Methods in Enzymology, in press

Calculating sedimentation coefficient distributions by direct modeling of sedimentation velocity concentration profiles

Julie Dam[&] and Peter Schuck[*]

[*]Protein Biophysics Resource, Division of Bioengineering & Physical Science, ORS, National Institutes of Health, Bethesda, Maryland, USA, [&]Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland, USA

keywords: analytical ultracentrifugation, size-distribution, hydrodynamics, sedfit maximum entropy, Fredholm integral equation, Lamm equation, protein interactions

***Address for Correspondence and Proofs:**

Dr. Peter Schuck
National Institutes of Health
Bldg. 13, Rm. 3N17
13 South Drive
Bethesda, MD 20892

Phone: (301) 4351950
Fax: (301) 4966608
Email: pschuck@helix.nih.gov

Abstract

The c(s) method results in high-resolution sedimentation coefficient distributions by deconvoluting the effects of diffusion broadening. It is a direct model for sedimentation data covering the complete sedimentation process, based on finite element solutions of the Lamm equation, and algebraically accounts for the systematic noise structure encountered in transport experiments. A continuous sedimentation coefficient distribution is expressed via a Fredholm integral equation and solved using maximum entropy regularization. The extent of diffusion for each species is approximated by hydrodynamic considerations and parameterized with a weight-average frictional ratio of the protein mixture under study. In addition to the sedimentation coefficient distribution, c(s) can be used to estimate protein molar mass, detect trace components of protein aggregates as well as other small and large molar mass impurities, and series of c(s) distributions obtained at different protein concentrations can be used for the study of homogeneous and heterogeneous protein interactions. The present paper provides an overview of the strategy and numerical approaches of the c(s) method, including special cases useful for certain protein studies, briefly discusses experimental requirements, and illustrates the main types of applications to non-interacting and interacting protein systems.

Introduction

Sedimentation velocity analytical ultracentrifugation can give rich information about the purity, molar mass, state of association, protein interactions, hydrodynamic shapes, conformational changes, size-distributions, among other properties of proteins (for recent reviews, see e.g. (1-3)). It is based on a conceptually very simple principle of applying a gravitational force to the protein solution and observing the resulting changes in the concentration distribution (4, 5). Sedimentation velocity has been continuously refined during the last eighty years, and historically has provided many important contributions to the development of both polymer chemistry and biochemistry, starting with the discovery of proteins being macromolecules and of well-defined sizes (6). Most of the theoretical foundations have long been laid, such as in 1929 by Lamm in Svedberg's laboratory the partial differential equation for macromolecular sedimentation and diffusion in the centrifugal field (7). However, the last several years have witnessed important developments, as it has become possible to solve the Lamm equation on laboratory computers and to use it now routinely to model experimental data from increasingly complex systems (8-16). This is in confluence with more precise and substantially larger experimental data basis, provided for example by the laser interferometry detection system of the commercial analytical ultracentrifuges (17).

One of the key problems in the interpretation of the macromolecular concentration distributions and their evolution is the separation of the effects of diffusion and sedimentation. This is particularly important in the study of proteins, where frequently the diffusional transport equals or exceeds the migration caused by the centrifugal force. This problem can be illustrated by considering the sedimentation profiles of a sample of a non-interacting protein species (Figure 1). Diffusion effects can be experimentally minimized by applying a high centrifugal field to increase the sedimentation rate so as to form a sharp sedimentation boundary, and an average sedimentation rate of the sample can be determined by simply measuring the velocity of this boundary (Figure 1A). However, it is

clear that this neglects an enormous amount of information contained in the shape of the evolving concentration distributions. It also does not permit the analysis of molar mass, and the characterization of different protein subpopulations. The simplest approach of taking diffusion into account is to interpret the spread of the boundary shape as if resulting from the diffusion of a single species (Figure 1B). In principle this permits the determination of the diffusion coefficient and the molar mass of the sedimenting species. However, in practice this approach frequently fails, because it requires the protein under study to be highly homogeneous. For heterogeneous mixtures where the root-mean-square displacement from diffusion is in the same order as the separation of species due to their different sedimentation rate (Figure 1C), such as mixtures of small or similar-sized proteins, a complex evolution of boundary shape and boundary location is observed in which the boundary spreading is due to the combined effects of diffusion and differential migration. In this situation, all known data transformations and extrapolation techniques frequently fail to provide either a good description of diffusion properties of the mixture or a resolution of the sedimentation coefficients of the different species (18).

Modern sedimentation velocity methods using solutions of the Lamm equation assume a specific model for macromolecules under study, calculate their sedimentation behavior and fit the model to the experimental data by least-squares techniques. The present paper focuses on the currently most general direct boundary model, which is that of a continuous distribution of non-interacting proteins. In the form of a sedimentation coefficient distribution c(s) (19), it addresses the problem outlined above and provides a robust estimation of the populations of species with different sedimentation rates, at a high resolution, and along with a description of the diffusion of the ensemble via a weight-average frictional ratio (18). Since its introduction in 2000, this approach has found application in many published studies (20).

The c(s) distribution is related to the well-known apparent sedimentation coefficient distribution

g*(s) of non-diffusing particles, which was previously introduced in the form g(s*) arrived at through a data transformation dc/dt (21, 22). In comparison with the g*(s) distribution, it will be illustrated that the consideration of diffusion in the c(s) distribution leads to a substantially increased resolution and provides a solution to the problem depicted in Figure 1, the deconvolution of multiple sedimenting and diffusing protein species.

The present paper is intended as an overview of this approach in theory and practice. First, we describe the numerical methods for the continuous sedimentation coefficient distribution analysis. They are applied in the software SEDFIT (www.analyticalultracentrifugation.com) for analysis of experimental analytical ultracentrifugation data. Second, several variations are described that can give molar mass distributions, or permit the use of different prior knowledge to constrain the model, such as for proteins with conformational changes or known molar mass, large particles, sedimenting co-solutes, compressible solvents and others. Next, a brief description of the experimental procedures providing the best data basis for the c(s) analysis is given followed by several examples of practical applications. Finally, its use for the study of interacting protein mixtures will be discussed.

Numerical Methods

Calculating sedimentation coefficient distributions from experimental data requires several numerical methods that address different aspects of sedimentation analysis. In the following, we will briefly review: 1) the numerical solution of the Lamm equation for a single species; 2) how the model of the sedimentation boundary model can be adapted to the special noise structure of ultracentrifugation data; 3) the description of a continuous distribution by a Fredholm integral equation and different regularization methods; 4) the approximation of the size-dependent diffusion coefficients.

Numerical Solutions of the Lamm Equation

One of the key components for the calculation of sedimentation coefficient distributions is the precise and efficient solution of the Lamm equation for a large range of s-values. The evolution of the spatial distribution of a single sedimenting species χ(r,t) in a sector-shaped solution column the centrifugal field can be described by the Lamm equation

$$\frac{\partial c}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left[rD\frac{\partial c}{\partial r} - sw^2 r^2 c \right] \qquad (1.1)$$

where *r* denotes the distance from the center of rotation, ω the rotor angular velocity, and s and D the macromolecular sedimentation and diffusion coefficient, respectively (7). s and D are related with the molar mass M by the Svedberg equation

$$D = \frac{sRT}{M(1-\bar{v}\,r)} \qquad (1.2)$$

(with the protein partial-specific volume $\bar{v}$, the solution density ρ, the gas constant R and the absolute temperature T) (4). A method for estimating D as a function of s for an ensemble of protein species with different sizes is described below.

The Lamm equation (1.1) can be solved by a finite element method, which was first introduced to ultracentrifugal sedimentation by Claverie (23), and recently generalized to permit greater efficiency and stability (11). The basic strategy consists in the approximation of the concentration χ(r,t) as a superposition

$$c(r,t) \approx \sum_{k=1}^{N} c_k(t)\, P_k(r,t) \qquad (1.3)$$

of N triangular elements P_k, which are defined as stepwise linear hat-functions on a grid of radial points $r_i(t)$

$$P_k(r,t) = \begin{cases} (r - r_{k-1})/(r_k - r_{k-1}) & r_{k-1} \le r \le r_k \\ (r_{k+1} - r)/(r_{k+1} - r_k) & r_k < r \le r_{k+1} \\ 0 & else \end{cases} \qquad for\ k = 2,\ldots, N-1 \tag{1.4}$$

(details on the elements at the boundary of the solution column can be found in (11)) and thereby reducing the solution of Eq. (1.1) to the determination of the evolution of the coefficients $\chi_k(t)$. These can be found by multiplication of Eq. (1.1) with the elements P_k and integration in radial coordinates from meniscus m to bottom b of the solution column:

$$\int_m^b \frac{\partial c}{\partial t} P_k(r,t)\, r dr = \int_m^b \frac{\partial}{\partial r}\left[r\left(D\frac{\partial c}{\partial r} - sw^2 r\, c \right)\right] P_k(r,t)\, dr \tag{1.5}$$

Integration by parts of the right-hand side (rhs), taking advantage that the inner parenthesis represents the total flux, which vanishes at both ends of the solution column, leads to

$$\int_m^b \frac{\partial c}{\partial t} P_k(r,t)\, r dr = sw^2 \int_m^b c\, \frac{\partial P_k(r,t)}{\partial r} r^2 dr - D\int_m^b \frac{\partial c}{\partial r}\frac{\partial P_k(r,t)}{\partial r} r dr \tag{1.6}$$

Insertion of Eq. (1.3) leads to

$$0 = \sum_j \frac{\partial c_j}{\partial t}\int_m^b P_j P_k r dr + \sum_j c_j \int_m^b \frac{\partial P_j}{\partial t} P_k r dr - w^2 s \sum_j c_j \int_m^b P_j \frac{\partial P_k}{\partial r} r^2 dr + D\sum_j c_j \int_m^b \frac{\partial P_j}{\partial r}\frac{\partial P_k}{\partial r} r dr \tag{1.7}$$

In the original Claverie approach (23), the hat functions P_k are constant in time, which reduces Eq. (1.7) to a linear equation system for the coefficients $\chi_k(t)$. With the abbreviations

$$\mathbf{B}_{jk} = \int_m^b P_j P_k r dr, \quad \mathbf{A}^{(1)}_{jk} = \int_m^b \frac{\partial P_j}{\partial r}\frac{\partial P_k}{\partial r} r dr, \quad \mathbf{A}^{(2)}_{jk} = \int_m^b P_j \frac{\partial P_k}{\partial r} r^2 dr \tag{1.8}$$

we can write Eq. (1.7) as a matrix-vector equation

$$\mathbf{B}\frac{\partial \vec{c}}{\partial t} = \left[w^2 s\mathbf{A}^{(2)} - D\mathbf{A}^{(1)} \right]\vec{c} \tag{1.9}$$

Although Eq. (1.9) can be integrated with a discrete time-interval Δt to give the evolution of $\vec{c}$, it is

advantageous to use a Crank-Nicholson scheme (24, 25). Here, the fluxes in the rhs are evaluated in the middle during the time-step Δt, which leads to a higher order accuracy of the time step (25). Substitution of $\vec{c}$ in the rhs by $\left(\vec{c}(t)+\vec{c}(t+\Delta t)\right)/2$ gives

$$\vec{c}(t+\Delta t)=\left(2\mathbf{B}-\Delta t\left(w^2 s\mathbf{A}^{(2)}-D\mathbf{A}^{(1)}\right)\right)^{-1}\left(2\mathbf{B}+\Delta t\left(w^2 s\mathbf{A}^{(2)}-D\mathbf{A}^{(1)}\right)\right)\vec{c}(t) \qquad (1.10)$$

which permits larger time-steps without loss of precision (9).

Unfortunately, the use of static elements P_k is numerically stable and efficient only if the sedimentation fluxes are small compared to the diffusion fluxes. For large particles with high s-value and small diffusion coefficient, the algorithm produces oscillations at the end of the solution column and in the vicinity of the sedimentation boundary, which can be overcome only by very time-consuming use of a very fine discretization in space and time. Therefore, a numerical method was designed that is not only stable in the limit of D = 0, but increases in efficiency for smaller diffusion coefficients. This can be accomplished in a natural way by introducing hat functions P_k on a moving grid adapted to the sedimentation process, where each radial grid-point (except the first and last) migrates as

$$\begin{aligned} r_k(t) &= r_{k,0}\,a(t-t_0) = r_{k,0}\exp\left\{s_G w^2 (t-t_0)\right\} \quad for \quad k=2,\ldots,N-1 \\ r_1(t) &= m\,, \quad r_N(t) = b \end{aligned} \qquad (1.11)$$

i.e. like a non-diffusing particle sedimenting in the gravitational field with the sedimentation coefficient s_G (with the notation α(t) to abbreviates this translation process). With the choice of the starting grid

$$r_{k,0} = m\left(b/m\right)^{(k-3/2)/(N-1)} \quad for \quad k=2,\ldots,N-1 \qquad (1.12)$$

the grid has the unique property that after a time interval of propagation

$$\Delta t_{swap} = \left[w^2 s_G (N-1)\right]^{-1} \ln(b/m) \qquad (1.13)$$

it is mapped precisely onto the starting grid

$$r_k\left(t_0+\Delta t_{swap}\right)=r_{k+1,0} \quad for \quad k=2,\ldots,N-2 \tag{1.14}$$

and the propagation $\alpha(t)$ reduces to a simple renumbering of the indices of r_k, which is computationally trivial. When solving the Lamm equation with the moving grid, the hat-functions are time-dependent

$$\frac{\partial P_k}{\partial t}=w^2 s_G \times \begin{cases} -r/(r_k-r_{k-1}) & r_{k-1}\le r\le r_k \\ r/(r_{k+1}-r_k) & r_k\le r\le r_{k+1} \\ 0 & else \end{cases} \qquad k=3,\ldots,N-2 \tag{1.15}$$

and the second term of Eq. (1.7) does not vanish. The abbreviation $\mathbf{A}^{(3)}_{kj}=\int_m^b\left(\partial P_j/\partial t\right)P_k r dr$ and the transformation of variables $r\left(r,t\right)=r/a\left(t-t_0\right)$ leads to

$$\mathbf{B}\frac{\partial\vec{c}}{\partial t}=\left[w^2\left(s\mathbf{A}^{(2)}-s_G\mathbf{A}^{(3)}\right)-\frac{D}{a\left(t-t_0\right)^2}\mathbf{A}^{(1)}\right]\vec{c} \tag{1.16}$$

This is a form analogous to Eq. (1.9), but generalized to a sedimenting frame of reference. At finite diffusion, an additional term $\alpha(t)^{-2}$ corrects the diffusion for the stretching of the reference frame. Again, it can be very efficiently solved with the Crank-Nicholson scheme, as the matrices are independent of time, except for the corner 2×2 submatrices if time-intervals other than Δt_{swap} are used (11). As designed, the rhs is zero for D = 0 and s = s_G, and thus Eq. (1.16) is trivial to solve, and it remains numerically stable and efficient at low D. Finally, for the limiting case of very large particles with negligible diffusion, an analytical solution of the Lamm equation can be used (26, 27):

$$\chi(s,r,t)=c_0e^{-2\omega^2st}\times\begin{cases}0 & for \quad r<me^{\omega^2st}\\ 1 & else\end{cases} \tag{1.17}$$

In combination, the static finite element approach by Claverie and the moving frame of reference solution, both in Crank-Nicholson scheme, allow the efficient and precise calculation of the

macromolecular concentration profiles in the centrifugal field for the complete spectrum of s-values. Because of greater freedom in the choice of the time-steps, the former is more efficient for small particles with small s and high D, while the latter is better for simulating sedimentation of larger particles with high s and small D. In practice, when calculating the kernels for the sedimentation coefficient distribution, an empirical threshold can be used to select the best algorithm.

The numerical algorithms described above can be adapted to account for unavoidable experimental complications that can cause measurable deviations from an 'ideal' sedimentation process Eq. (1.1). These include the acceleration phase of the rotor, which can be modeled by a time-dependent rotor speed in Eqs. (1.9) and (1.16) (28). It also includes the compressibility of the solvent, which even for aqueous buffers at high rotor speeds can produce density gradients that lead to small but significant retardation of the macromolecular migration (16). Compressibility can be accounted for by locally varying sedimentation coefficients, again approximated as superposition of hat-functions, and an extension of Eq. (1.17) for compressible solvents has been derived (16). The finite element method also allows to account for the sedimentation of co-solutes which change the local solvent density and viscosity and again lead to locally varying macromolecular sedimentation and diffusion coefficients (15). The approach can also be used without further complications to describe flotation (28) and alternative experimental configurations, such as analytical zone centrifugation (9, 29).

Accounting for the noise structure of sedimentation velocity data

Analytical ultracentrifugation data from sedimentation velocity experiments exhibit noise components that are not random. It is well-known that interference optical data can be entirely dominated by these signal offsets, but they are usually significant also for absorbance data. The systematic signal offsets are caused by the spatial imperfections of the optical components, and by

time-dependent vibrations and/or 2π phase shifts of the interference patterns, respectively. Accordingly, the systematic noise can be decomposed into a time-invariant component $a_{TI}(r)$ and a radial-invariant component $a_{RI}(t)$

$$a(r,t) = S(\{p\},r,t) + a_{TI}(r) + a_{RI}(t) + e(r,t) \tag{2.1}$$

(with the experimental data and random noise denoted as $a(r,t)$ and $\varepsilon(r,t)$ respectively, and with $S(\{p\},r_i,t_j)$ denoting any model for the sedimentation boundary, which may in general depend on a set of parameters $\{p\}$). Although other effects clearly exist, such as higher-order vibration modes, these two orthogonal components can describe the systematic components with a precision usually better than the random noise of the data acquisition. (An exception is shown in Figure 2, where some residual higher-order systematic contributions are visible.) They are also sufficient to eliminate the need for an optical reference in absorbance optical transport experiments, for example sedimentation velocity and analytical electrophoresis (30).

It is possible to account the noise components $a_{TI}(r)$ and $a_{RI}(t)$ algebraically (31). It is described here for the time-invariant baseline noise $a_{TI}(r)$, but it can be combined with an analogous procedure for radial-invariant noise $a_{RI}(t)$. If we abbreviate $a_{TI}(r_i)$ as b_i, we can calculate them by least-squares

$$\underset{\{p\},b_i}{\text{Min}} \sum_{i,j} \left[a(r_i,t_j) - \left(b_i + S(\{p\},r_i,t_j) \right) \right]^2 \tag{2.2}$$

leading to

$$b_i(\{p\}) = \bar{a}_i - \bar{S}_i(\{p\}) \tag{2.3}$$

where $\bar{a}_i = (1/N_s)\sum_j a(r_i,t_j)$ (the 'average scan') and $\bar{S}_i(\{p\}) = (1/N_s)\sum_j S(\{p\},r_i,t_j)$ (the 'average boundary model') with the total number of scans N_s. Insertion of this into the least-squares problem for the calculation of the remaining parameters $\{p\}$ of the boundary model leads to

$$\underset{\{p\}}{\text{Min}} \sum_{i,j} \left[\left(a(r_i, t_j) - \bar{a}_i \right) - \left(S(\{p\}, r_i, t_j) - \bar{S}_i(\{p\}) \right) \right]^2 \qquad (2.4)$$

This shows that the boundary parameters can be modeled relative to the 'average' radial profile as a reference. (It should be noted that this is different from the pair-wise differencing which is used in other approaches for sedimentation analysis (32), which leads to a stronger amplification of the random noise.) It also shows that the best-fit systematic noise estimate is easily calculated by least-squares, but will be dependent on the boundary model. $a_{TI}(r)$ and $a_{RI}(t)$ can slightly correlate with the sedimentation profiles at very small sedimentation coefficients. This correlation is equivalent to that introduced by a time-difference analysis (18) and can be reduced by acquiring data for a large the time interval, leading to a large boundary displacement (30). It can be seen from Eqs. (2.1) and (2.4) that modeling the sedimentation data is invariant under a transformation $\tilde{a}(r,t) = a(r,t) - \tilde{a}_{TI}(r) - \tilde{a}_{RI}(t)$. Therefore, the best-fit estimate of the systematic noise components can be subtracted from the raw data without changing their information content (as long as systematic noise components according to Eq. (2.1) are still considered). This can dramatically improve the possibility for visual inspection of the raw sedimentation data and the quality of their fit (Figure 2).

The central observation exploited for calculating high resolution sedimentation coefficients is that boundary spreading from differential migration and from diffusion is different, and that this difference can be extracted from the experimental data. To obtain reliable results, it is crucial to verify that the data are modeled within the random noise of the data acquisition. For testing the quality of a fit and if the residuals are within the statistical noise of the data acquisition, the overall rms error (and χ^2 statistics) can be used. The presence of remaining systematic deviations can be diagnosed with a runs test (33). However, high Z-values can be caused by both a systematic deviation of the model from the experimental data of the sedimentation boundary, or sometimes by optical imperfections, such as remaining higher-order vibrations in the data acquisition system. In

order to help distinguish these cases, a bitmap representation was developed, in which the residuals encode the grayscale of pixels, which are placed in a picture according to the radius and time coordinate of each data point. Visual inspection then permits the detection of correlation of the residuals with the migration of the sedimentation boundary, as opposed to static or periodic residuals pattern from optical imperfections. Although this is just an effective graphical representation of the residuals adapted to the data space of sedimentation velocity, it may be possible in future work to derive a quantitative measure for this correlation.

Distributions of Sedimentation Coefficients

A differential distribution of sedimentation coefficients c(s) can be defined as the population of species with a sedimentation coefficient between s and s+ds. Accordingly, integration of the peaks of c(s) can be used to calculate the weight-average s-value s_w of the sedimenting components, and to obtain their partial loading concentrations. The distribution can be related to the experimentally measured evolution of the local concentration profiles throughout the centrifuge cell a(r,t) by a Fredholm integral equation

$$a(r,t) = \int_{s_{min}}^{s_{max}} c(s)c(s, D(s), r, t)ds + a_{TI}(r) + a_{RI}(t) + e \qquad (2.5)$$

It states that the observed signal a(r,t) is a simple linear superposition of the contributions of all subpopulations c(s) at different s-values (ranging from a minimal value s_{min} to a maximal value s_{max}). Each species contributes to the radial and time-dependent signal with χ(s,D,r,t) as predicted by the Lamm equation for a macromolecule with sedimentation coefficient s and diffusion coefficient D. As will be outlined in detail below, we will estimate D through a functional dependence on s, assuming an average frictional ratio or hydrodynamic shape for all species. The noise contributions are described above, and will be omitted for clarity in the following.

In principle, one could attempt to directly solve Eq. (2.5) by discretization of the sedimentation coefficients in a grid of N s-values s_k from $s_1 \ldots s_N$, and approximating c(s) on this grid as a set of values $c(s_k)$ (short c_k, or in vector form $\vec{c}$). With data acquired at radius values r_i and at times t_j, Eq. (2.5) leads to the linear least-squares problem

$$a(r_i, t_j) \cong \sum_{k=1}^{N} c_k \, c(s_k, D_k, r_i, t_j) \qquad (2.6)$$

Given the large number of data points, it is advantageous to solve Eq. (2.6) by normal equations with Cholesky decomposition (34). Positivity of the c_k values can be ensured with the algorithm NNLS by Lawson and Hanson (34). The drawback of this direct approach is that there are in general many different distributions c*(s) that solve Eq. (2.5) nearly equally well and cannot be distinguished within the experimental precision of the data ε. This is a general, well-known problem of Fredholm integral equations (35-37), and is true, in particular, for cases were the kernel (the characteristic signal of a homogeneous subpopulation) is a smooth function (36). Notoriously difficult are exponentials, such as the decay of the autocorrelation function in dynamic light scattering, for which much of the numerical strategies outlined in the following was originally introduced (38). Fortunately, however, the large data space of sedimentation velocity and the characteristic features of the Lamm equation solutions $\chi(s,D,r,t)$ make the numerical determination of c(s) less problematic.

In order to achieve a stable solution, regularization must generally be used, which allows exploiting additional prior knowledge or prior assumptions. Two different types of prior have shown to be very powerful in many other biophysical disciplines. Tikhonov-Phillips regularization (TP) selects from all possible solutions that fit the data well the one with the highest smoothness, calculated as $\int (dc/ds)^2 \, ds$. Alternatively, the maximum entropy principle (ME) selects the one that

has the highest informational entropy or the minimal information content, given by $\int (c \log c)\, ds$ (39). The rationale is that according to Occam's razor, the solution with the highest parsimony is to be preferred. ME can be justified by Bayesian principles, and will find a solution that is most likely, given a prior expectation (which in the absence of other data is commonly taken to be a uniform distribution) (25). Instead of solving the linear least-squares problem Eq. (2.6), we simultaneously minimize the residuals of the fit and optimize the parsimony constraint $H(\vec{c})$

$$\underset{c_k}{\text{Min}} \left\{ \sum_{i,j} \left[a(r_i, t_j) - \sum_k c_k \chi(s_k, D_k, r_i, t_j) \right]^2 + \alpha H(\vec{c}) \right\} \qquad (2.7)$$

where $H(\vec{c}) = \sum_k c_k \log c_k$ for ME and $H(\vec{c}) = \vec{c}^T \mathbf{D}^T \mathbf{D} \vec{c}$ (with $\mathbf{D}$ denoting a second derivative matrix) for TP regularization (25). While the latter regularization method leads to a linear matrix equation, ME is a nonlinear problem which can be solved by the Levenberg-Marquardt method (25).

One important question is, however, how to balance the parsimony constraint with the quality of the fit. This can be done following as described first by Provencher in the context of the program CONTIN (40). The approach is based on the fact that all values of the regularization parameter $\alpha > 0$ increase the rms error (or χ^2 value) of the fit relative to the (generally instable) least-squares optimum ($\alpha = 0$). This allows use of a statistical criterion comparing the goodness of fit. The Fisher distribution predicts the probability of exceeding a ratio of $\chi^2(\alpha)/\chi^2(\alpha = 0)$. Therefore, α can be iteratively adjusted such that the probability of $\chi^2(\alpha)/\chi^2(\alpha = 0)$ corresponds to a predefined confidence level (usually 0.7). The degrees of freedom for the Fisher distribution are the number of data points (these are usually in the order of 10^5, and therefore the number of fitted parameters are negligible). Other methods for estimating the regularization parameter are possible (41). The approach described ensures that the parsimony constraints are effective to suppress peaks in the c(s)

distribution that are not warranted by the data, but to extract all the details that are reliable.

The bias introduced by the regularization is generally small. However, it should be noted that the TP and ME methods can slightly differ in their properties and results. ME has the property that it can produce sharper peaks than TP and in our experience gives better results when dealing with a mixture of discrete protein species (18), although we have observed that peaks in close vicinity sometimes appear to 'attract'. For broad and more continuous distributions, however, it is known to produce artificial oscillations, and TP appears to perform better in this case (18).

It is clear that this analysis relies on the ability to model the experimental data well. In practice, this is usually the case, but some care may have to be taken with regard to the choice of the boundaries s_{min} and s_{max} of the distribution. If they are chosen too narrowly, they may constrain the model and produce unreliable results. Beyond the inspection of the fit and its residuals, this can be diagnosed by an increase of c(s) toward the limits s_{min} or s_{max}. With regard to s_{max}, this situation can be indicative of the presence of large protein aggregates, and it is straightforward to increase the s_{min} value (in order to avoid high computational load when a large s-range is spanned, the grid of s-values can be chosen non-uniform, for example logarithmic). A value of $c(s_{min}) >> 0$ may be a result of either constraining s_{min} to a too high value, in which case it should be lowered, or a correlation with the baseline parameters. In our experience, the latter case is not detrimental to the reliability of the distribution at $s > s_{min}$, and therefore to be preferred.

Estimating the extent of diffusion

So far, we have described the numerical methods for efficiently solving the Lamm equation, combining these solutions to calculate a sedimentation coefficient distribution, and adapting this boundary model to the special noise structure of sedimentation velocity data. The remaining problem, which was deferred above, is to estimate the extent of diffusion for each species in the

distribution Eq. (2.5). The best approach will depend on what is known about the ensemble of macromolecules under study. In principle, three different levels of complexity could be considered. The simplest approach would be to assume D to be constant for all species. This is a strong assumption, but it can be appropriate, for example, for distributions of proteins of the same Stokes radius, such as ferritin and apoferritin, and D could be taken from a measurement by dynamic light scattering. In the other extreme, one could consider the diffusion coefficients to be distributed themselves, independently of the sedimentation coefficients, and instead of a one-dimensional sedimentation coefficient distribution c(s), a joint two-dimensional distribution c(s,D) could be used to characterize the sample. This is the most general approach, but unfortunately a single sedimentation velocity experiment does not provide sufficient information. It is possible, however, to characterize such a c(s,D) distribution by a global analysis of multiple sedimentation velocity experiments at different rotor speeds, or in global analysis with autocorrelation data from dynamic light scattering (42).

An intermediate strategy is to estimate the diffusion coefficients from a monotonous single-valued function of the sedimentation coefficient. This can be expressed via the frictional ratio f/f_0, which is the ratio of the translational friction coefficient of the molecule relative to that of a sphere of the same mass and density. It is well-known that the values of f/f_0 are only very weakly dependent on the macromolecular shape (43) (examples for commonly observed values of the hydrated frictional ratio are 1.2 – 1.3 for relatively globular proteins, 1.5-1.8 for asymmetric or glycosylated proteins, and larger values for very asymmetric or unfolded proteins or linear chains). Therefore, for a given sample, it can be a good approximation to fix f/f_0 to the weight-average frictional ratio $(f/f_0)_w$ of the macromolecules in the mixture. Using the Stokes-Einstein relationship and the Svedberg Equation (1.2) we can derive

$$D(s) = \frac{\sqrt{2}}{18p} kT\ s^{-1/2} \left(h\left(f/f_0\right)_w\right)^{-3/2} \left(\left(1-\bar{v}r\right)/\bar{v}\right)^{-1/2} \qquad (2.8)$$

(with k denoting the Boltzmann constant). This relationship can be used in Eq. (2.5) to calculate the Lamm equation solutions for each single species, followed by the determination of the best-fit sedimentation coefficient distribution c(s). The correct value for $(f/f_0)_w$ can be determined iteratively by non-linear regression, optimizing the quality of fit of the c(s) boundary model as a function of the $(f/f_0)_w$ value. Except for experiments with unusually poor signal-to-noise ratio or shortened observation time, $(f/f_0)_w$ is well-determined by the data. Very importantly, sub-optimal values for $(f/f_0)_w$ value are detrimental for the resolution of the c(s) distribution, but have relatively little effect on the location of peaks in c(s) (18): This is similar to the apparent sedimentation coefficient distribution g*(s) of non-diffusing particles, which experiences (approximately Gaussian) broadening from the unaccounted macromolecular diffusion, but still reports the correct s-value. To a much lesser extent, sub-optimally corrected diffusion in the c(s) distribution will cause a slight broadening of its peaks, but not a displacement.

In summary, we exploit three helpful properties for calculating the sedimentation coefficient distributions: First, the diffusion D depends only on the square root of s, which is a result of the sedimentation coefficient being much stronger size-dependent than the diffusion coefficient. The dependence D(s) can be expressed quantitatively through the frictional ratio f/f_0. Second, for mixtures of macromolecules of similar origin (i.e. folded proteins versus random polymer chains), f/f_0 does also not depend strongly on the detailed macromolecular shape. Third, if we approximate f/f_0 as a constant, the resulting c(s) distribution is robust against small errors in $(f/f_0)_w$. As a result, we obtain a sedimentation coefficient distribution c(s) which takes into account the diffusional spread for each species, and is therefore able to extract the heterogeneity and differential migration from the measured sedimentation boundaries.

Properties and variations of the c(s) distribution

An important question is the sensitivity and precision of the c(s) distribution, and the derived values for s_w and the partial loading concentrations. A safeguard against over-interpretation is the regularization, which provides only the simplest distribution that can model the data well. (It should be noted that this causes broadening of the peaks dependent on the noise and size of the raw data set; the peak width is therefore not always well-suited as a characteristics of the sedimenting particles, and it is generally also not a good measure of the error in the s-values.) Nevertheless, the distribution may contain a lot of details, since the data basis is very large, consisting in the order of 10^5 data points from the complete sedimentation process. Simulations show that from a set of profiles covering the complete sedimentation process at a signal-to-noise ratio of 200:1 (which can be readily achieved, e.g. at a loading concentration of 0.3 to 0.4 mg/ml of protein in the interference optics), minor peaks consisting of 0.2% of the total protein concentration can be reliably detected. The statistical accuracy of the calculated c(s) distribution can be assessed by Monte-Carlo simulations (25). Assuming that the best-fit boundary model from the c(s) fit is a good description of the data, and that the noise is normally distributed, a large number of synthetic data sets *j* can be generated and each subjected to the c(s) analysis. The resulting family of $c_j(s)$ curves can be analyzed in different ways. First, for any s-value s_k the limits of central 68% of $c_j(s_k)$ values from the Monte-Carlo simulation can be calculated. Applied to all s-values, this procedure generates a one standard deviation contour for c(s). For a detailed analysis with the intent to quantify the concentration and the s-value of sedimenting components, an integral approach is advantageous. If each $c_j(s)$ curve is integrated from s_1 to s_2, the resulting statistics reflects the uncertainty of the partial concentration and s_w-value of species sedimenting between s_1 and s_2.

Several variants of the c(s) distribution have also proven useful in practice. For the study of the protein sedimentation coefficient distribution in the presence of a small molecular weight compound

that contributes to the signal (e.g., a low concentration of nucleotides observed with the absorbance optics, or unmatched buffer salts in the refractometric optics), the sedimentation profiles can be characterized by a skewing baseline (from the approach of equilibrium for the small component) superimposed by the sedimenting boundaries of the protein components. In this case, the small component can frequently be described best as a species with discrete s and D-values, superimposed to but not part of the c(s) distribution (and thus not distorting $(f/f_0)_w$ and the regularization of the protein). Its sedimentation parameters can either be separately determined, or optimized in a non-linear regression of the experimental data, jointly with the other non-linear parameters governing the c(s) distribution, such as $(f/f_0)_w$ and the meniscus position of the solution column. Another variant of c(s) is useful when a certain peak (between s' and s") can be identified with a species of known molar mass: The subpopulation of species between s' and s" can be excluded from the estimation D(s) via Eq. (2.8), and instead the known molar mass can be used with the Svedberg equation to calculate D for these species. This additional constraint can be useful, for example, for the study of proteins with conformation changes (44).

Since in the c(s) distribution each s-value is assigned an estimated D-value, it is possible to transform c(s) into a differential molar mass distribution c(M). However, this requires caution – although the c(s) distribution is not strongly influenced by the best-fit value for $(f/f_0)_w$, this is not true for c(M). Here, a peak location will strongly depend on $(f/f_0)_w$ being a good estimate for that species. However, this can be fulfilled, for example, if it is known that the different species have the same hydrodynamic shape, or if the distribution consists of a single major peak (which will thus govern the weight-average $(f/f_0)_w$). Despite this caveat, the c(M) distribution can be a highly useful tool to obtain molar mass estimates of the sedimenting species (see example below, and e.g., (45, 46)).

The c(s) distribution can also be used to calculate apparent sedimentation coefficient distributions g*(s) of non-diffusing particles. This can be achieved either in the limit of a fixed very

large value of $(f/f_0)_w$ (> 10), or by substitution of the ordinary Lamm equation solutions with those for non-diffusing particles (Eq. (1.17) or its extension to compressible solvents (16). This g*(s) distribution has advantages over the distribution g(s*) derived by the dcdt method (21, 22, 47), since the artificial broadening resulting from the approximation of dt by Δt is absent, permitting the analysis of larger data sets (27, 48).

In other variations, the c(s) distribution was adapted to provide flotation coefficient distributions of spherical emulsion particles, making use the known size-dependence of the partial-specific volumes of the particles (28). Also, the modeling of data from analytical zone centrifugation (29) by c(s) has been implemented. In this configuration, slightly higher correlation of the Lamm equation solutions was observed (in particular if the lamella thickness is treated as an unknown parameter), and stronger regularization appears to be required.

If the c(s) analysis is applied to proteins that exhibit reversible interactions on the time-scale of sedimentation, the c(s) analysis does not reveal the populations of species sedimenting with a certain rate. This is because the fast reversible conversion of slower sedimenting and faster sedimenting protein complexes leads to an overall broadening of the sedimentation boundary, which can only be empirically modeled as an apparent superposition of sedimentation profiles of non-reacting species. In many cases, the chemical reaction results in artificial peaks in c(s) at s-values intermediate to those of the existing protein species, and in a bias of the $(f/f_0)_w$ value. Nevertheless, the c(s) analysis can in this situation still reveal a large amount of important information. First, if experiments are conducted at different loading concentrations, a shift in the peak positions can reveal the presence of an interaction on the time-scale of sedimentation. Second, the c(s) distribution can allow to distinguish the interacting components from those not participating in the reaction, such as small proteolytic degradation products, impurities, and/or irreversible protein aggregates, which can usually be well resolved from the native proteins and their reversible complexes.

Third, integration of c(s) over the range of s-values of the interacting species reveals a weight-average s-value s_w, which is fully equivalent to that from second moment methods, and can be used for interacting systems to study the binding isotherm $s_w(c^*)$ of the interacting species (48). (Interestingly, if the regularization is scaled as described above and provides a boundary model that is undistinguishable from that in the absence of regularization, it can be shown by second moment considerations that it does not affect the calculated s_w value, even though it may considerably change the c(s) distribution itself.) Because of radial dilution, the effective concentration c* for which the s_w-value is determined is generally smaller than the loading concentration, but because the c(s) distribution describes the entire sedimentation process, c* is higher than the plateau concentration (for a precise analysis, see (48)). Because the c(s) distribution can be based on a very large data base, good estimations of s_w can be accomplished at very low loading concentrations (with the loading signal only 2 – 3fold the noise of the data acquisition), which can be crucial for the interpretation of the binding isotherm $s_w(c^*)$ (48).

Experimental considerations

In contrast to the g(s*) distribution by dcdt (21, 22, 47) or the integral G(s) distribution from the van Holde-Weischet method (18, 49, 50), the c(s) analysis does not require constraints in the data analysis to subsets of the sedimentation process. For c(s), the resolution increases with increasing amount of data and observation time, and optimal are data sets comprising the complete sedimentation process from directly after start of the centrifuge up to the depletion of the smallest visible species. This allows the characterization of species over a large range of s-values in one experiment. No depletion at the meniscus and no established plateaus are required.

From theory, one could expect the presented approach to work best at high rotor speeds, where the resolution in s is highest, and the diffusional spread is smallest. In practice, however, the method

also works well at lower rotor speeds, except for very heterogeneous samples (with species of different frictional ratios) in the approach to equilibrium. In order to optimize the precision of the experimental data, a long temperature equilibration period (~ 1 hour) before start of the experiment is recommended, and an initial adjustment of the optics at a low rotor speed should be avoided. The acceleration of the rotor and the compressibility of the solvent can be accounted for when solving the Lamm equations. Because generally the meniscus position cannot be determined graphically with sufficient precision (it should be noted that it may not coincide with the peak of the characteristic optical artifact), it is usually required to be included as a non-linear fitting parameter that is optimized in a series of c(s) analyses (jointly with $(f/f_0)_w$). If the sedimentation data included regions of back-diffusion from the bottom of the solution column, or if a small molar mass species is included in the model, the bottom position of the solution column should also be determined by non-linear regression. In most cases, a fit of the data within the noise of the data acquisition can be achieved.

Examples of applications

The c(s) method has found application in numerous studies of proteins and their interactions (20). Here, we restrict the applications to only two examples. The first one will illustrate the use of c(s) for an oligomeric protein that is stable on the time-scale of sedimentation but exhibits impurities and microheterogeneity, and the second one will show the c(s) analysis of a heterogeneous interaction between two proteins that form different complexes.

Figure 3 shows data from the study of the oligomeric state of the extracellular domain of an NK receptor, expressed by drosophila cells. The molar mass calculated by amino acid composition is 43.7 kDa; however, the molar mass from MALDI was ~ 50.1 kDa, exhibiting several peaks ranging from 49.4 to 51.6 kDa as a result of differences in the extent of glycosylation. If the analysis is based on the assumption of the presence of a single sedimenting species, the spread of the sedimentation

boundary can be approximately modeled as if arising from the single-species diffusion (Figure 3A, dashed line). This leads to an s-value of 4.83 S and a molar mass estimate of 63.4 kDa, suggesting that the protein would be predominantly monomeric. However, the measured s-value exceeds the maximum possible s-value (4.32 S) of a monomeric protein.

In contrast, the c(s) analysis enables the decomposition of multiple sedimenting components and diffusional boundary spreading (Figure 3C). The best-fit frictional ration $(f/f_0)_w$ is 1.56, and the c(s) distribution as a main peak at 4.74 S (Figure 3E). Transformed into a c(M) distribution, the main peak corresponds to a species of 94.3 kDa, close to that of a dimer. As discussed above, although the determination of the molar mass is not generally possible, it provides a good estimate for the species sedimenting in a single main peak of the c(s) distribution. The dimeric state of the protein was confirmed independently by sedimentation equilibrium. However, while the molar mass estimates from c(s) are generally not as precise, more information is obtained here in comparison with sedimentation equilibrium: the c(s) distribution also shows the presence of impurities, which include ~ 5% of species of higher molar mass (estimated 150-200 kDa), and ~ 8% of smaller species including a fraction in the size range of the monomer (possibly misfolded and incompetent monomer). Moreover, from the model c(s) analysis, one can have access to the hydrodynamic shape of the molecules under study. For the given frictional ratio of 1.56, assuming hydration of 0.3g/g, we can calculate axial ratios of 7.2 and 8.0 for the hydrodynamically equivalent prolate and oblate ellipsoids, respectively. In the present example, however, a significant hydrodynamic friction is due to the glycosylation, and the true molecular shapes can be expected to exhibit lower geometric asymmetry.

Noteworthy is the highly significant improvement of the quality of fit from an rms deviation of 0.0084 fringes and substantial systematic errors in the single species model (Figure 3B) to 0.0039 fringes with very little systematic error (Figure 3D) in the c(s) analysis. This highlights the necessity

to obtain a fit of high quality, and it demonstrates that the details in the experimental profiles contain the information necessary to distinguish boundary spreading due to diffusion from boundary spreading from diffusion combined with heterogeneity. In the present case, the heterogeneity arises from both the impurities and the microheterogeneity in the extent of glycosylation. Qualitatively similar results were obtained with protein expressed in drosophila cells in the presence of tunicamycine, which suppresses the glycosylation. Like the fully glycosylated form, the c(s) analysis revealed some inhomogeneity of the protein sample, but with a main peak at 83 kDa, close to twofold the monomer mass measured by MALDI (44.9 kDa). Here, the dimeric state was confirmed independently in combination with dynamic light scattering.

The second example illustrates the study of interacting species when the chemical reaction is at the time-scale of sedimentation. The extracellular domain of a natural killer receptor NKR (15 kDa) is tested for binding in vitro with a class I MHC molecule (44 kDa) (51). These proteins were refolded from inclusion bodies expressed in E. coli and are unglycosylated. Interestingly, no interaction was measurable with a surface plasmon resonance biosensor, likely due to immobilization artifacts. By sedimentation velocity and sedimentation equilibrium the NKR molecule is at micromolar concentrations a stable dimer (showing a single peak in c(s) at 2.75 S; Figure 3B). At the same concentration, the MHC is monomeric with an s-value of 3.66 S, but exhibiting some low molar mass contaminant at 1.75 S (Figure 3B).

Figure 3A shows the sedimentation velocity data of a mixture of 43 μM NKR and 25 μM MHC (solid lines) and the best-fit c(s) boundary model (dashed line). The corresponding c(s) distribution has a large peak corresponding to the free NKR at 2.7 S, a smaller peak at 1.7 S corresponding to the contaminant of the MHC, and a bimodal peak in the range of 4 – 5 S (bold solid line in Figure 4B) indicating complex formation and unambiguously demonstrating an interaction between the molecules. Although c(s) appears to resolve two faster sedimenting components reflected in the

bimodal nature of the complex peaks, it should be noted that when chemical reactions are observed on the same time-scale as the sedimentation, the peaks in c(s) do not necessarily reflect populations of species at the given s-values (see above). This is due to the coupled sedimentation of the reacting species, which leads to a broadening of the sedimentation boundaries in excess of their individual diffusional spread. In the c(s) analysis, such a reactive boundary broadening results in a slight decrease in the quality of fit, and in a decrease of the best-fit $(f/f_0)_w$ value, which in the current example was 1.04, a value too small for a hydrated globular protein. A clear indication of the chemical reaction is the shift in the c(s) distributions at different loading concentrations and molar ratios. Also shown in Figure 4B are the results obtained from a series of experiments with a lower NKR concentration (10 μM fixed) and increasing concentrations of MHC (2.5 μM, 10 μM, 20 μM and 40 μM). While with excess NKR and MHC we find well resolved peaks at the s-values of the independently sedimenting molecules, the peaks of the complex are broad and shift from ~ 4.3 S to 5.5 S.

The range of complex peaks at higher s-values in the c(s) distributions suggests the hypothesis of two different complex species with a 1:1 and a 2:1 stoichiometry of MHC per NKR dimer. However, it cannot be rigorously concluded at this stage of the analysis because the c(s) peaks, when modeling a reaction boundary, do not correspond to s-values of sedimenting species. Further, the largest observed s-value of 5.5 S is smaller than the maximal s-value of 6.3 S for a smooth and compact molecule of the molar mass corresponding to a 1:1 complex (MHC/NKR dimer). However, the available crystallographic structure corresponds to a complex of 1 NKR dimer symmetrically binding 2 MHC molecules. Although this structure was obtained by symmetry of one NKR monomer bound to one MHC molecule, the dimeric state of the NKR is unambiguous and therefore the 2:1 complex appears relevant. Hydrodynamic modeling of the 1:1 and 2:1 complexes based on the crystallographic structure (52) leads to predicted s-values of 4.7 S and 6.3 S, respectively. On this

basis, the largest observed s-value of 5.5 S exceeds that of a 1:1 complex and appears to reflect a mixture of populations of 1:1 and 2:1 complexes in a fast reversible interaction. This illustrates the difficulty of interpreting the peak positions of c(s) in the presence of interactions on the time-scale of sedimentation, but also demonstrates how qualitative characteristics can be diagnosed from c(s), in particular in conjunction with crystallographic data.

A rigorous thermodynamic analysis of the interaction is possible on the basis of the isotherms of weight-average sedimentation coefficients as a function of protein concentration (2, 48, 53). These can be obtained from integrating the c(s) distribution (48). The integration can exclude the small contamination at 1.75 S, as this peak is clearly resolved and this species does not participate in the interaction. This underlines that the precise determination of s_w benefits from the high resolution of the c(s) distribution. The s_w data as a function of total MHC concentration are shown in Figure 4C (solid circle for NKR at 43 μM, and open squares for the data with NKR of 10 μM). Theoretical models for the s_w isotherms can be derived based on the laws of mass action and mass conservation, both for a single site model and for a model with two identical sites for MHC per NKR dimer. In contrast to ordinary binding isotherms, the isotherms of weight-average s-value typically exhibit a maximum at an optimal concentration and molar ratio, and decrease if either protein is present in excess concentration. These isotherms were fitted globally to the experimental data, using the s-values of the 1:1 and 2:1 complex (within constraints set by hydrodynamics) and the binding constants as unknown parameters. While the single-site isotherms cannot describe the data well, the isotherms for a two-site model provide an excellent fit, with a best-fit K_D of 1.7 μM (0.2 – 1.7 μM) and 4fold (2 – 57fold) negative cooperativity for (Figure 4C, bold solid and dashed line), and best-fit s-values for the 1:1 and 2:1 complex of 4.30 S and 6.18 S respectively. From these numbers, we find that the 2 binding sites of the NKR dimer are not saturated in our experimental conditions, considering that the highest s-value observed was only 5.5 S. Although a higher saturation of the

complex would be desirable in the analysis of the isotherm, non-ideal sedimentation prevents the use of higher protein concentrations in the present case. Nevertheless, because the s_w isotherms of the two-site models exhibit broader peaks than is possible with single-site isotherms, the available data clearly demonstrate the existence of both 1:1 and 2:1 complexes in solution. Interestingly, although sedimentation equilibrium clearly showed the existence of a 1:1 complex, it was consistent with but did not permit the unambiguous detection of a 2:1 complex. More details on the analysis of this interacting protein system can be found in (51).

Discussion

In recent years, the development of fast finite element methods for solving the Lamm equation in combination with algebraic techniques for the detection and elimination of systematic noise in sedimentation velocity experiments have enabled high resolution, diffusion deconvoluted sedimentation coefficient distributions c(s). They are solutions of Fredholm integral equations of the first kind, calculated with maximum entropy or Tikhonov-Phillips regularization, and based on approximations of the relationship between the diffusion and sedimentation coefficient D(s) derived from hydrodynamic considerations. The current review provides a summary of the computational approaches, together with a discussion of experimental requirements and examples for the application to mixtures of non-interacting and interacting proteins.

In comparison with existing methods for calculating sedimentation coefficient distributions, perhaps the most basic difference is that c(s) is a direct model for the complete data from the sedimentation process. In contrast, previous approaches include the transformation of time-derivative dc/dt to form an apparent sedimentation coefficient distribution g(s*) (21) (which does not distinguish diffusion and sedimentation), and the integral sedimentation coefficient distribution G(s)

which is obtained via an extrapolation procedure to infinite time (49). Although the latter does provide a correction for diffusion, it can only resolve species that exhibit well-separated sedimentation boundaries. Both methods require specific experimental configurations and restricting the analysis to a subset of the available data. A detailed theoretical analysis of the relationships between c(s), g*(s), and G(s) and a comparison of their results for different theoretical and experimental model systems can be found in (18).

The use of numerical solutions of the Lamm equation in c(s) permits a more flexible and general approach, analogous to continuous parameter distributions well-known in other biophysical techniques, such as dynamic light scattering (38, 39) (implemented, for example, in the program CONTIN), fluorescence anisotropy (54), and more recently optical affinity biosensing (55). The common computational aspect of these continuous distributions is the description of the experimental data via a Fredholm integral equation, and the use of a Bayesian principle to calculate the most parsimonious distribution consistent with the data. In comparison with other techniques where the kernel (the characteristic function of a single species) is frequently exponential, the sedimentation coefficient distributions can achieve relative high resolution due to the very low correlation between different Lamm equation solutions. However, the c(s) distribution does depend on additional global parameters (e.g., $(f/f_0)_w$ in the hydrodynamic approximation of D(s)), which require optimization by non-linear regression. The latter step can be avoided in a more general approach with the model of a two-dimensional size-and-shape distribution (e.g., c(s,D)), but this requires global analysis and additional data from dynamic light scattering and/or sedimentation velocity experiments at different rotor speeds (42).

The c(s) method has already found applications in a variety of protein studies (20). As illustrated, c(s) can be used in the determination of the protein sedimentation coefficients and molar mass, detect trace components of protein aggregates, small and large molar mass impurities (Figure 2

and 3), and series of c(s) distributions obtained at different protein concentrations can be used for the study of homogeneous and heterogeneous protein interactions (Figure 4) (48).

Acknowledgment

We thank Drs. Roy Mariuzza and Klaus Karjalainen for their support and contributions to the studies on the NK receptor systems.

References:

1. Lebowitz, J., Lewis, M. S., and Schuck, P. (2002) *Protein Sci* **11,** 2067-79.
2. Rivas, G., Stafford, W., and Minton, A. P. (1999) *Methods: A Companion to Methods in Enzymology* **19,** 194-212.
3. Laue, T. M., and Stafford, W. F. I. (1999) *Annu. Rev. Biophys. Biomol. Struct.* **28,** 75-100.
4. Svedberg, T., and Pedersen, K. O. (1940) The ultracentrifuge, Oxford University Press, London.
5. Schachman, H. K. (1959) Ultracentrifugation in Biochemistry, Academic Press, New York.
6. Elzen, B. (1988) Scientists and rotors. The development of biochemical ultracentrifuges, Dissertation, University Twente, Enschede.
7. Lamm, O. (1929) *Ark. Mat. Astr. Fys.* **21B(2),** 1-4.
8. Philo, J. S. (1997) *Biophys. J.* **72,** 435-444.
9. Schuck, P., MacPhee, C. E., and Howlett, G. J. (1998) *Biophys. J.* **74,** 466-474.
10. Schuck, P., and Millar, D. B. (1998) *Anal. Biochem.* **259,** 48-53.
11. Schuck, P. (1998) *Biophys. J.* **75,** 1503-1512.
12. Stafford, W. F. (1998) *Biophys. J.* **74(2),** A301.
13. Demeler, B., Behlke, J., and Ristau, O. (2000) *Methods in Enzymology* **321,** 36-66.
14. Behlke, J., and Ristau, O. (2002) *Biophys Chem* **95,** 59-68.
15. Schuck, P. (2003) *Biophys. Chem.* in press.
16. Schuck, P. (2003) *Biophys. Chem.* in press.
17. Yphantis, D. A., Lary, J. W., Stafford, W. F., Liu, S., Olsen, P. H., Hayes, D. B., Moody, T. P., Ridgeway, T. M., Lyons, D. A., and Laue, T. M. (1994) *in* Modern analytical ultracentrifugation (Schuster, T. M., and Laue, T. M., Eds.), pp. 209-226, Birkhäuser, Boston.
18. Schuck, P., Perugini, M. A., Gonzales, N. R., Howlett, G. J., and Schubert, D. (2002) *Biophys J* **82,** 1096-1111.

19. Schuck, P. (2000) *Biophys. J.* **78,** 1606-1619.

20. http://www.analyticalultracentrifugation.com/references.htm.

21. Stafford, W. F. (1994) *Methods Enzymol.* **240,** 478-501.

22. Philo, J. S. (2000) *Anal. Biochem.* **279,** 151-163.

23. Claverie, J.-M., Dreux, H., and Cohen, R. (1975) *Biopolymers* **14,** 1685-1700.

24. Crank, J., and Nicholson, P. (1947) *Proc. Cambridge Philos. Soc.* **43,** 50-67.

25. Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992) Numerical Recipes in C, University Press, Cambridge.

26. Fujita, H. (1962) Mathematical Theory of Sedimentation Analysis, Academic Press, New York.

27. Schuck, P., and Rossmanith, P. (2000) *Biopolymers* **54,** 328-341.

28. Perugini, M. A., Schuck, P., and Howlett, G. J. (2002) *Eur J Biochem* **269,** 5939-3949.

29. Lebowitz, J., Teale, M., and Schuck, P. (1998) *Biochem. Soc. Transact.* **26,** 745-749.

30. Kar, S. R., Kinsbury, J. S., Lewis, M. S., Laue, T. M., and Schuck, P. (2000) *Anal. Biochem.* **285,** 135-142.

31. Schuck, P., and Demeler, B. (1999) *Biophys. J.* **76,** 2288-2296.

32. Stafford, W. F. (2000) *in* Methods Enzymol. (Johnson, M. L., Abelson, J. N., and Simon, M. I., Eds.), Vol. 323, pp. 302-25., Academic Press, New York.

33. Straume, M., and Johnson, M. L. (1992) *Methods Enzymol.* **210,** 87-105.

34. Lawson, C. L., and Hanson, R. J. (1974) Solving least squares problems, Prentice-Hall, Englewood Cliffs, New Jersey.

35. Phillips, D. L. (1962) *Assoc. Comput. Mach.* **9,** 84-97.

36. Provencher, S. W. (1982) *Comp. Phys. Comm.* **27,** 213-227.

37. Hansen, P. C. (1992) *Inverse Probl.* **8,** 849-872.

38. Provencher, S. W. (1979) *Makromol. Chem.* **180,** 201-209.

39. Livesey, A. K., Licinio, P., and Delaye, M. (1986) *J. Chem. Phys.* **84,** 5102-5107.

40. Provencher, S. W. (1982) *Comp. Phys. Comm.* **27,** 229-242.

41. Hansen, P. C. (1998) Rank-deficient and discrete ill-posed problems., SIAM, Philadelphia.

42. Schuck, P. *manuscript in preparation*.

43. Cantor, C. R., and Schimmel, P. R. (1980) Biophysical Chemistry. II. Techniques for the study of biological structure and function, W.H. Freeman, New York.

44. Schuck, P., Taraporewala, Z., McPhie, P., and Patton, J. T. (2000) *J Biol Chem* **276,** 9679-9687.

45. Hatters, D. M., Wilson, L., Atcliffe, B. W., Mulhern, T. D., Guzzo-Pernell, N., and Howlett, G. J. (2001) *Biophys. J.* **81,** 371-381.

46. Benach, J., Chou, Y.-T., Fak, J. J., Itkin, A., Nicolae, D. D., Smith, P. C., Wittrock, G., Floyd, D. L., Golsaz, C. M., Gierasch, L. M., and Hunt, J. F. (2003) *J. Biol. Chem.* **278,** 3628-3638.

47. Stafford, W. F. (1992) *Anal. Biochem.* **203,** 295-301.

48. Schuck, P. *submitted*.

49. van Holde, K. E., and Weischet, W. O. (1978) *Biopolymers* **17,** 1387-1403.

50. Demeler, B., Saber, H., and Hansen, J. C. (1997) *Biophys J* **72,** 397-407.

51. Dam, J., and al., e.

52. Garcia De La Torre, J., Huertas, M. L., and Carrasco, B. (2000) *Biophys J* **78,** 719-30.

53. Correia, J. J. (2000) *Methods in Enzymology* **321,** 81-100.

54. Steinbach, P. J. (1996) *Biophys. J.* **70,** 1521-1528.

55. Svitel, J., Balbo, A., Mariuzza, R. A., Gonzales, N. R., and Schuck, P. (2003) *Biophys. J.* in press.

Figure Legends

Figure 1: Schematic principle of different strategies for sedimentation velocity data analysis. The concentration profiles at different times are the measured data (solid lines). Panel A depicts the determination of the boundary midpoints (dashed vertical lines), and the migration of the midpoint with time. If the midpoint is determined by the second moment methods (5, 48), the result is the weight-average sedimentation coefficient. Panel B: In addition to the displacement of the boundary midpoint, the spread of the boundary can be interpreted to result from macromolecular diffusion, and modeled with a Lamm equation solution (or an approximation thereof) (dashed line). If the sample under study consists of a single species, the resulting diffusion coefficient allows determination of the molar mass of the macromolecule. Panel C: For a two-component system, however, an additional boundary spreading can be caused by the differential migration of the sedimenting species (indicated by horizontal bars). For proteins, frequently, the separation of the species does not exceed their diffusional spread, causing a complex boundary shape. An 'apparent diffusion', if evaluated from the overall boundary shape, is not a meaningful parameter, and the existence of multiple species may not be obvious. The quantitative analysis of the data in Panel C requires distributions of Lamm equation solutions.

Figure 2: Components of the c(s) boundary model for experimental sedimentation velocity data. Panel A shows the experimental raw data from laser interferometry optics (of an IgG sample dissolved in phosphate buffered saline sedimenting at 40,000 rpm; only a data subset is shown). It is obvious that a substantial part of the signal is due to systematic noise contributions. The c(s) boundary model accounts not only for the sedimenting macromolecules, but also for a time-invariant baseline component (TI noise) and an orthogonal radial-invariant baseline (RI noise) (Eq. (2.5)), which are simultaneously fit to the data. The results of this fit are shown in Panels B to F. Panel B

shows the best-fit TI noise as a function of radius (dashed line, lower abscissa) and the RI noise as a function of time (solid line, upper abscissa). Besides the apparent 2π phase shifts, the RI noise contains smaller amplitude vibrations. Panel C is the calculated c(s) distribution (bold line, and in 20fold magnification as dashed line). It has a sharp peak for the major monomeric IgG component, but also shows the presence of small but significant populations of higher oligomers and aggregates. Panel D and E show the residuals of the fit, which has an rms deviation of 0.0023 fringes. Panel D are the superimposed radial profiles for all times, and Panel E is a bitmap representation of the residuals in the radius *vs*. time plane (with a linear grayscale covering –0.02 to + 0.02 from black to white). Vertical and horizontal structures are visible, which indicate the presence of small higher-order vibrations, but very little diagonal structure appears, which shows that there are no systematic residuals associated with the sedimenting boundary. Panel F: Since the data analysis is invariant under transformations that add systematic TI or RI noise, one can subtract the best-fit systematic noise components (Panel B) from the raw data. This provides an equivalent of the raw data but free of systematic noise.

Figure 3: Analysis of experimental sedimentation velocity data from the study of the oligomeric state of a glycosylated NK receptor fragment. Protein was dissolved in phosphate buffered saline, and sedimentation profiles were observed at a rotor speed of 55,000 rpm and a rotor temperature of 22ºC. The radial protein distribution was observed with the interference optics in time intervals of 30 sec. The partial-specific volume of the protein was estimated as 0.725 ml/g (based on the amino acid composition and the average extent of glycosylation as measured by MALDI). Panel A: Sedimentation profiles (every 20^{th} scan shown, solid lines) after subtraction of the systematic noise calculated with the best-fit single species Lamm equation model (dashed bold lines). Panel B: Superposition and bitmap representation of the residuals (rms error 0.0084 fringes). Panel C:

Sedimentation profiles (solid lines) after subtraction of the c(s) boundary model (dashed lines). Panel D: Superposition and bitmap representation of the residuals (rms error 0.0039 fringes). Panel E: c(s) distribution.

Figure 4: Study of the interaction between an NKR and an MHC molecule by sedimentation velocity and c(s) analysis. The experiments were conducted in phosphate buffered saline at a rotor speed of 55,000 rpm, and a rotor temperature of 20ºC. Panel A: Sedimentation profiles measured with a mixture of 43 μM NKR and 25 μM MHC (solid lines, every 10^{th} scan shown), and best-fit c(s) boundary model (bold dashed line), which converges to a best-fit $(f/f_0)_w$ value of 1.04. Best-fit systematic noise components are subtracted for clarity. Panel B: c(s) distributions obtained from the data of the mixture shown in Panel A (bold solid line), and from NKR and MHC alone (thin solid lines). Also shown are c(s) distributions from a series of experiments with NKR at 10 μM and MHC at 2.5 μM (long dash-dotted line), 10 μM (long dashed line), 20 μM (short dash-dotted line), and 40 μM (short dashed line). Panel C: Dependence of the weight-average s-value on MHC concentration with NKR at 43 μM (solid circle, single data point only) and at 10 μM (squares). The lines show the best-fit s_w isotherms for a 2:1 MHC per NKR dimer model (bold lines; solid line for NKR at 43 μM, dashed line for NKR at 10 μM), and for a 1:1 MHC per NKR dimer model (thin lines; short dash-dotted line for NKR at 43 μM, short dashed line for NKR at 10 μM). For more details on the experimental system and its analysis, see (51).

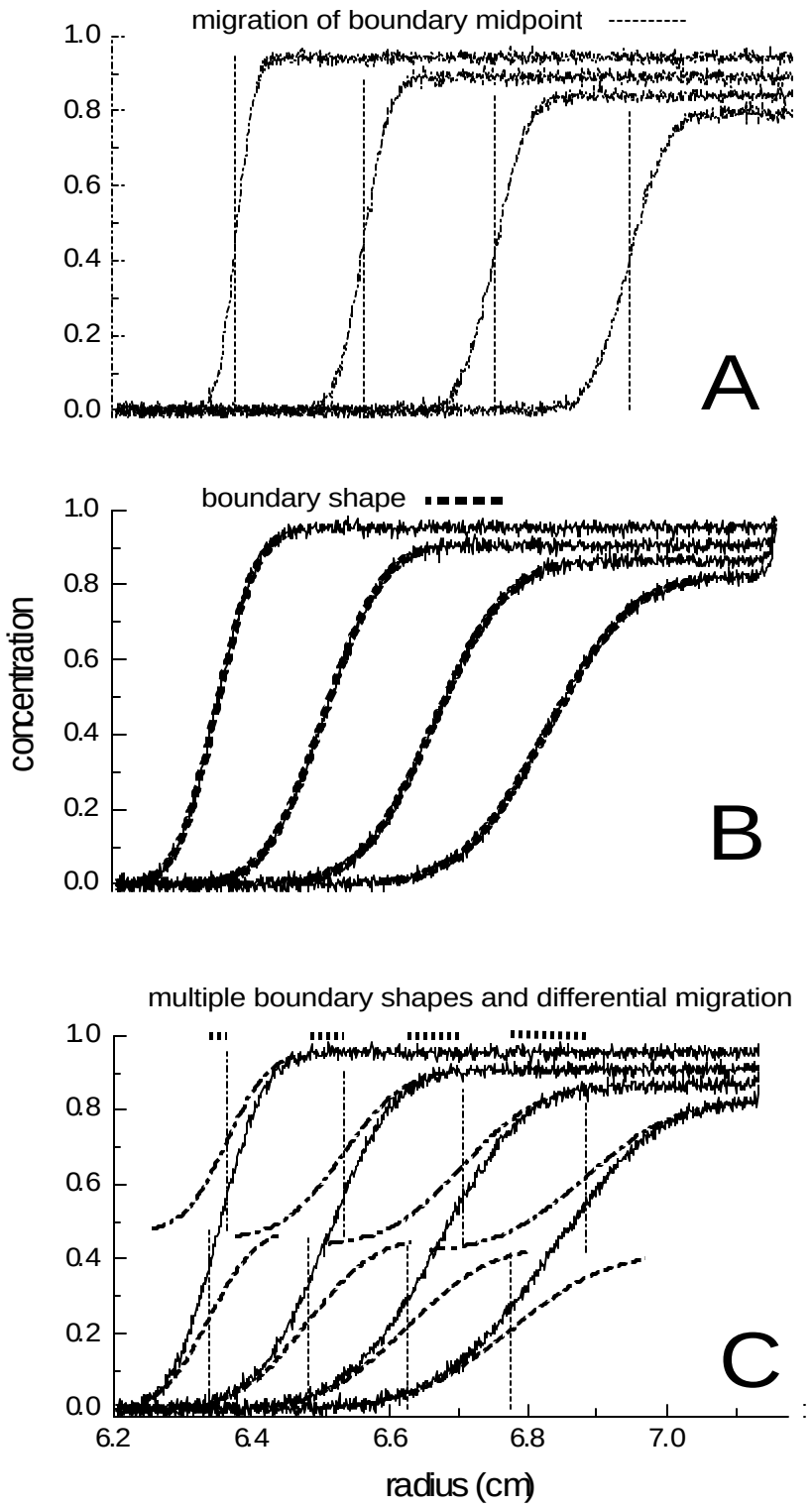
migration of boundary midpoint
1.0
0.8
0.6
0.4
0.2
0.0
A
boundary shape
concentration
B
multiple boundary shapes and differential migration
C
6.2
6.4
6.6
6.8
7.0
radius (cm)

Figure 1

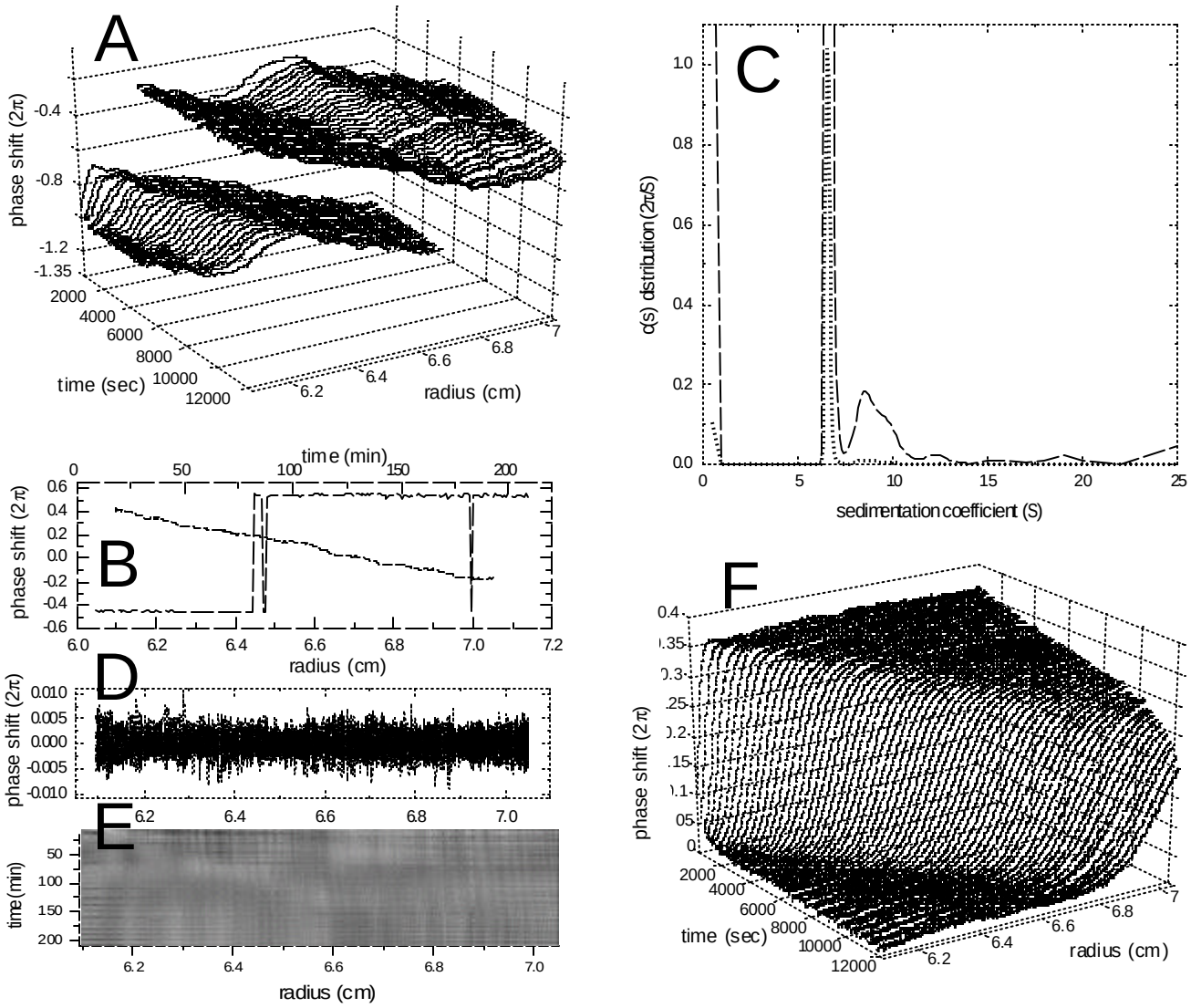
A
phase shift (2π)
-0.4
-0.8
-1.2
-1.35
2000
4000
6000
8000
10000
12000
time (sec)
6.2
6.4
6.6
6.8
7
radius (cm)
B
time (min)
0
50
100
150
200
phase shift (2π)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
6.0
6.2
6.4
6.6
6.8
7.0
7.2
radius (cm)
C
c(s) distribution (2πS)
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
sedimentation coefficient (S)
D
phase shift (2π)
0.010
0.005
0.000
-0.005
-0.010
6.2
6.4
6.6
6.8
7.0
E
time (min)
50
100
150
200
6.2
6.4
6.6
6.8
7.0
radius (cm)
F
phase shift (2π)
0.4
.35
.3
.25
.2
.15
.1
05
0
2000
4000
6000
8000
10000
12000
time (sec)
6.2
6.4
6.6
6.8
7
radius (cm)

Figure 2

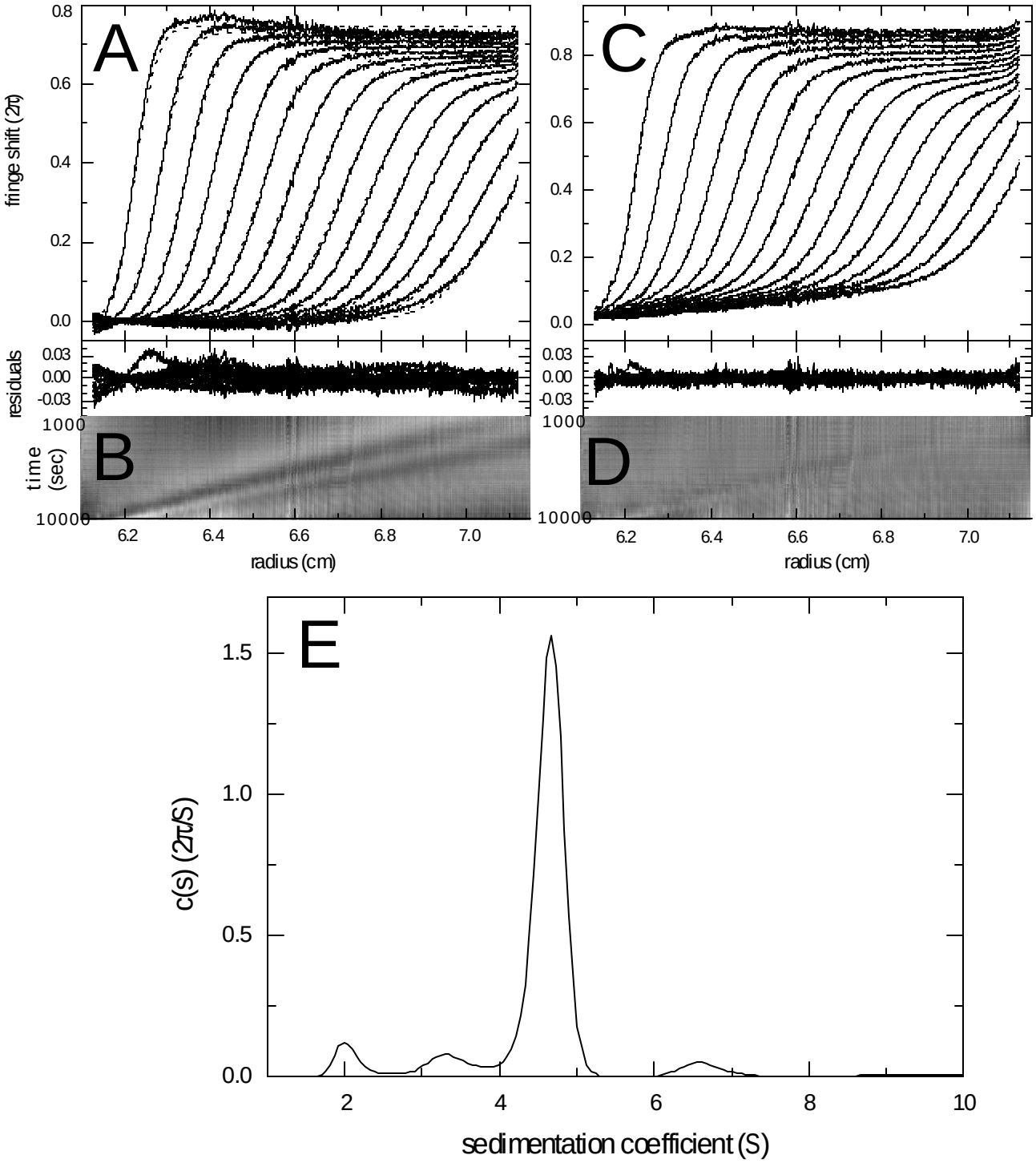
A
B
C
D
E
fringe shift (2π)
residuals
time (sec)
radius (cm)
radius (cm)
c(s) (2πS)
sedimentation coefficient (S)

Figure 3

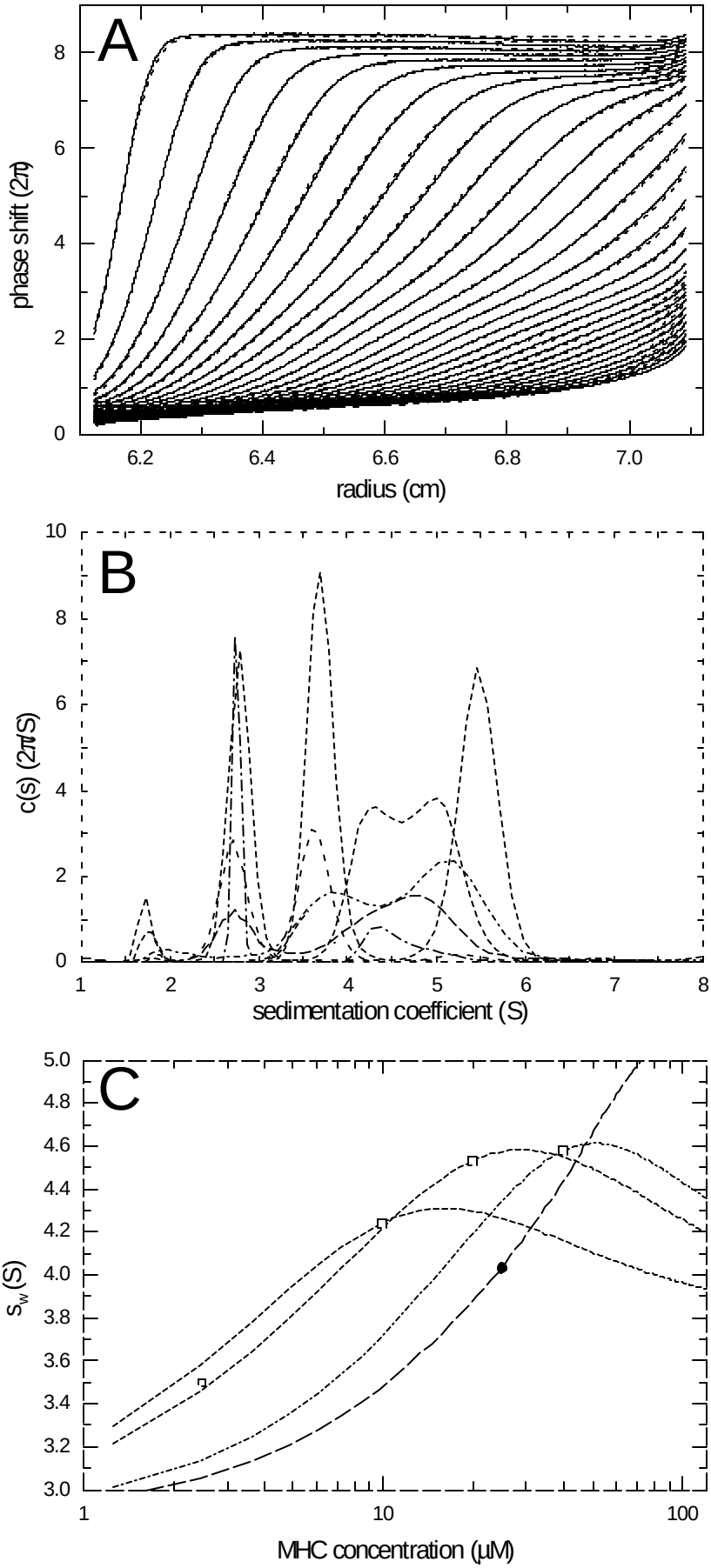
A
phase shift (2π)
radius (cm)
B
c(s) (2π/S)
sedimentation coefficient (S)
C
s_w (S)
MHC concentration (μM)

Figure 4