HHS Public Access

Author manuscript

Anal Biochem. Author manuscript; available in PMC 2016 August 15.

Published in final edited form as:

Anal Biochem. 2015 August 15; 483: 1–3. doi:10.1016/j.ab.2015.04.029.

A Histogram Approach to the Quality of Fit in Sedimentation Velocity Analyses

Jia Ma, **Huaying Zhao**, and **Peter Schuck***

Dynamics of Macromolecular Assembly Section, Laboratory of Cellular Imaging and Macromolecular Biophysics, National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, Bethesda, MD 20892, USA

Abstract

The quality of fit of sedimentation velocity data is critical to judge the veracity of the sedimentation model and accuracy of the derived macromolecular parameters. Absolute statistical measures are usually complicated by the presence of characteristic systematic errors and run-to-run variation in the stochastic noise of data acquisition. We present a new graphical approach to visualize systematic deviations between data and model in the form of a histogram of residuals. In comparison with the ideally expected Gaussian distribution it can provide a robust measure of fit quality and be used to flag poor models.

Keywords

sedimentation velocity; analytical ultracentrifugation

Sedimentation velocity (SV) analytical ultracentrifugation has re-emerged in the last decade as a popular and powerful physical tool for characterizing nanoscopic particles in a wide range of fields, including the study of biological macromolecules and their interactions [1–5]. This was contributed to, among other factors, by new instrumentation and extended detection limits [6–8], theoretical advances in the sedimentation of interacting systems [9,10], new sedimentation data analysis approaches [11–16], and new computational methods for hydrodynamic modeling [17–19]. Analysis strategies for the global analysis of hydrodynamic data and those of other techniques are expected to further enhance the utility of SV [20–22].

A critical step in the renaissance of SV has been an advance in the mathematical data analysis enabling the direct fitting of raw sedimentation velocity data with explicit models based on solutions of the Lamm equation [23], where macromolecular sedimentation parameters and/or distributions of parameters are calculated and/or refined in non-linear

Addresses for correspondence: Peter Schuck, National Institutes of Health, Bldg. 13 Rm. 3N17, 13 South Drive, Bethesda, MD 20892, Phone: 301-4351950, schuckp@mail.nih.gov.

optimization. An obvious criterion for the quality of fit, and the primary optimization objective, is the root-mean-square deviation (rmsd) between experimental data and model.

However, it is not always trivial to judge whether the final best-fit adequately describes the data, or whether extended models should be tested. One of the problems is that the rmsd (or χ^2) of the fit, is not necessarily a reliable absolute measure for the quality of fit, due to the common (and sometimes considerable) run-to-run variations in the level of stochastic noise of the data. The noise level can depend, for example, on the lamp emission intensity and buffer absorption properties at the acquisition wavelength when using the absorbance optical system, or on changes in the fringe contrast in the interference optical system, respectively. Thus, an important additional criterion for a satisfactory fit is the lack of systematicity of the residuals. Ideally they should be completely random; this has been quantified rigorously with a runs test [24], where the Z-value reports the number of standard deviations by which the runs of positive or negative residuals differ from the expectation for normally distributed residuals [24]. This is implemented as a default output of SV analyses in the software SEDFIT (https://sedfitsedphat.nibib.nih.gov/software/). But, unfortunately, when applied to SV, the Z-value is overly sensitive in practice. Here, it is useful only as a qualitative comparative measure of fit quality, since SV data are typically subject to considerably systematic errors from data acquisition. For example, even though algebraic noise decomposition techniques [25,26] can account explicitly for time-invariant and radial-invariant signal offsets, respectively, when using the interference optics, fluctuations in the radial baseline profiles can occur from vibrational modes or thermal distortions of the optical path that are not captured in this baseline model. Although the signals from these imperfections in the data acquisition are typically small compared to the macromolecular signal, they can still dominate the residuals in conjunction with a good model of the sedimentation process. This poses the question which criterion of goodness of fit can be used in practice, in addition to the overall rmsd, to examine in a robust way the quality of the sedimentation boundary model.

To this end, we have previously introduced and implemented in SEDFIT a picture representation of the residuals [27], where the time and radial dimension of the SV data are mapped to the row and column number of pixels, respectively, and the magnitude of the residual is mapped onto its grey scale (Figure 1C and G). This takes advantage of the superb sensitivity of the human eye to recognize patterns, and allows systematic misfits of the sedimentation boundary to be identified as diagonal features in a picture that would ideally be neutral grey, distinct from vertical and horizontal features that indicate imperfections in the TI and RI noise model, respectively [27,28]. Mapping residual values onto the color scale in bitmaps solves the problem that a simple overlay of the many radial residual curves for all scans at all times will conceal critical systematic misfits of the sedimentation boundary, even more so when modeling difference curves [14]. This bitmap representation was widely adopted and subsequently utilized also by the software SEDANAL [29] and ULTRASCAN [30].

Yet, even though the complete visual inspection of residuals with the bitmap overview is extremely useful and provides detailed insight in the origin of misfit, its interpretation is entirely qualitative and requires some degree of familiarity with the technique of SV for

empirical assessment. It would be desirable to achieve a simpler description that reduces the dimensionality of the residuals into a single curve or a single number that reflects in a robust way on their systematicity. Therefore, we have developed a residuals histogram, where the magnitude of deviation between best-fit model and data, usually comprising $10^4 – 10^5$ data points, are binned (aiming for ~ $10^2 – 10^3$ data points per bin) and compared to the frequency expected for normally distributed noise with the same rmsd (Figure 1D and H). The comparison with a normal distribution is further condensed into a number 'H', which is calculated as the sum of square differences between obtained and ideally expected frequencies for a normal distribution, normalized by the sum of squares of the ideal frequencies. These measures are complementary to the bitmap and overall rmsd in that they display the amplitude of residuals and frequency of larger than average deviations, which will be more than normally expected for boundary misfits.

As an illustration, Figure 1 shows two alternative fits of sedimentation velocity profiles of a bovine serum albumin sample. The *c*(*s*) model (left panels) results in a sedimentation coefficient distribution where the BSA monomer and dimer represents 95% of all signal, but there are also traces of higher order oligomers and smaller breakdown products (not shown). The rmsd is 0.0039 fringes, which is ~ 0.15% of the total signal, in our experience representing an excellent fit. However, there is a low level of remaining systematicity in the residuals, which may be recognized by a faint diagonal feature in the residuals bitmap (**C**), in addition to weak vertical and horizontal features indicating instabilities in the fringe pattern. With the 76,000 data points, the low level of systematicity causes the runs of residuals to be already 160 standard deviations removed from the expectation of a normal distribution. The residuals histogram is less detailed and shows frequencies deviating only little from the normal expectation with H = 0.6%.

For comparison, shown in the right panels is an alternative fit from modeling the data with only monomer and dimer species, not accounting for the remaining 5% of trace higher oligomers and breakdown products. (Similar simplified descriptions are often required to enable direct fits the coupled Lamm equations of interacting systems, which are intrinsically discrete models, as opposed to fits with sedimentation coefficient distributions followed by boundary structure analysis [31].) Even though the rmsd is more than twofold that of the good fit, the poor fit cannot be recognized either directly from inspection of the data fit (**E**) or the residuals overlay (**F**). The bitmap (**G**) clearly shows diagonal features corresponding to misfits co-localized with the data points of the moving boundary. In the residual histogram (**H**), these misfits are clearly highlighted as the substantial deviation from normalcy with an H-value of 3.2%.

We have added the residuals histogram and report of H-values in the default display after fits of SV data in the SEDFIT software. When modeling interference data, we found H-values > 1% to correlate well with inadequate fits. In particular, this value can highlight the effect of unaccounted low amplitude boundaries, such as of trace aggregates. However, in conjunction with data of lower signal/noise ratio it will be less sensitive to imperfections in the model and rather highlight statistical properties of the data acquisition. This, too, can offer useful insights in properties of SV data that are otherwise obscured. For example, in the application to fluorescence optical data the underlying noise distribution is sometimes

not Gaussian but exhibits tails. This makes the quantification by the H-value not as useful, but the symmetry and monotonicity of the residual distribution is still a good indicator for the quality of fit. In the meantime, the histogram has to be considered in the context of other measures for the quality of fit, such as the rmsd.

In conclusion, we found the residuals histogram to provide a useful tool to effectively evaluate statistical aspects of data acquisition and fit in a simpler way than bitmaps, robust to common instrumental imperfections, and requiring less technical experience.

Acknowledgments

We thank Dr. Rodolfo Ghirlando for critical reading of the manuscript. This work was supported by the Intramural Research Programs of the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health.

References

1. Harding SE, Rowe AJ. Insight into protein-protein interactions from analytical ultracentrifugation. Biochem Soc Trans. 2010; 38:901–7. [PubMed: 20658974]
2. Howlett GJ, Minton AP, Rivas G. Analytical ultracentrifugation for the study of protein association and assembly. Curr Opin Chem Biol. 2006; 10:430–436. [PubMed: 16935549]
3. Schuck P. Analytical ultracentrifugation as a tool for studying protein interactions. Biophys Rev. 2013; 5:159–171. [PubMed: 23682298]
4. Scott DJ. The shock of the old: hydrodynamics for the masses. Biochem Soc Trans. 2008; 36:766–70. [PubMed: 18631155]
5. Stafford, WF. Protein-protein and ligand-protein interactions studied by analytical ultracentrifugation. In: Shriver, JW., editor. Protein Struct Stability, Interact. Humana Press; Totowa, NJ: 2009. p. 83-113.
6. Walter J, Löhr K, Karabudak E, Reis W, Mikhael J, Peukert W, et al. Multidimensional Analysis of Nanoparticles with Highly Disperse Properties Using Multiwavelength Analytical Ultracentrifugation. ACS Nano. 2014; 8:8871–8886. [PubMed: 25130765]
7. MacGregor IK, Anderson AL, Laue TM. Fluorescence detection for the XLI analytical ultracentrifuge. Biophys Chem. 2004; 108:165–185. [PubMed: 15043928]
8. Zhao H, Mayer ML, Schuck P. Analysis of protein interactions with picomolar binding affinity by fluorescence-detected sedimentation velocity. Anal Chem. 2014; 18:3181–3187. [PubMed: 24552356]
9. Schuck P. Sedimentation patterns of rapidly reversible protein interactions. Biophys J. 2010; 98:2005–2013. [PubMed: 20441765]
10. Schuck P. Diffusion of the reaction boundary of rapidly interacting macromolecules in sedimentation velocity. Biophys J. 2010; 98:2741–51. [PubMed: 20513419]
11. Schuck P. Sedimentation analysis of noninteracting and self-associating solutes using numerical solutions to the Lamm equation. Biophys J. 1998; 75:1503–1512. [PubMed: 9726952]
12. Schuck P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. Biophys J. 2000; 78:1606–1619. [PubMed: 10692345]
13. Sontag CA, Stafford WF, Correia JJ. A comparison of weight average and direct boundary fitting of sedimentation velocity data for indefinite polymerizing systems. Biophys Chem. 2004; 108:215–230. [PubMed: 15043931]
14. Stafford WF, Sherwood PJ. Analysis of heterologous interacting systems by sedimentation velocity: curve fitting algorithms for estimation of sedimentation coefficients, equilibrium and kinetic constants. Biophys Chem. 2004; 108:231–43. [PubMed: 15043932]
15. Correia JJ, Stafford WF. Extracting equilibrium constants from kinetically limited reacting systems. Methods Enzym. 2009; 455:419–446.

16. Dam J, Velikovsky CA, Mariuzza RA, Urbanke C, Schuck P. Sedimentation velocity analysis of heterogeneous protein-protein interactions: Lamm equation modeling and sedimentation coefficient distributions c(s). Biophys J. 2005; 89:619–634. [PubMed: 15863475]
17. Aragon SR. Recent advances in macromolecular hydrodynamic modeling. Methods. 2011; 54:101–14. [PubMed: 21073955]
18. Ortega A, Amorós D, García de la Torre J. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. Biophys J. 2011; 101:892–8. [PubMed: 21843480]
19. Harding SE, Longman E, Carrasco B, Ortega A, García de la Torre J. Studying antibody conformations by ultracentrifugation and hydrodynamic modeling. Methods Mol Biol. 2003; 248:93–113. [PubMed: 14970492]
20. Perkins SJ, Nan R, Li K, Khan S, Abe Y. Analytical ultracentrifugation combined with X-ray and neutron scattering: Experiment and modelling. Methods. 2011; 54:181–99. [PubMed: 21256219]
21. Zhao H, Schuck P. Global multi-method analysis of affinities and cooperativity in complex systems of macromolecular interactions. Anal Chem. 2012; 84:9513–9. [PubMed: 23020071]
22. Ortega A, Amorós D, García de la Torre J. Global fit and structure optimization of flexible and rigid macromolecules and nanoparticles from analytical ultracentrifugation and other dilute solution properties. Methods. 2011; 54:115–123. [PubMed: 21163355]
23. Lamm O. Die Differentialgleichung der Ultrazentrifugierung. Ark Mat Astr Fys. 1929; 21B(2):1–4.
24. Straume M, Johnson ML. Analysis of residuals: criteria for determining goodness-of-fit. Methods Enzymol. 1992; 210:87–105. [PubMed: 1584056]
25. Schuck P, Demeler B. Direct sedimentation analysis of interference optical data in analytical ultracentrifugation. Biophys J. 1999; 76:2288–2296. [PubMed: 10096923]
26. Schuck P. Some statistical properties of differencing schemes for baseline correction of sedimentation velocity data. Anal Biochem. 2010; 401:280–287. [PubMed: 20206114]
27. Dam J, Schuck P. Calculating sedimentation coefficient distributions by direct modeling of sedimentation velocity concentration profiles. Methods Enzym. 2004; 384:185–212.
28. Zhao H, Brautigam CA, Ghirlando R, Schuck P. Current methods in sedimentation velocity and sedimentation equilibrium analytical ultracentrifugation. Curr Protoc Protein Sci. 2013; 7:20.12.1.
29. Stafford WF, Sherwood PJ. SEDANAL users' manual. 2015
30. Demeler B. ULTRASCAN III manual. 2015
31. Zhao H, Balbo A, Brown PH, Schuck P. The boundary structure in the analysis of reversibly interacting systems by sedimentation velocity. Methods. 2011; 54:16–30. [PubMed: 21315155]

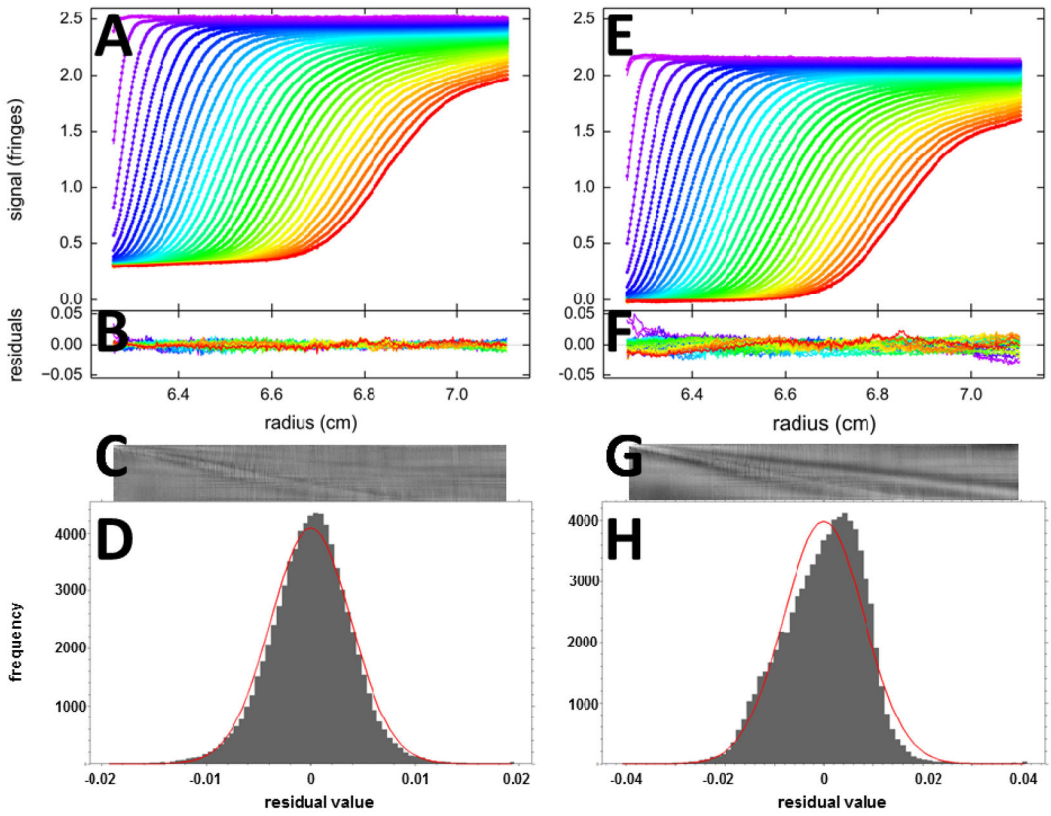

Figure 1.

Fits with two different models for SV data of BSA sedimenting at 50,000 rpm: (**A – D**) A *c*(*s*) analysis leads to a good fit with rmsd of 0.0039 fringes (Z-value 160); (**E – H**) a two discrete species model leads to an inadequate fit with rmsd of 0.0081 fringes (Z-value 224). The top panels (**A**, **E**) show the raw data (points, only every 2nd scan shown) and best-fit profiles (lines). Panels **B** and **F** are an overlay of the residuals, and **C** and **G** are the corresponding residual bitmaps (with time indicated by row number, and radius by column number of pixel). The residuals histogram is shown in panels **D** and **H**, with the grey bins indicating the frequency of residuals of a certain magnitude, and the red solid line the ideally expected distribution of residuals for normally distributed noise with the same rmsd, leading the H-numbers of 0.6% and 3.2%, respectively.