WE MAKE LIVES BETTER

UT HEALTH SCIENCE CENTER™ SAN ANTONIO

Characterization of shape, mass and reversible interactions in complex protein mixtures by analytical ultracentrifugation

Borries Demeler and Emre Brookes, Department of Biochemistry, UTHSCSA

Abstract

We present a survey of recent developments in the field of hydrodynamic analysis. New methodologies for extracting ever more detail from sedimentation velocity experiments are shown. Using several biopolymer systems, we demonstrate the utility of Analytical Ultracentrifugation (AU) to investigate complex mixtures, protein aggregation, and multi-component assemblies by extracting information about mass, conformation, reaction coefficients, rate constants and composition [1]. The methods developed by us employ a 2-dimensional spectrum analysis to sweep the parameter space for conformational and molecular weight signals produced by solutes in the sedimentation experiment [2, 3]. A genetic algorithm approach optimizes this solution by applying parsimonious regularization [4] and Monte Carlo analysis [5]. Self- and heteroassociating systems can be analyzed by a new genetic algorithm for fitting discrete nonlinear models to obtain molecular weight, equilibrium constants and reaction constants. To improve statistics, multiple experiments can be analyzed globally. High-performance computing (HPC) solutions are applied to reduce computation time [6, 7].

Introduction

Many biomedical research projects investigating the molecular basis for diseases or fundamental biochemical mechanisms focus on the understanding of dynamic interactions between molecules. AU is the method of choice to investigate macromolecular systems in the solution phase. AU provides information about the hydrodynamic properties of macromolecules by exposing them to a large centrifugal force field and measuring the macromolecular sedimentation and diffusion transport over time. AU data are modeled by finite element solutions of the flow equation [1] and analyzed on a supercomputer [6, 7]. Macromolecular concentration, buffer pH, reduction potential, composition and ionic strength can be adjusted to match physiological conditions. In addition, molecules can be examined in the presence of cofactors, ligands and drugs. The concentration distributions of the analytes in the AUC cell are recorded during the experiment and interpreted by the analysis software to derive sedimentation and diffusion coefficients, molecular weight, relative amounts of each analyte and its relative shape.

Our methods are implemented in the Ultrascan software [8], a comprehensive data analysis toolkit for AU experiments. We have recently added several new high-resolution data analysis methods to this software. The unique innovation in these tools is their utilization of parallel computing technology to accelerate time consuming computation. In addition, we have developed a convenient web-based user interface permitting access to remote supercomputers using standard Internet browser software. Our parallel implementation of algorithms on a supercomputing grid provides significant speedup and permits analysis at levels of resolution which were not practical in the past. The data flow between the various UltraScan modules is shown in Figure 1.

Methodology

Our approach consists of performing a sequence of optimization steps to arrive at a high-resolution description of the composition represented by a sedimentation velocity experiment. Our fitting procedures consist of finding the correct values for n, c_i, s_i and D_i during the minimization process, which can be stated as follows:

$$\boldsymbol{M} = \sum_{l=s_{min}}^{s_{max}} \sum_{m=f_{min}}^{f_{max}} c_{l,m}\, L\left(s_l, D\left(s_l, k_m\right)\right) \qquad Min \sum_{i=1}^{r} \sum_{j=1}^{t} \left[\boldsymbol{M}_{ij} - \boldsymbol{b}_{ij}\right]^2$$

where our model $\boldsymbol{M}$ represents a superposition of n ASTFEM Lamm equation solutions L [1], which are parameterized by the sedimentation coefficient, s, and the frictional ratio k. $\boldsymbol{b}$ is the vector of experimental data points over time t and radius r, and the solution is given by the minimum of the $l2$-norm, which is solved using the NNLS algorithm [9] or the genetic algorithm optimization method. NNLS reports non-negatively constrained amplitudes c, or zero for solutes that are not present in the solution. For reacting systems, $\boldsymbol{M}$ is replaced by discrete nonlinear models incorporating reaction terms.

1. Initialization: The s-value range is initialized using the enhanced van Holde – Weischet method [9], and the frictional ratio is typically set to values between 1 – 4 (Figure 1).

2. 2DSA, TI noise elimination, and meniscus fitting: In the next step, we perform a single-pass 2DSA analysis, and simultaneously eliminate time invariant noise [10] and fit the meniscus position by typically evaluating a 10-point meniscus grid and fitting the position vs. χ^2 to a 2nd order polynomial (Figure 2).

3. 2DSA Monte Carlo Analysis: In the next step we perform a 100 iteration Monte Carlo analysis using the 2DSA method. This approach amplifies the signal contained in the data and provides a refined view of the parameter surface (Figure 3).

4. GA analysis and parsimonious regularization: In this step, regularization is applied to eliminate false-positive solutes identified in the 2DSA analysis. The GA search space is initialized with the results from steps 2 or 3 (Figure 4).

5. GA Monte Carlo Analysis: Now a Monte Carlo analysis is performed to obtain statistical descriptions of each fitted parameter. This provides the necessary confidence intervals and facilitates data interpretation (Figure 5).

6. Global Multi-Speed Analysis: A third GA Monte Carlo can be used to further improve the confidence intervals of the data obtained in step 5. By globally analyzing multiple speeds, the signal-to-noise ratio is significantly improved and ultimate resolution can be obtained (Figures 6 + 7).

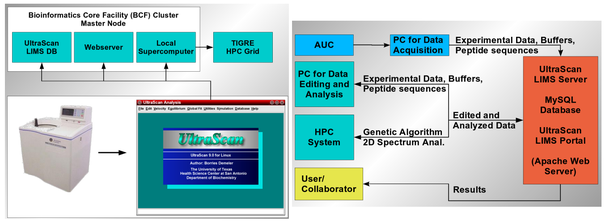


Figure 1: Data flow between between Instrument, UltraScan, LIMS and HPC system

Experimental Data: A mixture of Lysozyme and a 208-bp linear DNA fragment were mixed at about 40:60 optical density and sedimented at 40,000 rpm. The resulting data were analyzed with our approach. Analysis results for 2DSA and GA MC are shown in Figure 2, fitting statistics are summarized in Table 1..

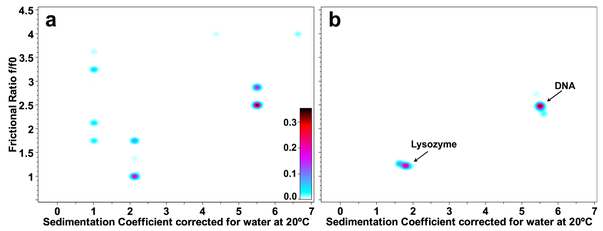


Figure 2: 208 bp DNA/Lysozyme mixture analyzed by low resolution 2DSA (a) and high resolution 2DSA Monte Carlo analysis (b). Color gradient shows optical density units and applies to both plots.

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

Table 3: Results for high resolution 2DSA Monte Carlo analysis of 208 bp DNA/Lysozyme mixture. Values in round parenthesis are 95% confidence intervals, values in square brackets refer to known molecular weights.

Simulated Data: We show a 5-component aggregating system, simulated with realistic noise representing an aggregating system in an end-to-end fashion with heterogeneity in shape and molecular weight. The parameters for the simulated data are shown in Table 1. Simulation was done for 60 and 20 krpm and 60 scans. Results for each step of our analysis are shown in Figure 4. Monte Carlo statistics are shown in Table 2.

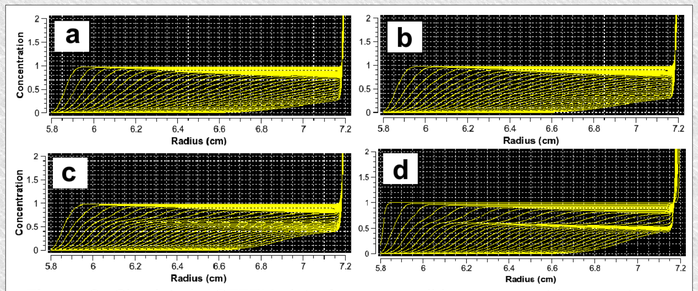


Figure 3: Simulated ASTFEM data for a reversible monomer-trimer system. Differences in sedimentation profiles are observed for different kinetic rate constants: a) k_{off} = 1/sec; (b) k_{off}=1x10^{-3}/sec; (c) k_{off}=1x10^{-4}/sec; (d) non-interacting

References

[1] Cao, W and B. Demeler. Modeling Analytical Ultracentrifugation Experiments with an Adaptive Space-Time Finite Element Solution for Multi-Component Reacting Systems. Biophys. J. (2008) 95(1):54-65

[2] Brookes, E., Boppana, R.V., and B. Demeler. (2006) Computing Large Sparse Multivariate Optimization Problems with an Application in Biophysics. Supercomputing '06 ACM 0-7695-2700-0/06

[3] Brookes, E., W. Cao and B. Demeler. A 2-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. Eur. Biophys. J., (in press)

[4] Brookes, E and B. Demeler. Parsimonious Regularization using Genetic Algorithms Applied to the Analysis of Analytical Ultracentrifugation Experiments. GECCO Proceedings ACM 978-1-59593-697 4/07/0007 (2007)

[5] Demeler, B. and E. Brookes. Monte Carlo analysis of sedimentation experiments. Colloid Polym Sci (2008) 286(2) 129-137

[6] Brookes, E. and B. Demeler. Parallel computational techniques for the analysis of sedimentation velocity experiments in UltraScan. Colloid Polym Sci (2008) 286(2) 138-148

[7] Demeler, B. (2008) A Teragrid Science Gateway for the analysis of hydrodynamic data: UltraScan Laboratory Information Management System (USLIMS), http://www.teragrid.org/programs/sci_gateways/

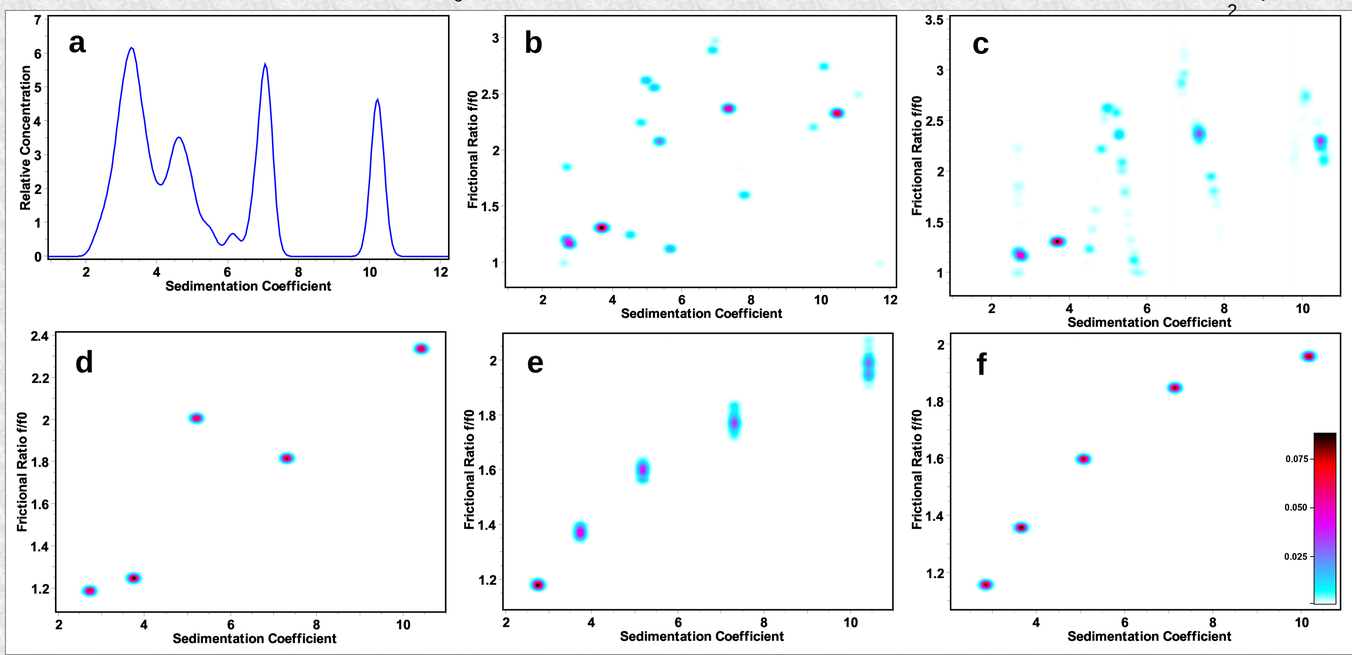
[8] Demeler, B. UltraScan A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments. Modern Analytical Ultracentrifugation: Techniques and Methods. D. J. Scott, S.E. Harding and A.J. Rowe. Eds. Royal Society of Chemistry (UK) (2005) 210-229

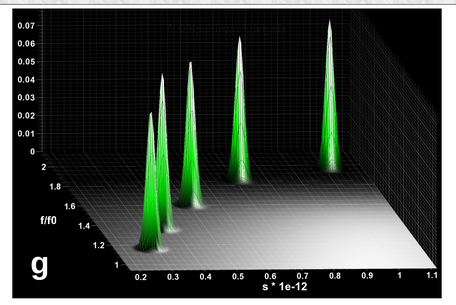
[9] Lawson, C. L. and Hanson, R. J. Solving Least Squares Problems. (1974) Prentice-Hall, Inc. Englewood Cliffs, New Jersey

[10] Demeler, B. and K. E. van Holde. Sedimentation velocity analysis of highly heterogeneous systems. (2004) Anal. Biochem. Vol 335(2):279-288

[11] Schuck, P. and B. Demeler. Direct Sedimentation Boundary Analysis of Interference Optical Data in Analytical Ultracentrifugation. (1999) Biophys. J., 76:2288-2296

Acknowledgments: This research has been supported by NIH grant RR-022200 to B.D.





Solute	Partial Concentration (OD)	Molecular Weight (kDa)	Shape (f/f_0)
1	0.1	25	1.2
2	0.1	50	1.4
3	0.1	100	1.6
4	0.1	200	1.8
5	0.1	400	2.0

Table 1: Aggregating 5-component system

Figure 4: Simulated 5 component system – step-by-step analysis, showing increasing precision and accuracy of the results with each step in the analysis. The color scale is proportional to optical density, with a value of 0.1 representing the maximum. All panels except **a** and **g** have the same color scale.

Solute	*Molecular Weight (kD)*	*Partial Concentration*	*Frictional Ratio, f/f_0*
1	24.26 (24.20, 24.33) [25]	0.0972 (0.0966, 0.0982) [0.1]	1.21 (1.21, 1.21) [1.2]
2	48.04 (47.74, 48.46) [50]	0.102 (0.101, 0.104) [0.1]	1.41 (1.40, 1.42) [1.4]
3	100.2 (97.96, 101.8) [100]	0.0995 (0.0982, 0.101) [0.1]	1.65 (1.63, 1.67) [1.6]
4	198.0 (194.2, 200.8) [200]	0.0996 (0.0989, 0.101) [0.1]	1.84 (1.82, 1.86) [1.8]
5	385.3 (380.4, 394.0) [400]	0.100 (0.100, 0.101) [0.1]	2.01 (1.99, 2.04) [2.0]

Table 2: Monte Carlo Results from a global genetic algorithm optimization using multi-speed data. The results demonstrate remarkable agreement with the original target model. Round brackets: 95% confidence intervals; square brackets: target value. All values rounded off to 3 or 4 significant digits.

a) van Holde – Weischet G(s) distribution used for initialization of s-value range.

b) 2DSA analysis of 60 krpm data. The solution includes false positives but covers well the range of the target solution.

c) Monte Carlo Analysis using 2DSA. As can be seen, the Monte Carlo analysis amplifies the signal to noise ratio, and reduces the contribution of stochastic noise.

d) GA analysis of results obtained in Figure 4b/c. Using parsimonious regularization, false positives are eliminated from the 2DSA solution, and 5 solutes are evident.

e) GA Monte Carlo analysis of results obtained in Figure 4d. Most of the variation is in the frictional ratio, not the sedimentation coefficient, because of high speed used here.

f) Sedimentation coefficient distributions of global multi-speed GA Monte Carlo analysis. Results have extremely narrow standard deviation due to the additional information from the low speed data (20 krpm).

g) 3-D representation of the same data shown in Figure 4 f.