Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytical Biochemistry 354 (2006) 238–246
ANALYTICAL BIOCHEMISTRY
www.elsevier.com/locate/yabio

Improved methods for fitting sedimentation coefficient distributions derived by time-derivative techniques

John S. Philo *

Alliance Protein Laboratories, Thousand Oaks, CA 91360, USA

Received 8 January 2006
Available online 11 May 2006

Abstract

Time-derivative approaches to analyzing sedimentation velocity data have proven to be highly successful and have now been used routinely for more than a decade. For samples containing a small number of noninteracting species, the sedimentation coefficient distribution function, $g(s^*)$, traditionally has been fitted by Gaussian functions to derive the concentration, sedimentation coefficient, and diffusion coefficient of each species. However, the accuracy obtained by that approach is limited, even for noise-free data, and becomes even more compromised as more scans are included in the analysis to improve the signal/noise ratio (because the time span of the data becomes too large). Two new methods are described to correct for the effects of long time spans: one approach that uses a Taylor series expansion to correct the theoretical function and a second approach that creates theoretical $g(s^*)$ curves from Lamm equation models of the boundaries. With this second approach, the accuracy of the fitted parameters is approximately 0.1% and becomes essentially independent of the time span; therefore, it is possible to obtain much higher signal/noise when needed. This second approach is also compared with other current methods of analyzing sedimentation velocity data.

Keywords: Sedimentation velocity; Analytical ultracentrifugation; Time-derivative analysis; Sedimentation coefficient; Diffusion coefficient; Numerical methods; Least-squares fitting

The time-derivative or dc/dt method for sedimentation velocity analysis, pioneered by Walter Stafford at the Boston Biomedical Research Institute [1,2], has proven to be very useful and currently is used by a large fraction of laboratories performing sedimentation velocity experiments. One important advantage of this method is that it removes the time-invariant systematic noise in the raw scans (baseline distortions and window effects) algebraically in a manner that is model independent. A second important feature of the time-derivative method is that it is possible to fit individual peaks in the $g(s^*)$ distribution to Gaussian functions and thereby derive the concentration (from the peak area), the sedimentation coefficient (from the center position), and the diffusion coefficient (from the width of the Gaussian) for individual species [3,4]. With both the sedimentation and diffusion coefficients determined, one also can obtain the molecular mass from their ratio using the Svedberg equation. This can provide a simple means to answer questions about stoichiometry (is the primary species a monomer, dimer, or tetramer?) and homogeneity (is the $g(s^*)$ distribution consistent with a single species?).

Another significant advantage of the time-derivative approach is that it is rapid, easy to use, and intuitive, and it quickly provides estimates of the precision of the estimated sedimentation coefficients and molecular masses. That is undoubtedly a major reason for its continued use and popularity despite the emergence of other data analysis approaches for analyzing mixtures that can be superior in parameter accuracy and/or signal/noise ratio and that may be more fundamentally rigorous. One such alternate analysis approach is whole boundary modeling, where multiple raw scan files are globally fitted using solutions of the

* Fax: +1 805 388 7252.
E-mail address: jphilo@mailway.com.

0003-2697/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2006.04.053

Lamm equation, as pioneered by Holladay in 1980 [5]. A number of software packages implementing that approach are now available for desktop personal computers, including SVEDBERG by Philo (in 1994) [6], LAMM by Behlke and Ristau (in 1997) [7], ULTRASPIN by Demeler and Saber (in 1998) [8], SEDFIT by Schuck and coworkers (in 1998) [9], and SEDANAL by Stafford and Sherwood (in 2004) [10]. Many of these programs also now provide more complex models, including reversible associations and/or solution nonideality effects. The time-derivative approach is complementary to those methods, rather than necessarily directly competing with them, and often is used as a first-stage data analysis to suggest models and/or provide starting parameters for those other approaches. Similarly, there are also newer methods for deriving sedimentation coefficient distributions, including the least-squares $g^*(s)$ method [11], the $c(s)$ method [12], and the improved van Holde–Weischet method [13]; however, in the latter two of these methods, the diffusion information is removed and therefore so they cannot provide molecular mass information about multiple components.

Although the fitting of $g(s^*)$ distributions using Gaussians to obtain molecular properties certainly is quite useful, it still has some significant shortcomings. The purpose of this article is to propose and test some revised algorithms to address those problems. One issue is the limited accuracy of the derived molecular parameters. As discussed in some detail in a previous article [14], even for theoretically perfect data for a single species, there are errors of up to approximately 8% in the diffusion coefficient and mass for moderate-sized proteins (~150 kDa) and systematic errors of 2–10% in the apparent sedimentation coefficients for smaller proteins ($\lesssim$ 40 kDa). Such errors in apparent diffusion coefficient or mass are trivial if one needs to know only whether the protein is a monomer or a dimer, but they are quite significant for purposes such as assessing homogeneity of protein pharmaceuticals. Similarly, the errors in sedimentation coefficients can be sufficiently large to preclude use for purposes such as estimating axial ratios [15].

In practice, a more significant problem arises as more and more scans are used in the analysis to obtain sufficient signal/noise ratio, increasing the time span from the first to last scan used in the analysis. The longer time span makes the $\Delta c/\Delta t$ curve calculated from each pair of scans a much poorer approximation of the true derivative, dc/dt, and that produces broadening of the peaks. As discussed in detail previously [14], this effect often can lead to errors in diffusion coefficients for single species of 15% or more. These peak-broadening effects are particularly troubling when trying to resolve multiple species both because one needs the highest possible signal/noise to do this and because the degree of broadening will be much greater for the higher mass components. In previous work, it has been demonstrated that when trying to resolve properties for minor species such as covalent or irreversible oligomers, it can be quite advantageous to use prior knowledge to reduce the number of fitting parameters [6,16]. For example, we have found it to be useful to constrain the mass of the oligomer to be an appropriate integer multiple of the monomer mass. This approach, however, is also seriously compromised by the peak-broadening effects.

In this article, two new approaches to reduce these limitations of the current methods are described. First, an approach to reducing the effects of large time spans by calculating correction terms derived from a Taylor series approach is briefly described, and its limitations are discussed. Then a second approach is described based on directly simulating data for all scans used in the analysis. It is shown that this works much better than the Taylor series approach. This second approach can improve the accuracy of the derived hydrodynamic parameters by more than an order of magnitude and permit an improvement in signal/noise ratio of up to approximately 10-fold while maintaining an accuracy of approximately 0.1% for the hydrodynamic parameters.

Materials and methods

Sedimentation velocity data were obtained using absorbance scans at 280 nm in a Beckman Optima XL-A analytical centrifuge. Numerical simulations of sedimentation velocity experiments were done using the Claverie finite element method as described previously [14]. The calculation of dc/dt and $g(s^*)$ distributions was done using a new version of a Visual Basic .NET program named DCDT+ that implements the new algorithms described below.[1] Comparisons were done using version 1.13 of DCDT+, version 9.3b of SEDFIT, and version 4.1b of SEDPHAT [9]. Nonlinear least-squares fitting within DCDT+ employed a modified Gauss–Newton method, as described previously [6]. The fitting of Gaussians to least-squares $g^*(s)$, or ls–$g^*(s)$, distributions from SEDFIT was done using version 7.0 of ORIGIN (OriginLab, Northampton, MA, USA).

Results

In the standard approach to fitting $g(s^*)$ distributions, each species is fitted using a Gaussian function. The use of a Gaussian is based on the Faxén approximate solution to the Lamm equation [4]. The Faxén solution does not provide a highly accurate description of the concentration distribution in the cell, $c(r,t)$, and this is one reason why fitting to Gaussians does not always give accurate values for the diffusion coefficient. In practice, however, the number of scans and range of time needed to obtain sufficient signal/noise ratio will generally produce systematic errors from peak broadening that are significantly larger than those related to the fact that the peaks are not exactly Gaussian in form. Indeed, the use of time-derivative

[1] The program description and instructions for downloading can be found at www.jphilo.mailway.com/dcdt+.htm.

analysis has generally required a fairly severe trade-off between using more scans to improve signal/noise and a loss of accuracy due to peak broadening. Can this situation be improved?

Approach A: Using Taylor series to reduce the effects of large time spans (peak broadening)

Reducing the loss of accuracy at longer time spans first requires understanding the source of the errors. It is important to recall that the Stafford algorithm [2] involves subtracting scans in pairs, where for a total of $2N$ scans, scan 1 is paired with scan $N+1$, scan 2 is paired with scan $N+2$, and so forth. In general terms, the inaccuracies arise because as the time span between the scans in each pair grows longer (and hence the boundary movement increases), the approximation that $\Delta c/\Delta t \cong \mathrm{d}c/\mathrm{d}t$ becomes less and less accurate. Let us assume that the time interval between scans is a constant value Δt. Then for each scan pair, the algorithm calculates that at each radial position r,

$$\frac{\Delta c}{\Delta t}(r,t_i) = \frac{1}{N\Delta t}\left[c\left(r, t_i - \frac{N\Delta t}{2}\right) - c\left(r, t_i + \frac{N\Delta t}{2}\right)\right], \tag{1}$$

where t_i is the mean time for the ith pair of scans. If we do a Taylor series expansion to evaluate the function $c(r,t)$ relative to its true value at the time t_i, this becomes

$$\frac{\Delta c}{\Delta t}(r,t_i) = \frac{1}{N\Delta t}\begin{bmatrix} c(r,t_i) - \frac{N\Delta t}{2}\frac{\partial c}{\partial t}(r,t_i) + \frac{1}{2}\left(\frac{N\Delta t}{2}\right)^2\frac{\partial^2 c}{\partial t^2}(r,t_i) + \frac{1}{6}\left(\frac{N\Delta t}{2}\right)^3\frac{\partial^3 c}{\partial t^3}(r,t_i) + \cdots \\ -c(r,t_i) - \frac{N\Delta t}{2}\frac{\partial c}{\partial t}(r,t_i) - \frac{1}{2}\left(\frac{N\Delta t}{2}\right)^2\frac{\partial^2 c}{\partial t^2}(r,t_i) + \frac{1}{6}\left(\frac{N\Delta t}{2}\right)^3\frac{\partial^3 c}{\partial t^3}(r,t_i) - \cdots \end{bmatrix}, \tag{2}$$

which simplifies to

$$\frac{\Delta c}{\Delta t}(r,t_i) = \frac{\partial c}{\partial t}(r,t_i) + \frac{1}{6}\left(\frac{N\Delta t}{2}\right)^2\frac{\partial^3 c}{\partial t^3}(r,t_i) + \frac{1}{120}\left(\frac{N\Delta t}{2}\right)^4\frac{\partial^5 c}{\partial t^5}(r,t_i). \tag{3}$$

This expansion makes it apparent that the errors are related to the higher derivatives of $\partial c/\partial t$ and that the first-order errors are proportional to the square of the time span $N\Delta t$. Furthermore, this suggests that we should be able to explicitly correct for these errors during fitting by using Eq. (3) in calculating $\mathrm{d}c/\mathrm{d}t$ and $g(s^*)$ for each species.

There is actually one additional significant (but smaller) source of error related to large time spans, however, that arises when we average together the data from each pair of scans to calculate an overall average data set for $\mathrm{d}c/\mathrm{d}t$ or $g(s^*)$. This averaging is done not at constant values of radius but rather at constant values of $s^* = (1/\omega^2 t)\ln(r/r_m)$. Thus, the average experimental data are calculated from

$$\left.\frac{\Delta c}{\Delta t}\right|_{\mathrm{avg}}(s^*,t_o) = \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta c}{\Delta t}(r^*,t_i), \tag{4}$$

where t_o is the overall mean time for this set of scans and r^* is the radial position corresponding to the value of s^*. If we again use a Taylor series expansion around the mean time t_o, Eq. (4) becomes

$$\left.\frac{\Delta c}{\Delta t}\right|_{\mathrm{avg}}(s^*,t_o) = \frac{\Delta c}{\Delta t}(s^*,t_o) + \frac{(\Delta t/2)^2}{2N}\frac{\partial^3 c}{\partial t^3}(s^*,t_o)\sum_{i=1}^{N}(2i-N-1)^2, \tag{5}$$

where Eq. (4) is used to calculate $\Delta c/\Delta t$ at t_o and the $\partial^3 c/\partial t^3$ derivative in Eq. (5) must be evaluated at constant s^* (rather than at constant r as in Eq. (4)).

Code was written to implement Eqs. (4) and (5) using the analytical time derivative of the modified Fujita–MacCosham function [16] to evaluate $\partial c/\partial t$, with the higher derivatives being calculated numerically. The resulting $\mathrm{d}c/\mathrm{d}t$ versus s^* results were either used directly to fit the experimental average $\mathrm{d}c/dt$ data or fed into the usual iterative procedure within the Stafford algorithm to calculate the theoretical $g(s^*)$ distributions used in fitting.

Tests at large time spans

Simulated data sets covering large time spans were constructed to evaluate how well this Taylor series correction is able to derive the correct molecular parameters and loading concentration despite the inevitable broadening of the peaks. Stafford proposed a "rule of thumb" (www.bbri.org/dcdt/Rule.pdf) to calculate the maximum time span between the first and last scans used in the analysis, $\Delta t_{\max}$, before substantial broadening occurs. This rule has been revised over time, but currently the value to be used when the data are being fitted to derive diffusion coefficients or masses is given by

$$\Delta t_{\max} = \frac{80\cdot t}{\sqrt{M}\cdot(\mathrm{RPM}/1000)}, \tag{6}$$

where t is the arithmetical mean time of the first and last scans (in s), M is the mass (in kDa), and RPM is the rotor speed. This formula assumes that the boundary is near the midpoint of the cell, a partial specific volume, $\bar{\nu}$, of 0.725 ml/g, and a solvent density, ρ, of 1 g/ml.

Because this maximum time span is directly related to the sample molecular mass, for any given set of scans, this rule can also be used to calculate the maximum molecular mass that can be present in the sample without significant broadening, $M_{\max}$:

$$M_{\max} = \left(\frac{80\cdot t}{\Delta t\cdot(\mathrm{RPM}/1000)}\right)^2\frac{0.275}{(1-\bar{\nu}\cdot\rho)}. \tag{7}$$

This Taylor series correction has been tested against simulated data sets corresponding approximately to ovalbumin, immunoglobulin G (IGG),[2] or an approximately 75 kDa, $s = 5$ S, $D = 6$ Fick (F) species, while systematically varying the number of scans used in the analysis. The behavior for the different masses is similar, so only the results for the

[2] *Abbreviations used:* IGG, immunoglobulin G; F, Fick (where $1\ \mathrm{F} = 10^{-7}\ \mathrm{cm^2\ s^{-1}}$); TRAP, *trp* RNA-binding attenuation protein.

ovalbumin simulations ($s = 3.55$ S, $D = 7.89$ F) are presented. Fig. 1 summarizes the relative accuracy obtained for s, D, M, and the loading concentration, c_o, from fits to either the dc/dt or $g(s^*)$ data using the current algorithm or with this Taylor series correction approach. These graphs show that the new approach does indeed substantially improve the accuracy of the fitted parameters. Over the range up to $M/M_{max} = 8$, where the curves are fairly linear, the improvement in accuracy is approximately 25-fold for s, 50-fold for D, 40-fold for M, and 20-fold for c_o. For values of M/M_{max} greater than approximately 10, the Taylor series correction becomes less effective and the errors grow significantly more rapidly; nonetheless, an accuracy for D and M of approximately 3% can be maintained out to approximately $M/M_{max} = 20$, whereas holding that level of precision with the standard algorithm requires $M/M_{max} < 0.8$. Fig. 2A illustrates the range of scans corresponding to this approximately 3% level of precision with and without the Taylor series correction. Fig. 2B shows the large variation among the dc/dt curves that arises over the long time span usable with the Taylor series approach, including large changes in amplitude and peak width as well as small shifts of the peak position.

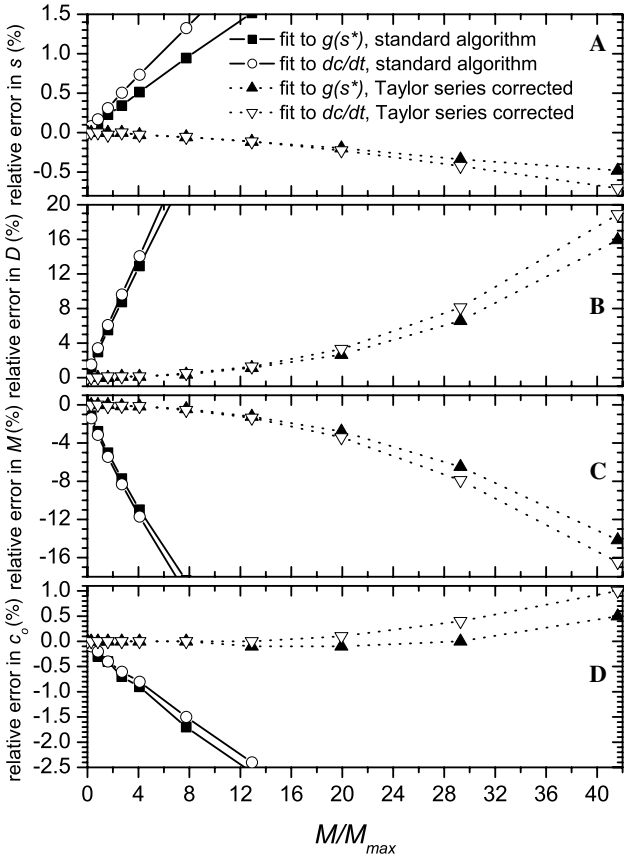

Fig. 1. Variation in parameter errors with the time span (number of scans) used in the analysis (plotted vs. the ratio of true mass to M_{max} as calculated using Eq. (7)). The errors are calculated relative to the values obtained using simulations of very rapid scans (no peak broadening). These results are from noise-free simulations for ovalbumin (~42 kDa) at 60,000 rpm with scans recorded every 210 s. Analyses were done starting with 4 scans chosen when the boundary is near the midpoint of the cell and then adding equal numbers of earlier and later scans up to a total of 32 scans. Panel A shows the percentage error in sedimentation coefficient from fitting to either $g(s^*)$ (closed symbols) or dc/dt (open symbols) data using either the standard algorithms (solid connecting lines) or the new algorithms with Taylor series correction (dotted lines). Panels B, C, and D give the relative errors in diffusion coefficient, mass, and loading concentration, respectively.

Holding accuracy over much larger time spans should permit significant increases in signal/noise ratio without loss of accuracy. Assuming a constant scan rate, the signal/noise should increase approximately as $N^{3/2}$ or equivalently as $(M/M_{max})^{3/4}$, so that by allowing M/M_{max} to increase 20- to 50-fold while maintaining equivalent accuracy, the Taylor series correction theoretically should permit increases in signal/noise by factors of 9–19. Another benefit of a larger time span is that this also increases the upper limit of sedimentation coefficients covered in the analysis (allowing coverage of more species), even when the midpoint of the range remains fixed.

However, it should be noted that accuracy of the fitted parameters is not the only criterion for a "good" fitting procedure. Particularly for multispecies fits, it is also quite

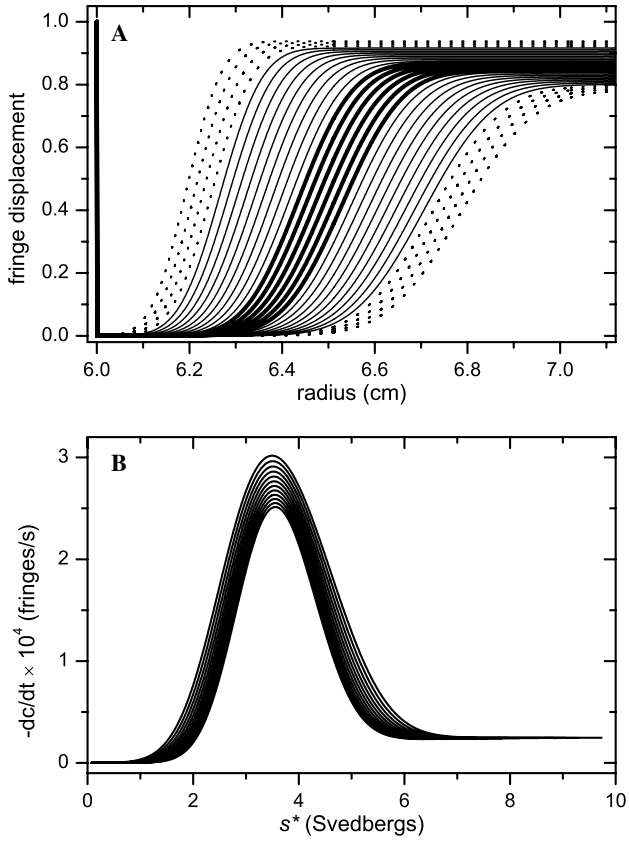

Fig. 2. (A) Simulated velocity data for ovalbumin. The six heavy lines show the maximum range of scans that allows 3% accuracy of D or M values with the standard algorithms (somewhat under the maximum to comply with the rule of thumb). The 24 solid lines show the range that still allows 3% accuracy using the Taylor series correction method. The additional 8 dotted curves show the full range covered in Fig. 1. (B) dc/dt curves from 24 scans.

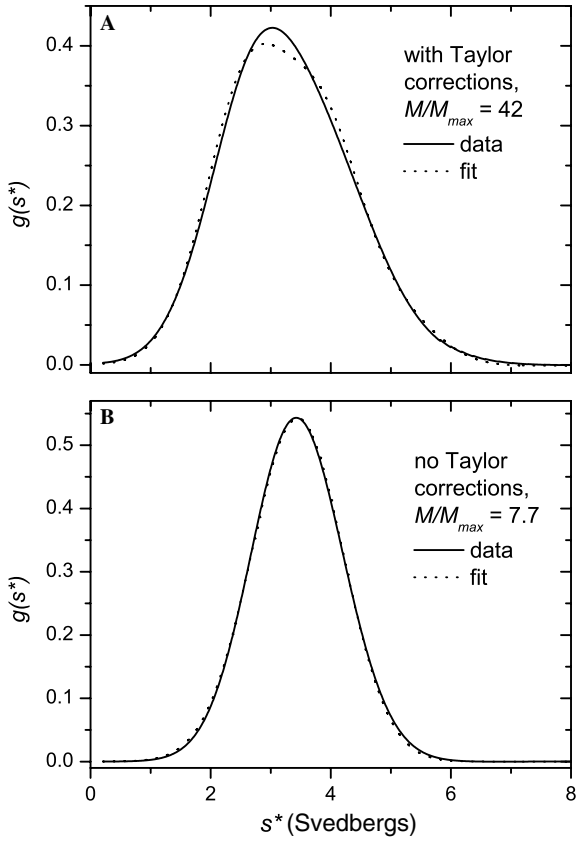

Fig. 3. (A) The $g(s^*)$ data (solid line) from 32 scans of the ovalbumin simulation (the full range shown in Fig. 2A), corresponding to $M/M_{max} = 42$ in Fig. 1, are overlaid with the fit to those data using the Taylor series corrections (dotted line). Note that although the fitted parameters are still fairly close to the correct values (e.g., the relative error in D is 16%), the shape of the fitted curve deviates significantly from that of the experimental one. (B) Data and fit for 16 scans but without the Taylor correction. Although the error in fitted parameters is actually higher than in panel A (e.g., the relative error in D is 21%), the shape of the fitted curve is quite similar to that of the actual $g(s^*)$ distribution.

important that the theoretical $g(s^*)$ curves accurately match the shape of the experimental ones. Unfortunately, in the regime above $M/M_{max} = 20$, even though the errors in D and M may be acceptable, the shape of the fitted curves begins to deviate significantly from that of the experimental ones, as illustrated in Fig. 3. Without a good match in shape between experimental and fitted curves for single species, the resolution and reliability of multispecies fits would be severely compromised.

Approach B: Extending the time span through direct simulation of the corresponding whole boundaries

Although this Taylor series correction approach clearly provides some significant improvements, it would be desirable to reduce the effects of large time spans even further, particularly with regard to obtaining theoretical curves that accurately match the shape of the experimental ones. To do this, it seemed best to directly mimic the data processing applied to the raw data. That is, the idea is to generate theoretical scans (signal vs. radius) corresponding to the actual times of the experimental scans and then to replicate the data processing used on the raw data to obtain a corresponding theoretical curve for either $g(s^*)$ or the average dc/dt data. To obtain high-quality theoretical sedimentation boundaries, a very accurate approximate solution to the Lamm equation described by Behlke and Ristau [17] was chosen. The actual function used is the first three terms in Eq. 28 of that article, that is, the terms that describe the moving boundary. This function was shown to give errors for sedimentation coefficient of less than 0.2%, errors for D or M of less than 0.5%, errors for proteins of 2 kDa or greater, and errors of only 0.1% or less for all parameters for proteins larger than 10 kDa (the range where dc/dt analysis usually is applied) [17].

This approach does indeed result in significantly better accuracy for fitted parameters when using long time spans than does the Taylor series correction approach. Indeed, with this approach the returned parameters are nearly independent of the time span. For example, when applied to the ovalbumin simulations used to generate Fig. 1, even at the longest time spans covered there ($M/M_{max} = 42$), the absolute errors in s, D, M, and c_o are only -0.01%, 0.06%, -0.06%, and 0.01%, respectively, for fitting to either the dc/dt or $g(s^*)$ data. That is, with this approach, the errors due to all sources, over any group of scans within the full range shown in Fig. 2, are completely negligible compared with the precision of any real experiment.

Using this algorithm, it is actually possible to use the full span from the time the meniscus region is just cleared until the plateau region is about to disappear (Fig. 4A). Such a large time span leads to rather extreme broadening of the dc/dt curves and saturation of their amplitude (Fig. 4B) and, consequently, a severely distorted $g(s^*)$ distribution (Fig. 4C). Nonetheless, the fitted parameters maintain excellent accuracy, with absolute errors for all parameters of less than 0.06%. Furthermore, the shape of the theoretical $g(s^*)$ curve (or of the dc/dt curve [not shown]) is an excellent match for the experimental data, with maximum residuals of less than 0.08% of the peak amplitude.

Another approach that can be used to calculate $g(s)$ distributions using fairly broad time spans is the ls–$g^*(s)$ method developed by Schuck and coworkers and implemented in SEDFIT. It has been shown that these ls–$g^*(s)$ distributions can also be fitted as Gaussians, at least when the time span is narrow [11]. How does this ls–$g^*(s)$ approach compare with approach B? Table 1 summarizes results from both methods when applied to 6, 12, 18, or 24 scans from the simulation shown in Fig. 2. The range up to 24 scans corresponds approximately to the recommended maximum[3] for ls–$g^*(s)$ of two to three times the six to eight scans that normally would be used with the dc/dt method. These results show that the ls–$g^*(s)$ distribution systematically underestimates the true

[3] See www.analyticalultracentrifugation.com/lsgofs_distribution.htm.

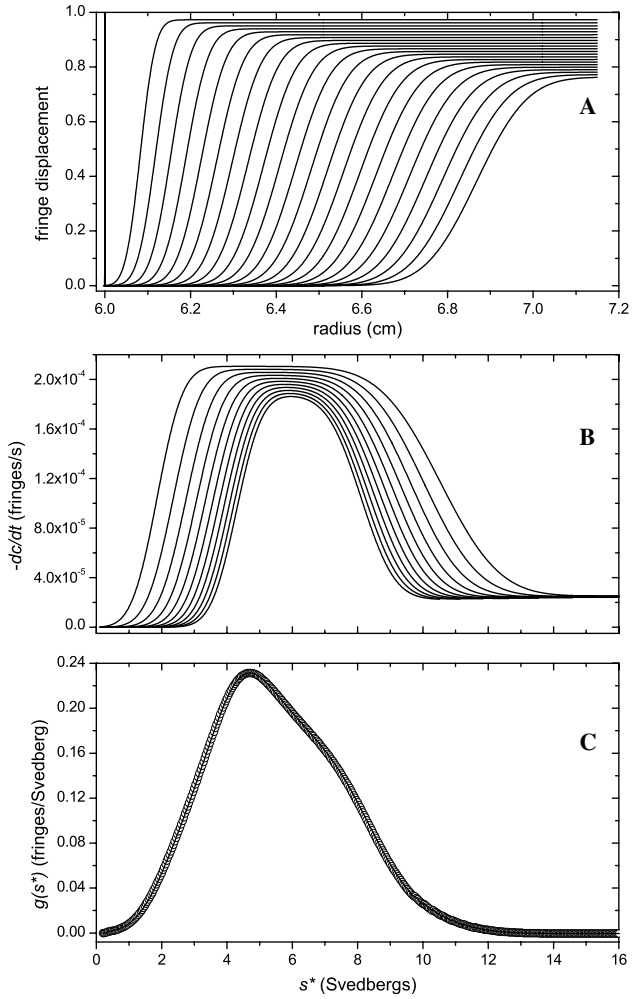

Fig. 4. Example of the very wide time span that can be used with the whole boundary simulation algorithm (approach B) while still maintaining excellent accuracy. (A) Simulated velocity data for IGG ($s = 6.2$ S, $D = 3.9$ F, $c_o = 1$ fringe), with every other scan shown. (B) Corresponding dc/dt curves, with every other curve shown. (C) Corresponding $g(s^*)$ data (circles) and fitted curve (line), with every fifth point shown. The returned values from the fit are $s = 6.1998$ S, $D = 3.913$ F, and $c_o = 0.99995$ fringe.

sedimentation coefficient by up to 1.5% when only a few scans are used and produces a 2–5% underestimate of the diffusion coefficient throughout this range of scans.[4] The loading concentration is overestimated by 1–2%. The accuracy of approach B is better by one to two orders of magnitude with a worst case error of 0.1% in any parameter. Perhaps more significant is the fact that for 24 scans the ls–$g^*(s)$ distribution falsely implies that the sample is heterogeneous. As can be seen in Fig. 5, the ls–$g^*(s)$ distribution shows a false half-peak extending from 0.8 S down to 0.1 S (the lower limit allowed in the analysis). Moreover, even if that feature is recognized as a data analysis artifact and ignored, there is still a distinct shoulder on the left side

[4] Note that the errors in D and M will be significantly larger (>10%) if the calculations are done using the harmonic mean time (the standard value used in the dc/dt algorithms) rather than the simple mean that was employed here.

Table 1
Comparison of results from fitting the least-squares $g^*(s)$ distributions from SEDFIT to a Gaussian versus fitting $g(s^*)$ data using approach B for a different number of scans from the ovalbumin simulation shown in Fig. 2

Number of scans used in the analysis	Results from Gaussian fits to ls–$g^*(s)$	Results from fitting $g(s^*)$ using approach B
6	$s = 3.497$ S [−1.5%][a] D[b] $= 7.472$ F [−4.9%] M/M_o[c] $= 1.0359$ [+3.6%] $c_o = 1.0084$ [+0.8%]	$s = 3.549$ S [−0.0%] $D = 7.900$ F [+0.1%] $M/M_o = 0.9986$ [−0.1%] $c_o = 1.0004$ [+0.0%]
12	$s = 3.509$ S [−1.1%] $D = 7.656$ F [−3.0%] $M/M_o = 1.0187$ [+1.8%] $c_o = 1.0198$ [+2.0%]	$s = 3.549$ S [−0.0%] $D = 7.899$ F [+0.1%] $M/M_o = 0.9987$ [−0.1%] $c_o = 1.0003$ [+0.0%]
18	$s = 3.525$ S [−0.7%] $D = 7.734$ F [−2.0%] $M/M_o = 1.0130$ [+1.3%] $c_o = 1.0218$ [+2.2%]	$s = 3.549$ S [−0.0%] $D = 7.898$ F [+0.1%] $M/M_o = 0.9989$ [−0.1%] $c_o = 1.0002$ [+0.0%]
24	$s = 3.538$ S [−0.3%] $D = 7.652$ F [−3.0%] $M/M_o = 1.0277$ [+2.8%] $c_o = 1.0197$ [+2.0%]	$s = 3.550$ S [−0.0%] $D = 7.894$ F [+0.1%] $M/M_o = 0.9993$ [−0.1%] $c_o = 1.0001$ [+0.0%]

Note. The group of scans chosen was symmetric around the time the boundary reached the middle of the cell (keeping the mean time for the group fixed as the number of scans increased). The true values from the simulation were $s = 3.55$ S, $D = 7.89$ F, and $c_o = 1$ fringe.

[a] Values in square brackets are percentage errors (rounded to nearest tenth).

[b] Diffusion coefficients were calculated from the width of the fitted Gaussian using the mean $\omega^2 t$ value for the scans.

[c] Apparent mass (from the s/D ratio) relative to the true value.

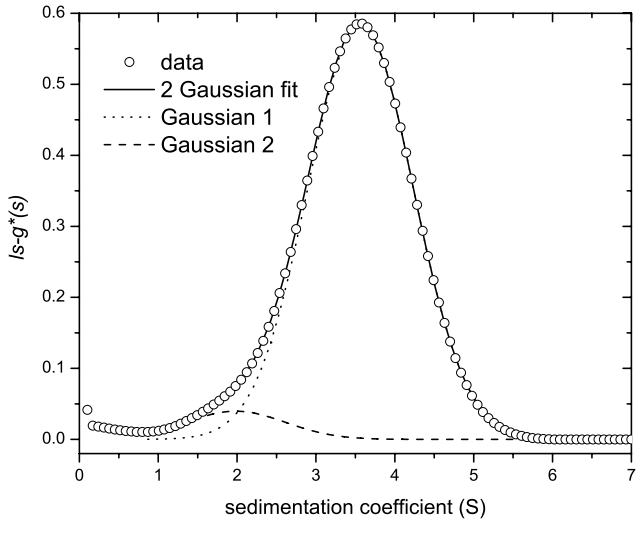

Fig. 5. Least-squares $g^*(s)$ analysis of 24 scans from the simulation for pure ovalbumin shown in Fig. 2, which incorrectly indicates sample heterogeneity. The open circles show the ls–$g^*(s)$ distribution, and the solid line is the fit of those data as two Gaussians (shown as dashed or dotted lines).

of the main peak, one that can actually be fitted as second Gaussian component centered at 1.96 S, as shown in the figure.

One of the useful properties of fitting individual peaks in $g(s^*)$ distributions is that it is possible to limit the range of sedimentation coefficients for the fit to exclude species that are not of interest. It is difficult to impose similar limits in whole boundary modeling methods because the fitting region for such methods normally spans a radial region within the cell and the radial values that correspond to any particular sedimentation coefficient are, of course, different for each scan used for whole boundary modeling. One real experiment that illustrates the utility of this approach was analysis of the *trp* RNA-binding attenuation protein (TRAP), a protein known to form tightly associated 11-mer ring structures both in crystals [18] and in solution [19]. The sample of TRAP from *Bacillus stearothermophilus* used in this experiment (a demonstration run at a workshop) also contained aggregates or larger structures sedimenting at approximately 8 to 14 S. By limiting the fit to the central portion of the major peak (as shown in Fig. 6) the influence of any minor components on the derived properties for the major component will be minimized. This fit returns a molecular mass for this complex of 90.8 kDa [95% confidence, 88.0–93.6 kDa], which corresponds to 11.0 [10.7–11.4] times the 8.242-kDa monomer sequence mass, exactly as expected.

An alternative approach that can be used for analyzing samples containing species that are not of interest is the so-called hybrid discrete/continuous model in SEDPHAT, where the species not of interest are modeled as a continuous distribution. When that model was applied to this TRAP experiment, the mass returned for the complex was 86.1 [84.7–87.0] kDa, corresponding to a stoichiometry of 10.44 [10.28–10.55]. Thus, that approach suggests that the complex is 10-mer rather than 11-mer, although 11-mer is not ruled out. Furthermore, when using that approach in SEDPHAT the computation of the confidence interval required 8.5 h, even though only 100 rounds of Monte Carlo simulation were used rather than the recommended 1000 rounds, compared with only approximately 1 s using approach B.

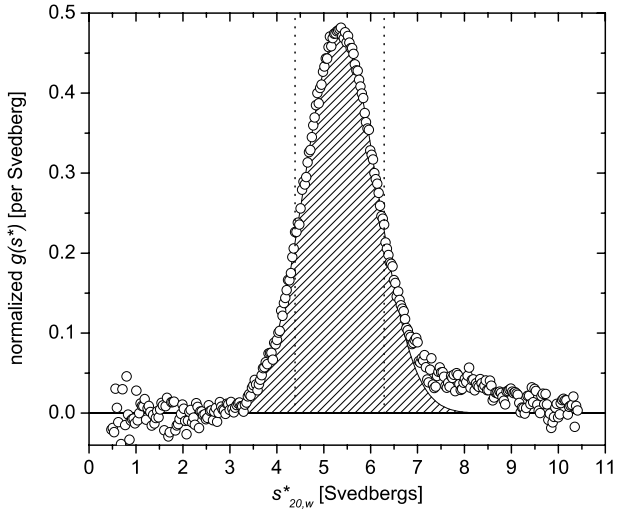

Fig. 6. Analysis of the stoichiometry of the major component of a sample of the TRAP from *B. stearothermophilus*. Because this sample contains aggregates or larger assemblies (this is evident from the tail on the right-hand side of the main peak), the fitting was limited to sedimentation coefficients corresponding to the central portion of the main peak, as indicated by the vertical dotted lines. The fitted major component is shown as the cross-hatched area. That fit shows that this major species is an 11-mer, as expected [18].

The ability to use more scans without losing accuracy can be quite helpful, particularly for detecting minor components or when the signal/noise ratio is low. For example, simulations were done for an IGG sample containing 5% of a trimer aggregate run at 45,000 rpm using absorbance optics (1 OD loading concentration). At the fastest scan rate possible when running three samples simultaneously, even using only four scans (the minimum) to compute $g(s^*)$ results in significant broadening of the trimer peak (and low signal/noise). Thus, when using the conventional algorithms, and even by fitting to the dc/dt data (which are more accurate [14]), as shown in Table 2, it is likely that this aggregate would be mistakenly identified as a dimer rather than as a trimer. With this new algorithm, however, the minor component can be consistently assigned as a trimer, and this can be done with increasing statistical confidence as the number of scans is increased. If the full data range of this same simulation is fitted using the $c(s)$ method in SEDFIT, the proportions and sedimentation coefficients

Table 2
Fits of dc/dt data from simulations for an IGG sample (1 AU loading concentration) containing 5% of a trimeric aggregate ($s = 10.54$ S, $D = 2.21$ F)

Number of scans used in the analysis	Results for minor component from standard analysis	Results for minor component using approach B
4	$D = 2.61$ F [2.06–3.19][a] M/M_o[b] $= 2.56$ [2.03–3.15] $s = 10.48$ S [10.37–10.59] 5.05%[c] [4.62–5.48]	$D = 2.31$ F [1.68–3.22] $M/M_o = 2.83$ [2.03–3.89] $s = 10.48$ S [10.36–10.59] 5.02% [4.43–5.62]
6	$D = 3.09$ F [2.69–3.51] $M/M_o = 2.27$ [1.97–2.58] $s = 10.53$ S [10.46–10.60] 4.99% [4.72–5.26]	$D = 2.39$ F [1.92–2.98] $M/M_o = 2.77$ [2.23–3.47] $s = 10.52$ S [10.45–10.60] 4.97% [4.62–5.32]
8	$D = 3.80$ F [3.49–4.13] $M/M_o = 1.92$ [1.76–2.08] $s = 10.55$ S [10.50–10.60] 5.02% [4.84–5.20]	$D = 2.44$ F [2.05–2.87] $M/M_o = 2.72$ [2.31–3.23] $s = 10.53$ S [10.48–10.59] 4.96% [4.74–5.19]
12	ND	$D = 2.25$ F [1.88–2.65] $M/M_o = 2.96$ [2.51–3.54] $s = 10.54$ S [10.50–10.58] 4.87% [4.73–5.01]

Note. The scan interval was 210 s (roughly the fastest possible scan rate when running three samples). To give a realistic noise level, random noise of 0.005 OD rms was added to the simulated data.

[a] Values in square brackets are 95% confidence intervals.

[b] Apparent mass of the minor component relative to that obtained for the major component in this analysis.

[c] Fraction of minor component (apparent loading concentration relative to total for major + minor component).

for the major and minor components are returned with high accuracy. However, if that $c(s)$ distribution is transformed to a $c(M)$ distribution, or if the $c(M)$ distribution is fitted directly, the mass of the minor component (the trimer) is returned as 2.22 times that of the major component. Thus, if either of those approaches is used, the trimer could easily be mistaken as a dimer. The discrete species model in SEDPHAT was also applied to this simulation. That approach returned an apparent mass for the minor component of 2.38 times that of the major component [95% confidence, 2.37–2.39] and thus also failed to correctly identify the minor component as a trimer rather than a dimer. Moreover, obtaining the best fit using SEDPHAT required 145 s versus only 3 s using DCDT+. Calculation of the confidence intervals took 7 s in DCDT+; doing this using only 100 Monte Carlo rounds in SEDPHAT (rather than the recommended 1000 rounds) took 950 s.

This antibody mixture example also illustrates another important point. Although approach B does ensure that the returned hydrodynamic parameters are accurate independent of the number of scans, N, it cannot counteract the peak broadening that still occurs. Because of that peak broadening, the precision of the D or M values improves only quite slowly as N increases once significant peak broadening begins (much more slowly than the theoretical $N^{3/2}$ dependence for the signal/noise ratio of the $g(s^*)$ or dc/dt curves).

Approach B has also been tested in real experiments for samples at very low concentrations. Aliquots of a homogeneous monoclonal antibody were run at concentrations of approximately 35, 2.3, and 1.2 μg/ml with scanning at 230 nm. With the standard algorithms, even using only 6 scans will violate the rule of thumb (Eq. (6)), and for the lowest concentration sample (0.012 total absorbance, <0.5 μg total protein) the signal/noise ratio with 6 scans only barely permits discerning that there is a peak around 6 S (Fig. 7A). Using approach B with 20 scans, however, we obtain fivefold better signal/noise (Fig. 7B) and by fitting obtain a sedimentation coefficient of 6.37 [6.29–6.46] S and a mass of 170 [124–233] kDa, consistent with the expected value of approximately 150 kDa for a monomer. The full time span of this same experiment was also analyzed using the discrete species model in SEDFIT. That approach gave similar best fit parameters of 6.32 S and 135 kDa. However, the 95% confidence interval returned for the molecular mass by SEDFIT was 28–1369 kDa; thus, that method was unable to confirm that this sample is indeed a monomer.

Perhaps more significant, an experiment was done at 1.2 μg/ml using a degraded antibody sample that contains approximately 20% total minor components at approximately 4, 5, and 9 S, that is, species that will not be well resolved from antibody monomer. The $c(s)$ distribution from SEDFIT for this sample, shown in Fig. 6C, does not resolve those minor components and gives only a single broad peak. The standard time-derivative analysis for this degraded sample using 6 or 8 scans does imply a rather low

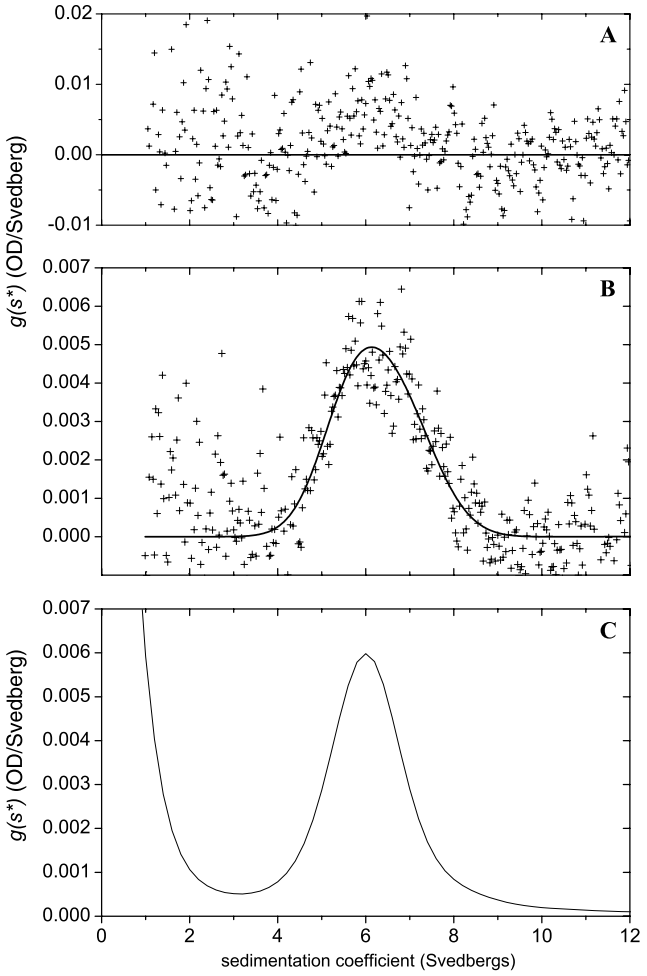

Fig. 7. Application of approach B and $c(s)$ analysis to monoclonal antibody samples at a very low concentration (1.2 μg/ml). (A) The $g(s^*)$ distribution for a highly homogeneous sample obtained using 6 scans, the maximum allowed by the rule of thumb. (B) The distribution for this same sample obtained using 20 scans along with the theoretical fit as a single species using approach B. (C) The $c(s)$ distribution for a degraded sample that contains approximately 20% total minor components at approximately 4, 5, and 9 S.

apparent mass of 60–80 kDa, suggesting that the boundary is broader than expected due to some sort of heterogeneity. However, given the low signal/noise ratio and the known systematic shifts to lower mass estimates with increased time span when using standard time-derivative analysis, the true heterogeneity in this material would likely be missed. In contrast, using approach B, we can be more than 99% confident that the apparent mass is too low to be consistent with the known monomer mass and, therefore, that some other species must be present.

Discussion

Although the Taylor series approach to minimizing errors associated with long time spans works fairly well, approach B is significantly better at longer time spans and produces theoretical curves that match the experimental data more accurately. The primary drawback to

approach B is that theoretical boundaries must be simulated for every scan and every data point in the distribution; thus, it becomes fairly computationally intensive when the number of scans is high, particularly for interference data. However, this computational burden certainly is not intractable. For example, the current implementation requires approximately 2 s per iteration (~12 s for convergence) for a two-species fit to $g(s^*)$ derived from 50 interference scans using a 2-GHz Pentium IV. Thus, approach B seems much superior to approach A overall; consequently, only approach B will be implemented in future program releases.

One aspect of using longer time spans that has not yet been discussed is how that affects the estimated errors on the dc/dt and $g(s^*)$ data and, hence, whether to use these error estimates as weighting values for the fits. The error estimates for individual data points are derived from the variations between the dc/dt curves from all of the scan pairs, a procedure that works well when the dominant noise is random noise from the optical systems. However, when the time span grows large, there are large systematic variations among the dc/dt curves, as is evident in Figs. 2B and 4B. These systematic variations can lead to gross overestimates of the error bars on individual points (often worst near the top of a peak in the distribution) and, consequently, false underweighting of such points during a weighted fit. Therefore, in the current work, nonweighted fits have been used; indeed, we have found that the results are considerably less accurate if weighted fits are used.

An alternative approach to weighting fits to $g(s^*)$ distributions is to calculate theoretical weights based on a reasonable assumption that the noise in the dc/dt data is uniform with sedimentation coefficient. The transform from dc/dt to $g(s^*)$ involves division by the sedimentation coefficient. These theoretical weights take this into account and assign low weights to the data at low sedimentation coefficients, which are indeed always significantly noisier. This approach makes theoretical sense and seems to work well, and it too will be provided in future program releases.

References

[1] W.F. Stafford III, Boundary analysis in sedimentation transport experiments: a procedure for obtaining sedimentation coefficient distributions using the time derivative of the concentration profile, Anal. Biochem. 203 (1992) 295–301.

[2] W.F. Stafford III, Boundary analysis in sedimentation velocity experiments, Methods Enzymol. 240 (1994) 478–501.

[3] W.F. Stafford III, Rapid molecular-weight determination by sedimentation velocity analysis, Biophys. J. 70 (1996) MP452.

[4] W.F. Stafford III, Sedimentation velocity spins a new weave for an old fabric, Curr. Opin. Biotechnol. 8 (1997) 14–24.

[5] L.A. Holladay, Simultaneous rapid estimation of sedimentation coefficient and molecular weight, Biophys. Chem. 11 (1980) 303–308.

[6] J.S. Philo, Measuring sedimentation, diffusion, and molecular weights of small molecules by direct fitting of sedimentation velocity concentration profiles, in: T.M. Schuster, T.M. Laue (Eds.), Modern Analytical Ultracentrifugation, Birkhauser, Boston, 1994, pp. 156–170.

[7] J. Behlke, O. Ristau, Molecular mass determination by sedimentation velocity experiments and direct fitting of the concentration profiles, Biophys. J. 72 (1997) 428–434.

[8] B. Demeler, H. Saber, Determination of molecular parameters by fitting sedimentation data to finite-element solutions of the Lamm equation, Biophys. J. 74 (1998) 444–454.

[9] P. Schuck, C.E. MacPhee, G.J. Howlett, Determination of sedimentation coefficients for small peptides, Biophys. J. 74 (1998) 466–474.

[10] W.F. Stafford, P.J. Sherwood, Analysis of heterologous interacting systems by sedimentation velocity: curve fitting algorithms for estimation of sedimentation coefficients, equilibrium, and kinetic constants, Biophys. Chem. 108 (2004) 231–243.

[11] P. Schuck, P. Rossmanith, Determination of the sedimentation coefficient distribution by least-squares boundary modeling, Biopolymers 54 (2000) 328–341.

[12] P. Schuck, Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling, Biophys. J. 78 (2000) 1606–1619.

[13] B. Demeler, K.E. van Holde, Sedimentation velocity analysis of highly heterogeneous systems, Anal. Biochem. 335 (2004) 279–288.

[14] J.S. Philo, A method for directly fitting the time derivative of sedimentation velocity data and an alternative algorithm for calculating sedimentation coefficient distribution functions, Anal. Biochem. 279 (2000) 151–163.

[15] W.F. Stafford III, Analysis of reversibly interacting macromolecular systems by time derivative sedimentation velocity, Methods Enzymol. 323 (2000) 302–325.

[16] J.S. Philo, An improved function for fitting sedimentation velocity data for low-molecular-weight solutes, Biophys. J. 72 (1997) 435–444.

[17] J. Behlke, O. Ristau, A new approximate whole boundary solution of the Lamm differential equation for the analysis of sedimentation velocity experiments, Biophys. Chem. 95 (2002) 59–68.

[18] X. Chen, A.A. Antson, M. Yang, P. Li, C. Baumann, E.J. Dodson, G.G. Dodson, P. Gollnick, Regulatory features of the *trp* operon and the crystal structure of the *trp* RNA-binding attenuation protein from *Bacillus stearothermophilus*, J. Mol. Biol. 289 (1999) 1003–1016.

[19] D. Snyder, J. Lary, Y.L. Chen, P. Gollnick, J.L. Cole, Interaction of the *trp* RNA-binding attenuation protein (TRAP) with anti-TRAP, J. Mol. Biol. 338 (2004) 669–682.