An Improved Function for Fitting Sedimentation Velocity Data for Low-Molecular-Weight Solutes

John S. Philo
Protein Chemistry Department, Amgen Inc., Thousand Oaks, California 91320 USA

ABSTRACT Many traditional approaches to the analysis of sedimentation velocity data work poorly with data for low-molecular-weight solutes, which have sedimentation boundaries that are severely broadened by diffusion. An approach that has previously had some success is to directly fit these broad boundaries to approximate solutions of the Lamm equation that directly account for the high diffusion. However, none of the available approximate solutions work well at times both early and late in the run, or give boundary shapes that are highly accurate, especially for species of molecular weight < 10,000. An improved fitting function has been developed to overcome some of these limitations. The new function adds two correction terms to the Fujita-MacCosham solution. The optimum coefficients for these new correction terms were determined by a least-squares approach. The accuracy and limitations of fitting with this new function were tested against synthetic data sets obtained by finite-element methods, for analysis of samples containing either single species or several noninteracting species. We also compare the strengths and weaknesses of this method of analysis, and its ability to work with noisy data, relative to recently developed time-derivative methodologies.

INTRODUCTION

Sedimentation velocity can be a powerful tool for analysis of the size and shape of macromolecules in solution, and for analysis of samples containing many species. The recent advent of improved analytical ultracentrifuges has brought about a resurgence of use of sedimentation techniques, and this, along with on-line digital data acquisition, has also fostered development of new and improved methods of data analysis (Hansen et al., 1994). Driven largely by the biotechnology industry, there is also recent interest in using sedimentation velocity to characterize proteins such as cytokines and growth factors with relatively low molecular weights (~5000 to 40,000) and to provide information about their conformation, molecular weight, and homogeneity. In addition, a number of important small structural modules with M_r ~5000-15,000 have now been shown to occur in many proteins (EGF modules, SH2, SH3, and PTP domains, PH domains, etc.). Such modules are often involved in protein-protein interactions and/or key signaling pathways, and it may therefore be useful to characterize their conformation (and possible changes in conformation after binding peptide ligands) by sedimentation velocity.

Unfortunately, the large diffusion coefficient of such low-molecular-weight solutes causes them to produce very broad sedimentation boundaries, even at the highest rotor speeds. Such broad boundaries make it very difficult to assess whether multiple species may be present. The high diffusion even makes it difficult to obtain an accurate sedimentation coefficient by traditional approaches such as the second-moment method, because there is only a very limited range of boundary movement during the time when the method is applicable (i.e., after the meniscus is clear but while there is still a plateau region). The newer time-derivative "dc/dt" techniques (Stafford, 1994) can certainly be applied in such situations, but as we will discuss below, it is necessary to make significant corrections to the sedimentation coefficients when this type of analysis is applied to such small proteins.

One approach to overcoming the problems caused by high diffusion is to incorporate the diffusion coefficient, and its effect on boundary shape, directly into the analysis by fitting the raw data to an appropriate approximate solution of the Lamm equation, with both the sedimentation and diffusion coefficients as fitting parameters. This approach was first applied to single data sets and single species some time ago (Holladay, 1980), using a fitting function with approximations that are only accurate early in the run. More recently we rediscovered a similar approach and extended it using a global analysis of many data sets, and showed that it is applicable to samples containing a small number of noninteracting species (Philo, 1994). However, the Faxen-type approximate solution of the Lamm equation that was employed (equation 2.94 from Fujita, 1975, which we call the "Fujita function") does not treat the effects of restricted diffusion at the meniscus and essentially assumes that the meniscus is rapidly cleared, forming an infinitely sharp boundary at the start of the run. This is certainly not a good approximation for very low M_r solutes, for which the meniscus is cleared very slowly, and even for higher M_r species it restricts that method to use at times later in the run when the effects of the meniscus are smaller. Furthermore, because of the effects of the meniscus, the shape of this function is not an accurate representation of the shape of the boundaries, which is of concern because the ability of this method to detect and resolve the presence of more than one

Received for publication 26 June 1996 and in final form 7 October 1996.
Address reprint requests to Dr. John S. Philo, Amgen Inc., Protein Chemistry 14-2-D, 1840 DeHavilland Dr., Thousand Oaks, CA 91320-1789. Tel.: 805-447-4641; Fax: 805-499-7464; E-mail: jphilo@amgen.com.

0006-3495/96/01/435/10 $2.00

species is based on improved matching of the boundary shapes as more species are included in the analysis.

This paper describes the development of a new fitting function that gives improved accuracy at low molecular weights, which is applicable both early and late in the run, and which is reasonably fast to compute. Its performance is then compared to that of the previous function for both single and two-species fits. The possibility of resolving even more species is then explored, as is the ability of this technique to cope with noisy data. Other strengths and weaknesses of the direct boundary fitting method are discussed, as well its advantages and disadvantages relative to time-derivative analysis methods.

METHODS

The nonlinear least-squares fitting techniques and the finite-element computations of simulated data sets were carried out as previously described (Philo, 1994). All finite-element simulations were done using a calculation time increment of 1 s and with the cell divided into radial increments of 0.003 cm. Time-derivative analysis was done using the program DCDT provided by the National Analytical Ultracentrifugation Facility, and the resulting $g(s^*)$ distributions were then transferred into Microcal ORIGIN 4.0 for fitting of peaks to Gaussians.

RESULTS AND DISCUSSION

Tests of existing solutions of the Lamm equation

Because the existing direct boundary-fitting method using the Fujita function is primarily limited by the effects of the meniscus, initially two other approximate solutions of the Lamm equation were considered as alternative fitting functions, both of which correctly treat the boundary conditions imposed by the meniscus: the Holladay solution (Holladay, 1979), and the Fujita-MacCosham solution (Fujita and MacCosham, 1959). Each of these was tested against simulated, noise-free data obtained using the Claverie finite-element method (Claverie, 1975).

As expected, both of these functions work well at times early in the run (where the approximations used in their derivation are very good). They were then tested on data from simulations for a species with a sedimentation coefficient, s, of 2 S and a diffusion coefficient, D, of 10 F (1 F = 10^{-7} cm^2/s), which corresponds to a protein with $M_r \sim 18{,}000$, using data from both early and late times in the run. In this situation, although both of these solutions give accurate values for s, they both give values for D that are significantly in error (about a 6% underestimate with Fujita-MacCosham, and about a 12% overestimate with the Holladay solution), and they also both give significant and systematic deviations from the correct boundary shape. Neither of these solutions is therefore an acceptable alternative.

Development of an improved fitting function

Rather than attempting to find an entirely new solution to the Lamm equation, it seemed reasonable to try to extend the range of application of one of the existing solutions that works well at early times by adding additional terms that would increase the accuracy of the approximation at later times. Because the Fujita-MacCosham solution appears to be the superior of the two at later times, and because it is also more rapid to compute, it was selected as the basis for potential improvement. This solution gives the concentration c at any time t and radial position r in terms of dimensionless parameters $\tau \equiv 2s\omega^2 t$, $x \equiv (r/r_o)^2$, $\epsilon \equiv 2D/s\omega^2 t$, and $z = \ln(x)$, as

$$c = \frac{c_o e^{-\tau}}{2}\left\{1 - \mathrm{erf}\left[\frac{\tau - z}{2\sqrt{\epsilon\tau}}\right] + \frac{2}{\sqrt{\pi}}\left(\frac{\tau}{\epsilon}\right)^{1/2}\exp\left(\frac{-(\tau - z)^2}{4\epsilon\tau}\right) + \left(1 + \frac{\tau + z}{\epsilon}\right)\left(1 - \mathrm{erf}\left[\frac{\tau + z}{2\sqrt{\epsilon\tau}}\right]\right)\exp\left(\frac{z}{\epsilon}\right)\right\}, \quad (1)$$

where c_o is the loading concentration, r_o is the meniscus position, and erf( ) is the error function. This solution was obtained under the approximations that $\tau \ll 1$ and that $\exp(-z) \cong 1$.

Upon fitting individual simulated boundaries, as the run progressed it appeared that the error in the returned D value grew approximately linearly with τ. This suggested that a better approximation could be obtained by adding a new correction term (or terms) of order τ, but how could the correct term(s) be found? A time-honored method for solving differential equations is, of course, to guess the correct solution, and we have partially adopted this approach. In this case, the fits suggested that we need to divide D by a term like $(1 + \alpha\tau)$, where α is a positive numerical constant, but then how can the correct value for α be determined? Rather than trying to mathematically solve for solutions of this form, a numerical approach was used instead, wherein α was left as a parameter whose value was to be determined by least-squares fitting.

Therefore the following approach was used to find an improved function. First, a term of the form $(1 + \alpha\tau)$ was added to Eq. 1 in a location that might correct its tendency to underestimate D at longer times in the run. Next, this new function was fitted in separate trials to groups of 8–9 simulated data sets representing proteins of ~6, 18, and 67 kDa at 60,000 rpm, which included scans from early in the run (well before the meniscus is cleared) until the leading edge of the boundary has nearly reached the cell bottom. During these fits the values of s, D, c_o, the unknown coefficient α, and a baseline offset were allowed to vary as needed to optimally fit the shape of the concentration profiles (the fits were not constrained to give the correct values for s, D, and c_o). Finally, each candidate function was evaluated for self-consistency. That is, if the new function truly represents a higher-order solution, then 1) the value of α that is returned by fitting should be essentially independent of the s and D values used in the simulation; 2) the returned values of s and D should closely match the correct values; and 3) the residuals of the fit should be significantly

reduced relative to those for the starting Fujita-MacCosham solution.

This procedure was indeed successful, but the shape of the candidate function was still not optimal for the lowest-molecular-weight species. The pattern of deviations suggested that a second small correction linear in D was also needed. In dimensionless parameters D is proportional to $\epsilon\tau$, so terms of the form $(1 + \beta\epsilon\tau)$ were also tested, with β as a coefficient to be determined by fitting. The doubly corrected function, which we call a "modified Fujita-MacCosham function," therefore becomes

$$c = \frac{c_0 e^{-\tau}}{2}\left\{\begin{array}{l} 1 - \mathrm{erf}\left[\dfrac{\tau - z}{2\sqrt{\epsilon\tau}}(1 + \alpha\tau)\right] \\ \quad + \dfrac{2}{\sqrt{\pi}}\left(\dfrac{\tau}{\epsilon}\right)^{1/2}\exp\left(\dfrac{-(\tau - z)^2}{4\epsilon\tau}(1 + \beta\epsilon\tau)\right) \\ \quad + \left(1 + \dfrac{\tau + z}{\epsilon}\right)\left(1 - \mathrm{erf}\left[\dfrac{\tau + z}{2\sqrt{\epsilon\tau}}\right]\right)\exp\left(\dfrac{z}{\epsilon}\right) \end{array}\right\}, \quad (2)$$

In this function the new $(1 + \alpha\tau)$ correction term alters the width of the first error function sigmoid (the dominant term in this expression, except for very early in the run) and thereby corrects the underestimate of D at longer run times, and the $(1 + \beta\epsilon\tau)$ correction term produces a much smaller correction in the exponential function of the second term.

When applied to the simulated data for the three different proteins, the optimum value for β was not truly constant, but ranged from 1.8 to 2.2. This term has almost no effect on the values of s or D, but its inclusion does significantly improve the shape of the function (i.e., it significantly reduces the residuals). Therefore it was decided to fix the value of β at 2. With this value for β, the simulated data for the three different proteins return optimum values for α of 0.2487 ± 0.0020, thus meeting the desired criterion of constancy.

The improvement obtained by using this new function instead of the Fujita function that was used previously (Philo, 1994) is shown in Table 1. The new function gives about an order of magnitude improvement in the accuracy of the D values, as well as in the residuals, for the lower molecular weights, while still providing good accuracy for s in all cases. For proteins above ~100 kDa, the two functions are about equivalent. Overall, the simulations indicate that by using this function it should be possible to obtain s and D with an accuracy of ~1%, even for quite small proteins. It should be noted that although we feel these results and this method of derivation justify the use of this modified Fujita-MacCosham function for this purpose, this function should probably not be regarded as a "solution" of the Lamm equation, but rather as a useful semiempirical fitting function.

Accuracy for resolving two species

One important reason for seeking a fitting function that more accurately represents the shape of the boundaries is the hope that this would improve the ability to detect and resolve the presence of multiple species. A real strength of the approach of directly fitting the entire boundary is that if it is possible to obtain a good fit to a single species model, and if the molecular weight implied by the derived values for s and D is consistent with the molecular weight known from an independent method, then this is strong evidence that the material is truly homogeneous. Conversely, the lack of a good fit as a single species (as manifested by large and systematic residuals) suggests the presence of additional species. (Note that although this method is not appropriate for actually characterizing interacting systems, the lack of a consistent fit as a single species is an appropriate test for homogeneity, even if the additional species are due to self-association.) However, if the shape of the fitting function is inaccurate, and given the presence of noise in real data, it can be very difficult to judge whether the deviations in a fit truly indicate an additional species or are just a consequence of the intrinsic limitations of the fitting proce-

TABLE 1 Comparison of fitting results using either the old Fujita function or the new modified Fujita-MacCosham functions on noise-free synthetic data sets covering the full range of boundary movement

True s and D	Fitted s and D using old function (% error)	Fitted s and D using new function (% error)	Residuals using old function (r.m.s. and maximum, % of loading c)	Residuals using new function (r.m.s. and maximum, % of loading c)
0.85 S, 13 F (6 kDa)	0.855 S (+0.6%) 11.15 F (−14.2%)	0.843 S (−0.8%) 12.86 F (−1.1%)	1.1% r.m.s., 6% max	0.032% r.m.s., 0.1% max
2 S, 10 F (18 kDa)	2.011 S +(0.6%) 9.48 F (−5.2%)	1.994 S (−0.3%) 9.99 F (−0.1%)	0.28% r.m.s., 2.3% max	0.036% r.m.s., 0.1% max
4.4 S, 6 F (67 kDa)	4.408 S (+0.2%) 5.95 F (−0.8%)	4.396 S (−0.1%) 6.05 F (+0.8%)	0.066% r.m.s., 0.6% max	0.015% r.m.s., 0.1% max
7.2 S, 1.2 F (550 kDa)	7.204 S (+0.1%) 1.22 F (+1.7%)	7.205 S (+0.1%) 1.22 F (+1.7%)	0.055% r.m.s., 1.1% max	0.087% r.m.s., 0.4% max

The simulations for the first three species are for a 60,000 r.p.m. rotor speed; that for the last is at 40,000 r.p.m.

dure. Furthermore, when there truly are multiple species present, if the shape of the fitting function is not accurate, the ability of curve fitting to accurately resolve multiple components is likely to be strongly compromised.

Samples of proteins that are normally monomeric are commonly contaminated with small amounts of dimer or higher oligomers that are not in association equilibrium with the monomer (e.g., a disulfide-linked dimer), and such samples are therefore appropriate candidates for analysis as multiple noninteracting species. Using the Fujita function, we have previously shown that it is possible to resolve <10% contamination of bovine serum albumin with such a dimer (Philo, 1994), but could this be done for a much lower-molecular-weight monomer, where the physical separation of monomer and dimer is very poor? Simulations were run of such a situation for a sample containing 0.9 AU of a monomer with $s = 1.5$ S, $D = 11$ F (~12 kDa), and 0.1 AU of a noninteracting dimer with $s = 2.25$ S, $D = 8.25$ F. To make the simulation more realistic, random Gaussian noise of r.m.s. amplitude 0.006 AU was added (typical of the photometric noise in a Beckman Optima XL-A).

The simulations indeed show that when using the Fujita function it is difficult to tell from the residuals that a second component is present. Fig. 1 shows the residuals using the Fujita function for the first and last scans in simulated runs for either a pure monomer (Fig. 1 *A*) or the mixture with 10% dimer (Fig. 1 *B*), both fitted assuming a single species is present. The actual first and last data sets for the mixture simulation are shown in Fig. 1 *E* to indicate the position of the boundaries at this time, and to show that this amount of dimer does not give any hint of a second boundary. Although the amplitude of the residuals is, as expected, clearly higher when dimer is present, the pattern of the residuals for the pure monomer sample (which arise from the incorrect shape of the Fujita function) is hardly distinguishable from the pattern when dimer is added (and this is true throughout the run). Thus it would not be obvious from the residuals that this sample is not homogeneous. With the Fujita function, the residuals still show some systematic deviations early in the run, even with a two-species fit (Fig. 1 *C*), whereas the residuals appear random when the new function is used in a two-species fit (Fig. 1 *D*).

These simulations also show that the MFM function gives a dramatic improvement in the ability to correctly resolve the properties of the two components in the mixture, as summarized in Table 2. The Fujita function does a very poor job in the two-species fit, implying that the mixture consists of about equal amounts of a 1.34 S and a 1.81 S species. In contrast, the modified Fujita-MacCosham function (which we will hereafter call the MFM function for brevity) does an excellent job of resolving monomer and dimer, deriving the correct s, D, and concentration for the monomer within ~1%, and the correct s and D for the dimer within ~2% (although, as might be expected, the 95% confidence interval for D of the dimer is quite large). Most importantly, the derived properties of the dimer are more than sufficiently accurate to correctly identify the species as a dimer. The

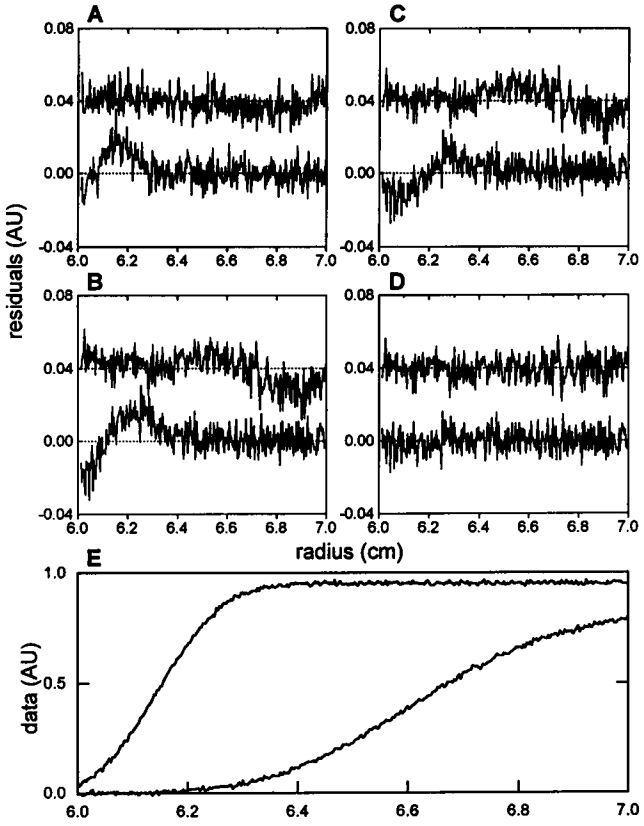

FIGURE 1 Residuals (experimental − fitted) from fits to simulated velocity experiments; in each panel the lower trace shows the residuals from the first data set used in the analysis, and the upper trace corresponds to the last data set (see text). (*A*) Single-species fit using the Fujita function to a simulation of a 1.5 S, 11 F monomer at 1 AU loading concentration. (*B*) Single-species fit using the Fujita function for a 90% monomer, 10% dimer simulation. (*C*) Two-species fit using the Fujita function on the monomer-dimer mixture simulation. (*D*) Two-species fit using the modified Fujita-MacCosham function on the monomer-dimer mixture simulation. (*E*) Sedimentation boundaries corresponding to the first and last data sets for the monomer-dimer mixture.

improved performance of the MFM function imposes at most an approximately twofold increase in computational time over the Fujita function. In current versions of the SVEDBERG program, this performance penalty is more than overcome by faster algorithms, such that the MFM function is now threefold faster than the Fujita function was in older versions. The two-species fits for Table 2 with the MFM function require ~40 s using a 90-MHz Pentium. (A version implementing this new function will be made available via http://www.bbri.harvard.edu/RASMB/rasmb.html.)

Tests of three-species fits

Recently there has been considerable interest in using sedimentation velocity techniques to obtain stoichiometry and conformation information about antigen-antibody (Hensley, 1996) and other protein-protein complexes, samples that typically contain more than two species. For example, Hensley et al. (1995) have used time-derivative analysis to re-

TABLE 2 Results from fitting simulations for a monomer-dimer mixture containing 0.9 AU of a monomer with s = 1.5 S, D = 11 F, and 0.1 AU of its dimer with s = 2.25 S, D = 8.25 F

Type of Fit	Results using Fujita function	Results using modified Fujita-MacCosham function
One species fit	1.571 S [1.569–1.573]	1.557 S [1.556–1.559]
	11.77 F [11.67–11.86]	12.52 F [12.41–12.58]
	1.0068 AU [1.0048–1.0083]	1.0012 AU [0.9994–1.0024]
Two species fit		
Species 1	1.339 S [1.326–1.351]	1.489 S [1.484–1.494]
	8.35 F [8.21–8.52]	10.92 F [10.77–11.07]
	0.475 AU [0.456–0.491]	0.888 AU [0.877–0.899]
Species 2	1.812 S [1.801–1.820]	2.200 S [2.169–2.230]
	11.11 F [10.79–11.40]	8.42 F [7.72–9.15]
	0.527 AU [0.512–0.547]	0.112 AU [0.101–0.123]

Values within square brackets are 95% confidence intervals.

solve and identify up to four different components in mixtures of interleukin-5 (IL-5), a dimeric protein, and the Fab fragment of an antibody directed against it. In such situations the binding of the complexes is often so tight, and the kinetics sufficiently slow, that the different species are effectively noninteracting during the course of the velocity run. Is the direct boundary-fitting approach also applicable in such situations?

Simulations were run to mimic an experiment carried out on a mixture containing 2 mol of Fab per IL-5 dimer (Hensley et al., 1995). This mixture contained three species in approximately a 2:1:1 ratio: a complex containing two Fab's and one IL-5 dimer (5.80 S, 4.32 F), a complex containing one Fab per IL-5 (4.49 S, 5.07 F), and free Fab (3.54 S, 6.59 F). For the simulations a total loading concentration giving 1 AU (about 150 μg/ml for scans at 230 nm) was assumed, with random noise of 0.006 AU r.m.s. First separate simulations for each individual species were run, both with and without added noise, to establish the "correct" values for the parameters (which differ slightly from the true values because of the approximations) and the maximum precision that could be obtained for the parameters, given the assumed signal/noise. These results are tabulated in Table 3.

Next a three-species fit was attempted on the simulation for the mixture. Somewhat surprisingly, this fit was able to converge starting from the default guesses of equal amounts of 2 S, 4 S, and 6 S species, even when all parameters were allowed to vary. These data, the fitted curves, and the residuals are shown in Fig. 2, *A* and *C*. As seen in Table 3, the results from this fit are quite good, with all of the sedimentation coefficients determined with an accuracy of better than 1%, the diffusion coefficients to an accuracy of better than 4%, and the fraction of each species to better than 1%. Furthermore, by comparing the parameter confidence regions from the mixture fit to those for an individual species fit at the same signal/noise, we see that the presence of the other species has reduced the precision of the s and D values by only two- to fourfold.

In any multispecies analysis, it is always desirable, when possible, to fix the properties of one or more species at independently determined values, to increase the reliability and accuracy of the remaining fitted parameters. In some cases it may also be necessary to fix some of the parameters even to get convergence of the fit (for example, as was the case in the multiple-Gaussian analysis of the experiment we are simulating; Hensley et al., 1995). Therefore we have tested how knowledge of only the s values, or of both s and D, for the largest and smallest species in this mixture would affect the results. Fixing these values does indeed somewhat improve the accuracy and precision of the remaining parameters, but not that dramatically. However, the rather small improvement in this case is a consequence of the good accuracy obtained when all of the parameters are fitted. If

TABLE 3 Tests of three-species fits

Data fitted and type of fit	Results for species 1	Results for species 2	Results for species 3
Each species simulated separately, no noise	5.797 S, 4.39 F, 0.500 AU	4.486 S, 5.11 F, 0.250 AU	3.535 S, 6.61 F, 0.250 AU
Each species simulated separately, with 0.006 AU noise added	5.797 [5.793–5.801] S 4.40 [4.34–4.46] F 0.5002 [0.4995–0.5010] AU	4.487 [4.480–4.494] S 5.07 [4.95–5.19] F 0.2499 [0.2492–0.2507] AU	3.536 [3.529–3.542] S 6.48 [6.34–6.63] F 0.2499 [0.2489–0.2508] AU
Simulations of mixture, all parameters varied	5.796 [5.779–5.811] S 4.42 [4.30–4.55] F 0.5019 [0.4899–0.5151] AU	4.473 [4.413–4.530] S 4.91 [4.59–5.22] F 0.2442 [0.2319–0.2573] AU	3.539 [3.515–3.563] S 6.82 [6.55–7.08] F 0.2542 [0.2419–0.2662] AU
Simulations of mixture, s for species 1 and 3 fixed at known values	Fixed at 5.797 S 4.41 [4.30–4.53] F 0.5008 [0.4975–0.5042] AU	4.474 [4.446–4.502] S 4.94 [4.61–5.28] F 0.2469 [0.2375–0.2569] AU	Fixed at 3.535 S 6.79 [6.44–7.15] F 0.2526 [0.2488–0.2604] AU
Simulations of mixture, s and D for species 1 and 3 fixed at known values	Fixed at 5.797 S Fixed at 4.39 F 0.5003 [0.4962–0.5043] AU	4.482 [4.462–4.504] S 4.93 [4.61–5.25] F 0.2465 [0.2377–0.2556] AU	Fixed at 3.535 S Fixed at 6.61 F 0.2532 [0.2460–0.2606] AU

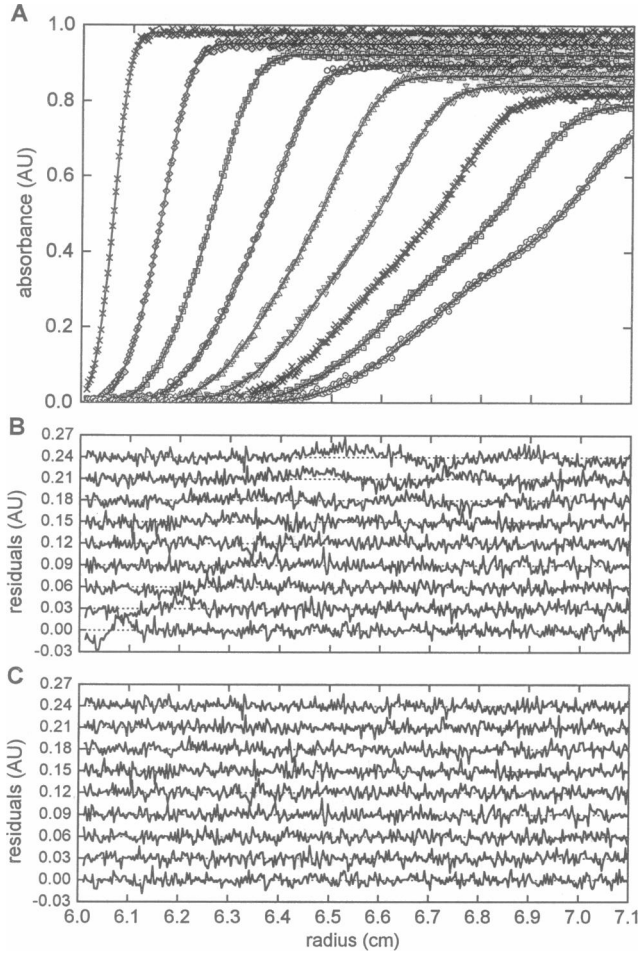

FIGURE 2 Simulated data for a mixture of three species (see text) and fitted curves (*A*), residuals for a two-species fit (*B*), and residuals for a three-species fit (*C*). The bottom curve in each residual plot is for the earliest data set, and each subsequent plot has been shifted upward by 0.03 AU for clarity.

the signal/noise were lower, or the physical separation poorer, an independent knowledge of at least the sedimentation coefficient of one or more of the species would almost certainly be required to obtain reliable values for the remainder. In this regard, it is also instructive to examine a two-species fit to these same data. A two-species fit finds species of 5.72 S, 4.94 F (close to the true values for the major component) and 3.92 S, 7.7 F (an s value about midway between those of the two smaller components). This fit actually reproduces the shape of the boundaries fairly well. Without any independent information about the number of species or their properties, the inadequacy of this two-species fit can really only be seen by the nonrandom pattern of the residuals, as shown in Fig. 2 *B*, and this pattern would not necessarily be apparent if the signal/noise were lower by about twofold. Fig. 2 *B* also illustrates the importance of including data both early and late in the run, because the middle three scans are fitted quite well with only two species.

The overall conclusion is that this technique should be capable of resolving at least three species (and possibly four if the properties of some species are known) from data at signal/noise levels routinely available with current commercial instrumentation. However, it will generally be important to have some independent information about the number of species, and highly desirable to be able to run one or more of them as individual species or in mixtures in which they are the dominant component. Moreover, one must always be cognizant of the underlying assumption that the species are effectively noninteracting on the time scale of the velocity run.

Can this method be applied when the sedimentation coefficient is concentration dependent?

One of the limitations of this direct fitting analysis is that the fitting functions do not take into account the possible concentration dependence of s (or D). With absorbance optics it is generally possible to acquire data at protein concentrations of <100 μg/ml, where for globular proteins the concentration dependence is usually negligible. (This is often done in the XL-A by scanning at 230 nm, a wavelength with excellent signal/noise, and where the absorbance is typically five- to sevenfold higher than at 280 nm.) However, in certain situations it may be necessary or desirable to work under conditions in which the concentration dependence is not negligible.

To obtain an estimate of how seriously this might affect the validity of the direct fitting analysis, simulations were run for a species with $s = 2$ S, $D = 10$ F, at concentrations of either 2 or 10 mg/ml, assuming this species had a concentration dependence given by $s_c = s_0 \times (1 - c \times 0.009)$, a magnitude typical of globular proteins (Laue et al., 1992). (Note that the Claverie routine did not incorporate a concentration dependence of D.) Fits to the 2 mg/ml simulation returned 1.960 S, 9.58 D, fits to the 10 mg/ml simulation returned 1.837 S, 8.36 D, whereas without concentration dependence the values returned were 1.994 S, 9.96 F. Thus the s values show reductions close to the expected factors of 1.018 and 1.09, and the D values are reduced almost exactly twice as much. At 2 mg/ml, the shape of the MFM function is still a reasonably good match to the boundaries, but at 10 mg/ml the maximum residuals exceed 2% of the loading concentration. Overall, these results suggest that this method is quite appropriate for determining raw, uncorrected sedimentation coefficients of single species with moderate concentration dependence, and that even the D values may be sufficiently accurate for many applications. However, because of the poorer match of the MFM function to the boundary shapes when there is a significant concentration dependence, the accuracy of multispecies analysis would be severely compromised.

Comparison with time-derivative analysis when applied to low-molecular-weight solutes

The "dc/dt" time-derivative analysis technique (Stafford, 1994) has recently become widely used for the analysis of sedimentation velocity data. Although the time-derivative method was developed primarily for analysis of interacting systems at low concentrations, it is often applied to obtain sedimentation coefficients for individual species, and more recently to obtain diffusion coefficients (Stafford, 1996; Hensley, 1996). For proteins smaller than ~40 kDa, some investigators have noted that sedimentation coefficients obtained via this method, either from the positions of peaks in the $g(s^*)$ distributions, or as weight-average values calculated from these distributions, are significantly smaller than those obtained from the direct boundary-fitting approach. This has led to some confusion about the applicability of both methods and has raised questions about which s values are "correct." It therefore seems important to make a direct comparison of these techniques and to explore this issue further.

Both methods were applied to simulations for a 2 S, 10 F (~18 kDa) species. Noise-free simulations were used, because for this purpose noise reduction is irrelevant. One difficulty in comparing these methods is that each method typically uses quite different data acquisition sequences. For time-derivative analysis a group of scans closely spaced in time is used (generally acquired fairly late in the run for optimal resolution of species), during which there is a modest amount of boundary movement. For the direct fitting approach scans taken at much greater time intervals and covering the full range of boundary movement are generally used. For this comparison, however, conditions optimized for time-derivative analysis were chosen, and the same set of eight closely spaced data sets (which span the time when the boundary crosses the midpoint of the cell) was used for both methods.

The $g(s^*)$ distribution from the time-derivative analysis is shown in Fig. 3. The peak of the $g(s^*)$ distribution occurs at 1.896 S, whereas the weight-averaged value obtained by integrating across this distribution is 1.867 S, i.e., results 5–7% below the true value. Also shown in Fig. 3 is the best fit of the $g(s^*)$ distribution to a Gaussian, which is centered at 1.883 S. The width of this Gaussian implies a diffusion coefficient of 9.91 F, in good agreement with the true value, although values of D differing by a few percent would be obtained at different times during the run (Stafford, 1996). By comparison, values of 1.993 S and 10.01 F are obtained when the MFM function is used to directly fit the same data sets. Thus if we simply directly compare the numbers, for small proteins the direct fitting approach is considerably more accurate. The underestimation of s values by $g(s^*)$ becomes worse for even smaller molecular weights, but is negligible above ~40 kDa (a range where both methods generally give accurate values for both s and D). However, because this is a reproducible, systematic property of the $g(s^*)$ curves, the s values from time-derivative analysis are not "wrong." It would be unfair not to point out that it is quite feasible to correct for this effect, for example by running simulations such as these to determine the size of the shift to construct a calibration curve. Therefore a knowledgeable and careful user of time-derivative analysis can correct for this shift and thereby obtain accurate s values even for low-molecular-weight species.

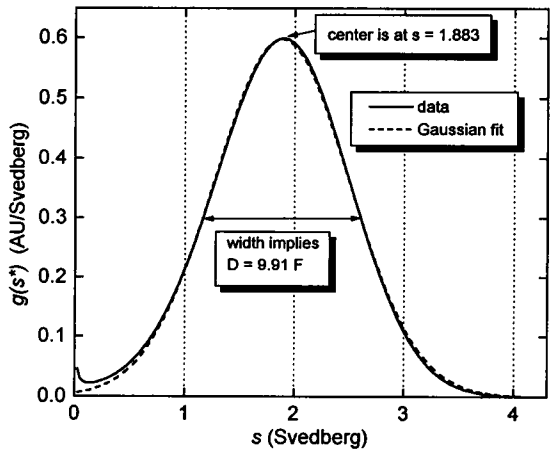

FIGURE 3 Results of time-derivative analysis on simulated, noise-free data for an s = 2 S, D = 10 F species, and a fit of the resulting $g(s^*)$ distribution to a Gaussian function.

It should also be noted that Fig. 3 demonstrates that although the Gaussian shape is a good approximation near the center of the $g(s^*)$ distribution, for low-molecular-weight species the shape differs significantly from a Gaussian in the "wings" of the distribution, especially when a boundary has not moved very far from the meniscus. The results above, where the Fujita and MFM functions were compared on a monomer-dimer mixture (Fig. 1, C and D, Table 2), suggest that the fact that the $g(s^*)$ curves for low-molecular-weight species differ in shape from a Gaussian may significantly limit the accuracy of multispecies fits of $g(s^*)$ distributions using Gaussian components.

Can direct boundary fitting be applied to noisy data?

One important reason why the time-derivative technique (Stafford, 1994) has rapidly become widely used is its powerful ability to remove systematic, time-independent noise from the data, and to reduce random noise by averaging. Because nonlinear least-squares analysis can, in many cases, also perform well at reducing the effects of noise, it seems worthwhile to compare and discuss the abilities of both methods to minimize the effects of both random and systematic noise. Because the focus of this paper is on paucidisperse samples, the goal is assumed to be to obtain accurate hydrodynamic parameters for individual species, and therefore the appropriate measure of noise reduction is improved precision (lower standard deviation)

of those parameters (as opposed to smoothing or noise reduction in the $g(s^*)$ distributions themselves).

For the purpose of comparing analyses of data that are limited by random noise (e.g., the intrinsic photometric noise of the optical system), data sets were simulated for a protein of 2.16 S and 6.72 D, corresponding to measurements on IL-5 done with custom Rayleigh interference optics (Hensley et al., 1995). In this case we will simulate the same experiment at the same protein concentration, but assume instead that it was done using absorbance scans at 230 nm, giving a total signal of 0.45 AU and 0.006 AU r.m.s. noise. Once again the direct fitting approach would optimally use data sets covering the full range of boundary movement, but in this case a more limited time range, optimal for time-derivative analysis, was chosen so that the two methods could use the same data.

Therefore, a group of 32 closely spaced data sets appropriate for time-derivative analysis (2-min interval) were created and analyzed by the dc/dt method. The resulting $g(s^*)$ distribution was then fitted to a Gaussian to derive best-fit values and standard errors for s and D. This fit gave values for s and D of 2.102 S and 6.99 F, with standard errors (estimated from the variance-covariance matrix) of ± 0.0028 S and ± 0.035 F. These same data were then analyzed by direct fitting using the MFM function, but only every fourth data set was used from among the 32. This latter fit gave values of 2.156 ± 0.0011 S and 6.73 ± 0.040 F. If only these eight data sets are used in the dc/dt method, the values 2.106 ± 0.0036 S and 6.98 ± 0.038 F are obtained.

Thus the direct fitting approach gives noise reduction (as measured by the standard errors) that is about equal to time-derivative analysis for D, and better for s, even when fewer scans are used. This simulation also shows that direct fitting can work well with the same scan sequences used for time-derivative analysis, although for direct fitting the precision of s and D would improve further if the scan interval were longer. In practice, probably the main situation where it may be advantageous to exploit this noise-averaging characteristic of direct boundary fitting arises for absorbance data, where the relatively slow scan speed can significantly limit the numbers of scans that can be acquired, especially when more than one sample is being run at one time. Thus, for example, if one needed to study a group of samples at very low concentrations, it may be beneficial to run seven samples simultaneously in the eight-hole rotor and analyze them with direct boundary fitting, whereas perhaps only one or two could be run simultaneously if time-derivative analysis were used. (This limitation would not apply for the Rayleigh optical system in the new Beckman XL-I because of its rapid data acquisition.)

However, it is important to emphasize that the type of "noise" that is often most important is not random photometric noise, but rather time-independent, systematic distortions in the data that are caused by the windows of the centrifuge cell (so-called window noise). This window noise is essentially completely eliminated by time-derivative analysis, but it poses a potentially much greater problem for the direct fitting approach.

It can be easily shown that noise "spikes" from dust or scratches on the windows, which affect the data over only a very limited radial distance, have very little influence on the results of direct fits. On the other hand, noise that affects broad regions of the cell (such as window distortion in refractometric scans, nonuniform window absorbance in uv scans, or protein deposits on the windows) are much more of a problem for accurate analysis by direct fitting. In such situations, it is often possible to reduce or eliminate the window noise by subtracting a "baseline file" from each data set used in the direct analysis. Such baseline files, which ideally exactly reproduce the window distortions, are generally best created by simply continuing the run and recording the baseline after the sample has been completely pelleted (but this is often not practical with low-molecular-weight species). It is also possible to obtain a baseline after a run by rinsing and refilling the sample channel with buffer, but such a baseline may not exactly reproduce the window noise if either the window distortions or the cell position is not reproducible, and/or if protein deposits are washed away. For absorbance data another method for obtaining a baseline is to record a scan at a wavelength where the sample does not absorb, but this approach has the drawbacks that the window absorbance may be wavelength dependent, and that such a scan will not correct for protein deposits on the windows. Some examples of these latter two methods for obtaining baseline files for absorbance data are shown in Fig. 4. Traces A–D show example baselines for moderately bad window noise obtained by scanning at wavelengths off the absorption peak (at 340 nm in this case) and after rinsing and refilling the cell and then scanning at the 230-nm measurement wavelength, for two different cells from the same run. Although for each cell the two types of baseline share many of the same noise features and overall trends, the differences between them are certainly significant (and not consistent from cell to cell), which illustrates some of the problems and limitations in obtaining an accurate baseline. Trace E shows an example that represents one of the worst cases of window noise we have seen for absorbance scans (probably the result of a fingerprint).

How much would "bad" distortions such as those in Fig. 4 *E* influence analytical results if they were not properly removed by a baseline file? To answer this, the trace from Fig. 4 *E* was first subjected to a nine-point adjacent-value smoothing (to reduce the random noise component but preserve the broad-scale distortions), and then this file was subtracted from the same simulated data for IL-5 used above for the photometric noise-averaging test. The analysis of the resulting distorted data gave 2.163 ± 0.0015 S and 6.56 ± 0.057 F, showing that this amount of window noise causes only minor changes in the best-fit parameters. Similarly, when this same window noise was added to the simulations used for Table 1, the maximum fractional change was $<0.5\%$ in s and $<1\%$ in D.

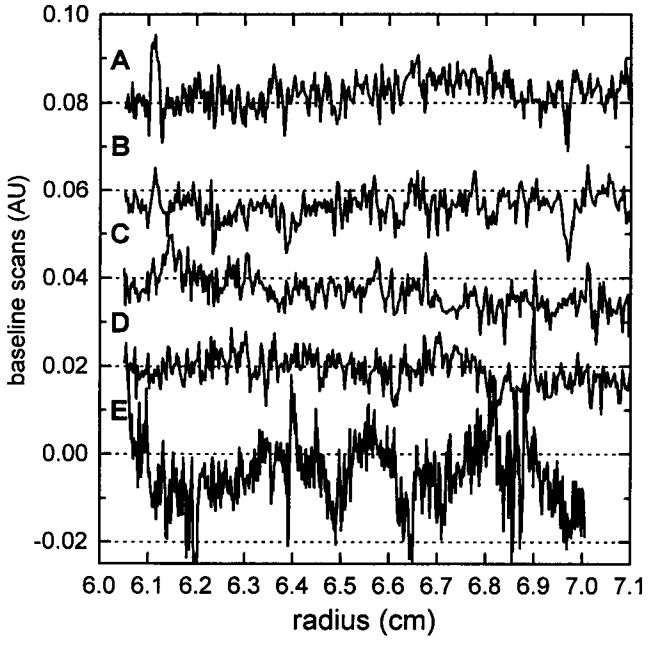

FIGURE 4 Examples of baseline files showing "window noise" from absorbance scans with the Beckman Optima XL-A. Traces A and C were recorded by rinsing and refilling the sample channels of two different cells after a run, reloading the cells, and then recording these scans at the run speed and using the same measurement wavelength as the actual data for the sample (230 nm for these data). Traces B and D are from the same cells, but were recorded at the end of the experimental run, without stopping the rotor, at 340 nm, where the protein absorbance is negligible. All four of these traces are averages of four individual scans that were later combined digitally to create one average scan. Trace E is a single scan like those in B and D but from a different run, showing a particularly bad example of window noise.

Thus for absorbance scans window noise is generally not a strongly limiting factor for direct fitting. However, for Rayleigh data the relative magnitude of window noise is generally much higher, and therefore window noise would probably strongly influence the results unless an accurate baseline file can be obtained and subtracted. We are exploring the possibility of using a scan of the plateau region, taken very early in the run, to provide a measure of the window distortions that can be used as a baseline for Rayleigh data, but validation of this approach must await access to an instrument with Rayleigh optics. In the absence of a baseline correction, the time-derivative approach is probably a better choice for analysis of Rayleigh data.

Other limitations to accuracy of the hydrodynamic parameters

The above results imply that direct fitting with the MFM function should routinely give a precision and accuracy of better than 1% for sedimentation coefficients, and a few percent for diffusion coefficients, for proteins that are 5 kDa or larger, and that data of this quality should be readily obtained at protein concentrations of ~100 μg/ml using absorbance data from the Beckman XL-A, even in the presence of window noise. One additional factor that could easily compromise this accuracy (no matter what the analysis method) is variations in viscosity due to changes in the sample temperature, and it is certainly important to allow sufficient time for temperature equilibration before starting a run. A second potential limit that is less obvious is the accuracy of the position of the meniscus. An error of only 0.006 cm in the meniscus position will typically produce an error of 1% in s, and such an error may represent a shift of only a single data point (or less) depending on the radial data density. Furthermore, it is not entirely clear how to determine the true correct meniscus position from the experimental data. Our usual practice is to define the meniscus position for absorbance data by the peak of the positive excursion (which is usually also approximately the center of meniscus region), but it is not obvious that this is the position that best corresponds to theory. With the direct boundary-fitting approach, it is possible to designate the meniscus position as another parameter to be determined by fitting. However, when the Fujita function is used, its poor representation of data near the meniscus means that fitting the meniscus position is problematic and is likely to produce inaccurate results. With the MFM function, at least for simulated data, if the meniscus is treated as an adjustable parameter, the fitted meniscus position is very close to the true one (usually within 0.0001 cm), and the values for other parameters are essentially unchanged. Therefore, fitting the meniscus position may now be a more reasonable and useful option (but certainly it will always be better to have an accurate experimental value).

In practice, our results show that the reproducibility of hydrodynamic parameters from this type of analysis is quite good, and that these and other sources of systematic error do not compromise the precision too severely. For example, the protein that we have measured the most times (five different runs over a 2-year period) gave peak-to-peak differences in s, D, and M_r (from s/D) of 0.9%, 5.0%, and 4.5%, respectively, and standard deviations of 0.4%, 2.1%, and 2.4%. Other proteins that we have measured on more than one occasion seem to give similar precision. The variations between runs are, however, generally significantly greater than those obtained for duplicate samples in the same run. Furthermore, these parameter variations between runs are also often outside the statistical 95% confidence interval from the fits, which suggests that some form of systematic noise is the true limiting factor. It also should be noted that, although our experience in applying the new MFM function to real systems is limited, the results to date suggest that the molecular weights obtained from s/D are a few percent lower than expected based on known molecular weights, which probably indicates that some unknown factor is causing the boundaries to be slightly broader than theory predicts, leading to a slight overestimate of D. This latter observation is also consistent with our earlier observation that the Fujita function, which theoretically should underestimate D, seems to give accurate molecular weights

(Philo, 1994). Thus all of our experiments seem to give boundaries slightly broader than expected, but the exact source of this broadening, and whether or not it is common to all instruments, remains to be determined. Nonetheless, it is quite clear that with these techniques D values and molecular weights with an accuracy of a few percent are easily obtainable. This accuracy is more than sufficient for many purposes, but this method is certainly not a replacement or substitute for sedimentation equilibrium.

This manuscript is dedicated to David Yphantis in honor of his 65th birthday. We thank Walter Stafford for helpful discussions and for the release of his program DCDT via the National Analytical Ultracentrifugation Center at the University of Connecticut.

REFERENCES

Claverie, J.-M. 1975. Sedimentation of generalized systems of interacting particles. I. Solution of systems of complete Lamm equations. *Biopolymers.* 14:1685

Fujita, H., and V. J. MacCosham. 1959. Extension of sedimentation velocity theory to molecules of intermediate sizes. *J. Chem. Phys.* 30:291–298.

Fujita, H. 1975. Foundations of Ultracentrifugal Analysis. John Wiley & Sons, New York. 70.

Hansen, J. C., J. Lebowitz, and B. Demeler. 1994. Analytical ultracentrifugation of complex macromolecular systems. *Biochemistry.* 33: 13155–13163.

Hensley, P. 1996. Defining the structure and stability of macromolecular assemblies in solution: the re-emergence of analytical ultracentrifugation as a practical tool. *Structure.* 4:367–373.

Hensley, P., C. C. Silverman, D. E. McNulty, M. L. Doyle, D. G. Myszka, T. G. Porter, and W. F. Stafford, III. 1995. The binding of an Fab to dimeric human interleukin-5: a solution interaction analysis using the time derivative method to interpret sedimentation velocity data. Beckman Instruments Discovery Seminar, Protein Society Annual Meeting, July 1995, available on the Beckman WWW site at http://www.beckman.com/biorsrch/sympo/dscvry/binding.htm.

Holladay, L. A. 1979. An approximate solution of the Lamm equation. *Biophys. Chem.* 10:187–190.

Holladay, L. A. 1980. Simultaneous rapid estimation of sedimentation coefficient and molecular weight. *Biophys. Chem.* 11:303–308.

Laue, T. M., B. D. Shah, T. M. Ridgeway, and S. L. Pelletier. 1992. Computer-aided interpretation of analytical sedimentation data for proteins. *In* Analytical Ultracentrifugation in Biochemistry and Polymer Science. S. E. Harding, A. J. Rowe, and J. C. Horton, editors. Royal Society of Chemistry, Cambridge. 90–125.

Philo, J. S. 1994. Measuring sedimentation, diffusion, and molecular weights of small molecules by direct fitting of sedimentation velocity concentration profiles. *In* Modern Analytical Ultracentrifugation. T. M. Schuster and T. M. Laue, editors. Birkhauser, Boston. 156–170.

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

Stafford, W. F., III. 1994. Boundary analysis in sedimentation velocity experiments. *Methods Enzymol.* 240:478–501.

Stafford, W. F., III. 1996. Rapid molecular-weight determination by sedimentation-velocity analysis. *Biophys. J.* 70:MP452.