Eur Biophys J (2010) 39:405–414
DOI 10.1007/s00249-009-0413-5

ORIGINAL PAPER

A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape

Emre Brookes · Weiming Cao · Borries Demeler

Received: 16 November 2008 / Revised: 22 January 2009 / Accepted: 29 January 2009 / Published online: 27 February 2009

Abstract We report a model-independent analysis approach for fitting sedimentation velocity data which permits simultaneous determination of shape and molecular weight distributions for mono- and polydisperse solutions of macromolecules. Our approach allows for heterogeneity in the frictional domain, providing a more faithful description of the experimental data for cases where frictional ratios are not identical for all components. Because of increased accuracy in the frictional properties of each component, our method also provides more reliable molecular weight distributions in the general case. The method is based on a fine grained two-dimensional grid search over s and f/f_0, where the grid is a linear combination of whole boundary models represented by finite element solutions of the Lamm equation with sedimentation and diffusion parameters corresponding to the grid points. A Monte Carlo approach is used to characterize confidence limits for the determined solutes. Computational algorithms addressing the very large memory needs for a fine grained search are discussed. The method is suitable for globally fitting multi-speed experiments, and constraints based on prior knowledge about the experimental system can be imposed. Time- and radially invariant noise can be eliminated. Serial and parallel implementations of the method are presented. We demonstrate with simulated and experimental data of known composition that our method provides superior accuracy and lower variance fits to experimental data compared to other methods in use today, and show that it can be used to identify modes of aggregation and slow polymerization.

Keywords Analytical ultracentrifugation · Sedimentation velocity · Molecular weight determination · Shape determination · Whole boundary fitting · ASTFEM method · NNLS method

AUC&HYDRO 2008—Contributions from 17th International Symposium on Analytical Ultracentrifugation and Hydrodynamics, Newcastle, UK, 11–12 September 2008.

E. Brookes · B. Demeler (✉)
Department of Biochemistry, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, MC 7760, San Antonio, TX 78229-3901, USA
e-mail: demeler@biochem.uthscsa.edu

E. Brookes
e-mail: emre@biochem.uthscsa.edu

W. Cao
Department of Mathematics,
The University of Texas at San Antonio,
One UTSA Circle, San Antonio, TX 78249, USA
e-mail: wcao@math.utsa.edu

Introduction

Sedimentation velocity experiments performed in an analytical ultracentrifuge provide results that can characterize hydrodynamic properties of biological macromolecules, such as sedimentation-, diffusion- and frictional parameters, as well as molecular weight. During the velocity experiment, solutes experience two transport processes, sedimentation in a centrifugal force field, and diffusional transport due to the development of concentration gradients. These processes can be measured by monitoring the concentration profile in the ultracentrifuge cell over time. Both transport processes are inversely proportional to the frictional properties of the sedimenting solute, and the sedimentation process is also directly proportional to the molecular weight of the particle. By modeling the entire concentration boundary in a sedimentation experiment it is

possible to simultaneously measure the sedimentation and diffusion processes for each solute. The methods commonly employed for sedimentation velocity analysis differ in terms of information content, resolution, their ability to provide diffusion coefficients and a direct measure of molecular weight, their applicability to heterogeneous systems, and their dependence on preconceived models entered by the user. As has been shown previously, an acceptable approximation for most systems is the model for a mixture of individual, non-interacting solutes described by the Lamm equation (Schuck 2003, Dam et al. 2005). For such a mixture of noninteracting solutes, the total concentration C_T of all solutes n in the ultracentrifuge cell can be represented by a sum of Lamm equation solutions L:

$$C_{\mathrm{T}} = \sum_{i=1}^{n} c_i L(s_i, D_i) \tag{1}$$

where c_i is the partial concentration, s_i is the sedimentation coefficient, and D_i is the diffusion coefficient of each solute i in the mixture, and L represents a solution of the Lamm equation (Lamm 1929) Eq. (2), which describes the sedimentation and diffusion transport of a single ideal solute in an analytical ultracentrifugation cell:

$$\frac{\partial C}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(s\omega^2 rC - Dr\frac{\partial C}{\partial r}\right), \quad m < r < b,\ t > 0 \tag{2}$$

where C is the concentration function of radius r and time t, s and D are the sedimentation and diffusion coefficients, and ω is the angular velocity. m and b are the radii at the meniscus and bottom of the cell. When fitting experimental velocity data the challenge then consists of finding the correct values for n, c_i, s_i and D_i. Because this fitting function is nonlinear with respect to fitting parameters c_i, s_i and D_i, an optimization approach capable of dealing with this nonlinearity needs to be employed. Several methods have been proposed to accomplish this: Iterative fitting methods using nonlinear least squares optimization were first proposed by Todd and Haschemeyer (1981), and later implemented by Demeler and Saber (1998), and by Schuck (1998). However, there are significant drawbacks to this approach: First, the correct model needs to be selected and verified by the user, which introduces considerable bias in the analysis. Secondly, although the method works well for simple systems of one or two well separated components, the nonlinear least squares fitting process tends to break down for more complicated systems that contain three or more components. The reason for this failure is based on the complexity of the error surface. Simple gradient descent methods fail to navigate the complex, multidimensional error surface and tend to become trapped in local minima, never converging to the global optimum and showing significant systematic deviations in the residuals. Another possibility is the presence of multiple minima with nearly identical residuals, or the inadequacy of the selected model which fails to consider additional signals present in the data. To address this convergence difficulty, Schuck proposed the C(s) method (Schuck, 2000), which implements a linearization of the problem and hence avoids the multidimensional search by iterative methods. Later an extension of this method was proposed by Brown and Schuck (2006) which added a regularized search over a coarse grid of both s and f/f_0. We reproduce here briefly the linearization idea behind these approaches. First, the sedimentation coefficient range presumed to be represented by the solutes in the experiment is divided into n, generally equidistant partitions, where n typically equals 50–100. Each partition represents one term in the sum shown in Eq. (1). The diffusion coefficient is treated as a constant and is parameterized with the sedimentation coefficient s and a given frictional ratio $k = f/f_0$ as shown in Eq. (3).

$$D = RT\left[N18\pi(k\eta)^{3/2}\left(\frac{s\bar{v}}{2(1-\bar{v}\rho)}\right)^{1/2}\right]^{-1} \tag{3}$$

where R is the universal gas constant, T the temperature, N is Avogadro's number, η and ρ are the viscosity and density of the solvent, and $\bar{v}$ is the partial specific volume of the solute. The value of k is maintained constant throughout Eq. (1), which reduces the nonlinear fitting problem to a linear problem where only the coefficients c_i need to be determined. For this task, a non-negatively constrained linear least squares analysis is applied (Lawson and Hanson 1974). This assures that the coefficients contain only positive values, or zero. For the C(s) analysis, a single-dimensional nonlinear search over k is generally added to this procedure in order to identify an approximate weight-average k for all solutes present in the mixture. The following concerns arise with this approach: While for a subset of experiments the weight-average approximation of the constant k may be sufficient, generality is sacrificed by treating k as a constant parameter, unless only a single component is present, or all species are spherical and the frictional ratio is equal to unity. Furthermore, if an average frictional ratio is used to transform the s-value distribution into a molecular weight distribution, it is generally true that the molecular weight of the most globular component will be overestimated, and the molecular weight of the most nonglobular component will be underestimated. As a consequence any one species found in the distribution may be assigned an inaccurate molecular weight. Frequently, heterogeneous mixtures may present heterogeneity not only in s, but also in k. Examples for such cases include molecules aggregating to long fibrils, where larger species gain considerable asymmetry. Other examples include mixtures of unfolded proteins, or mixtures of nucleic acids, or nucleic acid—binding protein systems. In such cases the

relatively broad boundaries for the most globular species are interpreted as heterogeneity by least squares fitting algorithms since multiple species with too small frictional ratios will fit better than a single species, causing a peak to split into multiple peaks. To address this issue, stochastic search algorithms have previously been explored, among them genetic algorithms by Brookes and Demeler (2006, 2007). Although the results provide convincing evidence that it is possible to resolve more than two components in a mixture with the same level of detail as direct boundary fitting methods afford, such stochastic methods require significantly greater computational effort, and implementation even on multi-core workstations is not very practical. The C(s, f/f_0) method can produce an improved description of the underlying parameters, however, it suffers from lack of resolution, large memory needs, and produces unnecessarily broad molecular weight distributions (Brown and Schuck 2006), and introduces false positives caused by noise in the data, and by failing to consider the entire parameter space in each minimization step. In this work we describe a two-dimensional spectrum analysis over parameters s and k which is suitable for the general case of noninteracting solutes, even when heterogeneity in both s and in k is present. The approach solves the minimization problem for the entire parameter space simultaneously at any desired resolution, and can be used on a single workstation in a serial implementation or in a parallel distributed computing environment for improved computational speed. The method also attenuates the signal of false positives by utilizing a Monte Carlo approach and simultaneously correcting for time- and radially invariant noise. The method provides a high-resolution description of both the shape and molecular weight domain by using a novel moving grid approach which allows the computation to proceed at any desired resolution without exceeding available memory. The coupled Monte Carlo method can then provide confidence limits for c_i, s_i, D_i, as well as the molecular weight of each solute present in the mixture.

Methods

Description of the method

Our approach for modeling experimental sedimentation data consists of building a two-dimensional grid of frictional ratios and sedimentation coefficients. For optimal results, the range of the s and f/f_0 domain should be initialized to match the range of possible values in the experimental system. For absorbance data, the range of s values can be conveniently initialized with the model-independent van Holde—Weischet method (Demeler and van Holde 2004). When significant time invariant noise exists, for example in intensity or interference data, the dC/dt approach by Stafford (1992) is preferred for initialization due to its superior time invariant noise handling capability. The frictional ratio provides a convenient way to parameterize the diffusion coefficient, which exhibits a well defined lower limit of 1.0 for a spherical molecule, and whose value range can be conveniently estimated (1–2 for globular proteins, 2–4 for non-globular molecules, >4 for very large, non-globular molecules such as linear DNA and fibrils). Using Eq. (3) we can now define a unique value for s and D at each grid point, and simulate the velocity experiment for a species with these parameters. For simulation of all Lamm equation models we use the adaptive space-time finite element solution proposed by Cao and Demeler (2005, 2008). We now build the sum:

$$C_T = \sum_{i}^{m} \sum_{j}^{n} c_{i,j} L[s_i, D(s_i, k_j)] \tag{4}$$

where s_i is the sedimentation coefficient at position i, k_j is the frictional ratio at position j, m is the number of grid points in the sedimentation domain, n is the number of grid points in the frictional ratio domain, and $c_{i,j}$ is the partial concentration of each simulated solute at grid point (i, j). In order to determine the values of $c_{i,j}$, we simulate each species i, j using unity concentration for h radial points r, and l time scans t. The minimization problem can then be stated as the task of finding the minimum for the l^2-norm:

$$\text{Min} = \sum_{r}^{h} \sum_{t}^{l} \left[E_{r,t} - C_{Tr,t}\right]^2 \tag{5}$$

where $E_{r,t}$ refers to the experimentally determined data points for h radial points r and l time scans t. This linear optimization problem can be expressed in matrix form:

$$Ax = b \tag{6}$$

where A is the matrix of finite element solutions, x the solution vector containing all coefficients $c_{i,j}$, and b is the vector of experimental data. In order to solve the minimization problem, we apply the NNLS algorithm (Lawson and Hanson 1974), which constrains the solution to values for $c_{i,j}$ which are either zero or positive, and hence avoids negative oscillations in the coefficients that would be observed in unconstrained general linear least squares minimization. Simultaneously, we algebraically account for time invariant and radially invariant noise contributions in the experimental data as described by Schuck and Demeler (1999).

Multi-stage refinement

A limitation of the approach described above is posed by the requirement for large amounts of computer memory demanded by the simultaneous solutions for $h \times l \times m \times n$

datapoints. The typical size for h is 500–800 points, for l it is 50–100, but these vectors could be as large as $h = 10^3$ and $l = 10^3$ when interference optics are used. Performing just a 10×10 grid search on such an array would require close to half a gigabyte of memory just for data storage of a single experiment. If multiple experiments are fitted globally, the need for memory increases approximately linearly. While this data size can result in prohibitive memory needs, the availability of more data is desirable for improving the signal to noise ratio, and ultimately the confidence limits of the results. Furthermore, for cases where broad distributions of s and f/f_0 are expected, a 10×10 grid as proposed by Brown and Schuck (2006) is insufficient to reliably describe the experimental parameter space. If the actual solute is not aligned with a grid point, false positives are produced (see "Results and discussion" below).

In order to address this problem, we introduce here a divide-and-conquer strategy for refining the original $m \times n$ grid into a grid of any desired resolution. Our approach is suitable for describing any size system even on computers with limited memory, but can also be implemented in a parallel high performance computing environment. The method which we term the multi-stage two-dimensional spectrum analysis (MS2DSA, or 2DSA for short) is based on a repeated evaluation of sufficient numbers of sub-grids regularly spaced over the entire grid such that the entire two-dimensional s and k space is covered by the simulation process. The algorithm proceeds as follows: The initial grid is partitioned into m regular intervals between $s_{\min}$ and $s_{\max}$ in the first dimension and n regularly spaced intervals between $k_{\min}$ and $k_{\max}$ in the second dimension (Fig. 1a). Finite element solutions are calculated for each grid point and the linear sum shown in Eq. (4) is formed. The least squares solution is computed with NNLS as shown in Eq. (5), and the solution vector containing all non-zero elements $c_{i,j}$ is saved in a storage vector $\boldsymbol{S}_1$ (indicating stage 1 of the multi-stage process) along with the corresponding grid positions from the original grid (Fig. 1c). For the first order refinement, this process is repeated three times by moving the entire grid to three different origins as follows: First, the grid is shifted in the first dimension by a small increment δs_a given by:

$$\delta s_a = \frac{s_{\max} - s_{\min}}{2am} \tag{7}$$

where a is the refinement's iteration number and m is the number of grid points over s. After performing NNLS, the non-zero elements $c_{i,j}$ and their grid positions are added to $\boldsymbol{S}$, and the process is repeated by shifting the original grid into the second dimension by a small increment δk_a given by:

$$\delta k_a = \frac{k_{\max} - k_{\min}}{2an} \tag{8}$$

where a is the iteration number and n is the number of grid points in the k domain. Again, NNLS is performed and nonzero elements are added to $\boldsymbol{S}$. In the fourth grid movement, we complete the square and shift the grid origin by $+\delta s$ and $+\delta k$ simultaneously. A schematic view of the grid generation by this algorithm is shown in Fig. 1. In order to achieve further refinement this process is repeated on the next smaller grid division until the desired resolution is obtained by further decreasing δs and δk according to Eqs. (7) and (8). Here we mean by iteration one full cycle of the four transformations of the grid origin explained above. At each grid position we populate the storage grid $\boldsymbol{S}_1$ by adding the non-zero elements of each NNLS calculation to $\boldsymbol{S}_1$. When the number of non-zero parameters in $\boldsymbol{S}_1$ matches the size of each individual subgrid, we perform a NNLS optimization on all parameters contained in $\boldsymbol{S}_1$. The output is stored in $\boldsymbol{S}_2$, forming the second stage of the multi-stage process. In each successive stage, we collect only the non-zero entries of the previous NNLS optimization. When the desired resolution is obtained, the final storage grid is once more processed by NNLS and the resulting elements of $\boldsymbol{S}_f$ are now representative of the solutes and their relative concentrations present in the sedimentation velocity experiment. In this process, it is important that the entire parameter space is covered by each grid. Clearly, each grid covers a slightly different parameter space, but the overall coverage remains at most within $2\delta s$ and $2\delta k$. To guarantee that the required parameter space is actually covered by each grid, we increase the original search space determined with the van Holde—Weischet analysis and the estimate for the minimum and maximum f/f_0 at both ends of each axis by δs and δk, respectively. This adds only an insignificant amount of extra space to be searched by the algorithm. Parallelization is achieved by distributing each subgrid simulation and NNLS fit to a different processor, collecting only the results for the storage grid. Communication between processors as implemented in UltraScan (Demeler 2005) is accomplished with the Message Passing Interface (Brookes et al. 2006, http://www.open-mpi.org/).

Simulation of grid elements

We use the ASTFEM solution proposed by Cao and Demeler (2008) to simulate Lamm equation solutions for each grid point. In order to reduce computational effort it is possible to take advantage of the invariance shown in Eq. (9), where a is a multiplier that covers the entire desired range of s and D values. The same solution can be used for different s and D values as long as the solution is calculated for the entire time range.

$$C(as, aD)_{r,t} = C(s, D)_{r,at}. \tag{9}$$

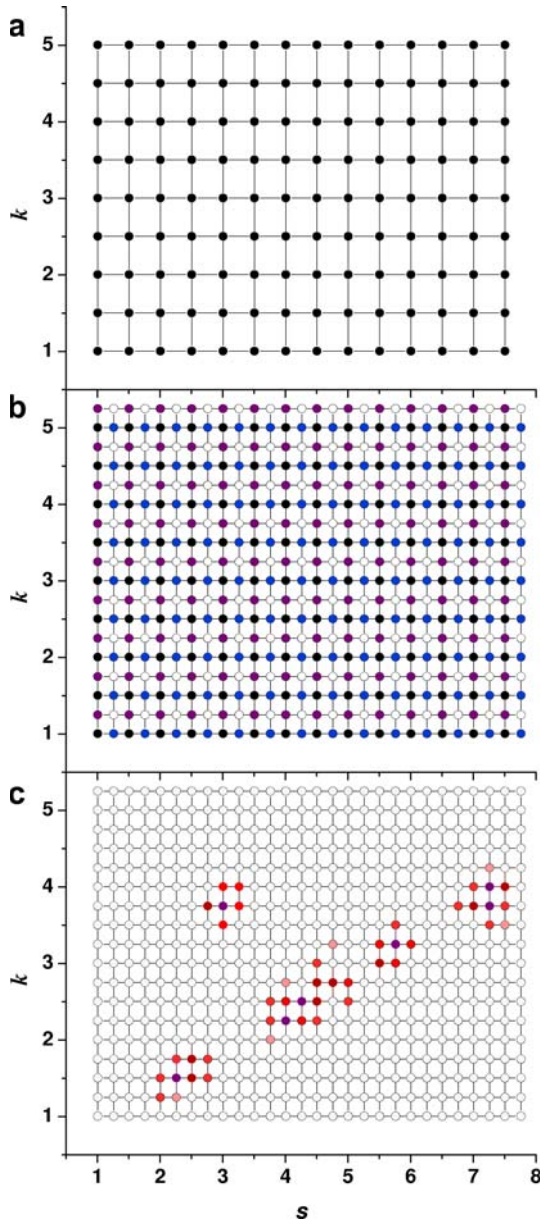

Fig. 1 **a** Initial grid spanning entire *s* and *k* parameter space with a sparse representation of each parameter dimension. **b** Grid evaluation points after one iteration of grid movements. *Black* initial grid. *Purple* grid displacement by δk. *Blue* grid displacement by δs. *White* grid displacement by δs and δk. **c** Typical storage grid $\boldsymbol{S}$ for a heterogeneous sample after one iteration of grid displacements; darkness of points indicates concentration level; white indicates zero concentration, pink indicates a small concentration, while dark purple indicates high concentration. Solutes get returned with discrete values of *s* and *k*

Iterative refinement

We have empirically shown that solving the iterative problem involving multiple low resolution sub-grids is equivalent to solving the high-resolution grid covering the same combined grid points if the following additional operation is performed: The non-zero grid points evaluated at the final state $\boldsymbol{S}_f$ are joined with each original grid in $\boldsymbol{S}_0$ and reprocessed. The analysis is then repeated until convergence is obtained (Brookes et al. 2006). This analysis produces a sparse parameter distribution with discrete solutes identified from the experimental data. Adding the sparse set of solutes obtained in $\boldsymbol{S}_f$ only marginally increases the size of grids in $\boldsymbol{S}_0$, and by judiciously choosing the original grid size any problem can be readily solved on a moderately equipped PC. It should be pointed out that the iterative refinement described here will not converge to *exactly* the same solution when time- or radially invariant noise corrections are performed simultaneously. However, differences are negligible and are much smaller than the noise level in a typical ultracentrifugation experiment.

Results and discussion

2DSA—Monte Carlo analysis of a 2-component system with heterogeneity in mass and shape

Due to the large number of fitting parameters, the solution obtained with the 2DSA method is overdetermined and uniqueness is not guaranteed. The higher the resolution, the larger the number of fitting parameters and a higher potential for degeneracy. To study the effect of a large number of fitting parameters on the solution, we have systematically evaluated the robustness of the solution as a function of resolution and number of fitting parameters. In this test, all fitting solutes represented by the fitting parameters are distributed over a regular grid with identical limits in both dimensions. Our test system consists of a globular protein (henn egg lysozyme) and an elongated molecule (a 208 bp linear fragment of double-stranded DNA), mixed in approximately equally absorbing amounts. This system was chosen because it illustrates the ability of the 2DSA to resolve a system that is heterogeneous in molecular weight and also heterogeneous in shape, and because the individual components are well studied and have known hydrodynamic properties and molecular weights. The mixture was run at 42,000 rpm in 200 mM NaCl and 25 mM TRIS buffer at pH 8.0 in standard 2 channel centerpieces. Velocity data were collected for 3 h and at 260 nm. Time invariant noise was subtracted as described in Schuck and Demeler (1999) and only stochastic noise remained in the data. The resulting data were fitted with the 2DSA method using 50 Monte Carlo iterations (Demeler and Brookes 2008), using the iterative refinement method with a maximum of 5 iterations. The limits of the frictional range was set from 1 to 4, and the limits of the sedimentation coefficient range was set from 1

to 10 s. The grid was built with the following 5 resolutions (s values x frictional ratio values x grid movings): 1. 100 (10 × 10 × 1); 2. 400 (10 × 10 × 4); 3. 10,000 (10 × 10 × 100); 4. 40,000 (10 × 10 × 400); 5. 90,000 (10 × 10 × 900). From the results, we plotted the RMSD of each fit, the mean and 95% confidence intervals for *s* and *k*, and the molecular weight and partial concentration for each species against the grid resolution. The results are shown in Fig. 2. From this analysis, we made the following observations:

1. The 2DSA is very robust and additional degeneracy introduced by increasing the resolution of the grid does not degrade the reliability of the solution. In fact, the opposite occurs, a higher resolution better defines the mean and reduces the 95% confidence intervals, and the results are more consistent with known values for these species. While the number of solutes increases with increasing number of fitting parameters, the relative positions of these solutes stay entirely confined to a narrow grid region in the parameter space, proving an extreme robustness against degeneracy of our approach. These results show that consideration of additional parameters has no effect on the detection of the actual signal present in the data.
2. A 10 × 10 grid suggested by Brown and Schuck (2006) is clearly insufficient to resolve even a moderate s-value range from 1 to 10 s and a *k* range from 1 to 4. Mean and 95% confidence intervals suggest a very poor description of the data at this resolution and clearly produce the wrong molecular weights for both species.
3. The 2DSA method shows very high precision and accuracy, reproducing faithfully the known molecular weights when adjusted for the appropriate partial specific volumes (0.724 ccm/g for lysozyme and 0.55 ccm/g for DNA).
4. The 95% confidence intervals obtained from the Monte Carlo approach clearly show a narrower range for DNA than for lysozyme. This effect can be explained by considering the basic signals contributing to this data: sedimentation and diffusional transport. The sedimentation signal is more pronounced for the larger component (DNA), and the diffusion signal will be markedly smaller when compared to the smaller, more globular lysozyme, producing a better resolution for the DNA than for the lysozyme. The shape or frictional ratio information is heavily influenced by the diffusion coefficient, which is derived from the shape of the boundary, or the boundary spread. Heterogeneity (or poor sedimentation resolution) has a similar spreading effect on the boundary, and spreading due to micro-heterogeneity can be misinterpreted as a diffusion coefficient that is too large. Therefore, when composition is poorly defined because of slow speed or slow sedimentation and large diffusion, the low confidence in the sedimentation coefficient translates into a uncertainty about diffusion and shape, which explains this difference in the 95% confidence intervals of DNA

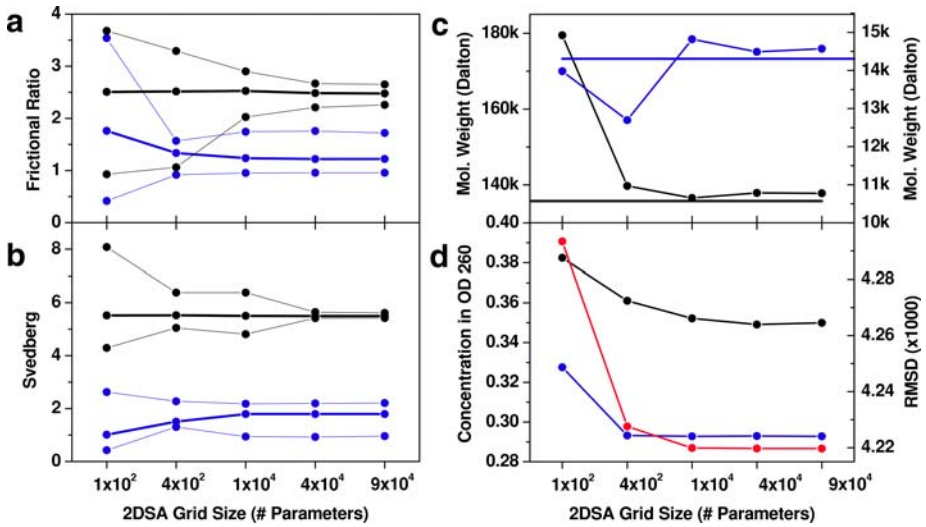

Fig. 2 2DSA Monte Carlo analysis of velocity data from a mixture of a 208 bp DNA fragment (*black lines*) and hen egg lysozyme (*blue lines*). *Heavy lines* indicate the mean, thin lines represent 95% confidence intervals for the parameter. The results for several parameters from multiple grid resolutions are compared. **a** Frictional ratio; **b** sedimentation coefficient (corrected to standard conditions); **c** molecular weight, horizontal lines indicate theoretical molecular weight based on sequence; **d** partial concentration and the residual mean square deviation of the fit (*red line*). Reliable results are obtained after a minimum of 10,000 iterations, higher resolutions do not improve the results significantly

and lysozyme. On the other hand, if the diffusion signal is low because of high rotor speed and short run times, and not much diffusional transport occurs, the uncertainty in shape arises from lack of time to let the sample diffuse before being pelleted. As is shown in "Global fitting of multi-speed data" this problem can be mitigated by globally fitting multiple speeds of the same sample.

5. In order to measure the effect iterative refinement has on the quality of the observed results, we also performed the same analysis without using the iterative refinement approach (data not shown). This approach showed identical trends as we observed in the optimization including iterative refinement, however, the results were less regular than those obtained when iterative refinement was employed. It can therefore be concluded that an additional benefit is derived from iterative refinement, especially when only a moderate grid resolution is used.
6. As additional parameters are added, an increased tendency to fit small frequency noise contributions is apparent, with a concentration of such points along the maximum frictional ratio boundary. Since the amplitude of these signals always remains within the noise level of the experimental data, and because their position is fixed at the upper frictional ratio limit, such solutes are easily identified and excluded. In addition, increasing the frictional ratio upper limit moves such noise contributions along with the upper frictional ratio boundary. We have introduced a Monte Carlo approach that effectively attenuates the relative signal from such noise contributions by amplifying intrinsic solute signal linearly, but amplification of stochastic noise only occurs with a factor of square root of two, which reduces the contribution of artifacts due to stochastic noise (Demeler and Brookes 2008). Pseudo-3D plots showing the difference between the lowest and highest grid resolution are shown in Fig. 3. The Monte Carlo results for lysozyme and DNA are shown in Table 1.

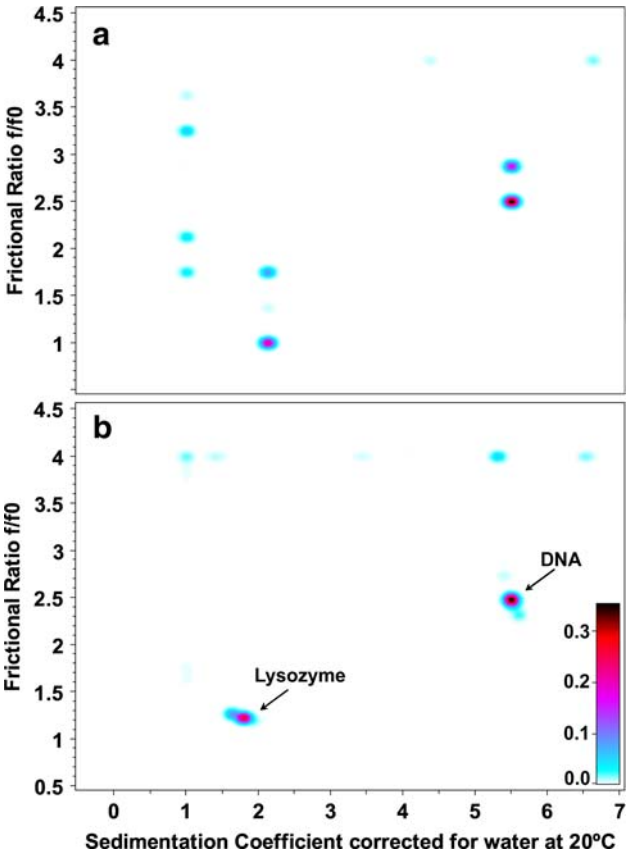

Fig. 3 Pseudo-3D plots for solute distributions for the 2DSA Monte Carlo results shown in Fig. 2 for the highest and lowest grid resolution examined. **a** Grid resolution of 100 solutes; **b** grid resolution of 90,000 solutes. At the low resolution the composition is poorly defined and solute peaks are split, at high-resolution both species are well defined in narrow regions without any significant peak splitting, noise contributions are well separated and identifiable at the upper frictional ratio fitting limit ($k = 4$). Globular shape of lysozyme and elongated shape of DNA is clearly reproduced by fitting result. The *color scale* represents the signal of each species in optical density units

Global fitting of multi-speed data

In an effort to better quantify the level of detail that can be obtained from a sedimentation velocity experiment when analyzed with the 2DSA method, we looked at ways to improve experimental signal. It is well known that improved information can be obtained from sedimentation equilibrium experiments when multiple speeds and multiple concentrations of the same data are measured and globally analyzed (Johnson et al. 1981). In this analysis approach, certain parameters such as molecular weight, and equilibrium constants can be treated as global parameters because they are invariant and governed by conservation of mass considerations. The similar approach can be used for velocity experiments. We have implemented a global 2DSA fitting method for non-interacting systems to globally fit experiments of samples with invariant composition. This approach imposes constraints on fits from all included data sets that require that all non-zero solutes obtained in the fit are present in the same relative ratio in all data sets. Different signals originating from dilutions or different optical systems or different centerpiece geometries are accounted for by scaling the amplitudes of all solutes with a different scalar multiplier for each datasets. The experiments can be performed at different speeds, or by different acquisition methods. Even data from different cell geometries can be fitted globally, such as experiments performed in band-forming Vinograd cells or standard 2-channel

Table 1 Statistics for the 2DSA Monte Carlo analysis of lysozyme and a 208 basepair DNA fragment

	Lysozyme	208 basepair DNA
Molecular weight (Dalton)	14,325 [14,306] (7,903, 18,790)	137,800 [135,725](120,860, 154,980)
Sedimentation coefficient (s, $s_{20,W}$)	1.783×10^{-13} (9.492×10^{-14}, 2.231×10^{-13})	5.498×10^{-13}(5.422×10^{-13}, 5.615×10^{-13})
Diffusion coefficient (cm^2/s, $D_{20,W}$)	1.085×10^{-6} (8.650×10^{-7}, 1.221×10^{-6})	2.156×10^{-7}(1.958×10^{-7}, 2.425×10^{-7})
Frictional ratio	1.22 (0.955, 1.72)	2.48 (2.26, 2.65)
Partial concentration	0.293 OD	0.350 OD

Values in curved parenthesis are 95% confidence intervals, values in square brackets are known molecular weights. Source: Lysozyme by mass spectrometry measurement: http://www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm, DNA molecular weight calculated with UltraScan (Demeler 2005) from sequence composition assuming a 0.75 ratio of Na^+/basepair bound (Manning 1969). *OD* optical density at 260 nm

centerpieces. We compared the information obtained from fitting data from a simulated system with known composition under four conditions: 10 krpm conventional centerpiece, 60 krpm conventional centerpiece, 10, 30, and 60 krpm conventional centerpiece, fitted globally, and 10, 30, and 60 krpm globally for both conventional and band-forming Vinograd experiments together. Our test system consists of equal concentrations of a linearly elongating aggregate with five noninteracting components: monomer (25,000 Dalton, frictional ratio: 1.2), dimer (50,000 Dalton, frictional ratio: 1.4), tetramer (100,000 Dalton, frictional ratio: 1.6), octamer (200,000 Dalton, frictional ratio: 1.8), hexadecamer (400,000 Dalton, frictional ratio: 2.0). Stochastic noise of 1% typical in a UV-absorbance XLA was added to all simulated data before fitting. All experiments were simulated to contain 70 equally spaced scans over a time period that was selected such that the total force exerted on the sample over the entire experiment was identical regardless of speed, and assured that all samples either pelleted or approached equilibrium. This led to 128 h and 12 min for 10 krpm, 14 h and 12 min for 30 krpm, and 3 h and 30 min for 60 krpm. In all cases a column of 14 mm was simulated extending from a meniscus of 5.8 to a cell bottom of 7.2 cm. The results show that the 2DSA—Monte Carlo method could in each case correctly map out the parameter space (Fig. 4). The difference between the analysis conditions was found in the resolution with which the individual components could be resolved. Specifically, we made the following observations from the data shown in the pseudo-3D plots:

1. The single speed analysis of the 10 krpm data using conventional centerpieces shows a poorly resolved band of signal, covering the correct range. Maxima can be detected near the expected positions in the 2D grid. Resolution in the horizontal dimension (molecular weight or sedimentation coefficient is worst from all conditions, but the frictional range is better defined than the high speed experiment (Fig. 4a).
2. The single speed analysis of the 60 krpm data using conventional centerpieces shows a more precise definition of the horizontal domain than the single speed 10 krpm run, but the frictional range is more poorly defined than in the low speed data, especially for the higher molecular weight species. This is presumably caused by lack of diffusion signal for the higher molecular weight species, which sediment quickly at this speed. Also, some peak splitting is observed for the higher molecular weight species (Fig. 4b).
3. A global, multi-speed analysis of data from 10, 30 and 60 krpm data using conventional centerpieces offers a slight improvement of the single speed experiments by eliminating the peak splitting of the medium sized species (100,000 Dalton). However, the high molecular weight species (400,000 Dalton) peak is still poorly defined in the shape domain, and the peak is still split (Fig. 4c).
4. A further improvement can be obtained by combining the data from the conventional centerpieces with band-sedimentation data performed at the same three speeds and globally fitting all six experiments (Fig. 4d). In this fit, all peak splitting has been resolved and all determined signals fit exceptionally well to the starting parameters, producing an optimal description of the data.

Summary

We have presented a novel algorithm for efficiently fitting sedimentation velocity data to high-resolution grids based on finite element solutions of the Lamm equation. This algorithm is suitable for serial calculation on a single processor or can be used in a parallel environment on a multi-processor machine or supercomputer. We have shown that low resolution grids as proposed by Brown and Schuck (2006) are insufficient to obtain reliable information from a two-dimensional approach. Another result of our study shows that globally fitting data from different speeds and different centerpiece geometries can further

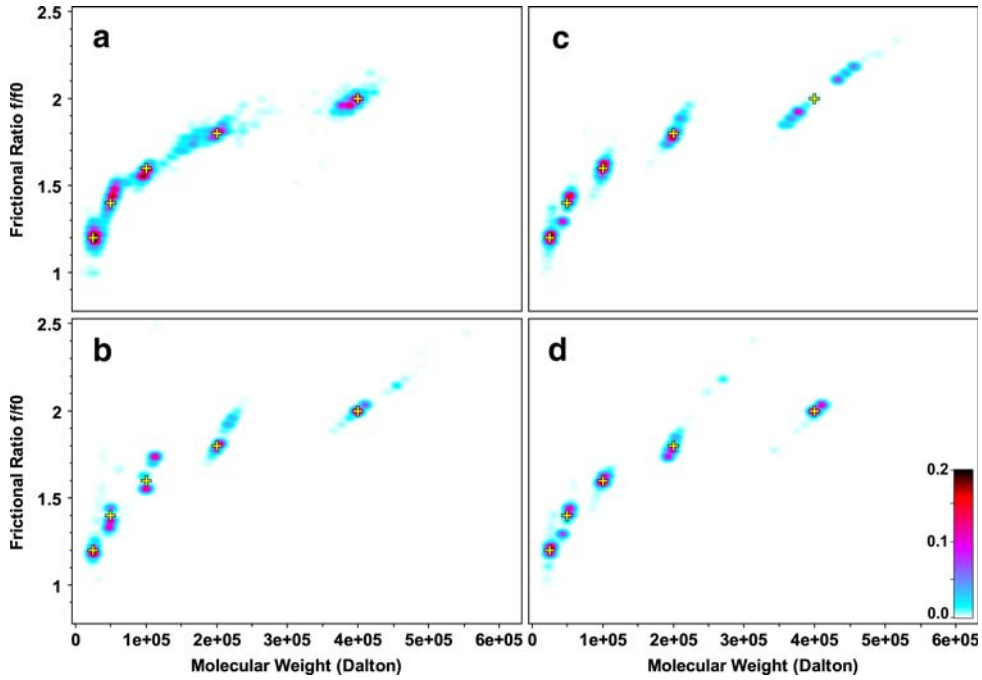

Fig. 4 Pseudo-3D plots for Monte Carlo 2DSA analysis results for a simulated five component system described in Sect. 3.2. **a** Single speed fit of data from conventional centerpiece (10 krpm); **b** single speed fit of data from conventional centerpiece (60 krpm); **c** global multi-speed fit of data from conventional centerpiece (10, 30 and 60 krpm); **d** global multi-speed fit of data from both conventional centerpiece combined with data from band-forming Vinograd centerpiece (10, 30 and 60 krpm for both centerpiece types). Improvement of the results is apparent in reduced peak splitting and improved confidence intervals in going from a → d. *Yellow crosses* indicate the positions of the known solutes that were simulated for the original data. The *color scale* represents the signal of each species in optical density units

incrementally enhance the resolution obtained with the 2DSA method. The global method shows also that further improvement of the results is most likely a function of signal quality, and can only be achieved by improving the detectors.

For non-interacting species, the 2DSA approach is general and model-independent, and does not depend on prior knowledge of the underlying model, for mixtures of rapidly equilibrating solutes the 2DSA approach can still provide approximations for solute distributions, although interactions coefficients such as equilibrium and rate constants can not be obtained by this approach. The 2DSA method can simultaneously resolve heterogeneity in shape and in molecular weight or sedimentation coefficients at very high-resolution, producing very well defined and narrow solute boundaries. The only user input required is a knowledge of the fitting limits, which can be determined with the van Holde–Weischet method (Demeler and van Holde 2004) or the dC/dt method (Stafford 1992). Because this method does not make any assumptions of constant frictional ratios for all species as the C(s) method does in SedFit (Schuck et al. 1998), the 2DSA is more rigorous and better able to also reliably resolve molecular weights, as long as the partial specific volume is known.

Acknowledgments This work and the development of UltraScan is supported by NIH Grant RR022200 (NCRR) to B. D.

References

Brookes EH, 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, Springer Progr Colloid Polym Sci 131:78–82

Brookes EH, Demeler B (2007) Parsimonious regularization using genetic algorithms applied to the analysis of analytical ultracentrifugation experiments. GECCO proceedings ACM 978-1-59593-697-4/07/0007

Brookes EH, Boppana RV, Demeler B (2006) Computing large sparse multivariate optimization problems with an application in biophysics. Supercomputing '06 ACM 0-7695-2700-0/06

Brown PH, Schuck P (2006) Macromolecular size-and-shape distributions by sedimentation velocity analytical ultracentrifugation. Biophys J 90(12):4651–4661. doi:10.1529/biophysj.106.081372

Cao W, Demeler B (2005) Modeling analytical ultracentrifugation experiments with an adaptive space-time finite element solution of the Lamm equation. Biophys J 89(3):1589–1602. doi:10.1529/biophysj.105.061135

Cao W, Demeler B (2008) Modeling analytical ultracentrifugation experiments with an adaptive space-time finite element solution for multi-component reacting systems. Biophys J 95(1):54–65. doi:10.1529/biophysj.107.123950

Dam J, Velikovsky CA, Mariuzza RA, Urbanke C, Schuck P (2005) Sedimentation velocity analysis of heterogeneous protein–protein interactions: Lamm equation modeling and sedimentation coefficient distributions c(s). Biophys J 89(1):619–634. doi:10.1529/biophysj.105.059568

Demeler B (2005) UltraScan—a comprehensive data analysis software package for analytical ultracentrifugation experiments. In: Scott DJ, Harding SE, Rowe AJ (eds) Modern analytical ultracentrifugation: techniques and methods. Royal Society of Chemistry, UK, pp 210–229

Demeler B (2008) UltraScan version 9.9—a multi-platform analytical ultracentrifugation data analysis software package: http://www.ultrascan.uthscsa.edu

Demeler B, Brookes E (2008) Monte Carlo analysis of sedimentation experiments. Colloid Polym Sci 286(2):129–137. doi:10.1007/s00396-007-1699-4

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. doi:10.1016/S0006-3495(98)77802-6

Demeler B, van Holde KE (2004) Sedimentation velocity analysis of highly heterogeneous systems. Anal Biochem 335(2):279–288. doi:10.1016/j.ab.2004.08.039

Johnson ML, Correia JJ, Yphantis DA, Halvorson HR (1981) Analysis of data from the analytical ultracentrifuge by nonlinear least squares techniques. Biophys J 36:575–588. doi:10.1016/S0006-3495(81)84753-4

Lamm O (1929) Die Differentialgleichung der Ultrazentrifugierung. Ark Mater Astr Fys 21B:1–4

Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs

Manning GS (1969) Limiting laws and counterion condensation in polyelectrolyte solutions: I. Colligative properties. J Chem Phys 51:933–942

Schuck P (1998) Sedimentation analysis of noninteracting and self-associating solutes using numerical solutions to the Lamm equation. Biophys J 75(3):1503–1512

Schuck P (2000) Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. Biophys J 78(3):1606–1619. doi:10.1016/S0006-3495(00)76713-0

Schuck P (2003) On the analysis of protein self-association by sedimentation velocity analytical ultracentrifugation. Anal Biochem 320(1):104–124. doi:10.1016/S0003-2697(03)00289-6

Schuck P, Demeler B (1999) Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. Biophys J 76(4):2288–2296. doi:10.1016/S0006-3495(99)77384-4

Schuck P, MacPhee CE, Howlett GJ (1998) Determination of sedimentation coefficients for small peptides. Biophys J 74(1):466–474. doi:10.1016/S0006-3495(98)77804-X

Stafford W (1992) Boundary analysis in sedimentation transport experiments: a procedure for obtaining sedimentation coefficient distributions using the time derivative of the concentration profile. Anal Biochem 203:295–301. doi:10.1016/0003-2697(92)90316-Y

Todd GP, Haschemeyer RH (1981) General solution to the inverse problem of the differential equation of the ultracentrifuge. Proc Natl Acad Sci USA 78(11):6739–6743