Globular or end-to-end? A two-dimensional Spectrum Analysis for Assessing Modes of Aggregation from Sedimentation Velocity Experiments

Emre Brookes, Dept. of Computer Science, University of Texas at San Antonio
Borries Demeler, Dept. of Biochemistry, University of Texas Health Science Center at San Antonio

Abstract

The mode of subunit polymerization, and the structure, molecular weight and relative abundance of the intermediates is of considerable interest to researchers in diverse fields. The determination of the molecular weight and shape distributions in solution has so far been difficult to obtain, especially when many different components are present together. We describe here the 2-dimensional spectrum analysis which is used to analyze sedimentation velocity experiments. The goal is the characterization of the mode of aggregation. Our approach is suitable for heterogeneous mixtures even with broad s and MW distributions, and we describe how global analysis of multiple velocity experiments can be used to improve the results. Using several example systems and simulated data we show how this method can be used to distinguish between an end-to-end fibril-type aggregation or a globular agglutination. The 2-dimensional spectrum analysis is particularly well suited for parallelization to help with the distribution of the required memory and the CPU load. We describe how a fine-grained resolution of the parameter space can be efficiently searched in parallel on a supercomputer.

Introduction

Several approaches have been described previously to obtain molecular weight and shape information from sedimentation velocity data. However, well known drawbacks exist with each approach. For example, whole boundary fitting with nonlinear least squares methods is unreliable for cases where more than 2-3 solutes are involved. For cases where heterogeneity in shape exists, the C(s) method reports incorrect molecular weight distributions. Genetic algorithms manage to overcome these shortcomings but at the expense of a large amount of compute effort necessitated by the large search space. Motivated by the need to devise a better initialization for the genetic algorithms, we developed the 2-dimensional Spectrum Analysis method, which is a general and nearly model-independent approach for modeling experimental sedimentation velocity data showing ***both*** heterogeneity in the sedimentation coefficient and also heterogeneity in shape. As the results presented here demonstrate, it is a rather useful analysis method in its own right. Our approach is to model experimental sedimentation velocity data with a superposition of finite element solutions of the Lamm equation, covering both the sedimentation and shape domains. Each solution describes a non-interacting solute as modeled by the highly accurate and efficient ASTFEM solution [CD05]. For such a mixture of noninteracting solutes, the total concentration C_T of all solutes n in the ultracentrifuge cell can be represented by a sum of Lamm equation solutions L:

$$C_T = \sum_{i=1}^{n} c_i \, L(s_i, D_i)$$

where c_i is the partial concentration, s_i the sedimentation coefficient, and D_i is the diffusion coefficient of each solute in the mixture. Using a non-negatively constrained linear least squares fitting approach [LH74] we solve the minimization problem of a 2-dimensional search over both sedimentation coefficient and frictional ratio ff_0.

Methodology

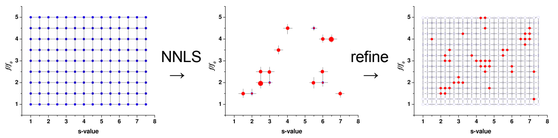
When fitting experimental velocity data the challenge then consists of finding the correct values for n, c_i, s_i and D_i during the minimization process. It is convenient to parameterize the diffuson coefficient using the frictional ratio, $k=ff_0$, and a corresponding sedimentation coefficient using the function:

$$D = \frac{RT}{N\,k\,6\pi\eta}\left(\frac{9\,s\,k\,\bar{v}\,\eta}{2(1-\bar{v}\,\rho)}\right)^{-0.5}$$

The minimization problem can then be stated as:

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

Where $\boldsymbol{M}$ is the model describing the mixture of solutes which can be heterogeneous in both shape and molecular weight, and $\boldsymbol{b}$ is the vector of experimental data points over time t and radius r, and the solution is given by the minimum of the l^2-norm. This is a linear problem with can be solved with the NNLS algorithm [LH74], which reports non-negatively constrained coefficients c or zero for solutes that are not present in the solution. It is possible to make reasonable estimates for the two variables s and $k=ff_0$ using model independent approaches. The s-value range can be well defined by using the enhanced van Holde – Weischet method [VW78, DV04], and the frictional ratio can be chosen based on an a priori estimate of the shape range, which is generally well described by a value between 1 – 4 for proteins, and 1 - 10 for extended or rod shaped molecules like linear DNA. A 2-dimensional mesh is then built where each node in the mesh represents one s, ff_0 pair:



Non-zero elements are saved, a new grid, slightly shifted, is calculated and the non-zero results are saved as well. This process is repeated until sufficient resolution has been obtained. After all mesh points have been evaluated, the non-zero elements are added back to all original grids and are recalculated until no further change is observed. The coefficients, c_i, then represent the relative optical signal each solute contributes to the total concentration. After refinement has been completed, a fine-grained mesh is obtained that only contains non-zero entries for components present in the solution, describing their respective sedimentation and frictional properties. Further refinement of the solution is possible by regularizing the degenerate solution space using either maximum entropy regularization or by using the using the genetic algorithm optimization [BD05]. In that case, the output parameters from the 2-dimensional spectrum analysis serve to define an initialization domain for the sedimentation and frictional parameter space for the genetic algorithm optimization.

Results:

We have applied the 2-dimensional spectrum analysis with time-invariant noise decomposition to several representative experimental systems to demonstrate the versatility and reliability of this method. We attempted to select systems that show strong heterogeneity in sedimentation coefficients and shape. In all cases variances obtained with the 2-dimensional spectrum analysis were lower than those obtained with any other method. To demonstrate the ability to describe a system displaying a strong heterogeneity in both sedimentation and shape we have chosen to analyse a restriction digest of a 2.7 kb plasmid. We found that the 2-dimensional spectrum analysis well describes the increase in frictional parameters with increasing molecular weight and also reproduces fairly accurately the molecular weight of the fragments in the solution, and their relative concentration, despite the large number of different sized fragments. In contrast, in order to represent a system heterogeneous in molecular weight but not in shape we chose two systems: 1. a molybdenum-iron compound ("$Mo_{27}Fe_{30}$, Keplerate") [L02] which aggregate into vesicle-like spherical clusters and 2. a colloidal dispersion of discrete spherical silica particles ("Bindzil 30/360") used as a flocculant. Both systems showed a clear heterogeneity in s, but complete homogeneity in the frictional parameters, with a frictional ratio of ~ 1.0 for all sedimenting species. This result is further confirmed with electron micrographs which show spherical shapes for both systems. It is well known that DNA in low salt buffer behaves like an extended, linear molecule in solution. With these data we have shown that shape and sedimentation properties of aggregating species can be well described by the 2-dimensional spectrum analysis. With the simultaneous determination of sedimentation and shape parameters it is therefore possible to identify if the system aggregates in a fibril, end-to-end fashion or a globular assembly.

References

[CD05] Cao W. and Demeler B. Modeling analytical ultracentrifugation experiments with an adaptive space-time finite element solution of the Lamm equation. (2005) Biophys J. 89(3):1589-602.

[BD05] Brookes, E. and Demeler B. Genetic Algorithm Optimization for obtaining accurate Molecular Weight Distributions from Sedimentation Velocity Experiments. Analytical Ultracentrifugation VIII, Progr. Colloid Polym. Sci. 131:78-82. C. Wandrey and H. Cölfen, Eds. Springer (2006)

[VW78] van Holde, K.E. and W. Weischet. Boundary Analysis of Sedimentation-Velocity Experiments with Monodisperse and Paucidisperse Solutes. 1978. Biopolymers 17:1387-1403

[DV04] Demeler, Borries and Kensal E. van Holde. Sedimentation velocity analysis of highly heterogeneous systems. 2004. Anal. Biochem. Vol 335(2):279-288

[LH74] Lawson, C. L. and Hanson, R. J. 1974. Solving Least Squares Problems. Prentice-Hall, Inc. Englewood Cliffs, New Jersey

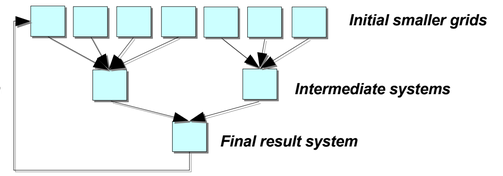
[L02] Liu, T. Supramolecular structures of polyoxomolybdate-based giant molecules in aqueous solution. 2002. J. Am. Chem. Soc. 124:37 10942-3

Acknowledgments

We would like to thank Karel Planken (Van 't Hoff Laboratory for Physical and Colloid Chemistry, Utrecht University, Netherlands) for kindly providing experimental data for Bindzil and the molybdenum-iron compounds, and Virgil Schirf and Mandy Rolando for the DNA experimental data. This research has been supported by NSF Grant DBI-9974819 and the San Antonio Life Science Institute with Grant #10001642

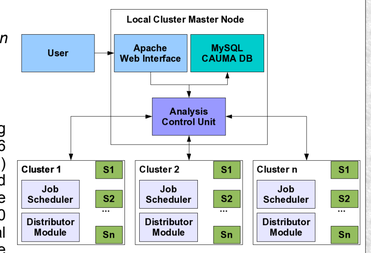
Parallelization:

Considerable improvement in information content is observed when multiple sedimentation experiments are analyzed simultaneously, especially for heterogeneous systems. For example, it is useful to combine several velocity experiments of the same sample conducted at multiple speeds in a combined fit. In such global multi-speed experiments the low-speed experiments produce a good signal on the diffusion coefficients, since the solutes have a lot of time to diffuse, while sedimentation and partial concentration signals are of the highest quality at high speed. A parallel supercomputer implementation of the 2-dimensional spectrum analysis can be used to globally fit multiple velocity experiments in an efficient manner by distributing the large memory and CPU needs over multiple processors which share in the computational load. Here we compare results from single speed experiments with those from multi-speed experiments. The standard method for single experiments is to model the proposed solutes and find the best fit constrained linear combination of each proposed solute to the experimental data. Extending this method to handle multiple experiments is simple. Each additional experimental data set can be concatenated to the end of vector b, increasing its length. For each proposed solute, additional models must be computed for each experiment. These are similarly concatenated in model $\boldsymbol{M}$ in the same order as the experimental data was concatenated to the end of vector $\boldsymbol{b}$. From this point, the computation proceeds identically to the single experiment case. In order to parallelize the 2-dimensional spectrum analysis, it is possible to split the solution grid into smaller sub-grids, process each grid separately, union the parameters with non-zero concentrations from each smaller grid into a new system of proposed solutes and repeat the standard methodology. In some cases, even this unioned system maybe too large, so it is necessary to further split the results in an inverted tree fashion:



In the iterative refinement approach, the final result system is unioned back into the initial smaller grids and the procedure is repeated until the final result is identical to the previous final result. Each smaller grid and intermediate union system can be placed on a separate processor and can thus run in parallel. Of course the dependence on previous results requires the intermediate and final unions can not begin until their antecedents have completed. This results in a processor utilization profile of high use during the processing of the initial smaller grids and finally one processor use during the computation of the final union. With large numbers of smaller grids, this arrangement provides excellent speedup compared to running the system serially on one processor. In our implementation of the parallelization of 2DSA, we have created a web interface to our system. The user uploads his experimental data to the web interface at http://bcf.uthscsa.edu/ultrascan. From there, data sets can be selected for analysis with the parallel version of 2DSA. Once the 2DSA parameters have been selected, the request is submitted to a queue. When the request is selected, it is given to the analysis control unit, which then sends the request to an available cluster. Within the selected cluster, a parallel program is started and each smaller grid is handed to an available processor. A master process keeps track of what is processing and collects up intermediate unions and hands those to available processors. When the final result is computed, the user is emailed with the results and the cluster is available for the next request.

Schematic for the parallel grid implementation of SA2D on the UTHSCSA cluster:

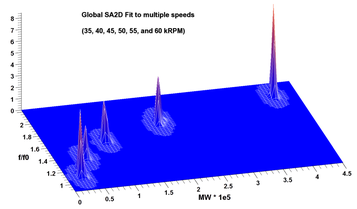


Experimental Results:

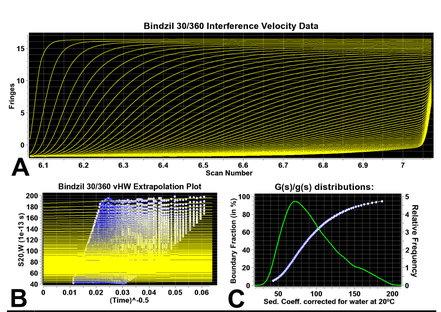
We simulated a 5-component, aggregating system using 1 % Gaussian noise for 6 different speeds (35, 40, 45, 50, 55, 60 krpm) and fitted the data using the SA2D method using 0.9 regularization. The data are analyzed either as a single speed run (50 krpm) or all speeds combined as a global multi-speed experiment and compared the results, which are tabulated below:

	Simulated		*50 krpm*		*35, 40, 45, 50, 55, 60 krpm*	
Solute	*MW*	*ff_0*	*MW*	*ff_0*	*MW*	*ff_0*
1	25 kDa	1.2	32.9 kDa	1.23	25.06 kDa	1.11
2	50 kDa	1.4			49.10 kDa	1.39
3	100 kDa	1.6	126.0 kDa	1.87	99.61 kDa	1.61
4	200 kDa	1.8	209.3 kDa	1.86	200.5 kDa	1.79
5	400 kDa	2.0	407.1 kDa	2.02	410.0 kDa	2.03

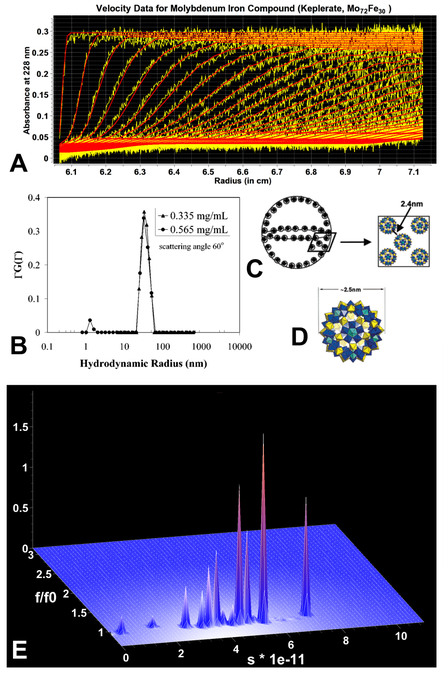
Discussion



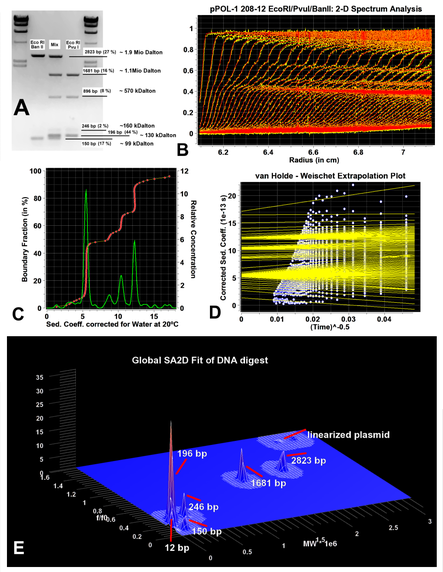
We have shown that global fitting of multi-speed sedimentation velocity data provides significant additional signal for an improved interpretation of the results when fitting with the 2-dimensional spectrum analysis, and when compared to single speed experiments. In addition, the approach described here makes use of parallel computing technology which significantly reduces computational time by exploiting intrinsic parallelisms in the 2-dimensional spectrum analysis method through distributing computational load as well as memory over multiple CPUs. Because of the sensitivity of the method even very small noise components will be considered in the model, and it is important to apply regularization to condition the results. In our experience a 0.9 – 0.95 regularization generally yields good results. By using multi-speed global fitting, the 2-dimensional spectrum analysis achieves a far higher resolution in both the sedimentation and frictional domain than other methods. By resolving appropriate frictional ratios for each species, this method leads to more reliable molecular weight distributions.



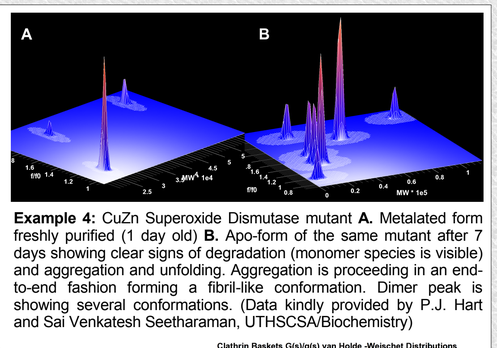
Example 1: Bindzil 30/360 (an aqueous dispersion of SiO_2 particles) shows heterogeneity in s, but not in shape. **A.** Experimental Rayleigh interference data. **B.** van Holde – Weischet extrapolation plot **C.** van Holde – Weischet integral distribution **D.** 2-dimensional Spectrum Analysis.



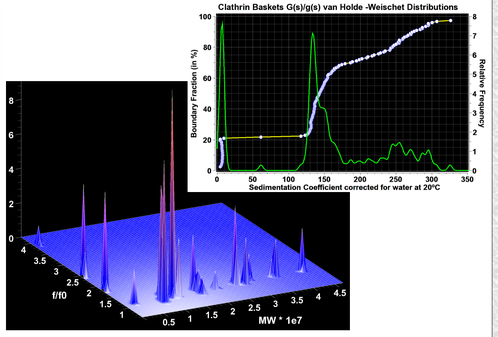
Example 2: Keplerate ($Mo_{72}Fe_{30}$) forms spherical micelles in solution (f/f0 = 1.0). It shows heterogeneity in s, but not in shape. **A.** SA2D fit of expt. data **B.** Light scattering data **C.** Schematic drawing of polymer assembly **D.** $Mo_{72}Fe_{30}$ monomer **E.** 2-dimensional Spectrum Analysis.



Example 3: DNA digest of pPOL-1 208-12: Each fragment is a highly elongated chain where frictional ratio strongly increases with molecular weight. **A.** 1% agarose gel of the digest **B.** SA2D fit of expt. data **C.** van Holde – Weischet Integral/differential distribution plot of velocity data **D.** van Holde – Weischet extrapolation plot **E.** Individual fragments can be distinguished in the 2-dimensional spectrum Analysis



Example 4: CuZn Superoxide Dismutase mutant **A.** Metalated form freshly purified (1 day old) **B.** Apo-form of the same mutant after 7 days showing clear signs of degradation (monomer species is visible) and aggregation and unfolding. Aggregation is proceeding in an end-to-end fashion forming a fibril-like conformation. Dimer peak is showing several conformations. (Data kindly provided by P.J. Hart and Sai Venkatesh Seetharaman, UTHSCSA/Biochemistry)



Example 5: Clathrin baskets. Sample shows a large heterogeneity of different sized baskets that assume a mostly globular form with a unity frictional ratio. The sample also displays several nonglobular species which represent the building block subunits required for assembly of intact baskets. (Data kindly provided by E. Lafer, UTHSCSA, Dept. of Biochemistry)