with an electrolyte is modeled by the solution of the one-dimensional convection-diffusion equation

\[
\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c_i}{\partial x^2} - v_i \frac{\partial c_i}{\partial x}
\]  

(1)

The solution \( c_i(x, t) \) denotes the one-dimensional concentration of species \( i \) with respect to space \( x \) and time \( t \). \( D_i \) represents the diffusion coefficient and \( v_i \) the electrophoretic velocity. We let \( v_i = \mu_i E \), where \( \mu_i \) represents the electrophoretic mobility of species \( i \), and \( E \), the applied electric field strength. Assuming a quasi-spherical particle, the diffusion coefficient and the electrophoretic mobility of species \( i \) are described by the Stokes-Einstein relations

\[
D_i = \frac{k_B T}{6\pi \eta R_i} \quad \mu_i = \frac{D_i q_i}{k_B T}
\]  

(2)

with the Boltzmann constant \( k_B \), the system temperature \( T \), the dynamic system viscosity \( \eta \), the hydrodynamic radius \( R_i \), and the net electric charge \( q_i \). The resulting partial differential equation,

\[
\frac{\partial c_i}{\partial t} = \frac{k_B T}{6\pi \eta R_i} \left[ \frac{\partial^2 c_i}{\partial x^2} - \frac{q_i E}{k_B T} \frac{\partial c_i}{\partial x} \right]
\]  

(3)

is also known as the Nernst-Planck equation under static electromagnetic conditions. The system is subject to the initial conditions

\[
c_i(x, 0) = \begin{cases} 
  c_i,0 & 0 \leq x \leq w_0 \quad \forall \text{ species } i \\
  0 & \text{ otherwise}
\end{cases}
\]  

(4)

With the initial zone width \( w_0 \), the detector location \( x = L_d \), and the boundary conditions \( c_i(\pm \infty, t) = 0 \quad \forall \text{ species } i \), this system admits a closed-form solution that can be stated as follows,

\[
c_i(L_d, t) = \frac{c_i,0}{2} \left[ \text{erf} \left( \frac{L_d - v_i t}{2\sqrt{D_i t}} \right) - \text{erf} \left( \frac{L_d - w_0 - v_i t}{2\sqrt{D_i t}} \right) \right]
\]  

(5)

and where \( \text{erf} (z) \) denotes the error function with argument \( z \). Except for the three parameters that are being estimated by the nonlinear least-squares optimization procedure \([13] [16]\), i.e. the electrophoretic velocity \( v_i \), the diffusion coefficient \( D_i \) and the initial relative concentration \( c_i,0 \) in the injected plug, all physical quantities are known prior to \( t = 0 \), and thus Eq. 5 is a function of time only.

For a complex mixture of \( N \) distinct molecular species subject to separation by capillary zone electrophoresis, the entire electropherogram measured at the detector location \( x = L_d \) is modeled as

\[
M(L_d, t) = \sum_{i=1}^{N} c_i(L_d, t)
\]  

(6)

The program uses nonlinear least-squares optimization to find the model \( M^*(L_d, t) \) that optimally approximates the data \( \Delta \), i.e. the electropherogram, in the least-squares sense. In particular, it finds the best estimates over finite intervals for \( v, D \) and \( c_0 \) for all \( N' \) detected species to minimize the \( L^2 \) norm of \( [M^*(L_d, t) - \Delta] \), i.e.

\[
\min_{v,D,c_0} |M^*(L_d, t) - \Delta|^2.
\]  

(7)

From the optimal model \( M^*(L_d, t) \), we then obtain the best estimates for the electrophoretic velocities and the diffusion coefficients for the \( N' \) detected species. With these quantities, we invoke the Stokes-Einstein relations, and compute numerical estimates for the hydrodynamic radii and the net electric charges for all detected species.
Operation and Features

In order to apply \textit{awkt} successfully, the CZE experiment must be conducted adhering to some well-defined experimental requirements. In particular, the initial width of the injected zone $w_0$, the distance from the capillary entrance to the detector location $L_d$, the electric field strength $E$, and the system temperature $T$ must be known to some acceptable level of accuracy. In fact, it is critical for the analysis, that the experiment is conducted in strictly isothermal conditions. Otherwise, the peak broadening characteristics cannot conceivably be anticipated with the currently employed model. However, the current model is amenable to modification. Any corrections to the current model may be added in the source code without affecting the overall structure of the program, as long as no additional unknown parameters are introduced. Furthermore, the current model can be completely replaced by another, potentially more complex model. It is also imperative to ensure that the background electrolyte concentration remains constant throughout the experiment.

The program has a number of user-adjustable data processing and analysis options. A list of the currently available options is shown below. For details and use please see the program documentation \cite{21}.

- peak detection
- hypothesis testing
- model selection
- single file and batch processing
- concurrent batch processing
- time-range selection
- noise filtering (Gaussian kernel smoothing)
- data and model visualization
- global baseline rectification
- local baseline rectification
- local baseline estimation

We give a brief high-level overview of the program operations in the default setting. The program loads the data and its associated experimental conditions from two separate files. A procedure attempts to detect as many peaks as possible given a preset detection sensitivity. Another procedure then, splits the data into frames that each contain at least one but potentially multiple peaks, again depending on a preset upper bound and peak separation tolerance. The objective is to have the fewest number of peaks possible in a frame because it vastly simplifies and speeds up the optimization process. Subsequently, given the supplied model derived from the Nernst-Planck equation, or any other suitable model for that matter, the data in the identified data frames are subject to nonlinear least-squares optimization. Since the peak detection procedure may have missed real peaks, we enable another algorithm to vary the number of detected peaks in order to potentially find a better model to the data, however only accept the new model if the value of the associated objective function has decreased sufficiently (user-adjustable). Finally, for each identified data frame we use its best model, i.e. the best estimates for the diffusion coefficients and electrophoretic velocities, together with the experimental conditions and the Stokes-Einstein relations, to compute the associated hydrodynamic radii and the net electric charges.

Software Requirements

The estimation tool \textit{awkt} is built on top of the \textit{Ceres Solver} \cite{16}, an open source C++ library for modeling and solving large, complicated optimization problems. In order to compile \textit{awkt} successfully, the current version requires that \textit{Ceres} and some of its dependencies such as the Eigen C++ template library \cite{17}, are installed on a user’s system. The \textit{Ceres Solver} can
awkt utilizes the C++ library persistenceId \cite{18} for peak detection. For visualization of the data and the computed models, the gnuplot-i C++ interface \cite{19} is employed. The corresponding source files for both projects are included in the awkt repository such that no further action is required.

For file and directory management, the Boost C++ library \cite{20} is used. This library is open source, and can also typically be installed via package manager on a Linux system. All other features of awkt are implemented using the capabilities of the C++ Standard Template Library. For more updated information on installation and requirement details, please refer to the documentation of awkt \cite{21}.

Summary

We introduce awkt, a fast, versatile and robust C++ parameter estimation tool. It is aimed at assisting in the characterization efforts of molecules, and in particular of globular proteins, using capillary zone electrophoresis. In order to apply the estimation tool successfully, CZE experiments are to be conducted with carefully established and maintained experimental conditions. The program can then utilize the generated electropherogram to estimate numerical values for the diffusion coefficient, hydrodynamic radius and the net electric charge of all molecular entities that are detected during the analysis process. From the estimated parameter values, associated physical quantities such as the molecular weight may also be amenable for estimation.

References

[1] Nicholas W. Frost, Meng Jing and Michael T. Bowser. Capillary Electrophoresis ANAL CHEM, 82 (12), 4682–4698, 2010

[2] Bernd Moritz, Volker Schnaible, Steffen Kiessig, Andrea Heyne, Markus Wild, Christof Finkler, Stefan Christians, Kerstin Mueller, Li Zhang, Kenji Furuya, Marc Hassel, Melissa Hamm, Richard Rustaudi, Yan He, Oscar Salas Solano, Colin Whitmore, Sung Ae Park, Dietmar Hansen, Marcia Santos, Mark Lies. Evaluation of capillary zone electrophoresis for charge heterogeneity testing of monoclonal antibodies J CHROMATOGR B, 983–984, 101–110, 2015

[3] Claire M. Ouimet, Cara I. D’amico and Robert T. Kennedy. Advances in capillary electrophoresis and the implications for drug discovery EXPERT OPIN DRUG DIS, 12 (2), 213–224 , 2017

[4] Pier Giorgio Righetti, Roberto Sebastianio, Attilio Citterio. Capillary electrophoresis and isoelectric focusing in peptide and protein analysis PROTEOMICS, 13 (2), 325–340, 2013

[5] Angelique Stalmach, Amaya Albalat, William Mullen, Harald Mischak. Recent advances in capillary electrophoresis coupled to mass spectrometry for clinical proteomic applications ELECTROPHORESIS, 34 (11), 1452–1464, 2013

[6] Xuemei Han, Yueju Wang, Aaron Aslanian, Marshall Bern, Mathieu Lavallée-Adam, and John R. Yates III. Sheathless Capillary Electrophoresis-Tandem Mass Spectrometry for Top-Down Characterization of Pyrococcus furiosus Proteins on a Proteome Scale ANAL CHEM, 86 (22), 11006–11012, 2014

[7] Yihan Li, Philip D. Compton, John C. Tran, Joanna Ntai, Neil L. Kelleher. Optimizing capillary electrophoresis for top-down proteomics of 30–80 kDa proteins PROTEOMICS, 14 (10), 1158–1164, 2014
[8] James W. Jorgenson and Krynn DeArman. Lukacs. Zone electrophoresis in open-tubular glass capillaries ANAL CHEM, 53 (8), 1298–1302, 1981

[9] Paul D. Grossman, Joel C. Colburn. Capillary Electrophoresis: Theory and Practice Academic Press, 1992

[10] Harold P. Erickson. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy BIOL PROCED ONLINE., 11, 32–51, 2009

[11] Reijenga, J.C., Kenndler, E. Computational simulation of migration and dispersion in free capillary zone electrophoresis, I: Description of the theoretical model. J CHROMATOGR A, 659, 403–415, 1994

[12] Michael S. Bello, Roberta Rezzonico and Pier Giorgio Righetti. Use of Taylor-Aris Dispersion for Measurement of a Solute Diffusion Coefficient in Thin Capillaries. SCIENCE, 266, No. 5186, 773–776, 1994

[13] Sandip Ghosal. Electrokinetic Flow & Dispersion in Capillary Electrophoresis ANNU REV FLUID MECH, 38, 309, 2006

[14] Ofer Dagan, Moran Bercovici. Simulation Tool Coupling Nonlinear Electrophoresis and Reaction Kinetics for Design and Optimization of Biosensors ANAL CHEM, 86, 7835–7842, 2014

[15] Jorge Nocedal, Stephen J. Wright. Numerical Optimization Springer Series in Operations Research and Financial Engineering, 2nd Ed.,Springer-Verlag New York, 2006

[16] Sameer Agarwal and Keir Mierle and Others. Ceres Solver – http://ceres-solver.org

[17] Benoît Jacob and Gaël Guennebaud and Others. Eigen C++ template library for linear algebra – http://eigen.tuxfamily.org

[18] Yeara Kozlov and Tino Weinkauf. Persistence1D – https://github.com/yeara/Persistence1D

[19] N. Devillard and Others. Gnuplot-i – http://ndevilla.free.fr/gnuplot/

[20] BOOST C++ Libraries – http://www.boost.org

[21] Christian Löffeld. awkt – https://github.com/christian-loeffeld/awkt