A Grid-Enabled Supercomputer Implementation for Fitting 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

We present an overview of the new web-based interface to the UltraScan software and the UltraScan Laboratory Information Management System (LIMS). We have developed a parallelized module with supercomputer and UltraScan LIMS interface to facilitate modeling of sedimentation velocity experiments with the 2-dimensional spectrum analysis, the genetic algorithm analysis and the Monte Carlo analysis. The interface brokers compute requests from the UltraScan software to NSF's Teragrid and supercomputing resources in the Texas Internet Grid for Research and Education (TIGRE) and allows a remote user to submit analysis requests through the UltraScan LIMS web interface. Analysis is processed on multiple remote clusters, either locally, or on Teragrid or on TIGRE, and the results are e-mailed back to the investigator for import into UltraScan, which can display the model in a new finite element model viewer program. This system dramatically reduces compute time for high-resolution finite element analysis and is available to the public.

Introduction

Ultrascan is a comprehensive data analysis toolkit for analytical ultracentrifugation 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.

To accomplish this goal, we have utilized a grid infrastructure developed by the Consortium for High Performance Computing across Texas termed *Texas Internet Grid for Research and Education* (TIGRE) and implemented it on a group of supercomputer clusters at the University of Texas Health Science Center at San Antonio. The grid infrastructure is linked to the UltraScan Laboratory Information Management System (UltraScan LIMS), and permits submission of data analysis jobs, tracking of the job queue, and storing analysis results in an online database. Navigation of job submission resources is accomplished through an authenticated web-based interface. Intermediate and final results, as well as manipulation of data stored in the LIMS is accomplished with the UltraScan GUI program.

The algorithms that can be remotely executed on the Texas supercomputing grid include the 2-dimensional Spectrum Analysis (2DSA [BBD06]), the Genetic Algorithm Analysis (GA [BD05]), and the Monte Carlo Analysis (MC). Upon completion, the user will receive an e-mail with the analysis results provided as an attachment. The attachment is in a format compatible with the UltraScan software.

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 [CD05], 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 [LH74]. It reports non-negatively constrained amplitudes c or zero for solutes that are not present in the solution.

1. Initialization: The s-value range is initialized using the enhanced van Holde – Weischet method [VW78, DV04], 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 [SD99] 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).

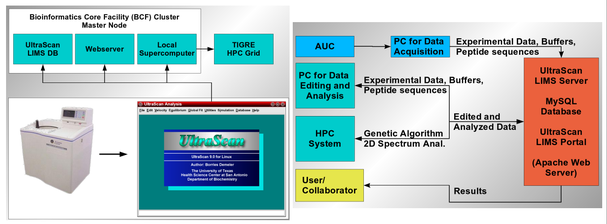
3. 2DSA Monte Carlo Analysis: In the next step we perform a 100 iteration Monte Carlo analysis using the 2DSA method. This approach amplifies the signal contained in the data and provides a refined view of the parameter surface (Figure 3).

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).

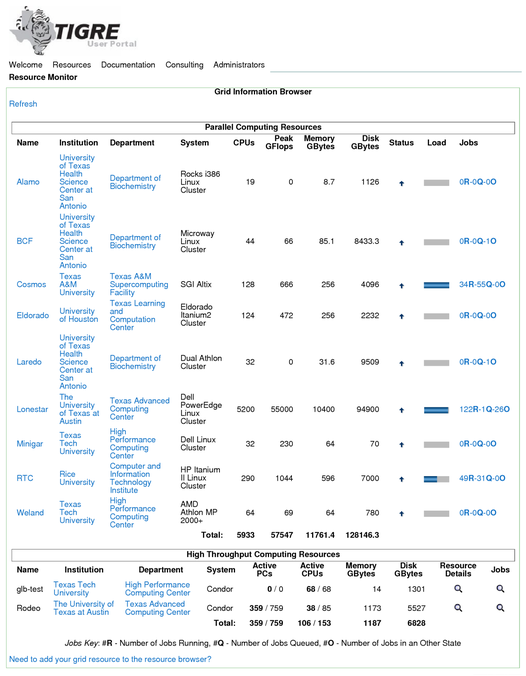
5. GA Monte Carlo Analysis: Now a Monte Carlo analysis is performed to obtain statistical descriptions of each fitted parameter. This provides the necessary confidence intervals and facilitates data interpretation (Figure 5).

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).

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 are shown in Table 2.



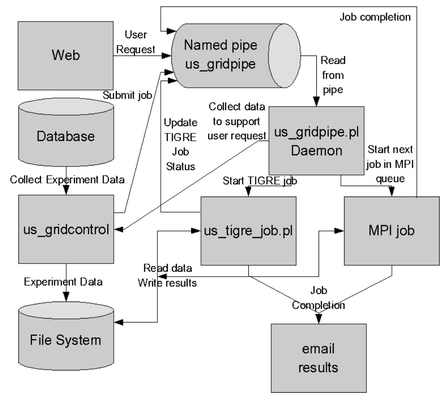
Data flow between between Instrument, UltraScan, LIMS and HPC system



TIGRE web portal showing grid-enabled resources from the Texas Internet Grid for Research and Education. System status and resource availability can be viewed from a central server.

Parallelization:

All parallelization in UltraScan is hand coded in C++ and uses the MPI message passing library for communication between processors. The submission and job distribution schematic is shown below:

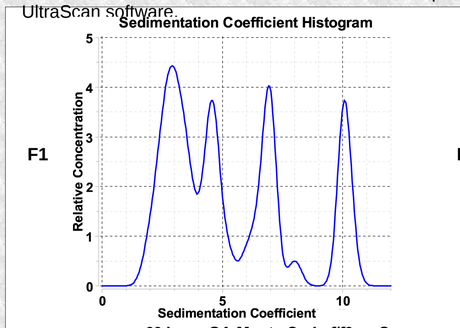


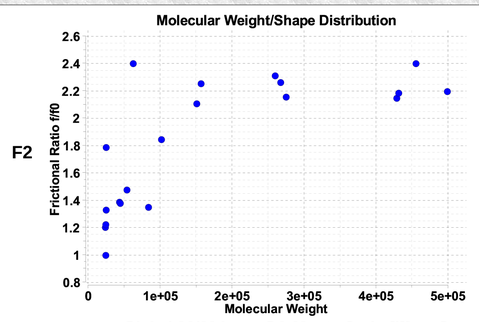
References

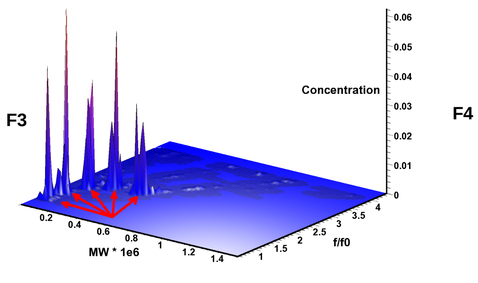
[CD05] Cao W. and Demeler B. Modeling analytical ultracentrifugation experiments with an adaptive space-time finite element solution of the Lamm equation. (2005) Biophys J. 89(3):1589-602
[BD05] Brookes, E. and Demeler B. Genetic Algorithm Optimization for obtaining accurate Molecular Weight Distributions from Sedimentation Velocity Experiments. Analytical Ultracentrifugation VIII, Progr. Colloid Polym. Sci. C. Wandrey and H. Cölfen, Eds. Springer (2006) 131:78-82
[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. Solving Least Squares Problems. (1974) Prentice-Hall, Inc. Englewood Cliffs, New Jersey
[BBD06] Brookes, E., Boppana, R.V., and B. Demeler. (2006) Computing Large Sparse Multivariate Optimization Problems with an Application in Biophysics. Supercomputing 2006
[SD99] Schuck, P. and B. Demeler. Direct Sedimentation Boundary Analysis of Interference Optical Data in Analytical Ultracentrifugation. (1999) Biophys. J., 76:2288-2296

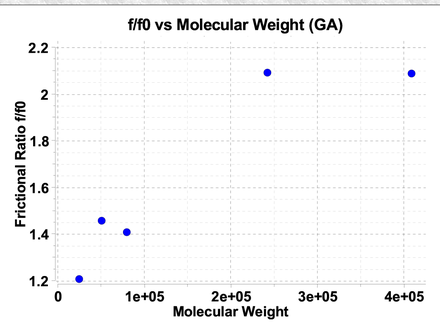
Acknowledgments

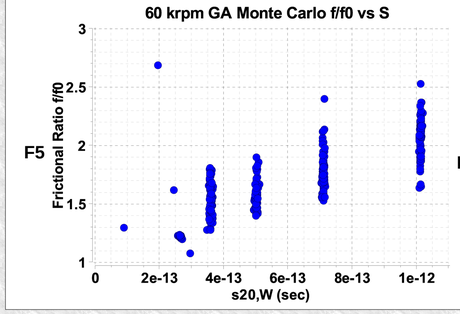
We would like to thank Josh Wilson, Yu Ning and Bruce Dubbs for contributions to the web interface code. This research has been supported by NSF Grant DBI-9974819 and the San Antonio Life Science Institute with Grant #10001642

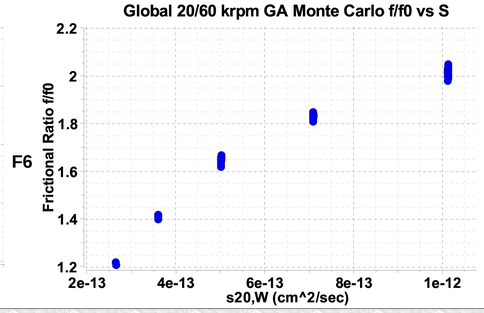


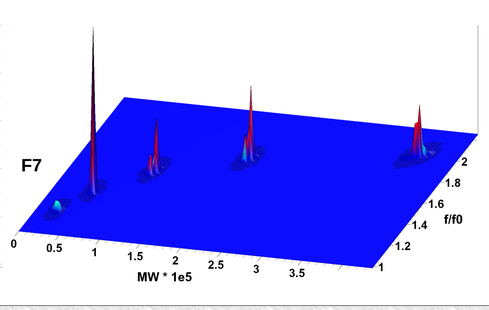












Solute	Partial Concentration (OD)	Molecular Weight (kDa)	Shape (f/f_0)
1	0.1	25	1.2
2	0.1	50	1.4
3	0.1	100	1.6
4	0.1	200	1.8
5	0.1	400	2.0

Table 1: Aggregating 5-component system

Figure 1: van Holde – Weischet G(s) distribution used for initialization.

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.

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

Figure 3: Monte Carlo Analysis using 2DSA. 5 groupings are clearly visible and indicated by red arrows. Grouping covers the known target range well. As can be seen, the Monte Carlo analysis amplifies the signal to noise ratio.

Figure 4: GA analysis of results obtained in Figure 2. Using parsimonious regularization, false positives are eliminated from the 2DSA solution.

Figure 5: GA Monte Carlo analysis of results obtained in Figure 4. Most of the variation is in the frictional ratio, not the sedimentation coefficient.

Figure 6: 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.

Figure 7: 3-D molecular weight/frictional ratio distribution of the same data shown in Figure 6.