Colloid Polym Sci
DOI 10.1007/s00396-007-1699-4

ORIGINAL CONTRIBUTION

Monte Carlo analysis of sedimentation experiments

Borries Demeler · Emre Brookes

Received: 25 March 2007 / Accepted: 15 May 2007

Abstract High resolution analysis approaches for sedimentation experiments have recently been developed that promise to provide a detailed description of heterogeneous samples by identifying both shape and molecular weight distributions. In this study, we describe the effect experimental noise has on the accuracy and precision of such determinations and offer a stochastic Monte Carlo approach, which reliably quantifies the effect of noise by determining the confidence intervals for the parameters that describe each solute. As a result, we can now predict reliable confidence intervals for determined parameters. We also explore the effect of various experimental parameters on the confidence intervals and provide suggestions for improving the statistics by applying a few practical rules for the design of sedimentation experiments.

Keywords Two-dimensional spectrum analysis · Genetic algorithms · UltraScan · Analytical ultracentrifugation · Molecular weight determination · Curve fitting

B. Demeler (✉)
Department of Biochemistry,
University of Texas Health Science Center at San Antonio,
San Antonio, TX, USA
e-mail: demeler@biochem.uthscsa.edu

E. Brookes
Department of Computer Science,
University of Texas at San Antonio,
San Antonio, TX, USA

Introduction

Analytical ultracentrifugation (AUC) experiments allow the researcher to study biological and synthetic polymers and macromolecules under native solution conditions by measuring the hydrodynamic and thermodynamic properties of the sample in a centrifugal force field. Molecules under the influence of this force field experience transport by sedimentation and diffusion. This transport determines how the solutes in the sample distribute in the centrifuge cell during the experiment. The observed signal is the combined concentration gradient of each solute in the solution. The concentration gradient evolves during the experiment and forms a moving boundary, and is commonly measured over multiple time points with either UV/visible absorbance, Rayleigh interference or fluorescence emission detectors. At the end of the experiment, an equilibrium gradient is formed, which balances sedimentation and back diffusion from the cell wall at the bottom of the cell.

The collected data are then fitted to a mathematical model describing the sedimentation and diffusion transport using least squares methods and parameters such as sedimentation and diffusion coefficients, molecular weights, partial concentrations, and the number of solutes present in the mixture are determined. Simultaneous knowledge of the sedimentation and diffusion coefficients can then be used to further calculate shape information. In this work, we will investigate the effect experimental noise has on the accuracy of these parameter determinations and present a method for measuring the magnitude of the effect. As a result, we can quantify the reliability of the parameter estimation and suggest ways in which data acquisition and experimental design can be optimized during routine sedimentation analysis.

One of the prerequisites for successful parameter estimation is the availability of a model that explains all signals in the data short of random noise, including systematic instrument error. For sedimentation velocity experiments, several high-resolution methods have been proposed over the past few years [1–7]. They all share the use of finite element solutions of the Lamm equation [8]. For equilibrium experiments, the Lamm equation simplifies to an exponential gradient that can be modeled analytically. Sedimentation velocity experiments can also be used to identify systematic noise. Such noise can be determined either algebraically [9], eliminated by differencing [10] or removed by baseline subtraction. Gaussian random noise cannot be eliminated and remains convoluted with the concentration signal and causes an uncertainty in the parameter estimate. Due to the random nature of the noise, each repetition of an experiment with the same sample will produce slightly different results. If unlimited resources were available, many repetitions of each experiment could be performed and provide a probability distribution for each fitted parameter. This distribution could be used to calculate a standard deviation and confidence intervals for each parameter. Clearly, repeating experiments is an unrealistic proposition. In this study, we propose the use of a suitable substitute, namely, the computer-generated simulation of repeated experiments. With the advent of fast computers and parallel-distributed computing technology, it is feasible to simulate the effect of random noise using random number generators. With this information in hand, it is then possible to determine a numerical value that describes the confidence the investigator can have in each fitted parameter of the model. This approach is termed a Monte Carlo analysis, reflecting the stochastic nature of our approach [11]. While the proposed methods are equally appropriate for use in real experimental systems, we will employ realistic simulations of experimental data instead of using actual experimental data for the purposes of this study because simulation offers the advantage that the exact input parameters are known and can be used to unequivocally identify error contributions from noise. All methods described in this work are implemented in the freely available UltraScan data analysis software, which can be downloaded from our website [12].

Description of the method

Realistic experimental data can be generated by adding noise of the same quality as that which is typically observed in an ultracentrifugation experiment to the Lamm equation sedimentation model. For velocity experiments, such a model requires the sedimentation and diffusion coefficients from any solute present in the solution and for equilibrium experiments, the molecular weights of each solute need to be known. Other model parameters also enter the equation, such as temperature, buffer density, buffer viscosity, cell geometry, rotor speed, and rotor acceleration.

To represent noise accurately, we need to take into consideration that the quality of the noise can vary and that certain instrument limitations contribute to the noise. For example, in the UV/visible absorbance detection method, the observed noise is proportional to the intensity of light striking the photomultiplier tube (PMT). This is a smooth function that increases with increasing absorbance, and it is dependent on the PMT response properties to intensity and the wavelength of the incident light. The precise physical basis of this function is not important, as long as we can determine the magnitude of the noise reproducibly for each absorbance and wavelength value. In addition to the random noise, the data often contain time- and radially invariant noise contributions that among other sources, result from contaminations in the optical path, such as scratches or deposits on the cell windows, and time-dependent variations affecting the detector. For interference optics, refractive index heterogeneities in the cell windows can cause a time-invariant noise contribution, and slight changes in the optical pathlength resulting from heating and cooling can cause time dependent, but radially invariant offsets. We can represent the noise contributions of a typical sedimentation scan by Eq. 1 where S_t is the total observed signal, L is the signal from the Lamm equation solution, which is to be determined, N_{ti} is the time-invariant noise contribution, and N_r is the random noise, which is some measured function of wavelength λ and intensity I. N_{ri} represents the radially invariant noise.

$$S_t = L + N_{ti} + N_r(\lambda, I) + N_{ri} \tag{1}$$

To determine the contribution of N_r, we employ the following approach: For each wavelength of interest, we collect several hundred velocity scans that cover the absorbance range typically used for experiments. The scans are fitted with a finite element solution of the Lamm equation using the two-dimensional spectrum analysis (2DSA) [3], and simultaneously, time and radially invariant noise is eliminated according to the procedures outlined in Schuck and Demeler [9]. Alternatively, we describe below a model-independent approach for removing just the radially invariant baseline offsets that can be applied before fitting.

To avoid contributions from incorrectly modeled data signals, only experiments producing perfectly random residuals (as judged by visual inspection of the residuals and their bitmaps) by this procedure are included. Then for each radial point two values are determined: (1) the magnitude of the residual at that point and (2) the value of the absorbance at that point. A plot of the residual's magnitude vs the log of the absorbance value is then

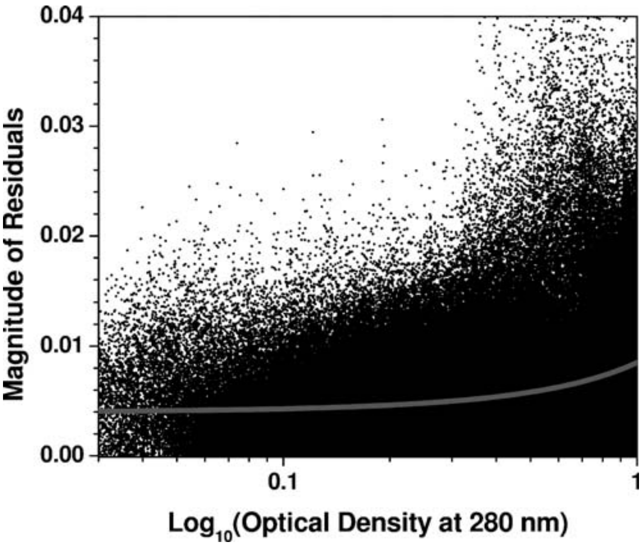

Fig. 1 Random noise contributions in an analytical ultracentrifuge between an optical density of 0 and 1 at 280 nm. The data can be fitted with an exponential function (shown in *gray*), which can be used as a standard deviation for the Box–Muller function (explained in text).

constructed and the data are fitted to an exponential growth function. This function can then be used to predict the magnitude of the random noise for a given instrument and wavelength at any absorbance value.

We have compared three UV absorbance systems in use at the Center for Analytical Ultracentrifugation of Macromolecular Assemblies (CAUMA) at 230, 260, and 280 nm. Even after thorough lamp cleanings, we have found differences in the intensity recorded on each machine and for each wavelength. Intensity variations due to changes in wavelength are to be expected due to the lamp emission profile. Variations between machines at the same wavelength were ascribed to subtle differences in lamp alignment and variations in equipment quality. This suggests that an individual N_r function should be calibrated for each machine. An example for an N_r plot taken at 280 nm for optical density values between 0.0 and 1.0 is shown in Fig. 1. As an alternative to the approach described above, the investigator can use a five-point running average (weighted by a Gaussian kernel) of the absolute magnitude of the residuals from the best fit to the data set, which is to be analyzed by the Monte Carlo approach. Both procedures can be used with equilibrium experiments as well, as long as time-invariant noise has been subtracted, which can be obtained from scans collected during the approach to equilibrium. For equilibrium experiments, we recommend using fits from the fixed molecular weight distribution model as a basis from which to determine the magnitude of residuals. This model is described in reference [13].

Finally, the new residuals are generated with the Box–Muller algorithm [14], which utilizes a random number generator to calculate a new random residual, given a mean and a variance. The variance is the value obtained from the exponential growth function or the running frame average. This function will produce a random variate with Gaussian distribution characteristic. The new residual is added to the best-fit solution of the original data, and the new data are fitted to produce a new parameter estimate. This process is repeated until a statistically significant number of parameter estimates have been obtained and reliable statistics can be determined for each parameter. We recommend at least a hundred repetitions or to continue the Monte Carlo iterations until the distribution statistics vary less than the noise level. A typical example for a molecular weight parameter distribution obtained from a Monte Carlo analysis of an equilibrium experiment is shown in Fig. 2.

After an equivalent sedimentation experiment has been simulated, we now ask the question of how reliably the input parameters are reproduced by fitting the simulated data with standard procedures. Just as in real experiments, each dataset generated with the same input parameters, but with different random noise contribution will produce slightly different values for the fitted parameters. The collection of fitted parameters over many Monte Carlo iterations provides the desired fitting statistics that allow us to specify the reliability of the estimated parameters. Among the statistics determined in UltraScan are the mean, the mode, the skew, the kurtosis, the standard deviation, and the 99% and 95% confidence intervals.

Application of the method

Our Monte Carlo approach can be applied to any system where experimental data are fitted by a least squares method to a parameterized mathematical model and wherever

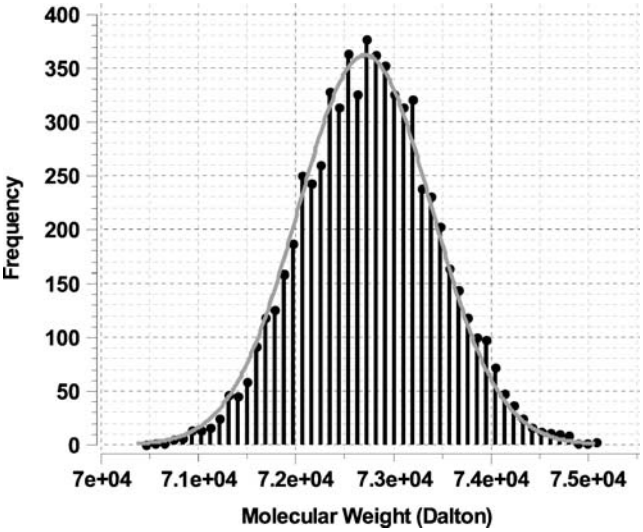

Fig. 2 A Monte Carlo analysis for a single species molecular weight fit. The frequency of each observation describes a distribution, which can be used to evaluate the statistics of the parameter confidence

the confidence intervals of the parameters are of interest. In this study, we provide examples for the application of our Monte Carlo approach to both sedimentation velocity and sedimentation equilibrium data. For modeling sedimentation velocity data, our laboratory has recently developed two novel approaches: (1) the 2DSA [3] and (2) the genetic algorithm analysis (GA) [2]. These methods provide much higher resolution than previous methods have afforded and permit determination of both molecular weight and shape distributions for mixtures of macromolecules even if the mixtures display heterogeneity in molecular weight *and* shape.

Both methods will provide sedimentation and diffusion coefficients for each solute in a sample mixture and a putative number of solutes and their respective concentrations, and both methods can be used in a Monte Carlo implementation. In practice, the resolution, accuracy, and precision of the parameters are strongly correlated with the signal-to-noise ratio. At the lower limit, the signal is lost in the noise and an accurate determination of the parameters is not possible. Without the knowledge of the signal-to-noise ratio for each solute in the mixture, the investigator is confronted with the nontrivial task to identify the reliability of a particular measurement, which leads to uncertainty in interpreting results from these methods. Until now, an independent approach to verify the results was missing and analysis results were inevitably subject to overinterpretation.

$$D = \mathrm{RT}\left[N 18\pi (k\eta)^{3/2}\left(\frac{s\overline{\nu}}{2(1-\overline{\nu}\rho)}\right)^{1/2}\right]^{-1} \tag{2}$$

Our first example illustrates the application of the Monte Carlo analysis to the modeling of a known sedimentation velocity system whose simulation includes realistic noise contributions. Our model system describes a mixture of an irreversibly aggregated species associated in a 1–2–4–8–16 stoichiometry. The monomer is 25 kDa in size and the system is aggregated in an end-to-end fashion and modeled by a non-interacting model. The parameters for this model are shown in Table 1.

$$I_{\mathrm{total},t} = \sum_{r=r_{\mathrm{m}}}^{r_{\mathrm{b}}} I_r \Delta r \tag{3}$$

Table 1 Aggregating five-component system

Solute	Partial concentration (OD)	Molecular weight (kDa)	Shape (f/f_0)
1	0.1	25	1.2
2	0.1	50	1.4
3	0.1	100	1.6
4	0.1	200	1.8
5	0.1	400	2.0

$$P(t) = \sum_{i=0}^{n} a_i t^i \tag{4}$$

Radially invariant noise removal In addition to the model dependent method proposed earlier [2], radially invariant noise can also be eliminated effectively by the following procedure: During editing of the data and before fitting the data, a data range excluding the meniscus and the high absorbance region at the bottom of the cell is selected by the user, and in the case of interference data, integral fringe shifts are adjusted. Next, for each scan the total signal $I_{\mathrm{total},t}$ over this range is determined by integrating between the limit near the meniscus, r_{m} and the bottom of the cell, r_{b} with respect to the radius (Eq. 3).The total signal for each scan is then plotted against the time t of each scan and fitted with a low-order polynomial. The idea here is that the change in signal is a smooth function of time that can be well-approximated with a polynomial function $P(t)$ with order n (Eq. 4). We then fit $\boldsymbol{At}-\boldsymbol{I}_{\mathrm{total}}$ using a general linear least squares approach where $\boldsymbol{A}$ is the coefficient matrix of amplitudes a_i in Eq. 4, $\boldsymbol{t}$ is the vector of scan times, and $\boldsymbol{I}_{\mathrm{total}}$ is the vector of total signals. All deviations of total signals determined in Eq. 3 are due to radially invariant baseline offsets present in a scan. Subtracting the difference between $I_{\mathrm{total},t}$ and $P(t)$ from each radial data point in the scan at time t will correct for any radially invariant baseline offsets.

Parameter space initialization To reduce the calculation time, it is helpful to restrict the parameter search space. Therefore, before applying the 2DSA or GA analysis, it is important to identify the parameter space over which a 2DSA or GA parameter refinement of the model will be attempted. Such an initialization is best accomplished by employing the model-independent van Holde–Weischet method [15], which readily provides diffusion-corrected $G(s)$ distributions from sedimentation velocity experiments. The diffusion coefficient, which is required for the solution of the Lamm equation, can be initially estimated by parameterizing the shape function using the frictional ratio, f/f_0, which is a measure of the globularity of a particle (Eq. 2 where N is Avogadro's number, k is the frictional ratio f/f_0, η and ρ are the viscosity and density of the solvent, repspectively, and s, D, M, and $\bar{\mathrm{v}}$ are the sedimentation and diffusion coefficients, the molecular weight, and the partial specific volume of the solute, respectively, R is the gas constant, and T is the temperature). A reasonable assumption can be made that the shape of the particle ranges somewhere between spherical ($f/f_0-1.0$) and rod-shaped ($f/f_0 \leq 4.0$) for most solutes. The s-value range determined with the van Holde–Weischet method and the assumption of particle shape define the limits of a two-dimensional parameter space covering s and f/f_0 values of interest.

Time-invariant noise subtraction We first simulated the system shown in Table 1 with the following parameters: 60,000 rpm, random noise with a residual mean square deviation (RMSD) of 0.005, and a loading concentration of 0.1 OD for each solute. Time-invariant noise with a magnitude typically observed in low-concentration interference experiments was added to the solution (Fig. 3). The data were fitted with the 2DSA analysis and, simultaneously, a time-invariant noise vector was calculated. After subtraction of the time-invariant noise, the data were compared to the same data without simulated time-invariant noise and found to be equivalent (data not shown). It should be noted that the effect of time-invariant noise on the fitting calculation is significant and it is therefore highly recommended that time-invariant noise should always be taken into account, even in absorbance experiments where such contributions are generally minor. In such cases, a noticeable improvement of the fitting accuracy can still be observed. A similar improvement in the determined parameters can be observed if radially invariant noise has been removed before fitting or is simultaneously removed during fitting (data not shown).

To illustrate this effect, we analyzed the data using the 2DSA method without taking into account time-invariant noise subtraction. The results of this fit are shown in Fig. 4. It can be seen from these results that the 2DSA analysis is very sensitive to time-invariant noise and that the results are far from the simulated parameters. Visual inspection of Fig. 4 does not show a single component identified correctly. Because the time-invariant noise calculation adds a significant amount of CPU time to a calculation, especially when the dataset has many scans, we have implemented in UltraScan a module that allows the user to subtract time- and radially invariant noise determined during 2DSA fitting from the experimental data. The data without time-invariant noise can then be processed by all other analysis methods without any further need for time-invariant noise corrections.

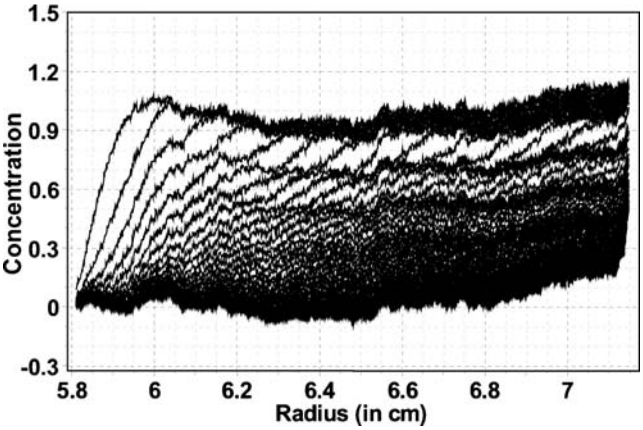

Fig. 3 Time-invariant noise contributions to the experimental data resulting from flaws in the optical path. Refractive index heterogeneities in the windows, window scratches or opaque deposits on windows can cause this problem. Such noise needs to be eliminated during or before fitting. Time-invariant noise is common in low-concentration interference experiments, but also occurs in absorbance experiments due to window contaminations

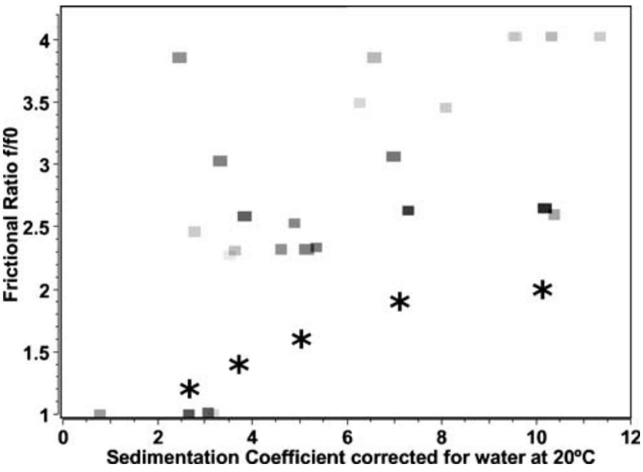

Fig. 4 2DSA analysis of the model system shown in Table 1 without removal of time-invariant noise. Fitting results bear no resemblance with the simulated model (*asterisk*), and the time-invariant noise causes major distortions of the results and invalidates the analysis, stressing the importance of time-invariant noise removal. Each *square* represents a single solute, the *darkness* indicates relative concentration

When the same data are analyzed with the 2DSA approach where simultaneously time-invariant noise has been accounted for, a subset of the initial two-dimensional parameter space is obtained and relative concentrations for all solutes are determined. Although the solution provides a good qualitative representation of the experimental data, the solution cannot be regarded as unique at this point because it is subject to degeneracy and false positives, which occur because of the random noise present in the experimental data and because the true solutes are not necessarily exactly aligned with the grid that was used in the 2DSA analysis (Fig. 5). As a result, the number of solutes found are much higher than in the original system (24 solutes in this analysis), and most of the signals arise from noise. In addition, as can be seen in Fig. 5, the distribution of the solutes detected with the higher amplitudes is in the vicinity of the expected values, however, there are a number of false positives with relatively small amplitudes that fall significantly outside of the expected range.

2DSA Monte Carlo approach The amplitude of most of the false positive signals is related to the magnitude of the random noise still present in the data. With a well-calibrated instrument and a well designed experiment, this noise should be small compared to the signal from the actual solutes present in the data (the random noise level is typically ~1–2% of the total signal for data from a well-calibrated UV/visible absorption system and even less for

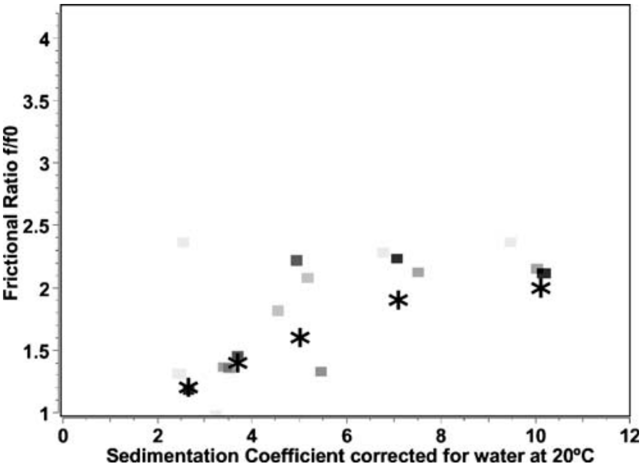

Fig. 5 2DSA of the model system shown in Table 1. Time-invariant noise has been properly removed during the fit and unlike the result shown in Fig. 4, the resulting data provide an approximated description of the correct parameter space (*asterisk*). However, the solution also produces a number of false positives, and the true number of solutes is not apparent. Each *square* represents a single solute, the *darkness* indicates relative concentration

interference data). When Monte Carlo analysis is employed, each iteration will contribute new false positives with small amplitudes. However, the actual signal stays constant in each iteration and hence can be more easily distinguished from the noise contributions. Multiple Monte Carlo iterations are performed to enhance this effect. Applying a Monte Carlo analysis in combination with the 2DSA method will generate a better description of the parameter space because each Monte Carlo iteration will amplify the actual signal as a linear function of the number of iterations, but the random noise contribution only amplifies proportional to the square root of two times the number of iterations.

The result of a 100-iteration 2DSA Monte Carlo analysis is shown in Fig. 6. Although this analysis resulted in an average of 25.86 solutes for each Monte Carlo iteration, a distinct signal of 5 separate solute groups is clearly apparent. The five groups are closely centered on the original solutes simulated from Table 1 and each group's outline includes the target value. As in the single 2DSA fit, signals resulting from noise contributions fall in part significantly outside of the expected range, but are now only inconsequential contributors to the overall signal and can be easily distinguished from the real signal. Still, the signals obtained from this analysis are not parsimonious and additional regularization and refinement is needed.

Parameter refinement and regularization Because the solution obtained in the 2DSA Monte Carlo analysis is overdetermined and not unique, we need to determine the most parsimonious solution. If we adapt Occam's razor for our problem, it can be stated as follows: the most parsimonious solution capable of producing nearly the same RMSD is the preferred solution. We know that for our system, a 5 solute solution is appropriate, but the 2DSA method identified 25.86 solutes on the average. This means that over 80% of all determined solutes are not necessary to describe the data, and that those 80% are due to experimental noise. We found that parameter refinement of the values obtained with the 2DSA is best achieved with a GA implementation, which employs parsimonious regularization (Brookes and Demeler [16]). In contrast to Tikhonov or maximum entropy regularization, parsimonious regularization will not smooth the parameter space, but instead identify only the most relevant solutes in the mixture. This is more consistent with a discrete composition usually present in biological systems. Any uncertainty in the identity of each solute is better expressed through statistical analysis by Monte Carlo methods than through regularization methods like Tikhonov and maximum entropy.

The implementation for the GA analysis is described in detail by Brookes and Demeler [2, 16]. Briefly, an evolutionary paradigm is used to optimize the solution. Populations of 100–200 individual parameter combinations initialized randomly within bounds obtained from the 2DSA analysis are "evolved" for the "survival of the fittest" parameter combination. Random number generators perform mutation, crossover, deletion, and insertion oper-

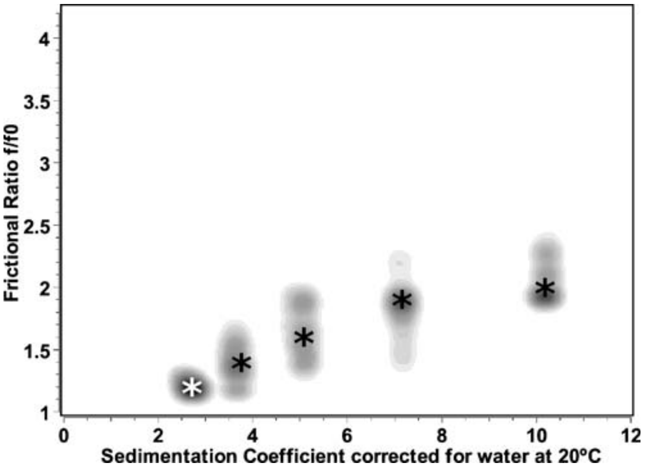

Fig. 6 Monte Carlo/2DSA analysis of the data shown in Fig. 5. As can be clearly seen in this figure, contributions due to random noise are minimized, and the amplitudes of the expected five solutes are amplified by the Monte Carlo analysis. Groups of 5 solutes can be easily distinguished where each group represents approximately 500 individual solutes from the Monte Carlo analysis, and relative concentration is indicated by the *darkness of the gray-shading*. The positions of the five solute groups in the two-dimensional plane correlate well with the original parameters used in the simulation (*asterisk*). Compared to Fig. 5, much higher confidence for each parameter can be obtained from these groups. The asymmetric shape of the confidence interval is due to the 60 krpm speed selection during simulation. While sedimentation is well-resolved, shape information from the diffusion signal is reduced

ations on the parameter combinations to generate new individuals for the next generation. Several selection criteria favor survival of the fittest solution. To guard against loss of parameter diversity, we coevolve multiple demes, each consisting typically of a population of 100 individual solutions, and permit only limited parameter migration between demes. Regularization is achieved by penalizing the fitness of population individuals in direct proportion to the number of solutes represented. Using the five groupings obtained in the 2DSA Monte Carlo analysis, we establish boundary defining parameter limits for a GA optimization. Even if additional groups are identified in this step, unnecessary solutes will be excluded during the GA evolution by employing parsimonious regularization. At this point, a GA Monte Carlo analysis can be used to identify the confidence limits of each parameter. These results are listed in Table 2, and in every case the target parameter value is well within the 95% confidence interval.

Sedimentation equilibrium experiments

Monte Carlo analysis applied to fits of sedimentation equilibrium experiments proves to be a valuable aid for interpreting fitting results. We have previously employed Monte Carlo analysis to determine the confidence intervals of molecular weights and equilibrium constants [17–20, among others]. In this study, we describe the application of Monte Carlo analysis to the model of a fixed molecular weight distribution. This model has been described by Demeler [13] and is reproduced here by Eq. 5 where $C(r)$ is the concentration observed at radius point r, k is the number of fixed M molecular weight solutes in the solution, a is the amplitude of each solute at the meniscus, r_m is the position at reference point m, ρ is the density of the solvent, R the gas constant, ω is the angular velocity, T is the temperature, $\bar{v}$ is the partial specific volume of the solute, and c is the zeroth-order baseline term. In this model, k typically is between 100 and 500 species, and each species is simulated with unity concentration. This model is linear in the amplitudes of each species and can be fitted with the non-negatively constrained least squares solution (NNLS) by Lawson and Hanson [21] to the experimental data.

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

In such a fit, only terms with positive amplitudes are returned, all others are set to zero and do not contribute to the signal. As a result, this method provides a model-independent view of the solute composition of the data, terms with non-zero amplitude represent solutes present in the solution. Just like 2DSA, this method is quite sensitive to noise and produces a degenerate solution. As demonstrated above, we can apply Monte Carlo to amplify the actual signal and suppress the effect of noise on the final molecular weight distribution. It is well-known that equilibrium experiments are inferior to velocity experiments when resolution of multiple species is desired. Part of the reason is that optimal information is only obtained over a fairly narrow speed range, which is different for each molecular weight. If molecular weights are too disparate, any speed used will be a compromise, which will provide less signal for some of the species in the mixture. Furthermore, signal separation of multiple species from the sum of exponential gradients is a difficult problem, and small solution columns in equilibrium experiments compound the lack of signal. Hence, the precision of the information obtained is inferior to what can be obtained in velocity experiments. But by how much? Simulation with Monte Carlo analysis provides important insights.

In Fig. 7, a Monte Carlo analysis of three different fits of data simulated with an RMSD of 0.005 is shown. Each simulation includes three loading concentrations and four speeds. Fig. 7a shows a Monte Carlo analysis of a single species model with 50 kDa molecular weight (speeds 11, 19, 27 and 35 krpm). The fixed molecular weight distribution fit included 100 terms in Eq. 5 equally distributed between 10 and 100 kDa. In Fig. 7b, a system with 2 components, 50 and 100 kDa, both simulated with equal absorbance is shown (speeds 12.0, 17.7, 23.3, and 29 krpm). This fixed molecular weight distribution fit included 100 molecular weight terms equally distributed between 10 and 150 kDa. In Fig. 7b, the same data are analyzed with Monte Carlo with an unconstrained two-

Table 2 Genetic algorithm Monte Carlo statistics for the aggregating five-component system

Solute	Partial concentration (OD)	Molecular weight (kDa)	Shape (f/f_0)
1	*0.10* (0.067, 0.12) [**0.1**, <1%]	*24.4* (23.8, 25.0) [**25**, 2.5%]	*1.21* (1.11, 1.29) [**1.2**, <1%]
2	*0.094* (0.077, 0.11) [**0.1**, 6%]	*55.8* (43.0, 70.8) [**50**, 10.4%]	*1.70* (1.32, 1.83) [**1.4**, 17.6%]
3	*0.10* (0.96, 0.11) [**0.1**, 3.2%]	*91.6* (68.9, 121) [**100**, 9.2%]	*1.59* (1.32, 1.86) [**1.6**, <1%]
4	*0.10* (0.085, 0.11) [**0.1**, <1%]	*166* (131, 260) [**200**, 20.5%]	*1.63* (1.43, 2.20) [**1.8**, 10.4%]
5	*0.10* (0.097, 0.102) [**0.1**, <1%]	*408* (315, 510) [**400**, 2%]	*2.08* (1.77, 2.43) [**2.0**, 4%]

Values reported in italics reflect the mode of the distribution; the 95% confidence intervals of each parameter are shown in parenthesis; boldfaced values in square brackets represent the target value; italic values in square brackets refer to percent error.

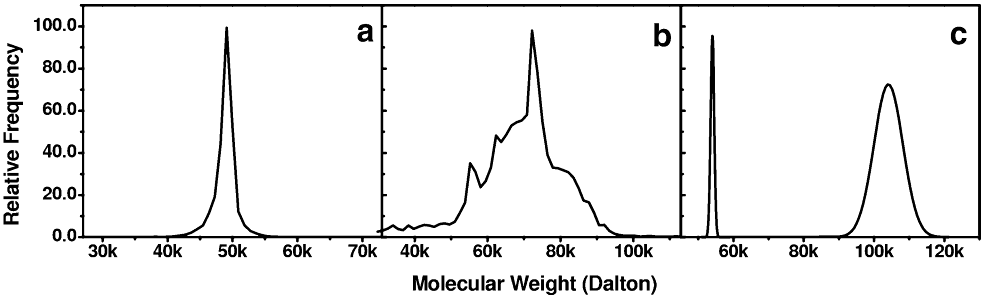

Fig. 7 Monte Carlo analysis of equilibrium data fits. **a** Single species model fitted with a fixed molecular weight distribution with 100 divisions between 10 and 100 kDa. **b** 2-Species model fitted with a fixed molecular weight distribution with 100 divisions between 10 and 150 kDa. **c** 2-Species model (prior knowledge) fitted with a nonlinear least squares fitting algorithm

species model fitted using a nonlinear least squares fitting approach. The results are revealing. While a single ideal species can be correctly identified (Fig. 7a), results from a two component model, when simulated under ideal conditions are much less reliable (Fig. 7b,c). The mode of the distribution in Fig. 7b falls at 72 kDa, approximately the midpoint between the 2 simulated solutes, and the distribution fails to resolve 2 species and shows instead a single, broad distribution with the majority of the signal between 50 and 100 kDa. When prior knowledge of a two-species model is used for the fit, the Monte Carlo analysis produces, as expected, a bimodal distribution centered around 54 and 103 kDa. While the 2-species equilibrium fit produces a bimodal distribution, the centers of the distributions are only within 10% of the expected values, and the 95% confidence intervals does not overlap with the known value for the smaller species. Simulation with the fixed molecular weight distribution model produces a broad distribution of molecular weights with limits close to the known values, but fails to resolve multiple components.

Discussion

We have demonstrated that Monte Carlo methods provide a useful approach for quantifying the effects of random noise on the parameter estimation from sedimentation experiments. We have further shown the importance of eliminating time-invariant noise from experimental data when sensitive methods like 2DSA and GA are employed for parameter estimation. Evaluation of the 95% confidence intervals from the GA Monte Carlo analysis reveals an interesting trend: The standard deviations for molecular weights and frictional ratios get significantly broader for larger species, but improve for the sedimentation coefficient. This can be attributed to reduced diffusion signal, and suggests that improvements in accuracy could be obtained by spinning at a lower speed. This will improve the diffusion signal by giving the sample more time to diffuse. This result has implications for the design of sedimentation velocity experiments. For optimal resolution of multiple species, we therefore recommend a high speed to emphasize the signal from sedimentation. Ongoing investigations in our laboratory focus now on software developments, which support global multispeed analysis for sedimentation experiments using GA, so that low speed diffusion information can be globally fitted for better shape and molecular weight determinations.

The comparison between results from the 2DSA analysis in Fig. 5 and the Monte Carlo analysis in Fig. 6 show that the benefit of signal-to-noise improvements are significant. Signal-to-noise improvements can also be obtained by following a few simple experimental design rules. The following factors will improve the signal-to-noise ratio: For velocity experiments, the largest possible column filling should be used and the full dynamic range of the optics should be exploited over which the instrument produces a linear response (for UV absorbance optics this range generally lies below 1.0 OD). Buffers that absorb should be avoided because their absorbance subtracts signal from the total available dynamic range. A list of suitable buffers can be found on our website at http://www.cauma.uthscsa.edu/buffer2.html. Furthermore, the fastest data acquisition rate setting should be used and delays between scans should be avoided except for all but the slowest sedimenting species. As explained above, sedimentation signal is emphasized by high speed, diffusion signal by low speed. Speed selection for high-speed experiments should represent a good compromise between the amount of data that can be collected during the sedimentation process and the resolution of the sedimentation coefficient that can be obtained.

Equilibrium experiments show a surprisingly broad confidence interval for just two-component mixtures when a model-independent method is used to fit the results. More importantly, even multiple speeds and multiple loading concentrations do not suffice to resolve multiple species. This analysis clearly shows that molecular weight determinations of heterogeneous mixtures by equilibrium sedimentation should be taken with extreme caution. Further simulation work is warranted to identify the effect of speed, loading concentration, molecular weight heterogeneity, and signal-to-noise.

Acknowledgments We would like to thank Mr. Jeremy Mann for the assistance with the supercomputing facility at the UTHSCSA Bioinformatics Core Facility. Funding from the National Science Foundation (Grant #DBI-9974819), the National Institutes of Health (Grant 1 R01 RR022200-01A1), and the San Antonio Life Science Institute (SALSI #10001642) is gratefully acknowledged.

References

1. Demeler B, Saber H (1998) Determination of molecular parameters by fitting sedimentation data to finite-element solutions of the Lamm equation. Biophys J 74(1):444–454
2. Brookes E, Demeler B (2006) Genetic algorithm optimization for obtaining accurate molecular weight distributions from sedimentation velocity experiments. In: Wandrey C, Cölfen H (eds) Analytical Ultracentrifugation VIII, Progr., Springer, Colloid Polym Sci 131:78–82
3. Brookes E, Boppana RV, Demeler B (2006) Computing large non-negative least squares-type problems with an application in biophysics. In: Supercomputing Conference Proceedings
4. Schuck P (2000) Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. Biophys J 78(3):1606–1619
5. Schuck P (1998) Sedimentation analysis of noninteracting and self-associating solutes using numerical solutions to the Lamm equation. Biophys J 75:1503–1512
6. Stafford WF, Sherwood PJ (2004) Analysis of heterologous interacting systems by sedimentation velocity: curve fitting algorithms for estimation of sedimentation coefficients, equilibrium and kinetic constants. Biophys Chem 108(1–3):231–243
7. Brown, PH, Schuck P (2006) Macromolecular size-and-shape distributions by sedimentation velocity analytical ultracentrifugation. Biophys J 90(12):4651–4661
8. Lamm O (1929) Die Differentialgleichung der Ultrazentrifugierung. Ark Mat Astron Fys 21B:1–4
9. Schuck P, Demeler B (1999) Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. Biophys J 76(4):2288–2296
10. Stafford WF (2000) Analysis of reversibly interacting macromolecular systems by time derivative sedimentation velocity. Methods Enzymol 323:302–325
11. Aster RC, Borchers B, Thurber CH (2005) Parameter estimation and inverse problems. Elsevier, New York, pp 35–36
12. Demeler B (2005) UltraScan website and software download. http://www.ultrascan.uthscsa.edu
13. Demeler B (2005) UltraScan: a comprehensive data analysis software package for analytical ultracentrifugation experiments. Modern analytical ultracentrifugation: techniques and methods. In: Scott DJ, Harding SE, Rowe AJ (eds) Royal Society of Chemistry (UK), pp 210–229
14. Box GEP, Muller ME (1958) A note on the generation of random normal deviates. Ann Math Stat 29:610–611
15. Demeler B, van Holde KE (2004) Sedimentation velocity analysis of highly heterogeneous systems. Anal Biochem 335(2):279–288
16. Brookes E, Demeler B (2007) Parsimonious regularization using genetic algorithms applied to the analysis of analytical ultracentrifugation experiments. Genetic and Evolutionary Computation Conference, London (in press)
17. Zhang Y, Zhang Z, Demeler B, Radhakrishnan I (2006) Coupled unfolding and dimerization by the PAH2 domain of the mammalian Sin3A corepressor. J Mol Biol 360(1):7–14
18. Belogrudov GI, Schirf V, Demeler B (2006) Reversible self-association of recombinant bovine factor B. Biochim Biophys Acta 1764(11):1741–1749
19. Comoletti D, Flynn RE, Boucard AA, Jennings LL, Demeler B, Schirf V, Newlin HR, Shi J, Südhof TC, Taylor P (2006) Gene selection, alternative splicing, and post-translational processing regulate neuroligin selectivity for beta-neurexins. Biochemistry 45 (42):12816–12827
20. Khakshoor O, Demeler B, Nowick JS (2007) Macrocyclic β-sheet peptides that mimic protein quaternary structure through intermolecular β-sheet interactions. J Am Chem Soc 129(17):5558–5569
21. Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs, NJ