Genetic Algorithm Optimization for obtaining accurate Molecular Weight Distributions from Sedimentation Velocity Experiments

Emre Brookes[*] and Borries Demeler[#]

[*] The University of Texas at San Antonio, Dept. of Computer Science

[#] The University of Texas Health Science Center at San Antonio, Dept. of Biochemistry

Corresponding Author: Borries Demeler

Abstract:

Sedimentation experiments can provide a large amount of information about the composition of a sample, and the properties of each component contained in the sample. To extract the details of the composition and the component properties, experimental data can be described by a mathematical model, which can then be fitted to the data. If the model is nonlinear in the parameters, the parameter adjustments are typically performed by a nonlinear least squares optimization algorithm. For models with many parameters, the error surface of this optimization often becomes very complex, the parameter solution tends to become trapped in a local minimum and the method may fail to converge. We introduce here a stochastic optimization approach for sedimentation velocity experiments utilizing genetic algorithms which is immune to such convergence traps and allows high-resolution fitting of nonlinear multi-component sedimentation models to yield distributions for sedimentation and diffusion coefficients, molecular weights, and partial concentrations.

Keywords: analytical ultracentrifugation, sedimentation velocity analysis, genetic algorithms

Abbreviations: GA=genetic algorithm, RMSD=residual mean square deviation, NNLS=non-negatively constrained linear least squares.

Introduction:

A commonly used approach for determining observable parameters from experimental sedimentation data involves building a mathematical model describing the transport of macromolecular species in the ultracentrifugation cell. In such a model, every feature of the system under study which results in an observable signal such as a change in absorbance is described by an adjustable parameter in the model. The parameters are adjusted by a fitting method that seeks to minimize the difference between the model and the experimental data in a least squares sense:

$$Min \sum_{i}^{n} \left[y_i - Y_i(\boldsymbol{P}) \right]^2 \qquad 1$$

Here, y_i represents each experimental data point i, and Y_i $(\boldsymbol{P})$ represents the corresponding data point simulated by the model as described by parameter vector $\boldsymbol{P}$. Models that are linear with respect to their parameters can be fitted by a generalized linear least squares approach in a single iteration, while models that are nonlinear with respect to their parameters have to be fitted in an approach that iteratively adjusts all parameters to an optimal value. The optimal parameter combination $\hat{\boldsymbol{P}}$ represents those parameters that minimize Equ. 1. For highly nonlinear problems, models with many parameters, and parameters that exhibit a high degree of correlation, the error surface described by Equ. 1 can be very complex and contain many local minima which tend to trap the optimization process. For such cases, conventional gradient descent methods such as Gauss-Newton, Levenberg-Marquardt, and quasi-Newton commonly will fail to converge to the global minimum, and will provide less than optimal solutions for the fit. Other problems associated with direct fitting of experimental data relate to the selection of the correct model function. Here, user input is required which introduces an unavoidable bias into the approach. To circumvent such caveats, other methods not relying on nonlinear least squares fitting have been explored. Not

unexpectedly, each method exhibits both advantages and shortcomings. We will briefly review three popular methods for sedimentation velocity experiments. Graphical transformations of sedimentation velocity data were introduced by van Holde & and Weischet [van Holde and Weischet, 1978] and later refined [Demeler & van Holde, 2004]. This approach yields model-independent, diffusion-corrected sedimentation coefficient distributions. However, diffusion coefficients and molecular weights, or frictional parameters can not be reliably obtained by this method. In another model-independent approach, Stafford [Stafford, 2000] introduced time-differencing of sedimentation velocity profiles to transform sedimentation velocity data to g(s) profiles. The main advantage of this approach is its capability to eliminate time-invariant noise contributions. Here too, molecular weights and diffusion coefficients are unavailable and sedimentation distributions are not deconvoluted from diffusion without further nonlinear least squares fitting. The C(s) method [Schuck, 2000] avoids a multi-dimensional nonlinear least squares fit by linearizing the problem over the sedimentation coefficient domain using a constant frictional ratio f/f_0 to provide a corresponding diffusion coefficient. In the latter approach, a linear combination of finite element solutions to the Lamm equation [Lamm, 1929] are fitted to the experimental data. The NNLS algorithm [Lawson & Hanson, 1974] is used to fit the linear coefficients of each term. The method finds all non-negative contributions, the remaining terms are set to zero. In an iterative approach, the fricitional coefficient (which makes this problem nonlinear) can be fitted as well, reducing the problem to a well-conditioned one-dimensional search. Because a frictional coefficient is available, diffusion coefficients can also be obtained from the fit and can be used to determine molecular weights. This approach works well for those systems where all components have very nearly the same frictional properties, and can be equally well described by a single frictional ratio. Such cases may be present when all components in the system are globular proteins. However, when components of appreciably different frictional ratios

are present, such as mixtures of globular proteins with different length aggregates, fibrils, extended DNA molecules, random coil conformations, or mixtures of any of the above, and complexes with intermediate frictional properties, a single frictional ratio will yield suboptimal results and is insufficient to describe the entire system accurately. When parameters from such a system are determined with the C(s) method, the determined molecular weights will be unreliable if only a single frictional coefficient is used to represent all components. Therefore, for such systems it is important to allow the frictional ratio or the diffusion coefficients to be adjusted independently for all components. Table 1 summarizes the results obtained from the analysis of a non-interacting mixture (Figure 1) of a linear 208 basepair DNA molecule and a small globular protein (lysozyme), with three different approaches: 1. a nonlinear least squares fitting approach of finite element solutions to the Lamm equations (overlays are shown in Figure 1), 2. the C(s) method, and 3. by the van Holde – Weischet method (results from all three methods are combined in Figure 2). Results from the nonlinear least squares fitting with finite element solutions of the Lamm equation allow parameters from both components in the system to be adjusted independently, and are in excellent agreement with the known molecular weights of the two components. As expected, two very different frictional ratios are observed. While the van Holde – Weischet method results in sedimentation coefficients closely matching the nonlinear least squares fitting results, molecular weight information or shape information is not available. When the C(s) method is used in conjunction with an adjustable frictional ratios, the ratio obtained can by design only represent the best average between the two species, and is therefore too high for lysozyme, and too low for DNA. When analyzed with no regularization, the C(s) method will introduce artifactual peaks around the lysozyme peak, and when regularization with an F-ratio of 0.95 is used, the peak is strongly broadened and the peak appears off-center. Furthermore, the C(s) analysis results in an RMSD 36% higher than the RMSD from the finite element fit. When the frictional ratio is

combined with the sedimentation coefficient distribution, this value results in incorrect molecular weights for both species. Even after correction with the appropriate partial specific volumes the molecular weight of lysozyme is predicted too high, while the molecular weight of the DNA molecule is predicted too low. Therefore, we conclude that for systems involving components with different frictional ratios the C(s) method is not the appropriate approach. In addition, using a constant frictional term for the description of all species in the cell further complicates the interpretation of the results when molecules with different densities are examined (for example, mixtures of nucleic acids and nucleic acid binding proteins). From the results shown in Table 1 it may appear that the nonlinear least squares fitting approach with finite element solutions of the Lamm equation will provide the best information possible. In the example shown here, a bimodal fit clearly results in a satisfactory solution, but this result will not be observed in the general case. The same approach will almost always fail for cases where more than 2 or 3 components are fitted and all relevant parameters are adjusted simultaneously (data not shown). The reason for this failure can be traced to the nonlinear least squares fitting approaches which are very sensitive to initial parameter guesses and the presence of local minima in the error surface of the fitting function. The problem becomes especially apparent when many nonlinear parameters are present, or the parameters are highly correlated. Such circumstances result in very complex error surfaces and often prevent convergence at the global minimum. Our goal is therefore to provide a method that combines a model-independent approach with the generality of directly and independently fitting each parameter's sedimentation and diffusion coefficient with an optimization method robust enough to reliably identify global minima in the error surface described by a multi-dimensional parameter space.

Proposed Method:

Here we propose an alternative approach to the global parameter optimization by nonlinear least squares fitting by introducing stochastic methods for parameter estimation. Stochastic methods do not rely on gradient descent in the error surface to find the best-fit solution, but instead introduce a random parameterization of the search space which allows the methods to probe for solutions by randomly placing parameter values within a constrained domain. Although these methods are computationally more expensive, they allow the solution to escape or ignore local minima and provide a higher likelihood of finding the global minimum. Several approaches employing stochastic parameter estimation have been explored, among them are genetic algorithms (GA) [Holland, 1975, Goldberg, 1989, Koza, 1992], Monte Carlo methods and simulated annealing methods [Kirkpatrick et al., 1983 and Cerny, 1985]. For our study, we chose to explore Monte Carlo methods and GAs. A Monte Carlo approach is based on randomly selecting parameters from the search space, simulating a model function based on these parameters, and evaluating the fitness function. However, there is no particular bias towards any solution, so if the search space is very large, the computational effort can be prohibitive because the entire space needs to be evaluated. When GAs are employed, a bias is introduced to the generic Monte Carlo approach by implementing a selection process analogous to natural selection in evolution. An initial random population of individuals is simulated, and each individual's fitness is evaluated and the population is allowed to evolve. Parameter vectors are treated as genes which can exchange or modify parameters (bases) by crossover with other parameter vectors or by mutation, insertion or deletion operators. Multiple populations (demes) can evolve independently, or experience a controlled migration rate, which allows for exchange of parameter information among multiple demes. Evolution of the best fit parameter combination within a population is controlled by a multi-generational selection process, which favors survival of individuals with a better fit. The survival pressure, migration rate, crossover frequency, mutation, insertion and deletion probability can be

independently controlled by random number operators, and each probability rate needs to be optimized for best efficiency. The fitness function is given by the l^2-norm of Equ. 1:

$$\frac{1}{n}\sqrt{\sum_{i}^{n}\left[y_i - Y_i(\boldsymbol{P})\right]^2} \qquad 2$$

Description of the Algorithm:

The algorithm proceeds as follows: To limit the search to a reasonable domain, the search space is initialized with a model-independent approach. We found that the van Holde – Weischet analysis [van Holde & Weischet, 1978, and Demeler & van Holde, 2004] best serves this requirement by providing a lower and upper limit of the sedimentation space, and by providing an approximate number of species in the system. Using the number of estimated components and the sedimentation coefficient range, the search space is partitioned accordingly, and a set of random individuals are initialized with sedimentation coefficients randomly selected between the limits projected by the van Holde – Weischet method. In order to form a complete parameter set for a given species *j*, a diffusion coefficient is needed as well. We chose to initialize diffusion coefficients based on a reasonable range of frictional coefficient ratios κ_j as a parameter to the sedimentation coefficient *s* for each species in the system. Equ. 3 shows the formula used to calculate the diffusion coefficient for species *j*:

$$D_j(s,\kappa_j) = RT\left(N\kappa_j 6\pi\eta\right)^{-1}\left(\frac{9 s_j \kappa_j \bar{\nu}\eta}{2(1-\bar{\nu}\rho)}\right)^{-1/2} \quad \text{and} \quad 1.0 \le \kappa_j \le \kappa_{max} \qquad 3$$

where *R* is the gas constant, *T* the temperature in Kelvin, *N* is Avogadro's number, η is the viscosity of the solvent, ρ is the density of solvent, and $\bar{\nu}$ is the partial specific volume of the molecule. Here, κ_j is assigned a random value in a reasonable range for the system under study. For

globular proteins, we suggest values between 1.0 and 3.0, for linear DNA fragments κ_{max} values as high as 15 are appropriate. The size of this initial set can vary between 50-500 individuals in each deme, and several demes can evolve simultaneously, depending on computational resources. Each pair of sedimentation and diffusion coefficients is used to calculate a finite element solution for the experimental conditions [Cao and Demeler, 2005] with a unity concentration factor. The NNLS algorithm [Lawson & Hanson, 1974] is used to solve Equ. 4, where $\boldsymbol{C}$ is the vector of coefficients c_j for each species j, and $\boldsymbol{L}$ is the matrix of Lamm equations used to model each individual component. The values of c_j correspond to the relative concentration of each species in the linear combination that forms an individual in the deme.

$$\boldsymbol{C}\,\boldsymbol{L}\left(s_j, \kappa_j\right) \approx \boldsymbol{B} \qquad 4$$

A fitness value is then computed for each individual in the population using the l^2-norm shown in Equ. 2. At this point, the first generation is completed and the GA is used to calculate the next generation. For each individual, progeny is generated by applying the GA operators on the parameter sets defining the individuals. Selection takes place by preferentially discarding individuals which display a poor fit, and maintaining the best individuals in the population. It is important to avoid dominance of one particular individual by always maintaining a range of fitness in the population, which will assure diversity in the parameter pool. The next generation is created by applying the GA operators to the selected individuals. Each operator is controlled by a probabilistic rate constant. Here, the deletion operator may delete a species from an individual, or a new species is added to an individual. Individuals containing fewer species are assigned a higher selection rate than individuals with more species and the same fitness value in order to reflect the lower computational cost of reproduction. Mutation operators may change the value of any one species' sedimentation or diffusion coefficient, and the crossover operator allows two parental individuals to randomly exchange information encoding one or more species. The fitness of the

individuals in the new generation is again evaluated, and the process is repeated until convergence is approached. Once the solution reaches an optimum, a nonlinear least squares optimization routine can be applied to quickly find the absolute global minimum in the vicinity of the GA solution. Once the solution is in the vicinity of the global minimum, any gradient descent will perform much more quickly in locating the global minimum, since the error surface is generally well conditioned in this much reduced parameter space. To alleviate the computational demand for solving this stochastic problem, we have implemented the GA on a parallel architecture using SUSE Linux 64 on a 44-node opteron cluster at the Department of Biochemistry at the University of Texas Health Science Center at San Antonio.

Results:

A typical problem that may be faced in a laboratory is the analysis of a DNA binding protein complex, with some free DNA and free protein present in the mixture. If the association kinetics are sufficiently slow, such a system can be simulated with a noninteracting model and will present a situation where all components will have different frictional ratios: A large frictional ratio for the free DNA, a globular frictional ratio for the free protein, and an intermediate frictional ratio for the complex. To verify the capability of the proposed method, we simulated such a system and added noise of comparable quality to that observed in the XL-A analytical ultracentrifuge to the solution. The parameters for this 3-species system are listed in Table 2. Data analysis results from the van Holde – Weischet method, the C(s) method, and the GA optimization are presented in Figure 3. As can be seen from this Figure, the results from the GA are nearly identical with the parameters used in the original simulation. We further compared the results from different speed simulations (20, 40 and 60 krpm, data not shown) and from globally fitting all three speeds simultaneously. Table 3 lists the results from these experiments and indicates the errors between the simulated parameters

and the parameters determined with the GA optimization. The results suggest that the highest speed provides the most reliable information for the partial concentration and sedimentation parameters, as long as enough signal for the fastest component can be collected. Adding slower speeds to a global multi-speed fit further improved the optimized parameters. To further test the capabilities of this method, we attempted to identify the components of an 8-species system, where the components cover a large range of molecular weights. Simultaneously, we wanted to challenge the resolving power of this method by simulating a system where two components may share the same molecular weight, but have markedly different frictional ratios. Such a condition may be presented by a system undergoing a conformational isomerization, an example may be a folded protein with a fraction of the protein in an unfolded or improperly folded state. A system displaying both heterogeneity in sedimentation and in frictional properties may be presented by aggregating proteins forming fibrils and other extended structures. The simulation parameters for the 8-component system are listed in Table 2, the results for the fit and errors are shown in Table 4, and a comparison of the partial concentrations and calculated molecular weights with those from the target distribution is shown in Figure 4. Here, the results suggest that the method can resolve well 8 components with disparate frictional ratios extending over a large sedimentation coefficient range. While sedimentation coefficients and partial concentrations are reproduced most reliably, the resolution of diffusion coefficients can be impaired when species with similar sedimentation coefficients but diverging diffusion coefficients are present. In all cases we compared evolutions spanning 100 generations and 500 individuals and 410 demes. At this point a well isolated solution had emerged in all cases, and further iterations did not improve the fit. It should be noted that the RMSD values from all individuals in each iteration provide a statistical sampling of the parameter space which can be used in place of Monte Carlo analysis results. An example of such a parameter distribution is shown in Figure 5 for the sedimentation coefficient distributions of the 8-component

system. We further observed a trend in the speed of convergence. For the case of the 3 component system, we noticed that convergence was most rapidly obtained in the 20 krpm and the global 20/40/60 krpm fits, followed by the 40 krpm fit, and then trailed by the 60 krpm fit. While all fits resulted in a convergence with essentially the same RMSD, the relative performance at different speeds suggests that different experimental speeds result in varying signal strengths that can be distinguished by the GA. The performance of a GA is measured by the rate of convergence to a target solution. Tuning the parameters controlling the GA has a major impact on performance. Let N be the population size times the number of generations. N is the total number of individuals tested throughout the run and is approximately proportional to the total running time. We evaluated the effect of several factors impacting the performance of the GA. First, we compared the performance of a small population and a large number of generations to the performance of a large population with a small number of generations. At the extremes, a population size of one with N generations would be useless, and a population size of N with one generation would be equivalent to random guessing. Neither case is optimal as can be seen from the results presented in Table 5. Next, we considered crossover and mutation rates together. In crossover, parts from two good individuals are taken and combined to create a new individual for the next generation. Mutation is applied by taking one individual and adding randomness to one or more parameters in this individual to create a new individual for the next generation. Naively, one may want high rates for both crossover and mutation operators, but our total population size is restricted, so high rates of mutation would hide any crossover benefit. Preliminary results indicate that crossover likely provides a benefit in finding solutions to sedimentation velocity experiments.

Discussion:

From this data it is clear that the GA method affords remarkable resolution in partial concentration,

sedimentation and diffusion coefficient determination for sedimentation velocity experiments. Therefore, the method excells at resolving molecular weights, even for systems with a relatively large number of components. In addition, GAs gain an increase in accuracy by globally fitting experiments conducted at multiple speeds. Insertion and deletion operators effectively and automatically control the selection of the appropriate model, removing the user bias associated with the selection of a fitting model, providing a model-independent and general approach for fitting sedimentation velocity experiments. Overall, the reproduced accuracy and resolution is unmatched by any other method available to us. We are currently investigating if GAs can also be applied to determine self-association properties and equilibrium constants, and if GAs can automatically determine the appropriate model for an interacting system. Future work in this area will focus on further reducing the search space and optimizing the GA itself. While the potential of this method for resolving individual species in a sedimentation velocity experiment are obvious from the presented data, GA calculations are computationally intensive and are best performed through parallel computation. Optimization of the method itself remains a requirement before this analysis method can be adopted on a routine basis. Although a poorly optimized algorithm will eventually arrive at the global optimum, convergence rates may vary drastically. Considering the computational expense of Lamm equation evaluations, a well-tuned GA is an important requirement. Our preliminary findings indicate that factors such as population size, the number of demes and generations, the rates for mutation, crossover, deletion, insertion, migration, and the method chosen for initialization all effect the convergence rate. Thus, we can consider these factors as parameters in a second optimization problem. Since the GA is a stochastic process, each parameter vector must be tested on multiple target systems. It is quite possible that performance tuning will suggest different search conditions for different systems, and that these parameters are dependent on the number of species, s and D distributions, and partial concentrations. Each

parameter needs to be evaluated with many trials using different random seeds. We are continuing to explore this parameter space. We further propose that this method could easily be extended to also include related experiments in a global analysis such as a combined analysis of sedimentation velocity and equilibrium experiments, and dynamic light scattering experiments. Here, the experiment can be described with a nonlinear least squares fitting model containing the same parameters used in the sedimentation velocity experiment (D and the ratio of s/D, which is proportional to molecular weight). In summary, we conclude that the GA approach excells in avoiding local minima convergence traps. From our analysis we conclude that the only limitation of the method is the signal contained in the data. As soon as changes in concentration needed for the determination of a parameter are masked by experimental noise, the limit of resolution has been reached. Highly correlated parameters such as multiple diffusion coefficients for molecules sedimenting with similar sedimentation coefficients have also proven to be difficult to resolve.

Model	$s_{20,W}$ *(x10^{-13})*	$D_{20,W}$ *(x10^{-7})*	f/f_0	*MW (kD)*	*Species*
Finite Element Fit	5.43	2.28	3.1	128.8 (130.7)	DNA
RMSD: 4.60x10^{-3}	1.71	10.2	1.29	14.6 (14.4)	Lysozyme
C(s) Fit,	5.19	3.67	2.29*)	76.65 (130.7)	DNA
RMSD: 6.26x10^{-3}	2.45	5.45	1.75*)	39.67 (14.4)	Lysozyme
van Holde – Weischet	5.35	N/A	N/A	N/A	DNA
	1.87	N/A	N/A	N/A	Lysozyme

Table 1

Comparison between sedimentation velocity methods. Data analysis parameters from the finite element and C(s) fitting method, and the van Holde – Weischet method. Molecular weights (MW) in parentheses indicate theoretical molecular weight.

*) frictional ratios are corrected for the partial specific volume of DNA and lysozyme, respectively.

Model	*Partial concentration*	*s*	*D*	*Molecular Weight (kD)*	f/f_0
3 species, 1	0.3	4.238×10^{-13}	7.343×10^{-7}	50	1.203
3 species, 2	0.2	5.943×10^{-13}	1.622×10^{-7}	198	3.765
3 species, 3	0.8	9.023×10^{-13}	1.872×10^{-7}	298	2.754
8 species, 1	0.1	2.481×10^{-13}	1.080×10^{-6}	20	1.112
8 species, 2	0.2	3.707×10^{-13}	6.423×10^{-7}	50	1.375
8 species, 3	0.3	4.587×10^{-13}	4.416×10^{-7}	90	1.644
8 species, 4	0.4	5.063×10^{-13}	3.147×10^{-7}	140	1.994
8 species, 5	0.3	1.053×10^{-12}	9.651×10^{-8}	945	3.435
8 species, 6	0.2	7.827×10^{-13}	1.514×10^{-7}	450	2.809
8 species, 7	0.15	1.760×10^{-12}	3.405×10^{-7}	450	1.249
8 species, 8	0.1	1.294×10^{-12}	1.502×10^{-7}	750	2.389

Table 2

Parameters used for the simulation of the 3- and 8-species velocity experiments. For the 8-species system, parameters were chosen to determine if components with the same molecular weight, but different shapes (species 6 and 7) and species with closely spaced sedimentation and diffusion coefficients (species 3 and 4) can be resolved. Also, the components were broadly spaced in *s* and molecular weight, to determine the limits of the method in its ability to resolve multiple species, even if they are closely spaced.

Parameter	*60 krpm*	*% Error*	*20/40/60 krpm*	*% Error*
Concentration 1:	0.2946	- 1.80 %	0.2989	- 0.37 %
Concentration 2:	0.2047	+ 2.35 %	0.1998	- 0.10 %
Concentration 3:	0.8009	- 0.11 %	0.8010	+ 0.13 %
Sed. Coeff. 1:	4.228×10^{-13}	- 0.24 %	4.236×10^{-13}	- 0.05 %
Sed. Coeff. 2:	5.915×10^{-13}	- 0.47 %	5.934×10^{-13}	- 0.15 %
Sed. Coeff. 3:	9.021×10^{-13}	- 0.02 %	9.016×10^{-13}	- 0.08 %
Diff. Coeff. 1:	7.239×10^{-7}	- 1.42 %	7.358×10^{-7}	+ 0.20 %
Diff. Coeff. 2:	1.899×10^{-7}	+ 17.01 %	1.6250×10^{-7}	+ 0.19 %
Diff. Coeff. 3:	1.890×10^{-7}	+ 0.96 %	1.875×10^{-7}	0.16%

Table 3

Parameters obtained from the average of 25 runs of 100 generations of the GA optimization for the simulated 3-species system shown in Table 2. The same system was simulated with a speed of 20 krpm (GA fit RMSD: 5.74×10^{-3}), 40 krpm (GA fit RMSD: 5.79×10^{-3}), and 60 krpm (GA fit RMSD: 5.84×10^{-3}). We show here the results from the 60 krpm experiment and the results from the global fit of multiple speeds (20, 40 and 60 krpm, GA fit RMSD: 4.05×10^{-3}). As can be seen from these results, not only is the overall percent error for parameter determinations smaller for a global multi-speed analysis, but the RMSD of the fit is smaller as well, indicating a better fit. The highest increase in accuracy results from the improvements in the estimation of the diffusion coefficients. We speculate that the longer run times in the low speed runs improve the signal for the diffusion and hence provide additional information to the global fit allowing a higher accuracy for diffusion determinations. Each run took approximately 7 minutes using 1 CPU.

Species	***Concentration***	***s***	***D***
1	0.100379 (- 0.38 %)	2.49×10^{-13} (+ 0.36 %)	1.062×10^{-6} (- 1.67 %)
2	0.198768 (- 0.62 %)	3.69×10^{-13} (- 0.46 %)	6.190×10^{-7} (- 3.63 %)
3	0.306232 (+ 2.08 %)	4.61×10^{-13} (+ 0.50 %)	4.190×10^{-7} (- 5.12 %)
4	0.394280 (- 1.43 %)	5.06×10^{-13} (- 0.06 %)	3.290×10^{-7} (+ 4.54 %)
5	0.198518 (- 0.74 %)	7.80×10^{-13} (- 0.34 %)	1.370×10^{-7} (- 9.51 %)
6	0.301060 (+ 0.35 %)	1.05×10^{-13} (+/- 0.0 %)	1.000×10^{-7} (+ 3.62 %)
7	0.099526 (- 0.47 %)	1.29×10^{-13} (- 0.46 %)	1.600×10^{-7} (+ 6.52 %)
8	0.151501 (+ 1.00 %)	1.75×10^{-13} (- 0.62 %)	5.340×10^{-7} (+ 56.83 %)

Table 4

Parameters obtained from the GA optimization for the 8-species system shown in Table 2 simulated with 60 krpm. At 60 krpm the errors for the sedimentation coefficient are much smaller than for the diffusion coefficient, indicating a stronger signal from sedimentation than diffusion. The resolution of the method is limited for diffusion coefficients when the species sediment very close together, even when the diffusion coefficients are far apart (species 8, see discussion). This multi-deme run took approximately 10 hours using 44 CPUs.

Population size	***Generations***	***Average RMSD***	***Best RMSD***
1000	1	0.01173	0.00965
100	10	0.00946	0.00902
50	20	0.00932	0.00906
20	50	0.00919	0.00903
10	100	0.00919	0.00901
5	200	0.00947	0.00901
2	500	0.04487	0.01776

Table 5

Performance observed for a fixed N (N = population size * generations). Targeted is the 3-species velocity experiment. For each population size and generations pair, 25 separate GA runs were performed with different random seeds. The lowest RMSD individual is chosen from each run giving 25 individuals from which the average and best RMSD are reported. The first case of a single generation can be considered a pure Monte Carlo method since no GA operators are applied. An intermediate population and generation size results in the most consistent performance improvement.

Figure 1:

Sedimentation velocity data of a 208 bp linear DNA fragment and lysozyme fitted to a finite element solution of a two-component non-interacting model. Such a mixture is representative of a system that exhibits two very different frictional ratios in a single experiment. Experimental datapoints are represented by open circles, the finite element solution is shown as continuous lines. Parameters for this fit are shown in Table 1.

Figure 2:

Sedimentation velocity analysis of the two-component system shown in Figure 1 by the C(s) method (regularization with F-ratio of 95%, solid line, no regularization: line with upside-down triangles, RMSD=0.0063), the van Holde – Weischet analysis (line with filled circles) and the direct boundary fit with the finite element method (stars, RMSD=0.0046). Due to the difference in frictional ratio, the C(s) method fails to provide reliable sedimentation coefficient distributions, while the van Holde – Weischet method approximates more closely the sedimentation coefficients observed in the finite element direct boundary fitting method.

Figure 3:

Sedimentation coefficient distributions reported by various data analysis methods when applied to the simulated 3-species system shown in Table 2 (60 krpm). Shown are the target values from the simulation (stars), the van Holde – Weischet analysis (dotted line), the C(s) analysis without regularization applied (solid line) and the results from the genetic algorithm (vertical bars). Residual bitmaps for the C(s) fit and the genetic algorithm fit are shown in insert on top left. The characteristic diagonal indicating systematic deviations that can be seen in the C(s) analysis bitmap is absent in the genetic algorithm fit, indicating that a more appropriate fit is obtained with the

genetic algorithm that allows for a variation in the frictional ratio.

Figure 4:

Molecular Weight distribution obtained from the genetic algorithm when fitting the simulated 8-component system listed in Table 2. The simulated target values are represented by stars, the values derived from the genetic algorithm optimization are shown as vertical lines. The vertical position of the stars and the height of the lines correspond to the partial concentration of each species. As can be seen from this graph, the genetic algorithm faithfully reproduces the number of components and the partial concentration of each component in the system, and in all but one cases closely matches the molecular weight of the target. The two targets at 450,000 dalton represent two species with the same molecular weight, but different frictional ratios and sedimentation coefficients. In this case, only one of the species could be resolved correctly.

Figure 5:

Monte Carlo parameter distributions derived from the 8-component genetic algorithm fit for the sedimentation coefficient distribution. Here the sedimentation coefficient distribution from the best-fit 5000 individuals are shown. The best definition is for well separated components in the center of the distribution which have the largest amplitude and the narrowest parameter spread. The best fit parameter combination is represented by the tips of each peak, and corresponds to an RMSD of 5.84×10^{-3}. The parameters at the bottom of the distribution are derived from those individuals with an RMSD similar to the RMSD of the worst individual in the top 5000 individuals. The area under each peak is exactly 5000 individuals, since all top 5000 individuals showed 8 species.

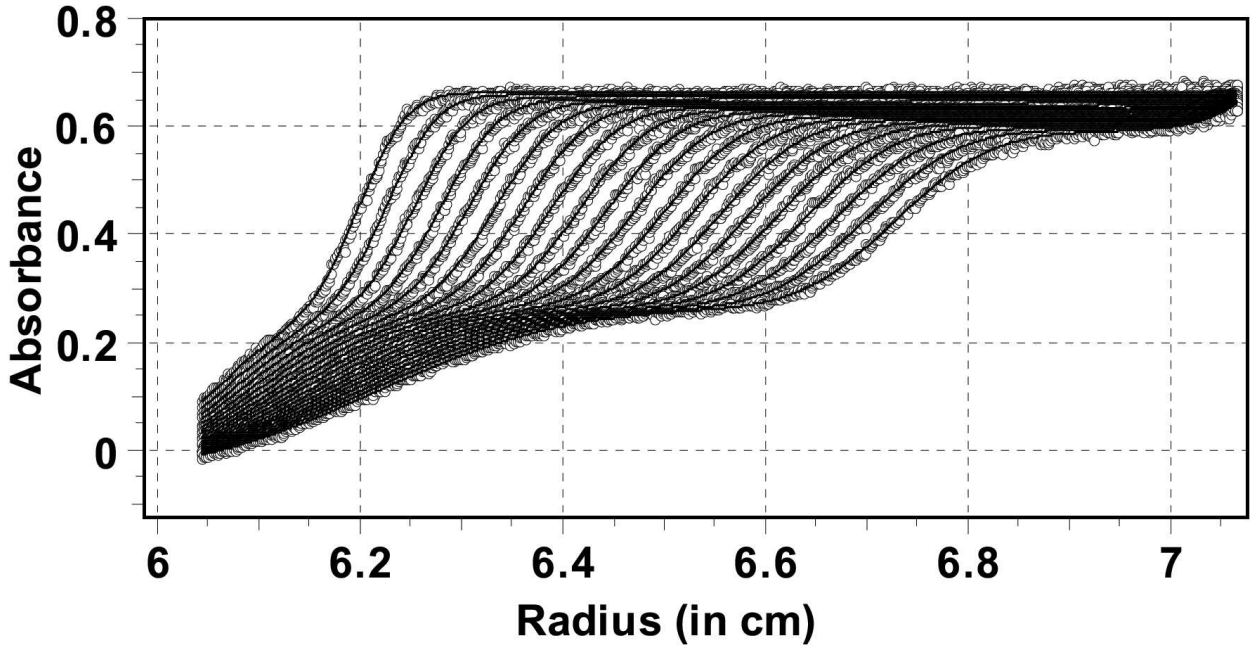

Figure 1

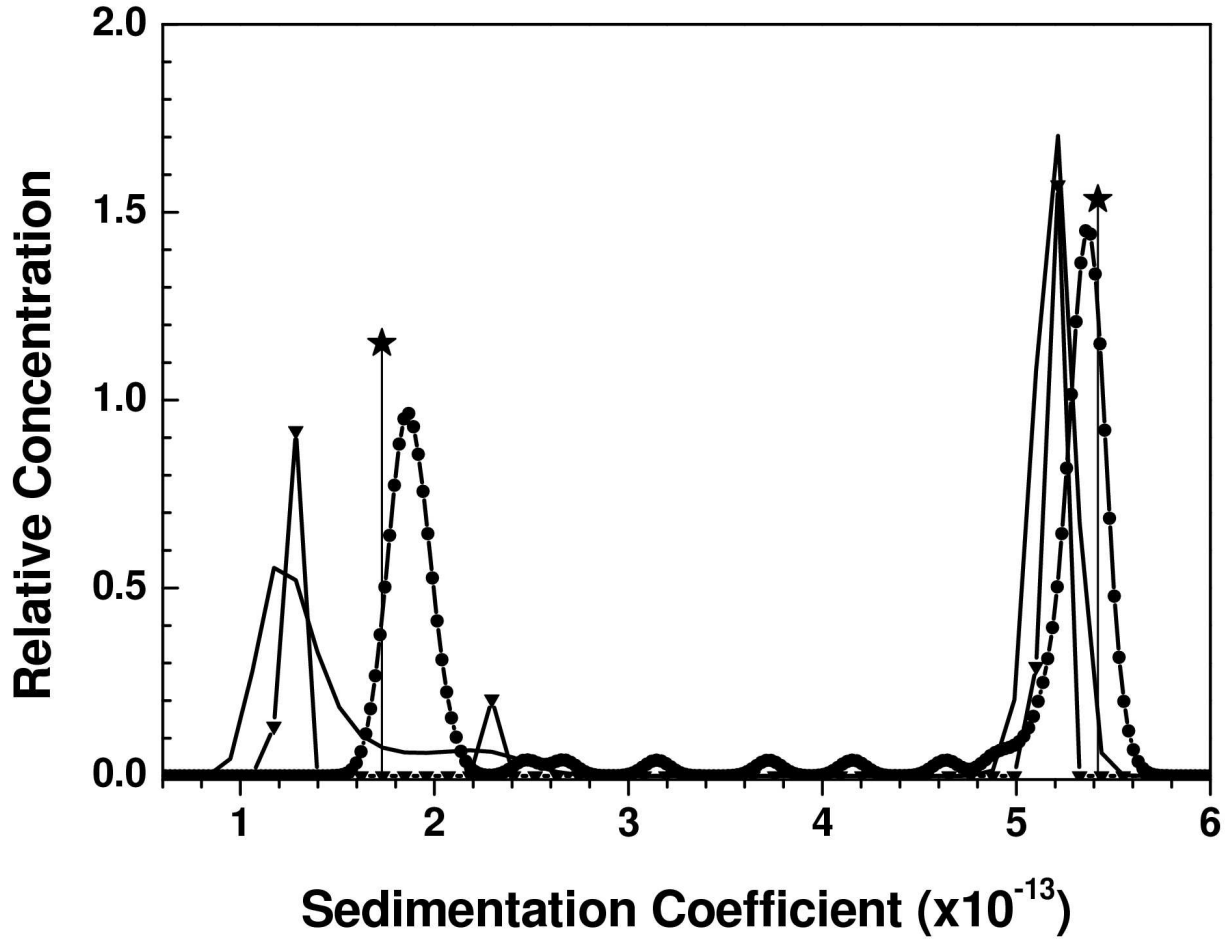
2.0
1.5
1.0
0.5
0.0
Relative Concentration
1
2
3
4
5
6
Sedimentation Coefficient (x10-13)

Figure 2

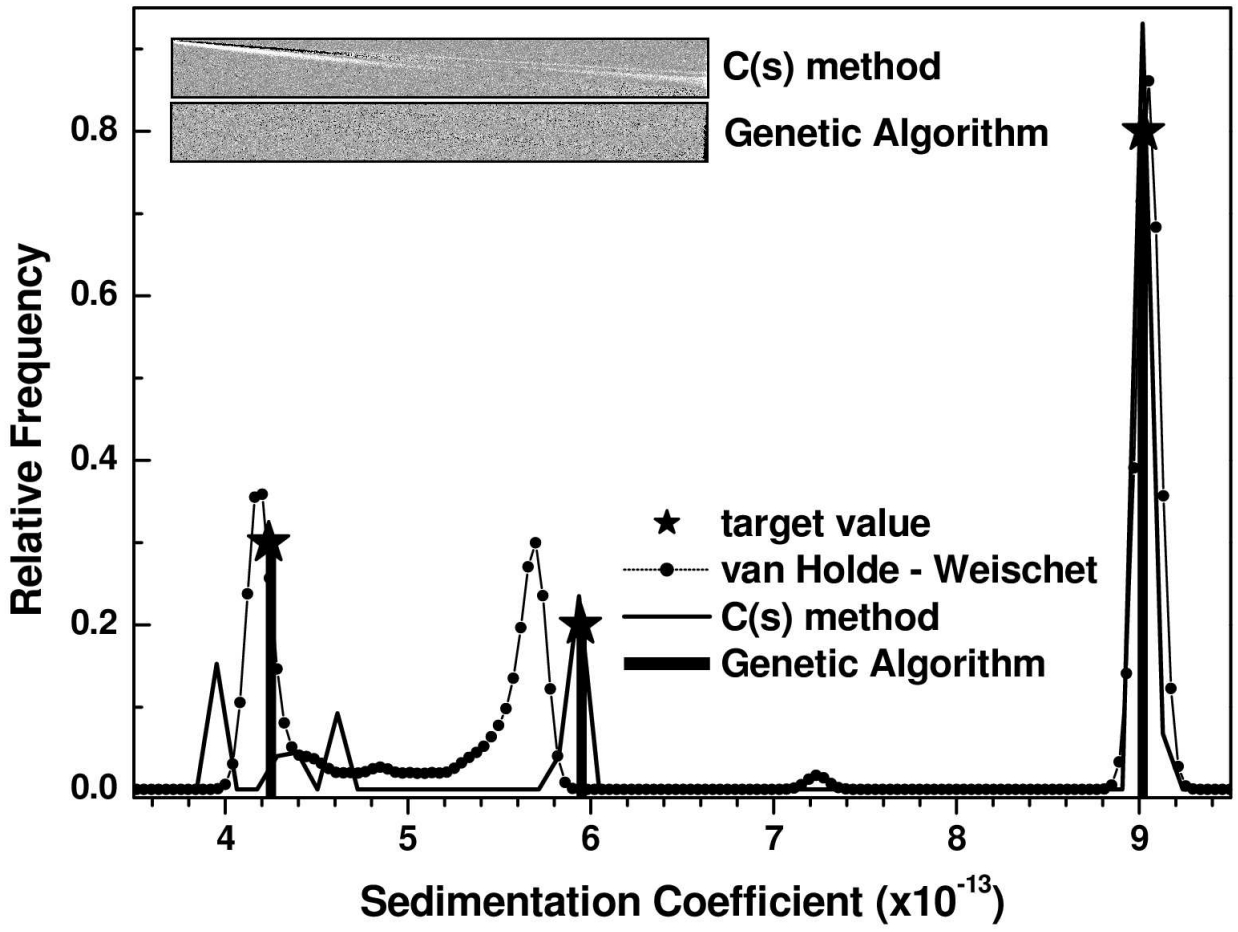
C(s) method
Genetic Algorithm
Relative Frequency
0.0
0.2
0.4
0.6
0.8
target value
van Holde - Weischet
C(s) method
Genetic Algorithm
4
5
6
7
8
9
Sedimentation Coefficient (x10-13)

Figure 3

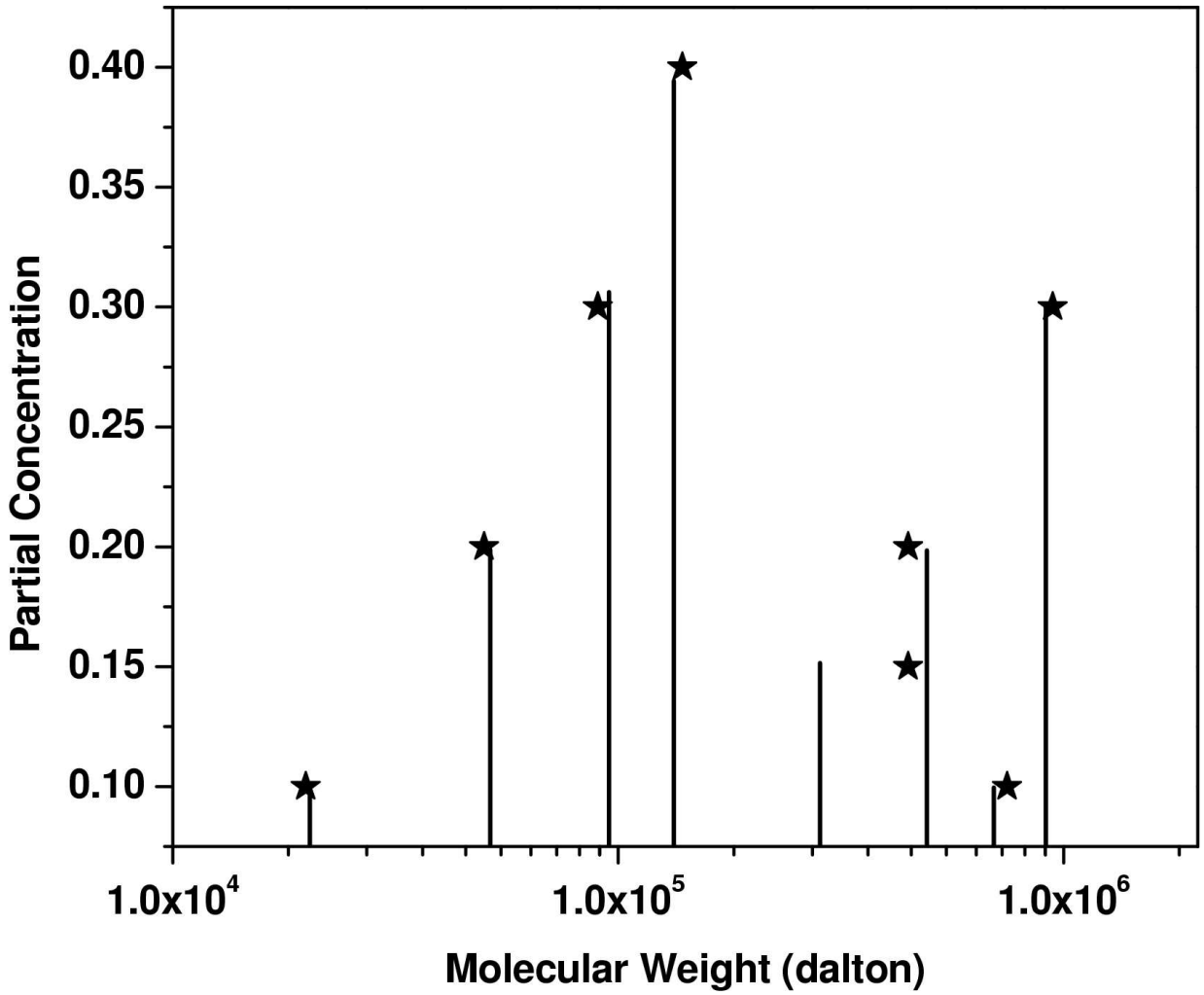
Partial Concentration
0.40
0.35
0.30
0.25
0.20
0.15
0.10
1.0x10^4
1.0x10^5
1.0x10^6
Molecular Weight (dalton)

Figure 4

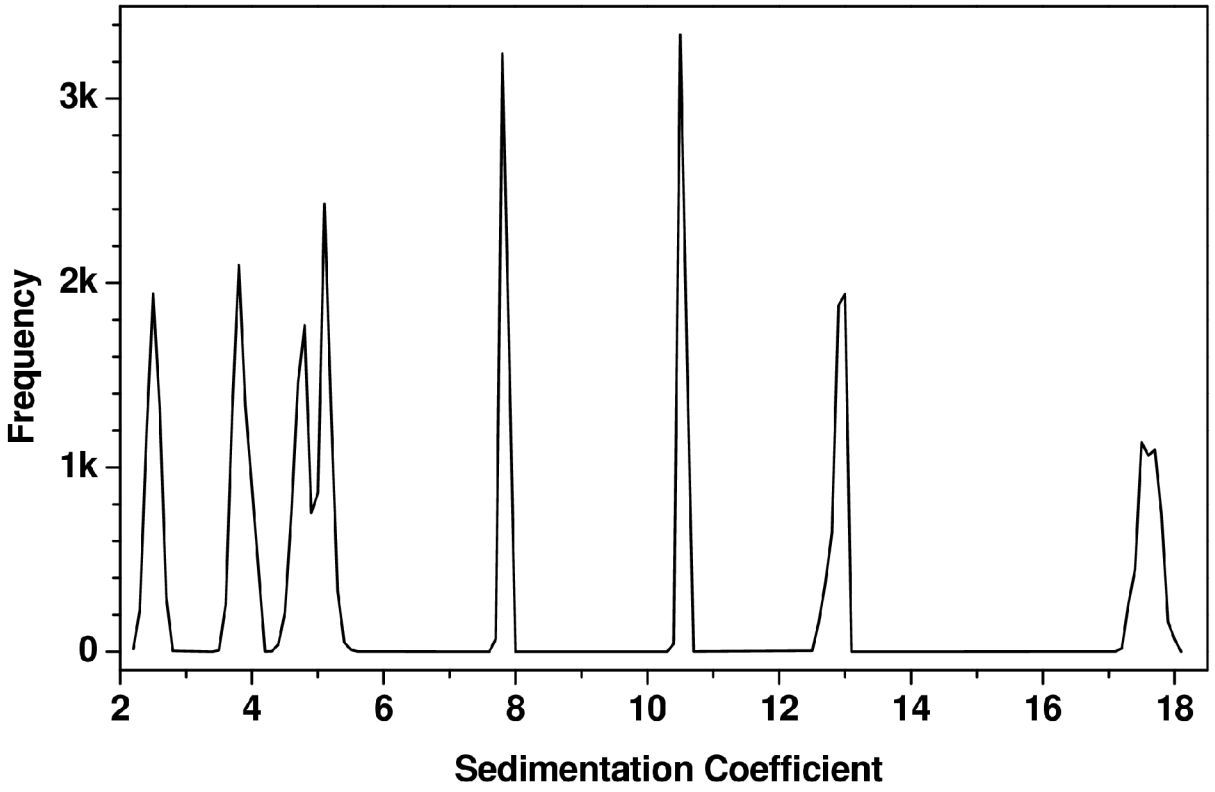
Frequency
3k
2k
1k
0
2
4
6
8
10
12
14
16
18
Sedimentation Coefficient

Figure 5

References:

Cao, W. and B. Demeler. Modelling Analytical Ultracentrifugation Experiments with an Adaptive Space-Time Finite Element Solution of the Lamm Equation (in press).

Cerny, V. (1985) Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm, J. Opt. Theory Appl., 45, 1, 41-51

Demeler, B. (2005). UltraScan version 7.1 – A Software Package for Analytical Ultracentrifugation Experiments. The University of Texas Health Science Center at San Antonio, Department of Biochemistry, San Antonio TX, 78229 USA. http://www.ultrascan.uthscsa.edu

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

Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.

Holland, J. H. Adaption in Natural and Artificial Systems. (1975). U. of Michigan Press

Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P. Optimization by Simulated Annealing. Science 220 (1983) 671-680

Koza, J. R. Genetic Programming: On the Programming of Computers by Means of Natural Selection, 1992, MIT Press, Cambridge, MA.

Lamm, O. 1929. Die Differentialgleichung der Ultrazentrifugierung. Ark. Mat. Astron. Fys. 21B:1-4

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

Schuck P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. Biophys. J. 78(3):1606-19, 2000

Stafford WF. Analysis of reversibly interacting macromolecular systems by time derivative sedimentation velocity. Methods Enzymol. (2000) 323:302-25.

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