CHAPTER 11

UltraScan – A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments

BORRIES DEMELER

Abstract

With the introduction of the Beckman XL-I/A in the early 1990s, digital data acquisition from the analytical ultracentrifuge has made it possible to readily analyze sedimentation data with a PC. UltraScan,[1] one of the several data analysis software packages available for such analysis, is a comprehensive, multi-platform software package designed to not only address the tasks associated with the interpretation of analytical ultracentrifugation (AUC) experiments, but also to provide guidance with the design of sedimentation experiments and to address data management challenges arising in multi-user facilities. The availability of data in digital format has led to a wealth of AUC data, which demand new approaches for dealing with the large amounts of data generated. To address this challenge, UltraScan now includes a laboratory information management system (LIMS) which is based on a relational database. In addition, UltraScan integrates many routines for designing, analyzing, interpreting, and displaying sedimentation equilibrium and velocity experiments in a user-friendly graphical interface to make sophisticated analysis methods approachable for a wide audience, including new and less-experienced users. In this publication, an overview of the design philosophy of the software and its algorithms is presented, and the various modules, methods, and their applications are discussed. Examples for each method are shown, and a guide for the experimental design and implementation is given.

1 Overview

The UltraScan software package is the result of a collaborative effort extending over more than a decade now with many contributors.[2] Over the years, many modules have been added to the software, increasing functionality and generality. This publication offers an overview of the capabilities of the current state of the software. The software has been created with the intent to provide a convenient and high-performance data analysis environment for AUC experiments. The software is programed in the platform independent C++/Qt language, and graphical versions for Linux, Unix, Windows, and Mac OS-X are available for free download.

The software addresses multiple issues related to the analysis: starting with the experimental design, the software offers modules that aid in the optimal design of the experiment. Once the data have been acquired, the data are edited with editing modules designed for different optical systems and centerpiece geometries and converted into a binary format suitable for rapid loading and analysis by a series of experimental analysis procedures. Sedimentation velocity analysis can be performed with several methods: determination of model-independent $G(s)$ distributions and partial concentrations can be accomplished with the van Holde–Weischet method, for samples with only a few discrete species, direct boundary fitting using finite element solutions of Lamm equation models can be used to determine s, D, molecular weight, frictional coefficients, and partial concentrations. The second moment analysis provides weight-average sedimentation coefficients and offers useful diagnostics for sedimentation experiments. Sedimentation equilibrium experiments can be analyzed with a global nonlinear least-squares regression analysis. Fitting statistics can be ascertained through either a bootstrap or Monte Carlo analysis. UltraScan includes a Beowulf interface, which allows this analysis to also be performed on a parallel Linux cluster to reduce computing time. Multi-wavelength analysis is complemented by a global extinction fitter suitable for determining intrinsic extinction coefficient distributions. Data analysis results can be directly exported in html format to a webserver, such that experimental results are available on the Internet.

To assist with data management tasks, an issue of concern especially for facilities with multiple machines and multiple investigators, an external relational database is incorporated into UltraScan. The database component of UltraScan has both a C++ implementation for direct database access from the software as well as a PHP interface for web-based access. In the relational database it is possible to link experimental data with associated information and not only store the data itself, but also the links to associated data. Examples for information that can be logically connected to the experimental data include: information describing details and the context of the experiment, the optical system used, the investigator, the cell type, the rotor, and the centerpiece geometry, sequence information of biopolymers (for purposes of estimating partial specific volume, molecular weight, extinction coefficients), buffer composition (for predicting hydrodynamic corrections) as well as data analysis results. PHP modules regenerate analysis reports on the fly by retrieving the relevant information through the Internet. An array of utility functions facilitates archiving of experimental data, calculation of ancillary constants from buffer composition and primary protein and nucleic acid sequences, molecular modeling, and common file operations.

2 Organization

The UltraScan software is organized into multiple modules that can be executed separately or simultaneously. Each module is a separate binary file linked against the main UltraScan library, which is dynamically loaded. This organization optimizes stability and memory needs, especially in a multi-user environment. Where appropriate, the code is multi-threaded to afford additional performance in a multi-processor environment. The software modules interface directly with several external programs. Database functionality is provided by MySQL, an Internet-enabled relational database. Context-specific help files, documentation, and data reports have been written in html and are accessed with an external browser such as Explorer, Netscape, or Mozilla. Data archives are generated with the public domain *tar* and *gzip* utilities to maintain cross-platform compatibility.[3] Web-based access is provided through the *Apache* webserver,[4] and a PHP[5] interface to the MySQL database.[6] Below, each module is discussed separately.

3 Modules

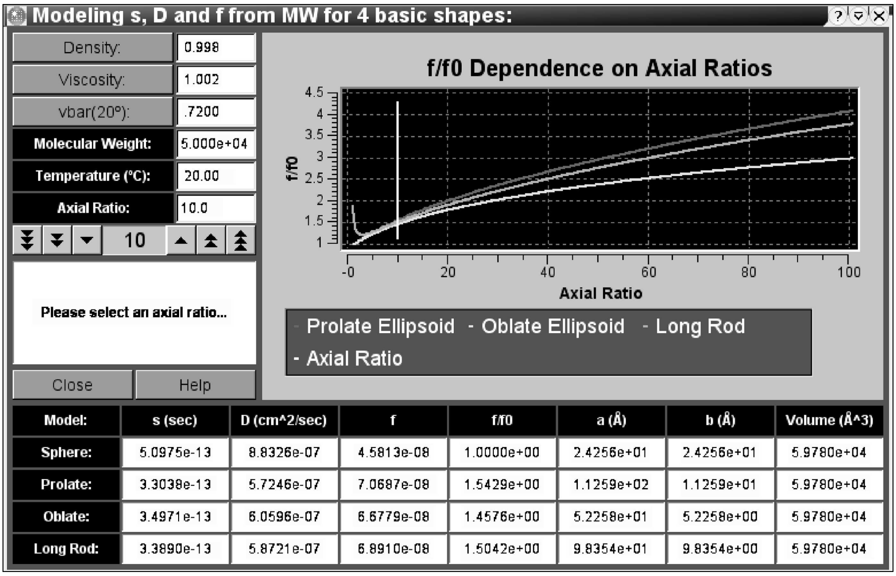
3.1 Experimental Design and Simulation Modules

An important part of AUC is the appropriate design of the run conditions, in particular the loading concentration, the rotor speed and the duration of the experiment, as well as wavelength selection. Suboptimal rotor speeds, incorrect run times, and scan times can lead to reduced information content of the resulting data. Incorrect sample concentration may lead to underrepresentation of a reversibly self-associating oligomeric species, while an incorrect wavelength selection can lead a noisy or nonlinear signal. Often, some information is known about the protein from Sodium dodecyl sulfate (SDS) gel electrophoresis, mass spectrometry, gel filtration, or sequencing before analytical ultracentrifuge experiments are performed. Such information can be exploited to optimize the design of the experiment. A good estimate of the monomer molecular weight as well as an approximation of the molecular shape can then be used to predict the sedimentation and diffusion coefficient for the molecule. A modeling module facilitates the prediction of these coefficients from molecular weight, partial specific volume, buffer conditions, and shape model. Coefficients can be predicted for a sphere as well as by specifying the axial ratio for an oblate or prolate ellipsoids, or a long rod model (Figure 1).

Once sedimentation and diffusion coefficients are available, finite element solutions of the Lamm equation[7] are used to simulate experimental conditions. From these simulations, optimized experimental parameters can be derived. For velocity experiments, the simulations are used to predict the highest possible speed compatible with the scan speed, which is dependent on the number of cells to be scanned, the optical system, and the number of desired scans. For equilibrium experiments, the program will predict appropriate speeds based on the reduced molecular weight σ of the samples. Speeds appropriate for equilibrium experiments can be obtained by substituting the appropriate values for σ in Equation (1):

$$\text{rpm} = \frac{30}{\pi}\sqrt{\frac{2RT\sigma}{M(1-\bar{v}\rho)}} \tag{1}$$

where R is the gas constant, T the temperature in K, M the molecular weight, ρ the density of the buffer, $\bar{v}$ the partial specific volume and σ is defined by



Modeling s, D and f from MW for 4 basic shapes:

Density: 0.998

Viscosity: 1.002

vbar(20º): .7200

Molecular Weight: 5.000e+04

Temperature (°C): 20.00

Axial Ratio: 10.0

Please select an axial ratio...

Model:	s (sec)	D (cm^2/sec)	f	f/f0	a (Å)	b (Å)	Volume (Å^3)
Sphere:	5.0975e-13	8.8326e-07	4.5813e-08	1.0000e+00	2.4256e+01	2.4256e+01	5.9780e+04
Prolate:	3.3038e-13	5.7246e-07	7.0687e-08	1.5429e+00	1.1259e+02	1.1259e+01	5.9780e+04
Oblate:	3.4971e-13	6.0596e-07	6.6779e-08	1.4576e+00	5.2258e+01	5.2258e+00	5.9780e+04
Long Rod:	3.3890e-13	5.8721e-07	6.8910e-08	1.5042e+00	9.8354e+01	9.8354e+00	5.9780e+04

Figure 1 *Module for the modeling of molecular parameters based on molecular weight, axial ratio, hypothetic molecular shape, partial specific volume, and buffer conditions. Hydrodynamic corrections can be imported from predefined buffer files and the partial specific volume of peptides can be estimated from the peptide sequence*

$$\sigma = \frac{M\omega^2(1-\bar{v}\rho)}{2RT} \tag{2}$$

It is recommended to perform equilibrium experiments with 4–5 speeds ranging in σ values between 1 and 4 for sample loading concentrations of 0.3, 0.5, and 0.7 OD. These conditions provide sufficient curvature and variation in the equilibrium gradient, exploit the entire linear range of the absorbance spectrum and provide sufficient variability in the gradient to improve global-fitting statistics. For reversibly self-associating systems, a selection of multiple wavelengths can effectively enhance the concentration range covered by the experiment, so that multiple oligomers are adequately represented in the signal. Next, finite element simulations are used to model the approach to equilibrium. The length of time required to achieve the equilibrium condition is given by satisfying the following equality:

$$k = \sum_{i=1}^{n}\left[L(s, D, r_i, t, C_1, \omega) - C_a \exp\left(\frac{M\omega^2(1-\bar{v}\rho)(r_i^2 - r_a^2)}{2RT}\right)\right]^2 \tag{3}$$

where k is a user-selectable constant determining the residual error (typically less than the error resulting from experimental noise), n the number of radial discretization steps, $L(s, D, r_i, t, C_1, \omega)$ is the Lamm equation solution, r_i the radius at the ith position, C_1 the loading concentration, t the time, and C_a, r_a are the concentration and the radius at a reference point in the cell, respectively. Mass conservation is guaranteed by the relationship

$$C_1(r_b - r_m) = \int_{r_m}^{r_b} C_a \exp\left[\frac{M\omega^2(1-\bar{v}\rho)(r^2-r_a^2)}{2RT}\right] \mathrm{d}r \qquad (4)$$

where r_m is the meniscus position and r_b the position at the bottom of the cell. Once the equilibrium condition in Equation (3) has been satisfied, the resulting concentration distribution is used to initialize the next higher speed and the process is repeated until all speeds have been simulated. The time required to reach equilibrium for each speed is recorded and can be used to program the analytical ultracentrifuge (allowing for additional time required to perform the actual scanning, as well as a separate scan spaced 4–6 h apart to verify that equilibrium has been reached).

3.2 Editing Modules

The process of editing sedimentation data consists of eliminating noisy data regions and to determine meniscus position and in the case of velocity data, determining a baseline and plateau estimates for each scan, and writing all data to a binary copy of the original data, which remain unchanged during editing. This binary representation can be loaded by any UltraScan analysis module, eliminating the need to re-edit experimental data for each analysis method individually. Another advantage of the binary data format is the significant increase in reading speed.

During editing, the software will extract all salient information from the file header, sort all scans according to scan number, cell, centerpiece channel, and wavelength and create a separate binary representation of all scans belonging to the same cell, centerpiece, and wavelength. A separate binary run information file is created to save the rotor speed, time, temperature, plateau, data range, wavelength, and the $\omega^2 t$ integral of each scan. Centerpiece geometry, rotor type, optical system, meniscus, average temperature, as well as rotor acceleration corrections are calculated and saved in this structure as well. If the database module is used, this file also includes the date of the experiment, the database links for the buffer composition file, as well as the links for the peptide and nucleic acid sequences from each channel, and the database name and address used for the experiment.

A special process is required for editing interference velocity data. Such data are often affected by several systematic imperfections, such as shifting baselines, integral fringe offsets, and time-independent noise. The following algorithm is employed to correct these artifacts: initially, all scans are aligned along the air-to-air region. Next, scans are shifted by integral fringe numbers until the numerical integral of each successive scan produces a monotonically decreasing function. In order to correct the baseline shift (‘breathing’), often found in interference data, this integral function is then fitted to a polynomial. Each scan’s residual of this fit is a reflection of the baseline offset variation in each scan (Figure 2). Subtracting the residuals from each corresponding scan corrects the variations in the baseline offsets. Finally, a provision is made to subtract baseline scans from each scan to correct for time-independent noise caused by heterogeneity in the refractive index of cell windows, which can be substantial for low concentration samples.

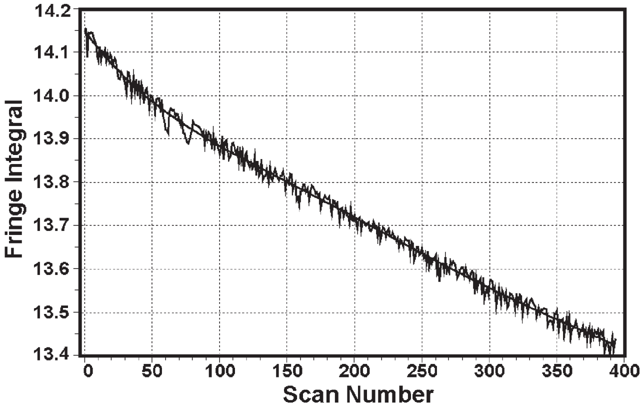


Figure 2 *Fringe integrals of 400 interference scans showing the variation in baseline offsets fitted to a polynomial function. The frequency of the fluctuation remains relatively constant over time suggesting a systematic cause for this fluctuation. Subtracting the residuals of this fit from the scans corrects the baseline offsets of the scans*

3.3 Velocity Analysis Modules

During sedimentation velocity experiments, the molecules under investigation are subject to two transport processes, sedimentation and diffusion. These processes depend both on molecular weight and the frictional properties of a molecule. The sedimentation coefficient is directly proportional to the molecular weight, and inversely proportional to the frictional coefficient. Diffusion is inversely proportional to the frictional coefficient. For simple systems of one or two components it is therefore possible to obtain both sedimentation and diffusion coefficients, which provide information about molecular weight and shape. However, if multiple components are present, the individual signal strengths from diffusion and sedimentation from each component are reduced, and often not sufficiently resolvable to identify with necessary certainty the shape and molecular weight contributions. In such cases it is still possible to reliably define a sedimentation coefficient distribution, but the corresponding diffusion coefficient distribution may remain hidden.

UltraScan offers methods for the analysis of either situation. A system with just a few noninteracting components can often be well described by whole boundary modeling. UltraScan uses finite element solutions of the Lamm equation[7] to perform this modeling. The finite element solutions are fitted with a nonlinear least-squares fitting algorithm to the concentration distributions from the experimental scans to obtain sedimentation coefficients, diffusion coefficients, partial concentrations, association constants for reversibly self-associating systems, as well as meniscus position and concentration dependency parameters. It is also possible to account for optical artifacts such as baseline drift and sloping plateaus in the model. An example for a fit of a noninteracting two-component system is shown in Figure 3.

Most nonlinear least-squares fitting algorithms rely on steepest descent calculations to find a set of optimal parameters that minimize the χ^2 condition. The calculation of

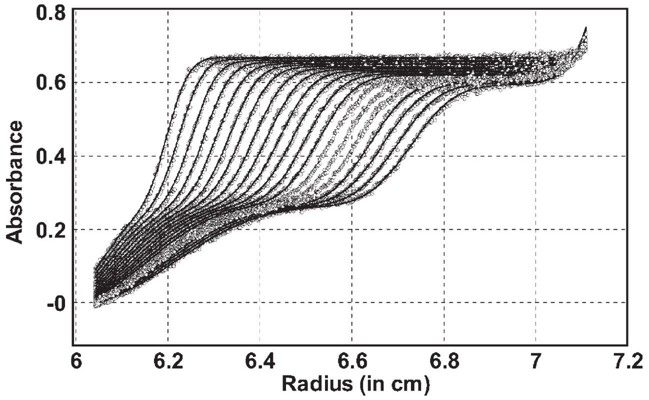


Figure 3 *Finite element fit of a sedimentation velocity experiment containing a two-component, noninteracting system. The continuous black lines show the finite element solution, the grey circles represent the experimental observations*

the steepest descent direction requires the calculation of a Jacobian matrix, which contains the elements of partial derivatives of the solution with respect to the estimated parameters. There exist numerous approaches to obtain the derivatives. An analytical evaluation of the derivatives provides the most accurate result, and facilitates convergence of the least-squares optimization.

However, for finite element solutions of the Lamm equation an analytical evaluation of the partial derivatives of all parameters is not available, and alternative methods have to be used. The finite element fitting modules offer two nonlinear least-squares fitting algorithms for whole boundary fitting of velocity experiments. The first method is based on an approach developed by Ralston and Jennrich[8] which uses first-order tangent approximations to estimate the partial derivatives. While this approach works well for most cases, for complex cases with many floating parameters the inherent error in the tangent approximation approach adversely affects convergence properties, and convergence can stall in a local minimum.

Better convergence properties can be obtained with the second optimization method, which is based on automatic differentiation.[9] In this approach, a tape of chain rule operations is recorded from the evaluation of the finite element solution, which is used to calculate the entries of the Jacobian matrix using the ADOLC C++ library[10] for automatic differentiation. The accuracy of the derivatives evaluated by automatic differentiation is equivalent to analytical solutions. A drawback to this approach is the large storage requirement for the tape, which can slow down computation on smaller computers.

For all systems, including those that are not well described by just a few discrete components, the van Holde–Weischet analysis[11] offers a model-independent graphical transformation of the data that results in diffusion-corrected sedimentation coefficient distributions. This method relies on the realization that while sedimentation is a transport process proportional to the first power of time, diffusion is a transport process proportional to the square-root power of time. In the limit of infinite time the effect on transport by diffusion is negligible compared to transport by sedimentation.

Extrapolation to infinite time can therefore provide diffusion-corrected sedimentation coefficient distributions. In this approach, apparent sedimentation coefficients are calculated from a fixed number of boundary fractions at each scan. Corresponding boundary fractions are then extrapolated to infinite time in a plot of apparent sedimentation coefficients *vs.* the inverse square root of time of the scan (Figure 4).

Special care has to be taken to account for boundary effects at the meniscus and the bottom of the cell, as well as differential radial dilution rates for different components in the system. These effects are taken into consideration in the enhanced van Holde–Weischet algorithm that is implemented in UltraScan. The details of this algorithm are described in ref. 12. The enhanced van Holde–Weischet analysis provides sedimentation coefficient distributions that can be displayed both as integral distribution plots $G(s)$ and as differential distributions $g(s)$ as shown in Figure 5. Given suitable estimates for the partial specific volume and the frictional ratio f/f_0, both distributions can also be transformed into molecular weight distributions. Weight-average sedimentation coefficients are also calculated from the sedimentation coefficient distributions.

An alternative method for obtaining weight-average sedimentation coefficients is provided by the second moment analysis. It provides weight-average sedimentation coefficients for each individual scan. However, only scans with a stable plateau and clear meniscus can provide reliable estimates of second moment weight-average sedimentation coefficients.

3.4 Equilibrium Analysis Module

Sedimentation equilibrium experiments always should include multiple equilibrium scans taken at different rotor speeds, different loading concentrations, and multiple wavelengths. Each individual scan contains a unique distribution of concentration

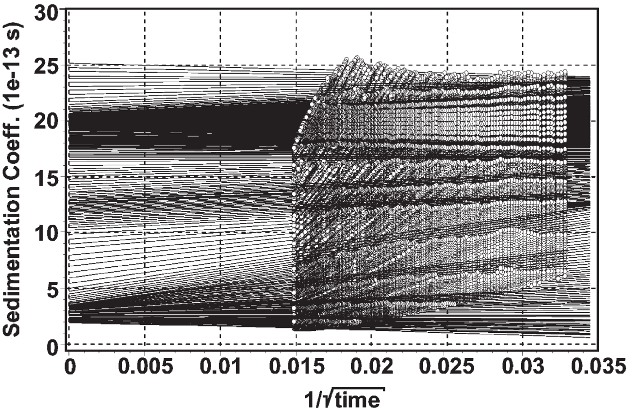


Figure 4 *van Holde–Weischet analysis of a sedimentation velocity data for a heterogeneous system containing multiple degradation products of a protease. Extrapolations to infinite time are made for increasing boundary fractions (bottom to top) and from early scans to late scans (right to left)*

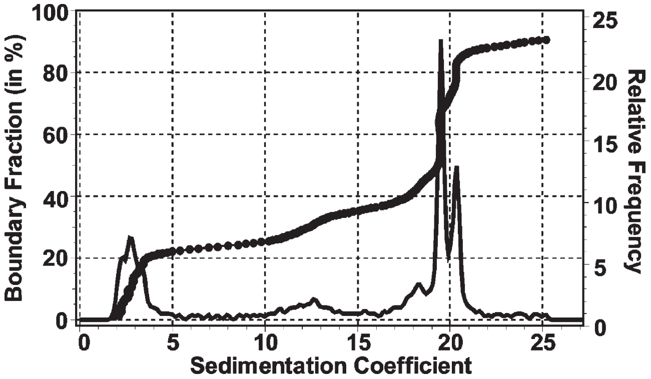


Figure 5 *Integral distribution plot (G(s)) and differential distribution plot (g(s)) for van Holde–Weischet analysis shown in Figure 4*

observations that reflect the composition of the sample. If the correct model is chosen for a global analysis approach, each observation has to satisfy the same global parameters such as molecular weight, extinction coefficients, and association constants in the model function. Furthermore, scans taken at different wavelengths exploit the varying extinction properties of the sample and allow analysis of a sample over a larger range of loading concentrations.

In a reversibly self-associating system a change in the loading concentration will change the ratio of larger oligomers *vs.* smaller oligomers or monomeric forms of the sample, thereby enhancing the signal of one species compared to another. A global analysis approach is essential for a reliable interpretation of data from multicomponent systems. Multiple experiments can be simultaneously analyzed by UltraScan through nonlinear least-squares fitting to a preset or user-defined global model as proposed in ref. 13. All models available in UltraScan are based on three basic forms:

(1) A noninteracting system:

$$C(r) = \sum_{i=1}^{k} \exp\left[\frac{\ln(a_i) + M_i\omega^2(1-\bar{v}_i\rho)(r^2-r_\mathrm{a}^2)}{2RT}\right] + c \tag{5}$$

(2) A reversibly self-associating system:

$$C(r) = \sum_{i=1}^{k} \exp\left[\frac{i\ln(a_1) + \ln\dfrac{iK_{1,i}}{(el)^{i-1}} + iM_1\omega^2(1-\bar{v}+\rho)(r^2-r_\mathrm{a}^2)}{2RT}\right] + c \tag{6}$$

AQ:2 (3) A reversibly hetero-interacting system for two components A and B:

$$C(r) = \exp\left[\frac{\ln(a_\mathrm{A}) + M_\mathrm{A}\omega^2(1-v_\mathrm{A}\mu)(r^2-r_\mathrm{ref}^2)}{2RT}\right]$$

$$+ \exp\left[\frac{\ln(a_{\mathrm{B}})+M_{\mathrm{B}}\omega^2(1-\nu_{\mathrm{B}}\rho)(r^2-r_{\mathrm{ref}}^2)}{2RT}\right]$$

$$+ \exp\left[\frac{\ln\left(a_{\mathrm{A}}a_{\mathrm{B}}K_{\mathrm{A,B}}\dfrac{e_{\mathrm{AB}}}{e_{\mathrm{A}}e_{\mathrm{B}}l}\right)+(M_{\mathrm{A}}+M_{\mathrm{B}})\omega^2(1-\left(\dfrac{\nu_{\mathrm{A}}+\nu_{\mathrm{B}}}{2}\right)\rho(r^2-r_{\mathrm{ref}}^2)}{2RT}\right] + c \quad (7)$$

where $C(r)$ is the concentration at radius r, k the maximum number of components or oligomer states, a the concentration of the component at the reference radius r_{ref}, M the molecular weight, $K_{1,i}$ the equilibrium constant for association state i, e the extinction coefficient, l the path length, c a baseline offset, ω the radial velocity, $\bar{v}$ the partial specific volume, and ρ the density of the buffer. UltraScan does not place any limits on the maximum number of scans that can be included in a global analysis. Each scan is allowed a separate entry for ρ and c.

Parameters can be floated or kept fixed. Parameters that can be floated include the concentrations at the reference point for each species, the association constants, the baseline offset, and the molecular weights or the partial specific volumes. Molecular weights, association constants, and partial specific volumes are considered global parameters, all other parameters are local. Global parameters are constrained to be identical for all scans described by the global model, local parameters are allowed to vary for each scan. The concentration a at the reference point r_{ref} is fitted as the natural log of the reference concentration. This effectively constrains the fit to positive values of the reference concentration only, avoiding spurious oscillations of the amplitudes of exponential terms that are common when the model is overdetermined. Molecular weights reported are corrected for density and partial specific volume. Equilibrium constants are displayed as association or dissociation constants, and are reported in molar units.

3.5 Optimization

The models used for fitting of equilibrium experiments shown in Equations (5)–(7) are nonlinear in the parameters and require iterative nonlinear least-squares fitting approaches. For highly nonlinear problems with many parameters the likelihood of convergence is quite dependent on the proper choice of initial parameter estimates. The closer the initial parameter estimates are to the least-squares solution, the greater is the likelihood that the solution will converge at the global minimum.

UltraScan will initialize all parameter estimates with reasonable guesses to improve the convergence properties. This is accomplished by linearizing the model for a single ideal species and fitting the amplitudes for each scan by general linear least squares. The molecular weight estimate is obtained by performing a line search over the nonlinear parameter of the model. For two- or three-component models the scans are divided into two or three equal sections and each section is fitted individually to generate an estimate for one of the components. This approach results in parameter estimates that automatically initialize all fitting parameters to reasonable values that facilitate a more stable convergence of the fit.

Parameter optimization is performed by one of the three optimization methods: for nonlinear least-squares optimization problems, the Levenberg–Marquardt method[14,15,16] and the quasi-Newton method are implemented. The Levenberg–Marquardt method applies a scaling factor to the diagonal of the information matrix to prevent it from becoming singular during optimization. This approach results in a robust method that is not too sensitive to the choice of initial guesses, and therefore useful for obtaining an initial fit.

However, for problems with many nonlinear parameters the Levenberg–Marquardt method has a tendency to converge in a local minimum, and a second approach is needed to find the global minimum. For cases where the solution is close to optimal the quasi-Newton method performs best, and it can help to overcome solutions that are trapped in a local minimum. The quasi-Newton method employs the BFGS formula (named after their developers Broyden,[17] Fletcher,[18] Goldfarb,[19] and Shanno,[20]) and a line search algorithm to update an approximation to the Hessian, which is needed to find the steepest descent direction for locating the global minimum for the least-squares solution. Due to the complexity of the error surface, multi-variate optimization often fails in the task of finding the global minimum solution, and the solution can get trapped in a local minimum. To alleviate this problem, UltraScan employs an automatic convergence algorithm that alternates between the Levenberg–Marquardt and quasi-Newton methods until a global minimum has been found, and neither method can improve the χ^2 value of the fit any further. An example for a global equilibrium fit for multiple speeds and loading concentrations is shown in Figure 6.

For linear least-squares problems, a nonnegatively constrained approach is used (the NNLS algorithm by Lawson and Hanson[21]). A linear problem is given when the molecular weights of each exponential term in Equation (5) are predetermined, and only the coefficients a_i of the linear combination of exponential terms have to be found by the fitting routine:

$$C(r) = \sum_{i=1}^{k} a_i \exp\left[\frac{M_i \omega^2 (1 \nu_i \mu)(r^2 - r_\mathrm{a}^2)}{2RT}\right] + c \tag{8}$$

AQ:3

where the baseline offset c is simply the zeroth-order term of this linear combination. In such a fitting problem, many terms can be used to account for all M_i present in the sample, and the range and spacing of all M_i species is provided by the user. The nonnegatively constrained Lawson–Hanson algorithm prevents coefficients a_i from turning negative, and coefficients of terms accounting for molecular weight species present in the system are fitted with a nonzero value, while those not present in the sample are assigned a value of zero. Contributions of each molecular weight species from all scans are summed to generate a molecular weight composition histogram which represents the relative concentration of each species in the sample. An example for such a molecular weight distribution is shown in Figure 7.

When studying biological systems the question often arises if the presence of molecular weight heterogeneity in a sample results from the presence of multiple, noninteracting components or from a reversibly self-associating oligomerization. UltraScan provides two diagnostic plots for global equilibrium fits to address this question: (1) a plot of the average molecular weight of all fitted data points *vs*. concentration, and (2) a plot of the average molecular weight of all fitted data points *vs*. the square of the

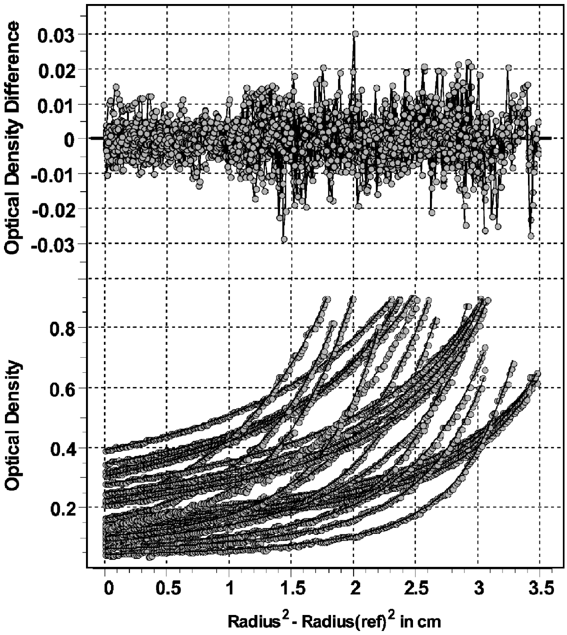


Figure 6 *Global equilibrium fit of 24 scans (six loading concentrations and four speeds). Circles represent experimental observations, solid lines represent the fitted model. Residuals are shown in the top panel. overlays, and the associated fitted model are shown in the bottom panel*

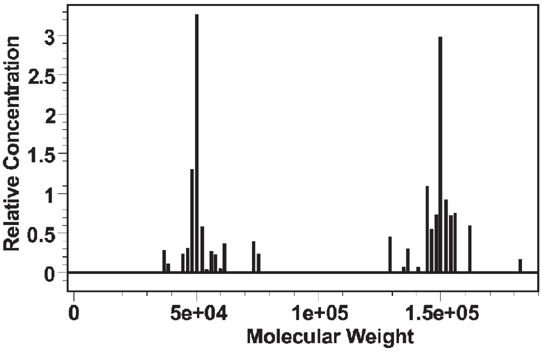


Figure 7 *Molecular weight distribution for a simulated equilibrium experiment containing approximately equal amounts of a 50 and 150 kDa species. The distribution was generated by globally fitting 24 scans for multiple speeds and loading concentrations with the nonnegatively constrained general least-squares fitting method by Lawson and Hanson*

radius. These plots are useful when the data is fitted with models derived from Equation (5) or (8). Since Equation (6) is a subset of Equation (5) or (8), reversibly self-associating systems may also be fitted with Equation (5) or (8), yielding equally good fits.

Due to mass action, for a reversible self-associating system, the ratio of oligomer concentration over monomer concentration stays constant according to

$$K_{1,i} - \frac{[iM]}{[M]^i} \tag{9}$$

Here, $K_{1,i}$ is the equilibrium constant for the ith association state of the monomer, M. Therefore, all scans will follow the same trace in plot (1), regardless of loading concentration or rotor speed. This is shown in Figure 8(a). For a noninteracting system the same plot will show a distribution that is not constrained by the equilibrium constant, and each speed or loading concentration will assume a different distribution. This is shown in Figure 8(c).

However, if the average molecular weights from the self-associating system are plotted against the square of the radius, concentrations at different radial positions will be sufficiently different at different rotor speeds and loading concentrations, and each scans will produce a different trace. This is illustrated in Figure 8(b). For the

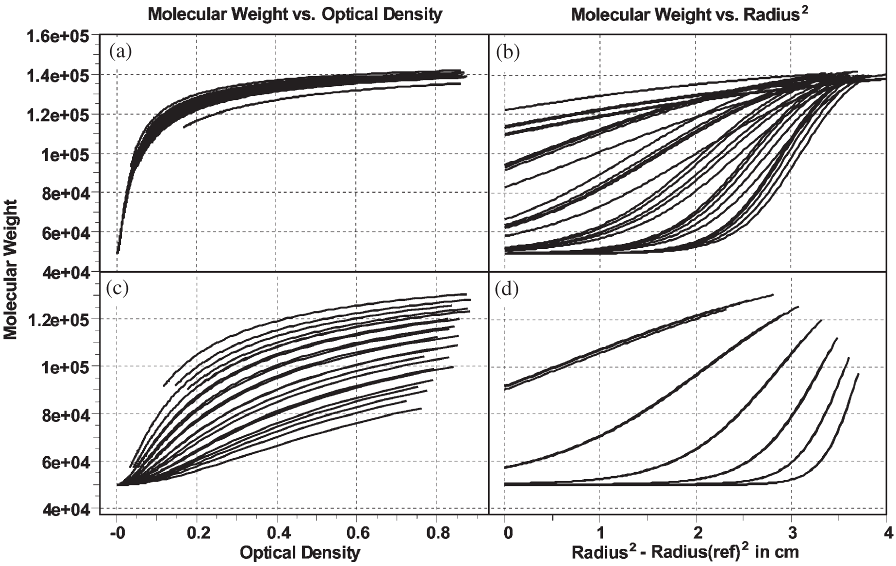


Figure 8 *Equilibrium diagnostic plots in UltraScan for two different simulated multi-component systems. The first system is a reversibly self-associating monomer–trimer system with a monomer molecular weight of 50 kDa (a), (b), and the second system is a two ideal noninteracting species model where component one is 50 kDa and component two is 150 kDa (c), (d). Six speeds and five loading concentrations were simulated for both systems. Both systems were fitted with the model shown in Equation (8) shown in (a), (c) is a plot of average molecular weight vs. concentration for each scan. In graphs (b), (d) a plot of average molecular weight vs. the square of the radius is shown. Note that all traces for a self-associating system overlay when molecular weight is plotted against concentration, but traces overlay for different loading concentrations when molecular weight is plotted against the square of the radius only for a noninteracting system. See text for a more detailed explanation of the diagnostics provided by these plots*

noninteracting system, molecular weight distributions are not the result of mass action, and different loading concentrations have no effect on the ratio of one species over the other. As a result, components will equilibrate solely according to the centrifugal force applied, and all scans at the same rotor speed will follow the same trace. An example for such a case is shown in Figure 8(d). Hence, visual inspection of plots (1) and (2) for systems fitted to a noninteracting model allows the investigator readily to distinguish between noninteracting and self-associating systems.

3.6 Extinction Fitting Module

For self-associating systems and hetero-associating systems the measurement of equilibrium experiments at different wavelengths can add important information to the analysis. For self-associating systems, different extinction coefficients at different wavelengths can be exploited to obtain measurements at markedly different concentrations. By bracketing a large concentration range in the experiment, the signal from both monomer and oligomeric species can be enhanced. For hetero-associating systems, the presence of chromophores at different wavelengths for each component can provide important constraints on the fit when multiple wavelengths are fitted in a global fit. When these constraints are included in the fit, it is important that extinction coefficients are correctly identified in the model for each wavelength and for each component to assure internal consistency of the model.

UltraScan offers a global extinction fitting module that generates intrinsic extinction profiles from wavelength scans at different concentration for each component in the fit. These profiles can be normalized with known extinction coefficients at one wavelength and imported into the fitting model such that the appropriate extinction coefficients are applied for each component in the fit. This also helps in cases where hypochromicity is an issue as long as the pure form of the associated species can be measured. The algorithm is applied as follows: for each component, wavelength scans are taken at 3–5 loading concentrations that are chosen so that each chromophore is represented at least once with an optical density between 0.5 and 1.0 absorbance units. All scans are then globally fitted to the following linear combination of Gaussian terms:

$$E_i(\lambda) = c_i \sum_{j=1}^{n} \exp\left[\frac{(-\lambda - a_j)^2}{2\sigma_j^2} \right] + b_i \tag{10}$$

where E_i the extinction profile of wavelength scan i, c_i the relative concentration of scan i, λ the wavelength, a_j is the position of chromophore j, σ_j the width of the peak produced by chromophore j, n the number of chromophores, and b_i a baseline offset for scan i. a_j and σ_j are global parameters required to be identical for all wavelength scans. Once a global solution has been found, the factor c_i is normalized with a known molar extinction coefficient at a desired wavelength and all other wavelengths are scaled accordingly. An example for a global extinction fit is shown in Figure 9. UltraScan provides the extinction coefficient at 280 nm as an estimate from peptide sequence according to the method of Gill and von Hippel,[22] which can be conveniently used to scale intrinsic extinction coefficient profiles to molar extinction coefficients.

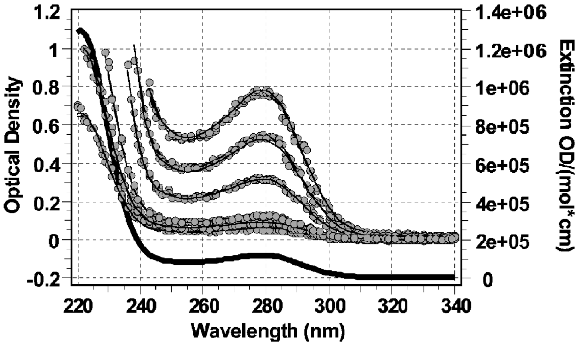


Figure 9 *Global extinction profile fir for a protein. Six separate concentrations in three replicate measurements were globally fitted (grey circles fitted with thin lines) resulting in an intrinsic extinction profile for the protein (heavy line). The extinction profile is calibrated at 280 nm with a known extinction coefficient derived from peptide sequence. The global extinction profile is plotted as the right Y-axis, the absorbance values from the wavelength scans are plotted on the left Y-axis*

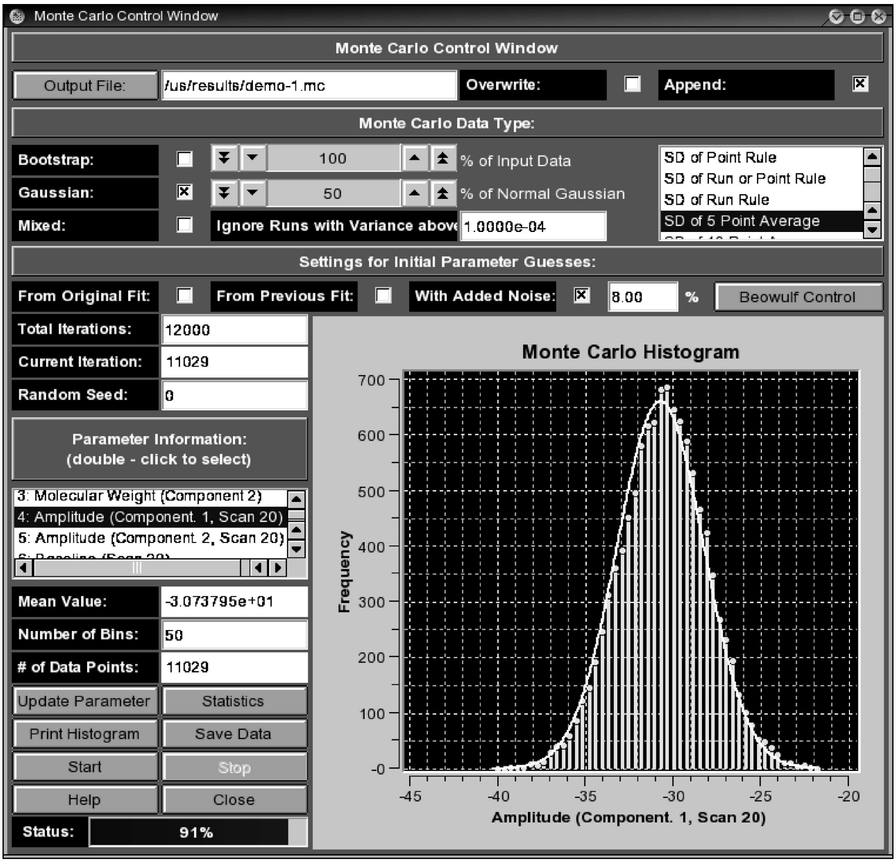
3.7 Monte Carlo Module

All experimental data fitted to a linear or nonlinear model contains experimental noise. The assumption in all fits performed by UltraScan is that the experimental noise is random and that all systematic deviations are accounted for in the model. If these assumptions are satisfied, the statistical confidence of the parameter estimates depends on the magnitude and distribution of the random experimental noise. One possibility for determining the confidence intervals of the parameter estimates would be to perform repeat experiments and fit them to the same model in order to generate a distribution of values for each floated parameter. This distribution could be used to obtain a statistical description of each parameter. Naturally, performing repeat experiments is impractical because of the time and expense considerations.

Instead, a Monte Carlo approach can be used to simulate a repeat experiment, where experimental noise is synthetically generated by a random number generator, and added to the best-fit model values. This will generate a new, synthetic dataset which can be refit with the same model. Repeating this process a sufficient number of times will provide the desired parameter distributions from which a statistical description of the parameters can be obtained.

Two different approaches for generating synthetic noise are possible: (1) generating Gaussian noise comparable to the residuals from the best fit to the experimental data, and adding it to the best fit; or (2) a bootstrap approach that randomly moves residuals from the best fit to the experimental data to a different position. UltraScan supports both the approaches. Because noise levels differ nonlinearly with absorbance AQ:4 (see Figure 10) a true representation of the random noise is often not achieved with the bootstrap approach, and the Gaussian noise generation is preferable.

In the Gaussian approach, the residuals in a frame of 5–10 points in the neighborhood of each experimental observation are averaged. The average then provides the standard deviation input for the Box–Muller function.[23] This function will return



Monte Carlo Control Window

Output File: /us/results/demo-1.mc
Overwrite: □
Append: ☒

Monte Carlo Data Type:

Bootstrap: □ 100 % of Input Data
Gaussian: ☒ 50 % of Normal Gaussian
Mixed: □ Ignore Runs with Variance abov 1.0000e-04

SD of Point Rule
SD of Run or Point Rule
SD of Run Rule
SD of 5 Point Average

Settings for Initial Parameter Guesses:

From Original Fit: □ From Previous Fit: □ With Added Noise: ☒ 8.00 %

Total Iterations: 12000
Current Iteration: 11029
Random Seed: 0

Parameter Information:
(double - click to select)

3: Molecular Weight (Component 2)
4: Amplitude (Component. 1, Scan 20)
5: Amplitude (Component. 2, Scan 20)

Mean Value: -3.073795e+01
Number of Bins: 50
# of Data Points: 11029

Status: 91%

Figure 10 *UltraScan Monte Carlo control window. A parameter distribution for each fitted parameter can be displayed and the analysis can be performed both locally or remotely on a parallel cluster (Unix only)*

Gaussian distributed random variables with a standard deviation that is commensurate with the original residuals. These values are added to the best-fit values predicted by the model function, which are then refit. This method assures that local changes in the size of the residuals are accurately reflected in the synthetically generated dataset. Five thousand Monte Carlo iterations are generally sufficient to obtain a reliable probability distribution of each fitted parameter. The Monte Carlo analysis returns a distribution of parameter values for each fitted parameter.

Once a sufficiently populated distribution has been obtained, the distribution is fitted to a Gaussian function, and the following statistics are reported: minimum and maximum, mean, median, skew, kurtosis, mode, standard error, standard deviation, variance, correlation coefficient, and 95 and 99% confidence intervals. The confidence intervals may be nonsymmetric. One problem with the application of the Monte Carlo analysis is the prohibitive computational expense, especially for large models with many parameters. To address this issue, UltraScan offers a Beowulf module that allows the Monte Carlo analysis to be performed on a cluster of appropriately configured Unix computers. The speed improvement is observed by performing the Monte Carlo analysis in parallel scales linearly with each added computational node. To assure that each node starts at a different point in the pseudorandom sequence generated by the random number generator, each node is initialized with a different random speed.

3.8 Utility Modules

UltraScan incorporates a number of utility functions to accomplish frequently performed tasks. The protein analysis module accepts peptide sequence information in Genbank format and calculates from it molecular weight, partial specific volume, and extinction coefficients at 280 nm. The partial specific volume ($\bar{v}$) is calculated by summing the weight fraction of partial specific volume contributions from each amino acid as reported by Durchschlag,[24] and the calculation of the extinction coefficient is based on the partial extinction contribution at 280 nm from the denatured amino acids tyrosine, tryptophane, and cysteine as reported by Gill and von Hippel.[22] A second utility provides for the calculation of hydrodynamic corrections resulting from density and viscosity contributions from a collection of commonly used buffer components according to methods outlined in the reference.[25] Values for density and viscosity are interpolated from polynomial fits to concentration data obtained from the Sednterp database files.[26] Molecular weights of RNA and DNA molecules can be calculated from the sequence, including the contributions from various counterions. Several file utilities assist the user in archiving experimental data, analysis results and experimental reports, to merge data files from different directories and to reorder the files into a single run, and to rename cell descriptions that have been incorrectly entered during data acquisition. A diagnostics module is also available that is useful for reviewing single files or identifying file errors, such as truncated or corrupted data acquisition files.

3.9 Database Modules

To facilitate the management of AUC projects and the large volume of data files collected and generated during analysis, UltraScan includes a LIMS that is designed to address the needs of a multi-user AUC facility. Many analytical ultracentrifuges are employed in multi-user environments, and managing multiple experiments, experimental data, and associated information quickly turns into a complex task. The UltraScan LIMS is based on an Internet-capable relational database with both an UltraScan interface and a web-based interface.

The LIMS addresses multiple objectives: first, it serves as a data repository for all data relevant to an AUC experiment. Second, it serves as a data retrieval and web-based presentation tool, and finally it assists multiple investigators to manage separate experiments. The first goal of the database is to provide logically linked storage for experimental project descriptions, experimental designs, the experimental data and analysis results, for peptide and nucleic acid sequence files, as well as for buffer composition files and images from gels or absorbance spectra. All experimental and result data are stored in a compressed format and are associated with investigator identifications, which facilitates data analysis and retrieval in multi-user environments.

Experimental data are committed to the database by linking an entire run, which may consist of multiple cells and channels in each cell, with an experimental project description and run profile, which are described by the investigator. Each channel is linked with up to three peptide or nucleic acid sequence files, one buffer file, the date of the experiment, the investigator information, as well as the name of the database, the run type (*i.e.*, equilibrium, velocity, diffusion, or wavelength experiment), and the optical measurement method (absorbance, interference, fluorescence, or intensity). Supplemental data needed for data analysis are stored in separate tables and logically linked to the experimental data.

During analysis, any required supplemental data are automatically retrieved from the database, processed on the fly and integrated into the analysis. For example, the appropriate buffer composition is associated with each centerpiece channel and retrieved during the analysis. Using the composition information, viscosity and density corrections are calculated and automatically applied in the analysis. Similarly, peptide sequence data are retrieved and used to calculate an estimate for the partial specific volume of the peptide, which is then applied to the analysis. All automatic values can be manually overridden by the user.

Centerpiece geometry, rotor type, and channel number are also stored in the database and associated with each dataset. This information is used to calculate a precise position for the channel bottom. Each rotor in use at a laboratory can be individually calibrated to determine speed-dependent rotor stretching. Precise information on rotor stretching and channel geometry are needed to determine the position of the channel bottom. A precise value for this position is needed to calculate accurate mass integrals for equilibrium experiments and to set boundary conditions for finite element solutions. Centerpiece geometry, rotor calibration, peptide, nucleic acid, and buffer information only have to be entered once and can be linked from any experiment instead of laboriously recalculating these values by hand for each new experiment.

After data analysis is completed, the LIMS searches the hard drive for all analysis results from each method applied during the analysis and commits the results to the database. An overview of the database structure and the relationships is shown in Figure 11. The relational database engine used in UltraScan is the open source MySQL database.[6] To facilitate data exchange and collaborations, the user can switch among multiple databases and access data from remote sites. Database access is restricted through username and password authentication to prevent unauthorized access to any private data.

3.10 Web-Based Modules

Several web-based PHP applets allow a remote user to interact with the database contents. After authentication, the investigator can enter project requests, peptide sequences, compose buffers, and upload related images. Next, the investigator can enter sample information and specify the type of experiment to be performed. Each item entered by the investigator receives a description which allows the investigator to later search and retrieve or review each item. In order to accommodate a multi-user

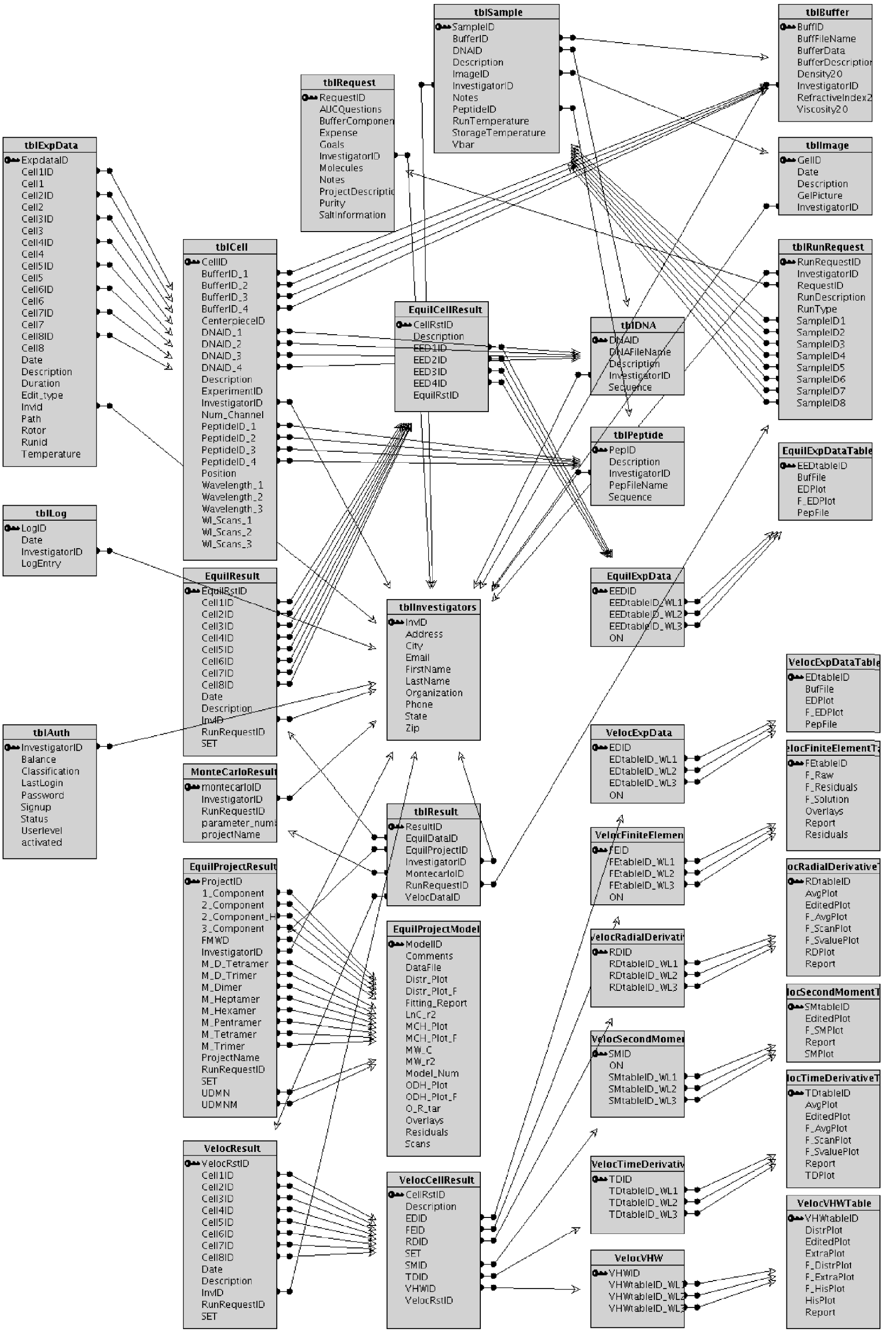


Figure 11 *Overview of the UltraScan LIMS database structure version 1.0 Relationships are shown by arrows*

environment, the LIMS offers different user levels: investigator, technician, supervisor, and data analyst. Each user level has a different authentication status allowing read-only, write, or no access to different data segments.

Once experimental results have been uploaded, the investigator can retrieve the results through the web interface and review or download them to his or her personal computer. A tracking system allows the LIMS users to track the status of each experiment (*i.e*., designed, scheduled, in progress, uploaded). In addition to the database-oriented data management, analysis results can also be presented as a web page and saved directly to a folder served by a webserver. UltraScan will generate an html encoded report file and provide convenient access to the experimental data and analysis results by linking graphs, data files, and analysis reports to the report file.

4 Summary

UltraScan represents a comprehensive software package addressing a wide range of experimental situations and providing multiple analysis methods for sedimentation velocity and equilibrium experiments. Design and modeling functions assist with the correct design of an experiment, and a Beowulf module allows computationally intensive tasks to be performed on a parallel computer. An Internet-based relational database is available for managing AUC experiments in a multi-user facility environment. The software is written in C++ using the portable QT toolkit. Software packages for Unix, Linux, Microsoft Windows, and Macintosh OS-X can be downloaded for free from http://www.ultrascan.uthscsa.edu.

Acknowledgments

I would like to thank Jeffrey Hansen for critically reviewing this Chapter. The UltraScan software development was supported by the National Science Foundation through grants NSF DBI 9724273 and NSF DBI 9974819, as well as through the Howard Hughes Medical Institute grant 76200-550802.

References

1. B. Demeler, UltraScan – A Software Package for Analytical Ultracentrifugation Experiments. The University of Texas Health Science Center at San Antonio, Department of Biochemistry, San Antonio TX, 78229 USA, 2004, http://www.ultrascan.uthscsa.edu
2. A listing of contributors and references can be found at: http://www.ultrascan.uthscsa.edu/references.html
3. The *gzip* and *tar* utilities are available from ftp://ftp.gnu.org/gnu
4. The *apache* webserver is available from http://www.apache.org
5. The *PHP* scripting language is available from http://www.php.net
6. The MySQL database engine is available from http://www.mysql.com
7. B. Demeler and H. Saber, *Biophys. J.*, 1998, **74**, 444–454.
8. M. L. Ralston and R. I. Jennrich, Technometrics. 1978, **20**, 7–14.
9. A. D. Griewank, Juedes, H. Mitev, J. Utke, O. Vogel and A. Walther, *ACM TOMS* 1990, **22**, 131–167.

10. A. Griewank and A. Walther, ADOL-C. Technische Universität Dresden, Institut für Wissenschaftliches Rechnen, D-01062 Dresden, Germany, 2004, http://www.math.tu-dresden.de/wir/project/adolc/
11. K. E. van Holde and W. O. Weischet, *Biopolymers*, 1978, **17**, 1387–1403.
12. B. Demeler and K. E. van Holde, *Anal. Biochem.*, 2004, in press.
13. M. L. Johnson, J. J. Correia, D. A. Yphantis and H. R. Halvorson, *Biophys. J.*, 1981, **36**, 575–588.
14. P. R. Gill, W. Murray and M. H. Wright, in *Practical Optimization*, Academic Press, London, 1981, 136–137.
15. K. Levenberg, *Quart. Appl. Math.*, 1994, **2**, 164–168.
16. D. Marquardt, *SIAM J. Appl. Math.*, 1963, **11**, 431–441.
17. C. G. Broyden, *J. Inst. Maths. Applics.*, 1970, **6**, 76–90.
18. R. Fletcher, *Comput. J.*, 1970, **13**, 317–322.
19. D. Goldfarb, *Math. Comput.*, 1970, **24**, 23–26.
20. D. F. Shanno, *Math. Comput.*, 1970, **24**, 647–656.
21. C. L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.
22. S. C. Gill and P. H. von Hippel, *Anal. Biochem.* 1989, **182**, 319–326.
23. G. E. P. Box and M. E. Muller, *Ann. Math. Stat.*, 1958, **29**, 610–611.
24. H. Durchschlag, in *Thermodynamic Data for Biochemistry and Biotechnology,* H.-J. Hinz, (Ed), Springer, New York, 1986, 45–128.
25. T. M. Laue, B. D. Shah, T. M. Ridgeway and S. L. Pelletier, in *Analytical Ultracentrifugation in Biochemistry and Polymer Science*, S. E. Harding, A. J. Rowe and J. C. Horton (eds), Royal Society of Chemistry, Cambridge, UK, 1992, 90–125.
26. J. Philo, Sednterp database files, personal communication.