A Re-Analysis of Christopher Chin’s CRP Data Using DCDT+ Version 2, and a Comparison to *c(s)* Analysis from SEDFIT.

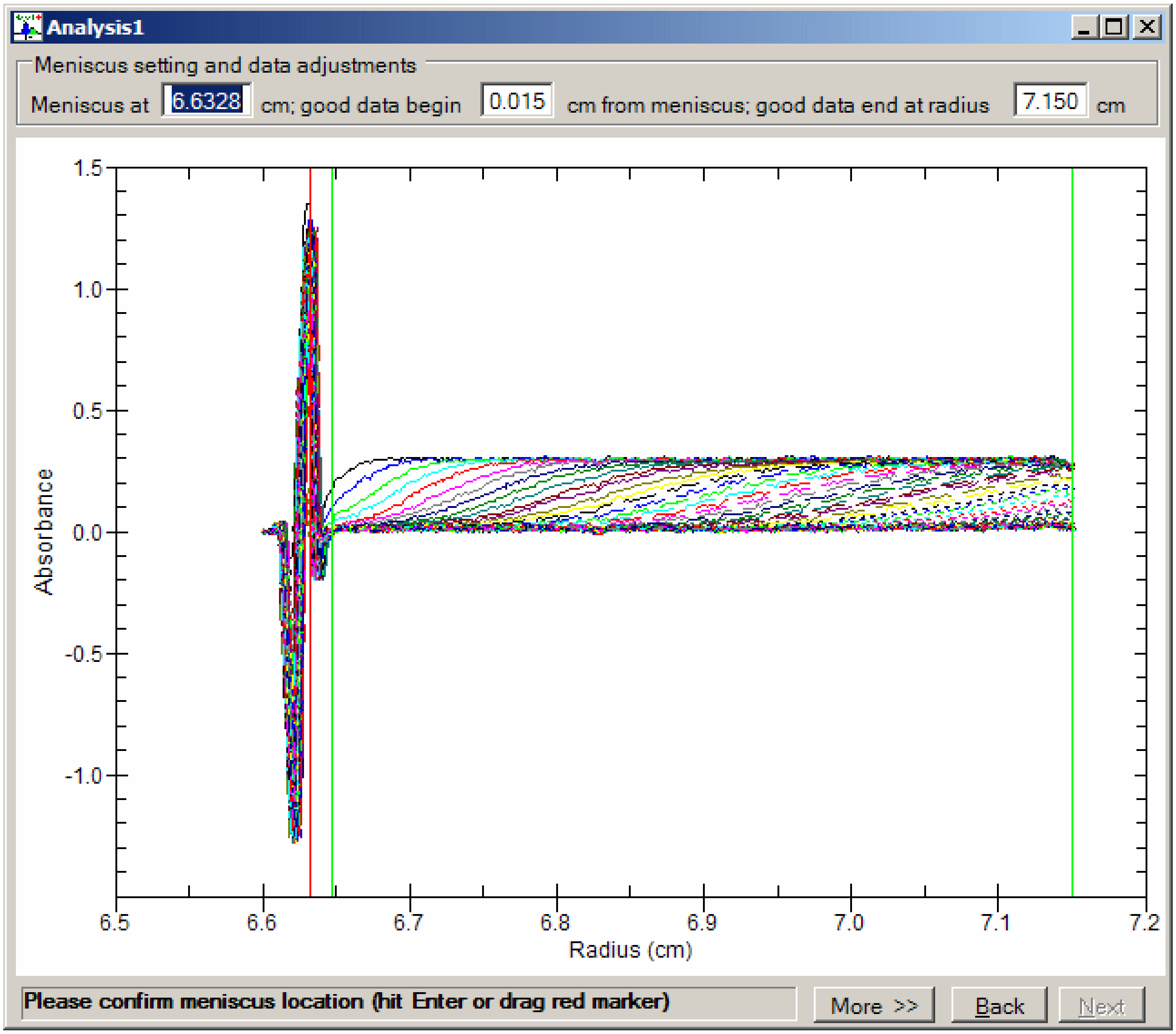
For this study/tutorial the data analysis in DCDT+ will first be done the new easy way, letting the program wizards do all the work. Only a few screen shots showing key intermediate steps will be shown here, but the program Help file includes several step-by-step tutorials with screen shots for each step.

In section B the same experiment will be analyzed by DCDT+ using the same scans that Chris had picked, to show that the new data fitting algorithm employed in version 2 gives the correct properties for the main peak even when the scan selection is very far from optimal.

In section C the same scans used by Chris for his ‘final’ SEDFIT *c(s)* analysis will be re-analyzed. This will show that the *c(s)* analysis is sensitive to the details of what data is being fitted, and that a more optimum choice indicates a second species is present (although it is not entirely clear whether this is real or an experimental artifact).

A) Analysis using the new automatic data selection feature

Version 2 of DCDT+ introduces a new paradigm for data loading and selection for *dc/dt* analysis. Unlike other *dc/dt* implementations, in this version generally one can simply load the entire run (scans 1-120 in this case) in one operation. The screen shot below shows the second step, the screen for setting the meniscus position and the region of the cell to use for data analysis. (Note that these screen shots show only a single analysis document, not the whole program window which can include multiple analysis documents and overlays of *g(s*)* curves from different samples.)



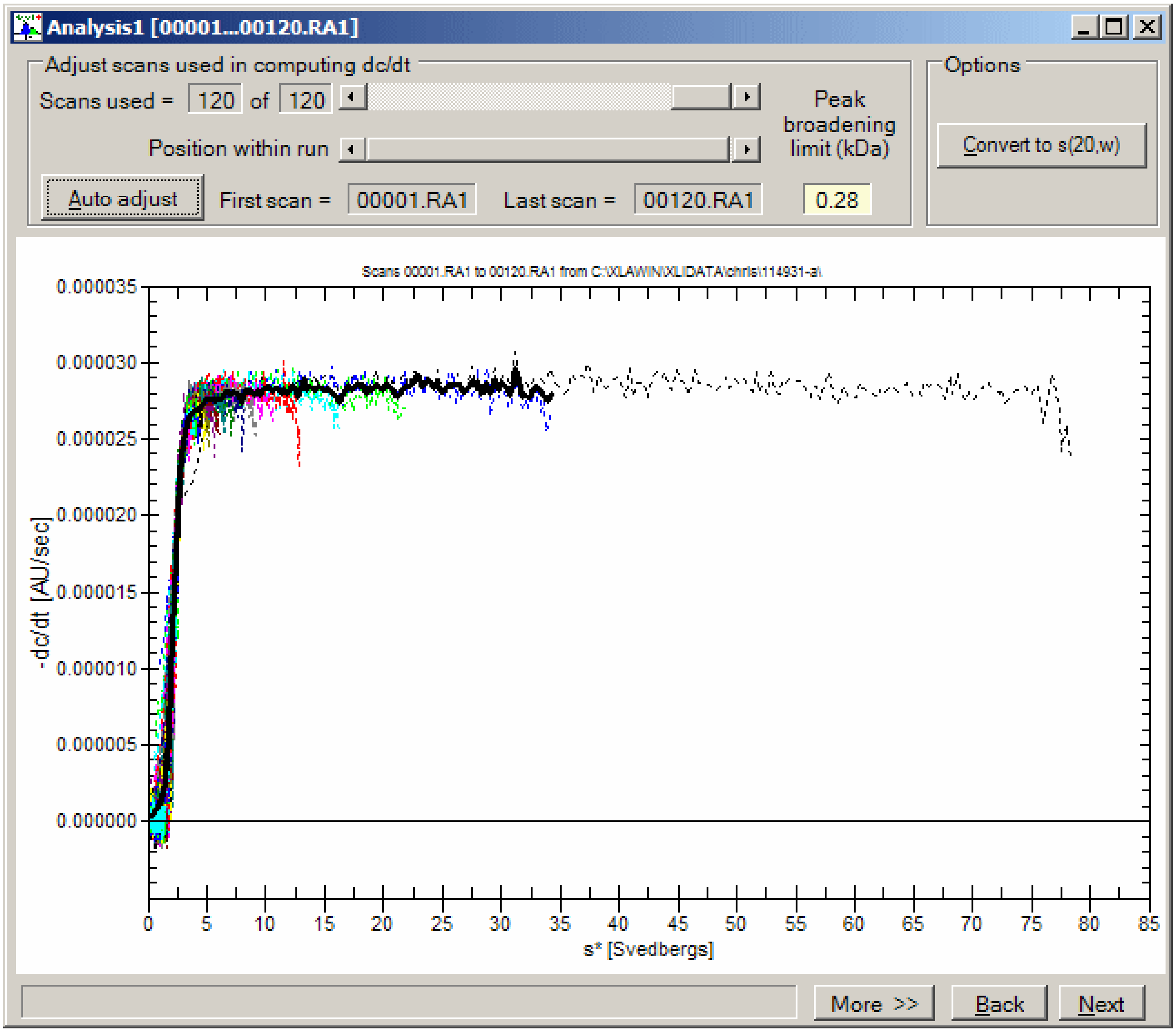
Meniscus setting and data adjustments

Meniscus at 6.6328 cm; good data begin 0.015 cm from meniscus; good data end at radius 7.150 cm

Please confirm meniscus location (hit Enter or drag red marker)

When you enter this step the graph shows an overlay of all scans that were loaded. A program wizard automatically locates the meniscus position for all 120 scans, averages those values, marks that position with the red vertical line, and displays that computed value in the highlighted text box at the top left. If the meniscus wizard did a good job, all you need to do is hit the Enter key to accept its choice. (Note that here the wizard has picked 6.6328 cm, while Chris had selected 6.6322 cm in his SEDFIT analysis.) If you need to manually adjust the meniscus, or the green lines marking the 'good' data region, you can simply click on the marker and drag it to a new position. The 'Next' button at the bottom then causes display of a different graph and set of controls within the same analysis document that allow you to select a subset of the total run to be used for analysis.

That selection of which scans to analyze can be made via the 2 slider controls located above the graph in the screen shot below. These slider controls allow you to 'tune' through the run, varying the number of scans to be used for the *dc/dt* analysis (top slider) and their position within the run (lower slider). As adjustments are made to the sliders the graph displays the *dc/dt* curve from each pair of scans as well as their average (the heavy black curve). This shot shows the initial state (all



Analysis1 [00001...00120.RA1]

Adjust scans used in computing dc/dt

Scans used = 120 of 120

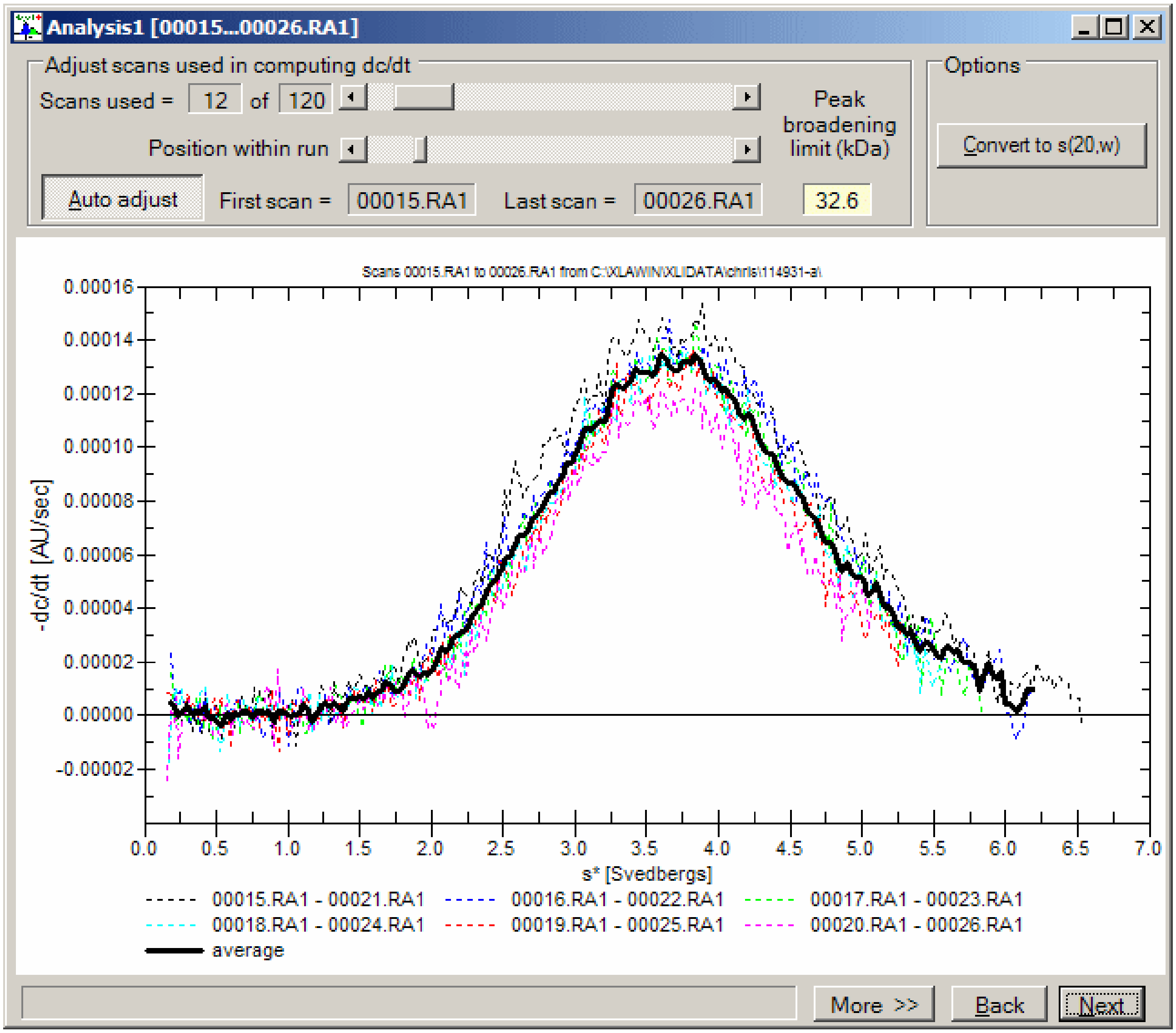
First scan = 00001.RA1 Last scan = 00120.RA1

Peak broadening limit (kDa) 0.28

Scans 00001.RA1 to 00120.RA1 from C:\XLAWIN\XLIDATA\chris\114931-a\

120 scans used to compute *dc/dt* curves), which gives mostly nonsense in this case since at scan 1 the boundary has not cleared the meniscus, and because the cell was essentially empty after the first 40 scans.

While these slider controls make manual scan selection much easier, the true easy way is simply to push the ‘Auto adjust’ button and let a program wizard select a workable group of scans. Obviously the wizard can’t anticipate whether you are particularly interested in species sedimenting slower or faster than the main boundary, so it may not pick scans that are quite right for your interests, but it will usually at least get things into the right range for characterizing the main component(s). In this case that ‘Auto adjust’ process takes only ~2 seconds and results in the following screen shot.



Analysis1 [00015...00026.RA1]

Adjust scans used in computing dc/dt

Scans used = 12 of 120

Position within run

First scan = 00015.RA1

Last scan = 00026.RA1

Peak broadening limit (kDa) 32.6

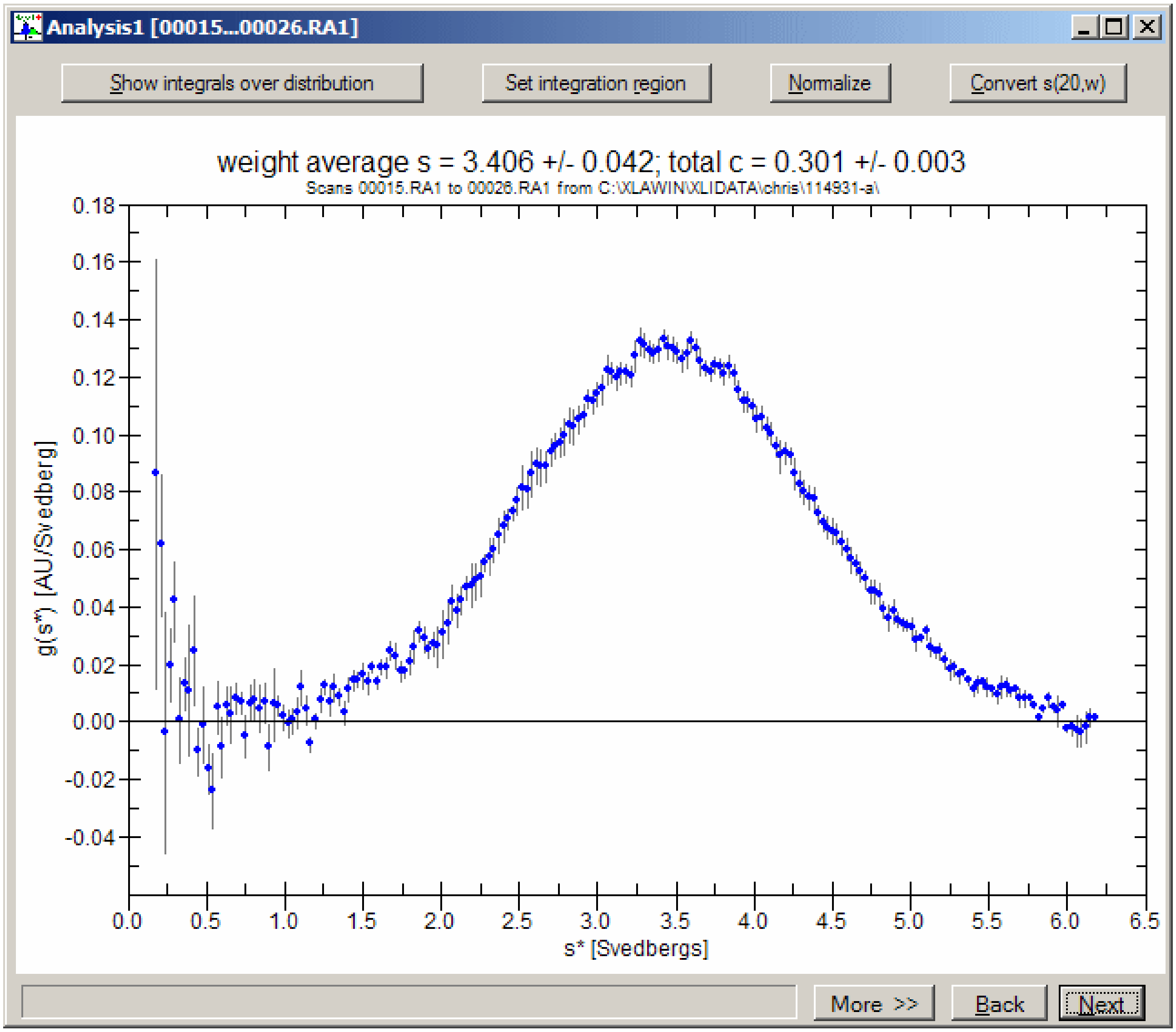
In this case the ‘Auto adjust’ wizard has selected a group of 12 scans starting with scan number 5. This group covers a time range where the boundary is about in the middle of the cell (usually about the right time in the run). The resulting 6 pairs of scans give *dc/dt* curves that are reasonably similar, and the average of those 6 (heavy black curve) gives a well-defined and symmetric peak near 3.5 S.[1] The wizard has selected 12 total scans as a compromise between using more scans to improve the signal/noise ratio and trying avoid excessive peak broadening that would result if there is too much boundary movement from the first to last scan that is used for analysis.[2] As indicated by the value displayed as the ‘peak broadening limit’, the 12 scans used here should produce substantial peak broadening only for species significantly larger than ~33 kDa.

[1] The buffer density and viscosity values and protein vbar that Chris used are not known so this study will stay in raw sedimentation coefficient units.

[2] What is actually calculated from each pair of scans is $\Delta c/\Delta t$. When the total boundary movement is too large the approximation that $\Delta c/\Delta t \approx dc/dt$ becomes poor and the peaks become artificially broadened.

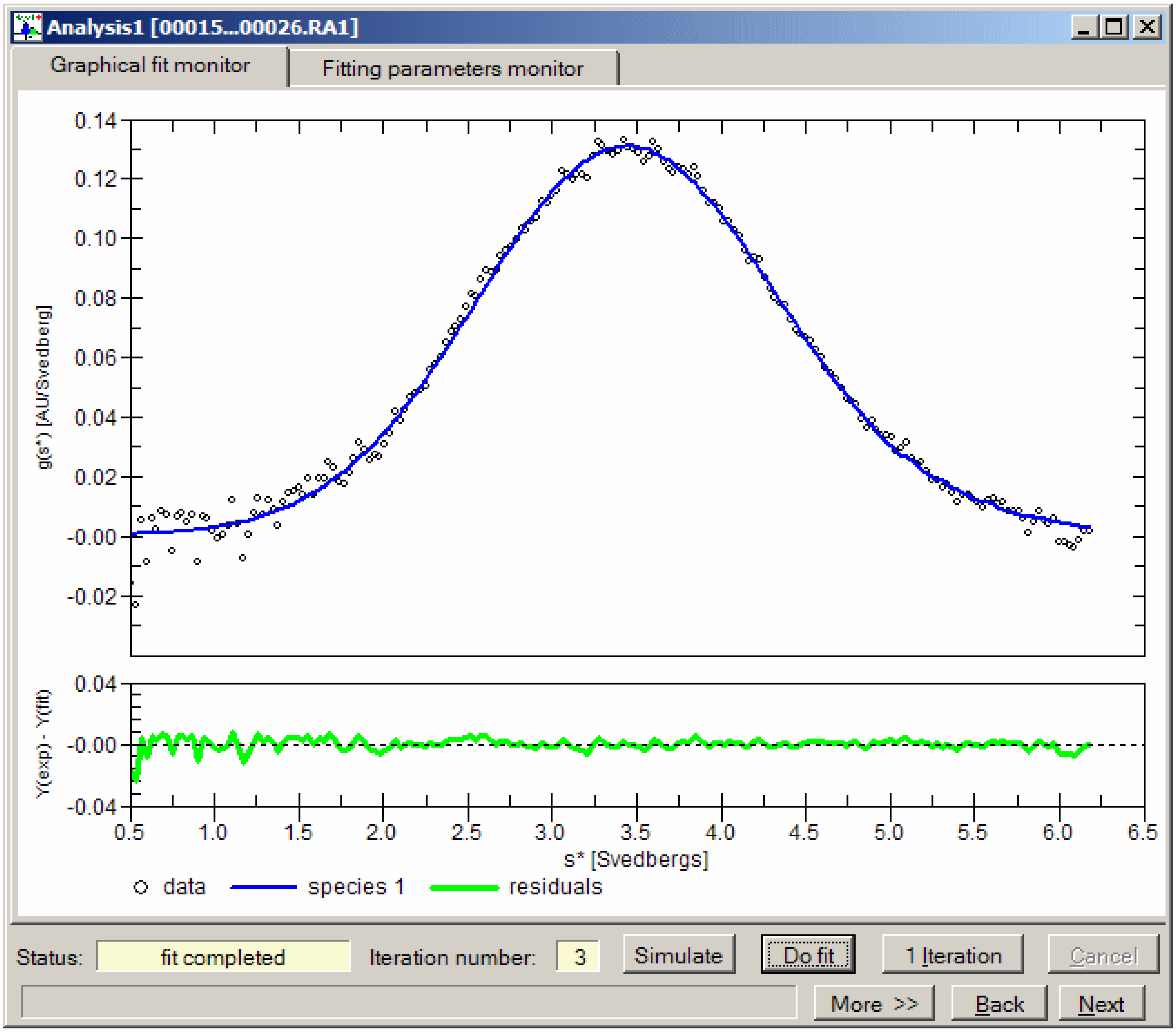
Clicking the Next button twice more produces the following *g(s*)* distribution. The error bars for each point in the above *g(s*)* distribution are indicated by gray vertical bars. Note that the distribution is rather broad here for a molecule of this mass (the resolution is low) because in this experiment the sample volume was quite low (the cell was only ~40% full).

To fit this distribution as a single species (the program default) all that is required is to push the Next button four times. The program wizards automatically locate the top of the peak and use that position to calculate initial guesses for the sedimentation coefficient and molecular mass.



Analysis1 [00015...00026.RA1]
Show integrals over distribution
Set integration region
Normalize
Convert s(20,w)
weight average s = 3.406 +/- 0.042; total c = 0.301 +/- 0.003
Scans 00015.RA1 to 00026.RA1 from C:\XLAWIN\XLIDATA\chris\114931-a\
g(s*) [AU/Svedberg]
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
s* [Svedbergs]
More >>
Back
Next

During fitting the current best fit is continuously displayed overlaying the fitted data (with curves for each peak in multi-species fit), along with a residual plot at the bottom. The screen shot below shows the fit monitor page at the end of the single-species fit (which gives a good fit of this distribution). When the fit converges the program advances to the next screen, shown below, to display the numerical results and allow printing of a full report. The fit gives a sedimentation coefficient of 3.518 S. Significantly, this approach immediately gives the molecular mass of this



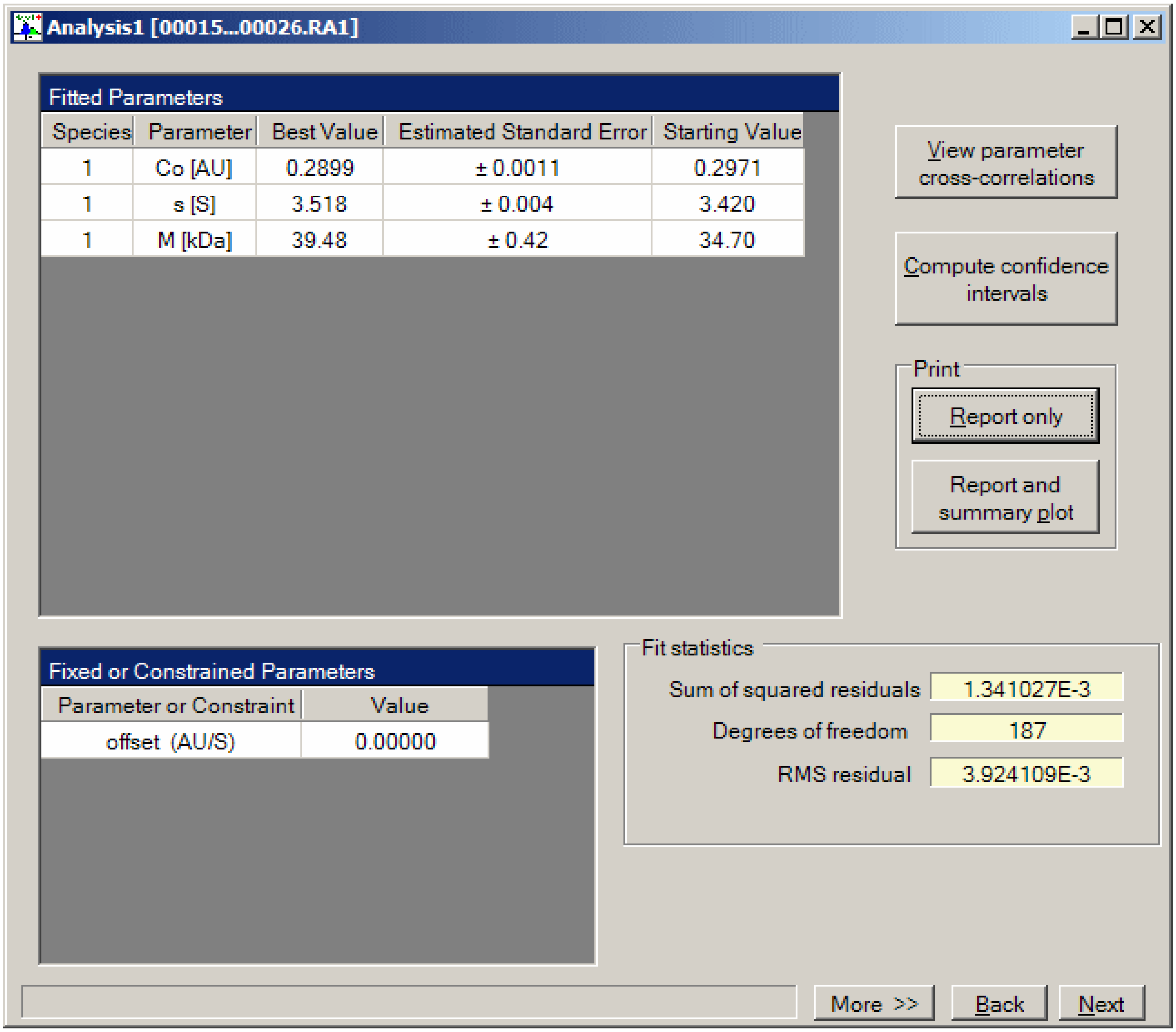
Status: fit completed

Iteration number: 3

species as 39 kDa.[3] This apparent molecular mass value can also give important clues about the homogeneity of the sample---if there is any unresolved heterogeneity (this sample isn't truly one species) this mass estimate will generally be below the expected value (because the heterogeneity broadens the boundary). Without knowing the actual monomer mass, vbar, and density for this sample, however, such inferences cannot be made.

This approach also immediately gives estimates for the precision of the fitted parameters. If this is a final fit and a more rigorous error analysis is desired, a button on this page will produce true confidence limits for each parameter (this requires only ~1 second in this case). That gives 95% confidence limits for *s* of 3.508 to 3.528 S, 38.51 to 40.47 kDa for *M*, and 0.2871 to 0.2925 OD for the loading concentration.

[3] The accuracy of this value cannot be evaluated since the true solvent density and protein vbar are not known. This value is calculated using the default vbar of 0.73 ml/g and the density of phosphate-buffered saline, 1.00535 g/ml.


Analysis1 [00015...00026.RA1]
View parameter cross-correlations
Compute confidence intervals
Print
Report only
Report and summary plot
More >>
Back
Next

Fitted Parameters

Species	Parameter	Best Value	Estimated Standard Error	Starting Value
1	Co [AU]	0.2899	± 0.0011	0.2971
1	s [S]	3.518	± 0.004	3.420
1	M [kDa]	39.48	± 0.42	34.70

Fixed or Constrained Parameters

Parameter or Constraint	Value
offset (AU/S)	0.00000

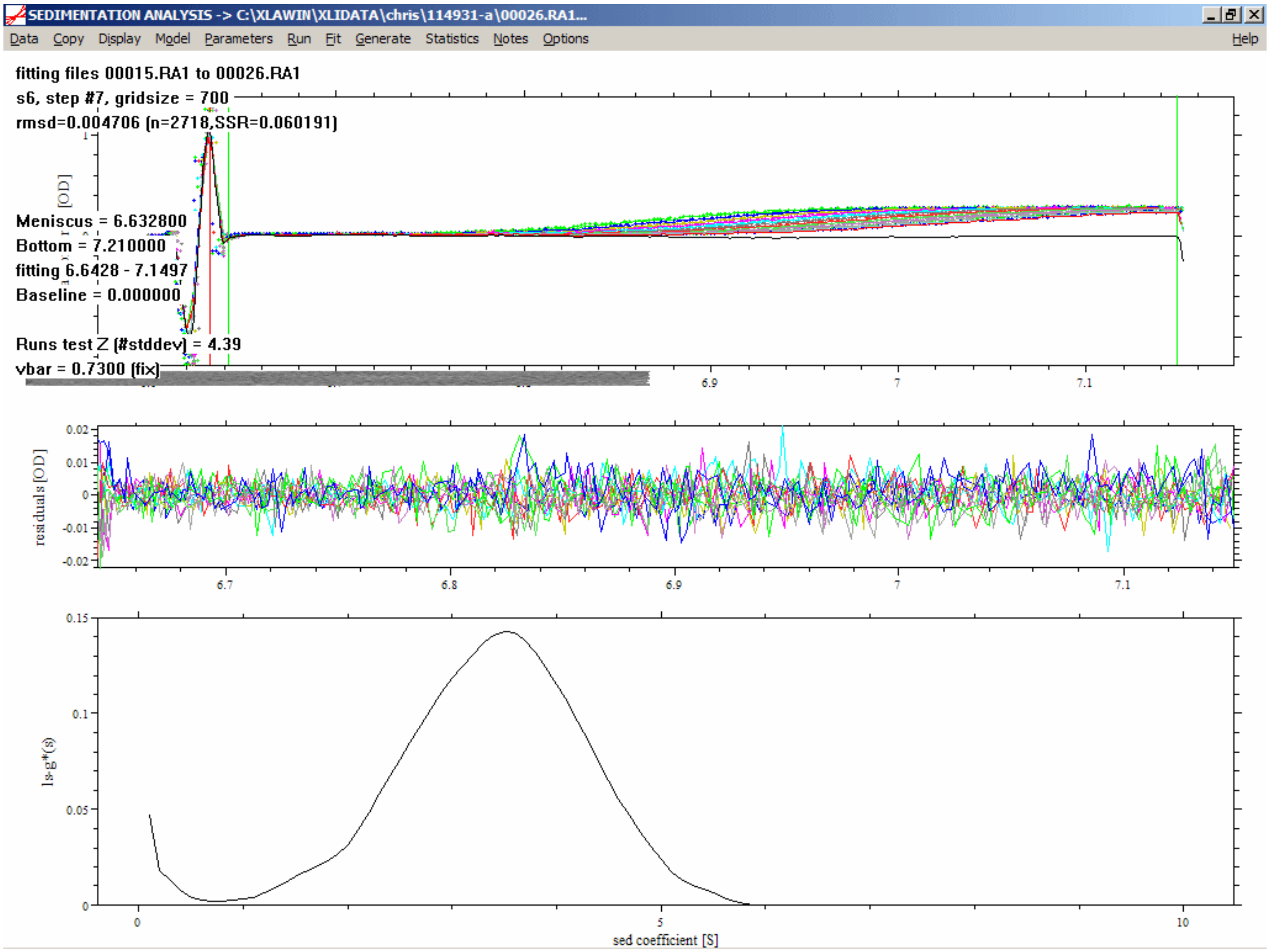
Fit statistics

Sum of squared residuals	1.341027E-3
Degrees of freedom	187
RMS residual	3.924109E-3

Thus this analysis gives results in excellent agreement with Chris Chin’s conclusion using scans 1-50 with SEDFIT that the major peak in his *c(s)* analysis has a sedimentation coefficient of 3.530 S,[4] even though the DCDT+ program wizards were allowed to select the meniscus position, the scans to be analyzed, and the initial fitting parameters, all automatically. **Once the 120 scans were loaded, all that was required to reach the result shown in the screen above was to hit the Enter key nine times to accept the program’s automatic choices. The complete analysis then requires less than 20 seconds.**

Note too that we have obtained more information than was available via the *c(s)* analysis, *i.e.* an error bar on the sedimentation coefficient, an estimate of the molecular mass of that species (with an error bar), and an error bar on the loading concentration.

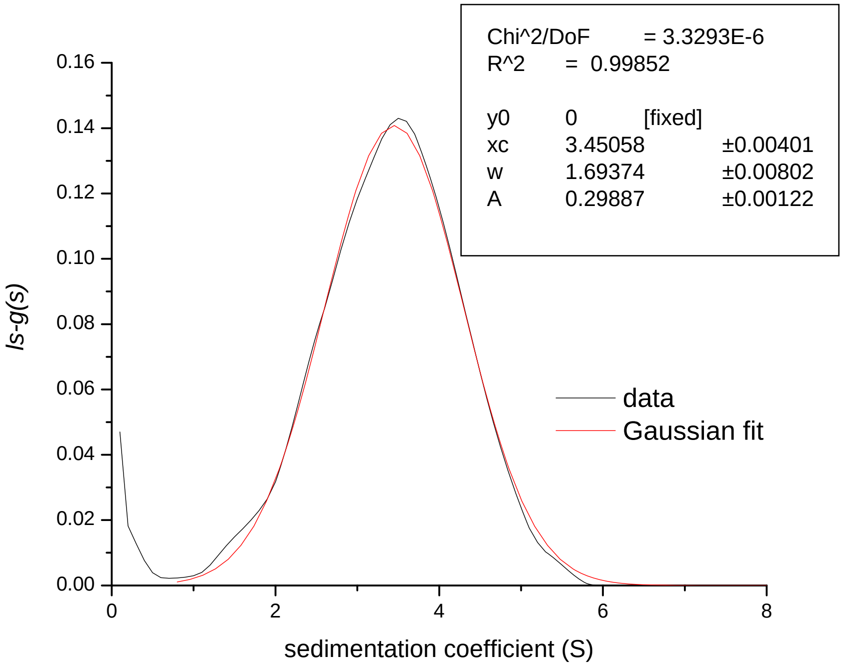
[4] We actually should not expect perfect numerical agreement of the sedimentation coefficients. Chris was actually fitting the *c(s)* curves to a Gaussian (although there is no theoretical reason for those peaks to be Gaussians) and reporting the center position of that Gaussian. Further, his analysis used a slightly different meniscus position.



```
SEDIMENTATION ANALYSIS -> C:\XLAWIN\XLIDATA\chris\114931-a\00026.RA1...
fitting files 00015.RA1 to 00026.RA1
s6, step #7, gridsize = 700
rmsd=0.004706 (n=2718,SSR=0.060191)
Meniscus = 6.632800
Bottom = 7.210000
fitting 6.6428 - 7.1497
Baseline = 0.000000
Runs test Z (#stddev) = 4.39
vbar = 0.7300 (fix)
```

It is also interesting to ask how this would compare to a least-squares *g(s)* or *c(s)* analysis of the same 12 scans using SEDFIT. The SEDFIT *ls-g(s)* screen shot is shown above. Even ignoring the half-peak near zero S, the shape of this *ls-g(s)* curve appears to non-Gaussian. The mean sedimentation coefficient of the main peak (by integration) is 3.414 S. By eye the main peak appears asymmetric and non-Gaussian. A fit of that peak as a Gaussian is shown below, which gives a center position of 3.451 S, and implies a molecular mass of 39.4 kDa.

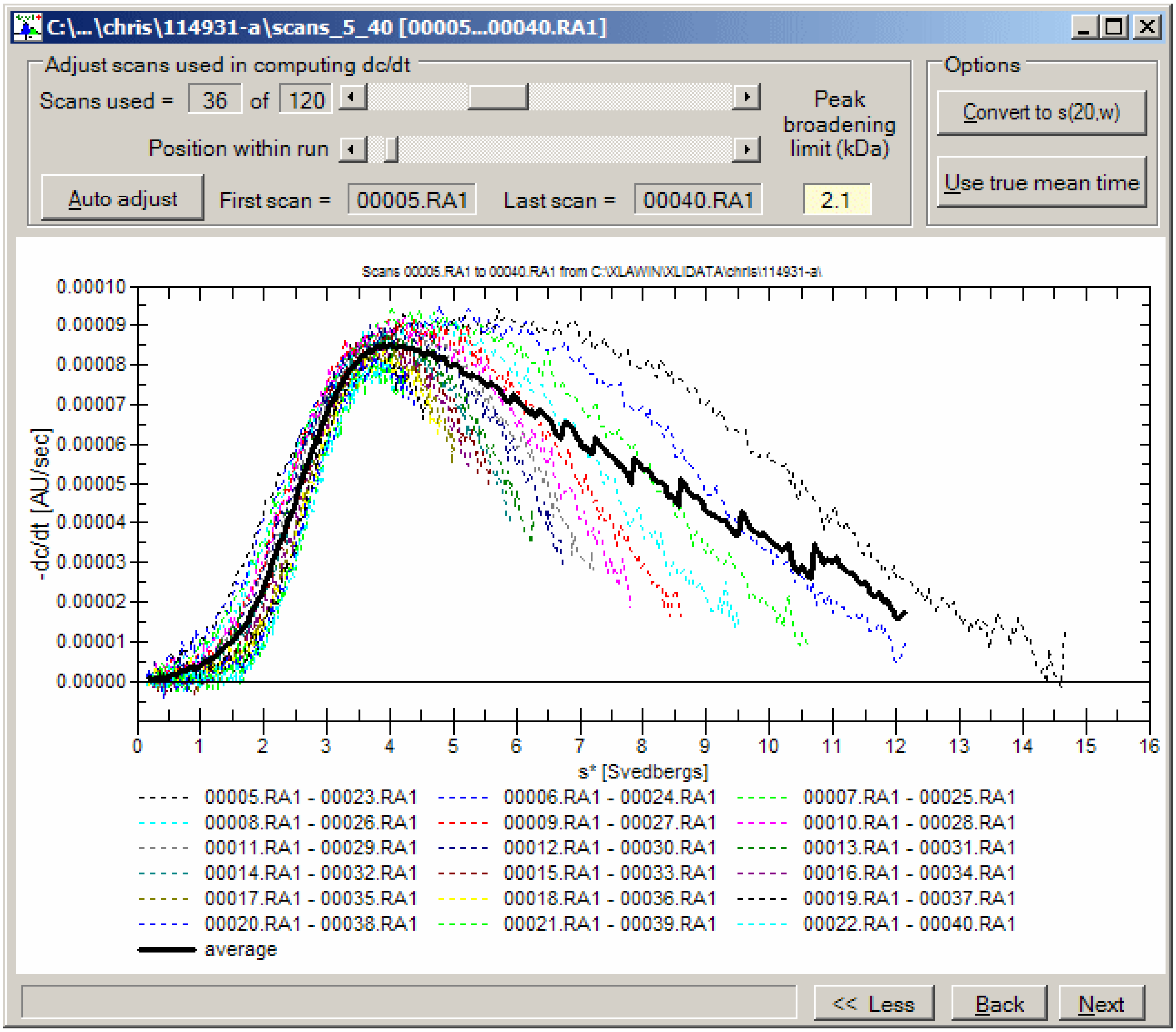
A fit of these same scans using the *c(s)* method in SEDFIT (not shown) gives a mean sedimentation coefficient of 3.526 S, in excellent agreement with DCDT+, but 2.2% higher than the *ls-g(s)* result.



B) DCDT+ v. 2 Analysis using scans 5-40 as Chris selected

Below the scan selection screen that corresponds to scans 5-40 as Chris had selected for his *dc/dt* analysis is shown.[5] (These scans essentially cover from the time the meniscus was first cleared until well after the boundary was gone.) This is much too large a time span (too many scans) for traditional *dc/dt* analysis, where typically scans covering only 15-20% of the total boundary movement are used. The overly-large range here is evident from the large divergence of the *dc/dt* curves from individual scan pairs, as well as the grossly distorted shape of the peak in the average *dc/dt* curve. A second important clue that something is wrong here is provided by the 'peak broadening limit' displayed at the lower right. This indicates that this scan selection is appropriate (roughly) for species of ≤ 2.1 kDa. (This diagnostic value is also provided in the earlier version of DCDT+ Chris was using.)

[5] For this analysis the meniscus position used by Chris in his SEDFIT analysis, 6.0322 cm, was used so direct comparisons can be made.



C:\...\chris\114931-a\scans_5_40 [00005...00040.RA1]

Adjust scans used in computing dc/dt

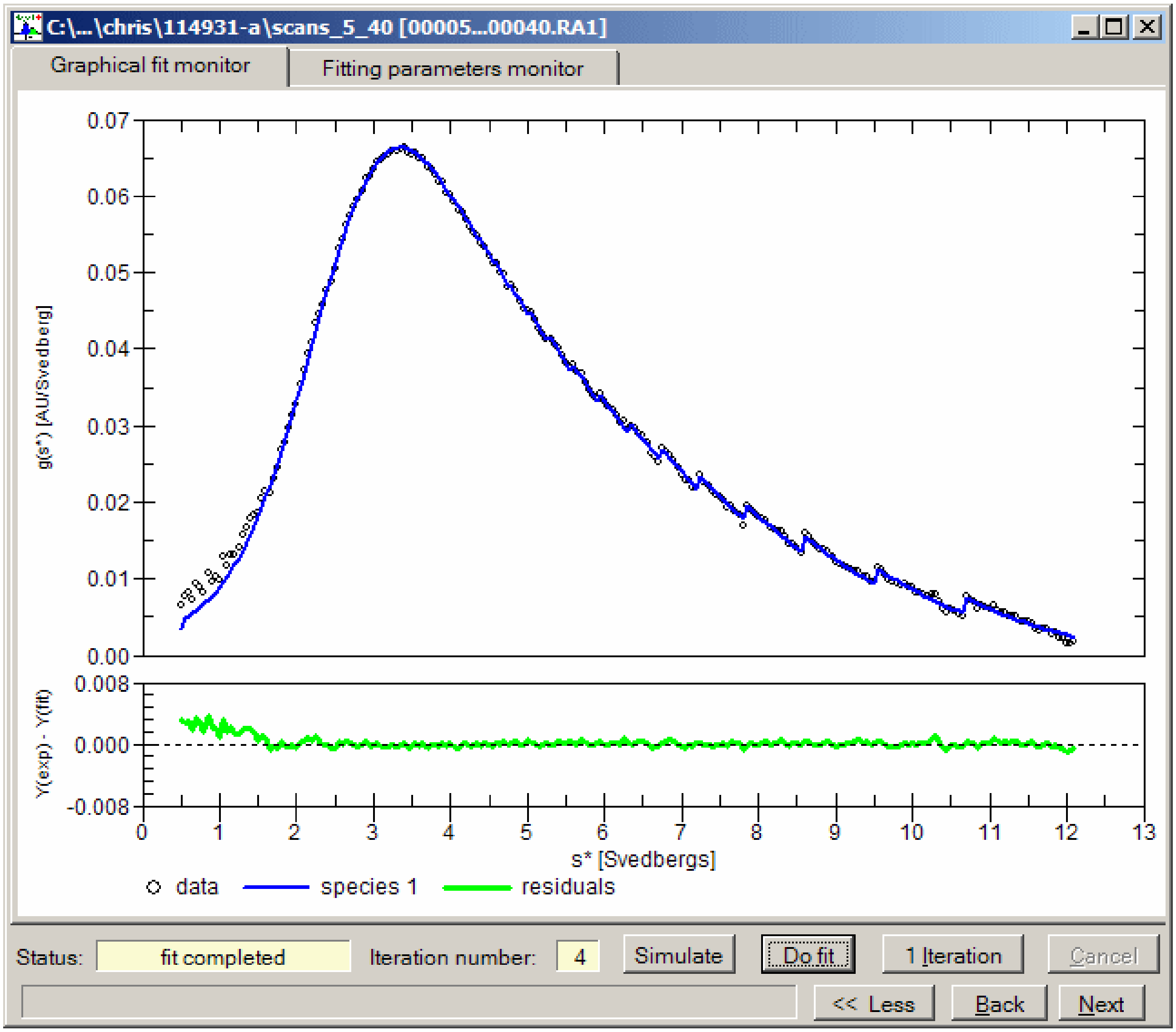
Scans used = 36 of 120

First scan = 00005.RA1   Last scan = 00040.RA1

Peak broadening limit (kDa) 2.1

However the new fitting algorithm introduced with version 2 of DCDT+ is explicitly designed to give a correct result, independent of whether the selection of scans for analysis has been ‘good’ or ‘poor’. The theoretical fit duplicates the distortions introduced into the *g(s*)* or average *dc/dt* data (whichever you chose to fit) when the selected scans cover too broad a time span, including the discontinuities in the high S region.[6] Thus a good fit can still be obtained even in this extreme case, as shown in the ‘Fit Monitor’ screen below. There is some systematic deviation from the experimental data below ~1.6 S, suggesting that a second slowly-sedimenting component may be present; that will be discussed further in part C) below.

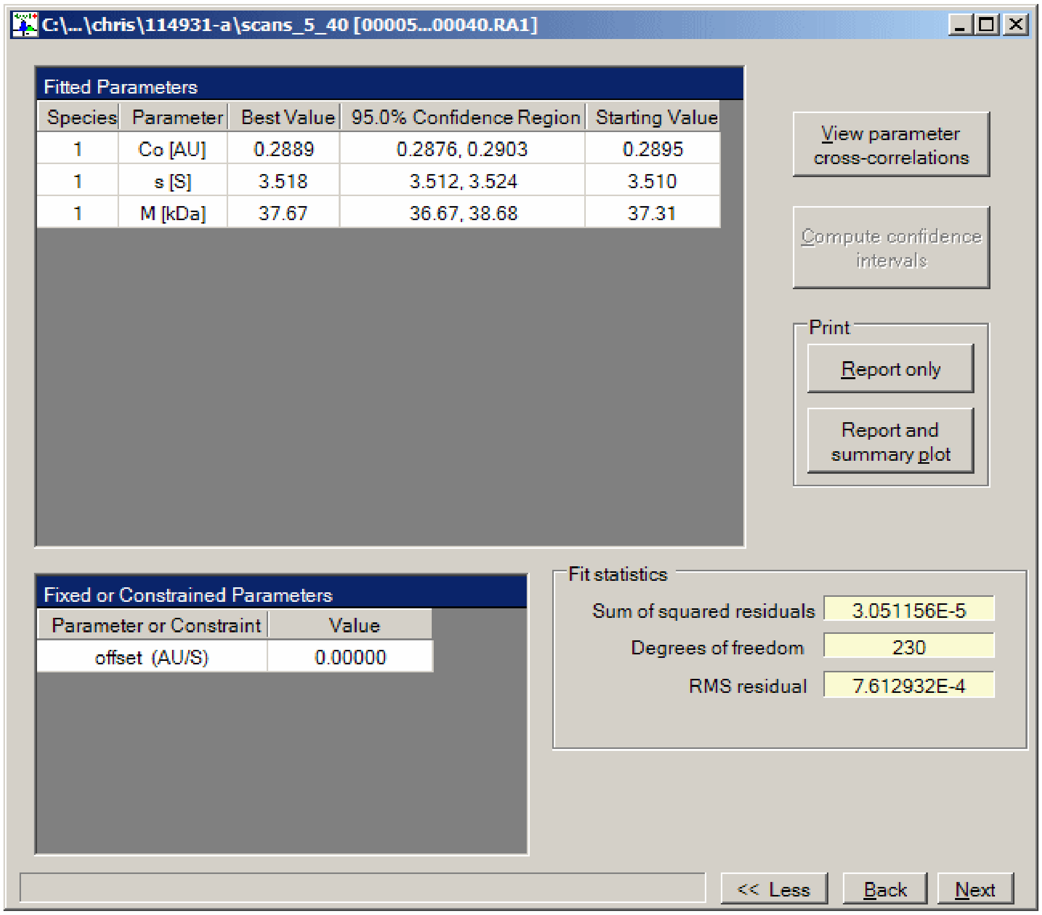
[6] These discontinuities occur at S values corresponding to the rightmost data point (highest radius) in the earliest scans. In this case only a single scan (the earliest one, scan 5 in this case) contributes to the average *dc/dt* curve in the region above ~10.7 S. The discontinuity at ~10.7 S then arises exactly where the second earliest scan starts to contribute, that at ~9.4 S is where the third earliest scan starts to contribute, *etc.* When a narrow range of scans is employed, as is usually done for *dc/dt* analysis, these discontinuities are quite small compared to the noise in the scans.



Status: fit completed

Iteration number: 4

The next screen shot gives the numerical results from this fit. The sedimentation coefficient is returned as 3.518 S [95% confidence 3.512 to 3.524]. The molecular mass value is somewhat lower than indicated by the 'automated' analysis in the section above (but the confidence intervals overlap), probably in part due to the influence of the systematic deviations at very low S values (the apparent mass rises if that region is excluded from the fit).


View parameter cross-correlations
Compute confidence intervals
Print
Report only
Report and summary plot
<< Less
Back
Next

C:\...\chris\114931-a\scans_5_40 [00005...00040.RA1]

Fitted Parameters

Species	Parameter	Best Value	95.0% Confidence Region	Starting Value
1	Co [AU]	0.2889	0.2876, 0.2903	0.2895
1	s [S]	3.518	3.512, 3.524	3.510
1	M [kDa]	37.67	36.67, 38.68	37.31

Fixed or Constrained Parameters

Parameter or Constraint	Value
offset (AU/S)	0.00000

Fit statistics

Sum of squared residuals	3.051156E-5
Degrees of freedom	230
RMS residual	7.612932E-4

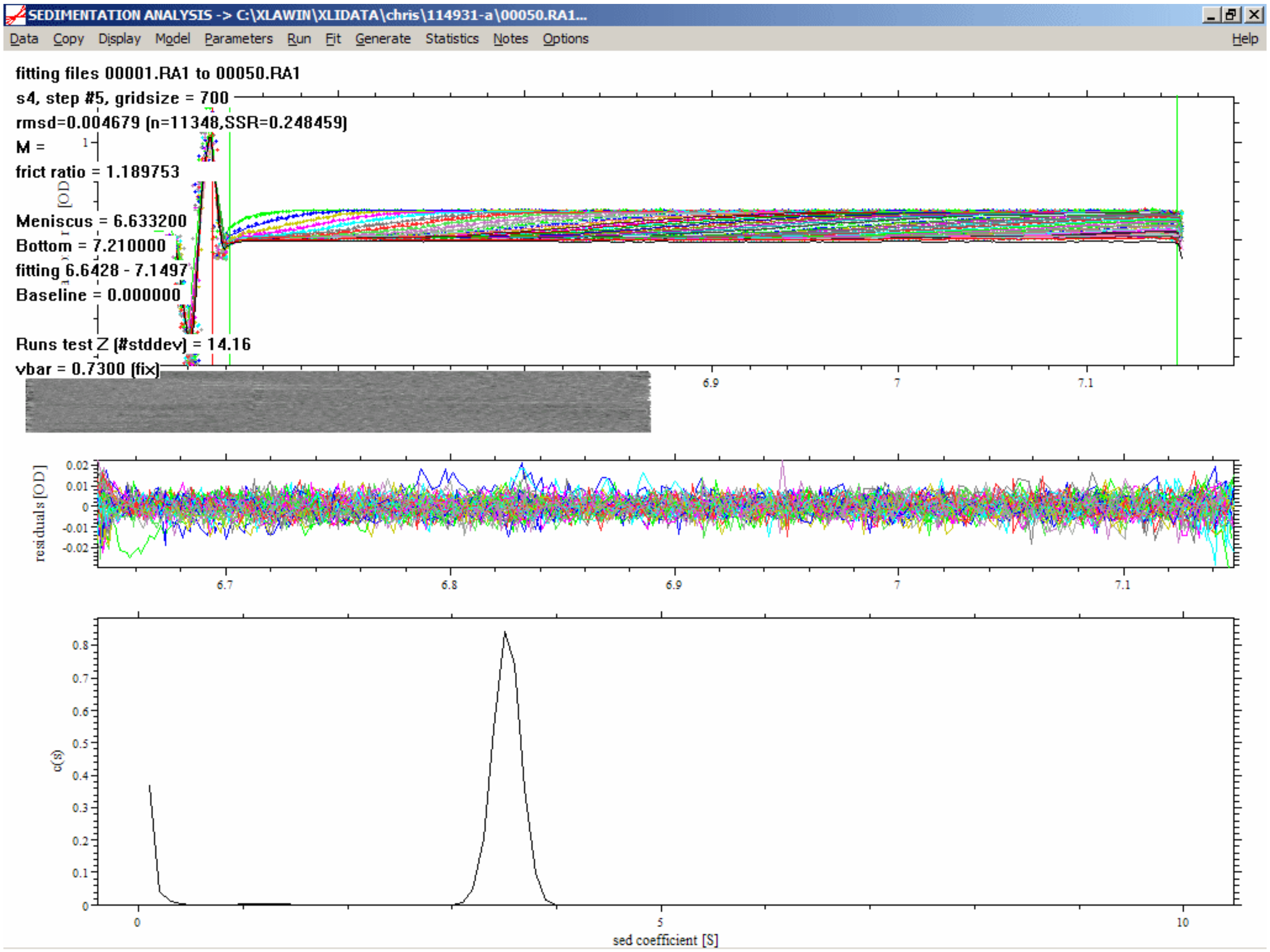
C) Analysis of scans 1-50 by the c(s) method

For Chris Chin's 'final' *c(s)* analysis he selected scans 1-50. Although Chris was trying to directly compare sedimentation coefficients from DCDT+ and SEDFIT, in DCDT+ he had requested internal conversion of the raw values to $s_{20,w}$ values (as can be seen from the axis labels on the graphs and the units on the fit results screen shots). It is unclear whether his results from SEDFIT were actually raw values or $s_{20,w}$ values.[7] Thus for that reason alone it is essential to repeat this analysis here.

The summary screen shot of my *c(s)* analysis is shown below.[8] The fitted f/f_0 ratio is expected to differ from that obtained by Chris because here the default values for density, viscosity, and vbar were used since the actual values are not known. The mean sedimentation coefficient of the main peak in this distribution (determined by integration) is 3.524 S, in excellent agreement with the

[7] As Peter Schuck recently discussed on the RASMB, SEDFIT generally gives results as raw sedimentation coefficients. It is possible though within SEDFIT to convert the *c(s)* graphs to the $s_{20,w}$ scale by using the first step of the 'Transform *s* distribution into *r* distribution' method under 'Size distribution options'. However after this transformation is done the *x* axis label does not change, so it impossible to tell from a screen shot whether this transformation has been carried out.

[8] This analysis was done using the same 6.6332 cm meniscus position as used by Chris, using 100 data points covering 0.1 to 10 S, and a regularization confidence level of 0.95.



SEDIMENTATION ANALYSIS -> C:\XLAWIN\XLIDATA\chris\114931-a\00050.RA1...

```
fitting files 00001.RA1 to 00050.RA1
s4, step #5, gridsize = 700
rmsd=0.004679 (n=11348,SSR=0.248459)
M =
frict ratio = 1.189753
Meniscus = 6.633200
Bottom = 7.210000
fitting 6.6428 - 7.1497
Baseline = 0.000000
Runs test Z (#stddev) = 14.16
vbar = 0.7300 (fix)
```

DCDT+ analyses.[9] If that peak is fitted with a Gaussian, as was done by Chris, this also gives a center position of 3.526 S (not shown). That value is just 0.1% lower than the 3.530 S value obtained by Chris, so clearly the *c(s)* results reported by Chris are actually raw sedimentation coefficients and thus should not have agreed with the $s_{20,w}$ values from his DCDT+ version 1.xx analysis.

However my *c(s)* distribution differs from that obtained by Chris in two significant ways. First, a second slowly-sedimenting component is detected at a mean sedimentation coefficient of 1.2 S (this peak is hard to see on the screen shot). This species represents 1.0% of the total sedimenting absorbance (excluding the half-peak at 0.1 S that is probably an artifact). Second, I did not obtain a half-peak at the upper limit of analysis (10 S). While many SEDFIT users apparently ignore such half-peaks, in my experience they often indicate the presence of real high molecular mass material that is sedimenting faster than the upper limit covered by the distribution.

[9] It is actually somewhat surprising the different methods give sedimentation coefficients that agree this closely for this particular experiment. Apparently there was some problem with the temperature control, and the temperature varied by 0.6° during the period the boundary reached about half-way down the solution column. Thus the sedimentation coefficient was actually varying quite significantly over the run, and therefore the average value will almost surely depend on exactly which scans are used, and will likely differ slightly from one analysis method to another.

One key difference between my *c(s)* analysis and that done by Chris is that Chris had left the position of the cell bottom marker (blue line) at 7.151 cm (the limit of the scan data, the position which gets set by default). (See the screen shot on page 5 of his report.) The actual position of the cell bottom is at a much higher radius; at 60,000 rpm the actual cell bottom is probably near 7.21 cm. The consequence of this incorrect cell bottom positioning was that the fits were very poor beyond 7.1 cm. At some point Chris obviously recognized that was a problem and therefore moved the right-hand data limit in to 7.08 cm.

For my analysis I moved the cell bottom to a fixed position of 7.21 cm and could then fit the data all the way to the end of the scans (I left the left-hand end of the fitting region at the same position Chris used). For these data there is no contribution from the accumulation of protein at the base of the cell (the sample is probably precipitating there) so the exact position of the bottom doesn’t really matter, as long as it is far enough to the right. One could also treat these data using the new ‘permeable bottom’ feature Peter Schuck added recently.

Is this second species real?

If my *c(s)* analysis is correct, then Chris Chin’s conclusion that his sample is a single species, based on his *c(s)* analysis, is technically actually wrong, although only if one is concerned about minor components at levels of a few percent. But is this peak near 1.2 S actually a real species?
One good indication that this species may be real is the fact that the analysis in part B) above showed systematic deviations in the region below 1.6 S. Indeed if one uses *dc/dt* analysis to examine groups of 10-14 scans starting around scan 20 there appears to be a shoulder or poorly-resolved peak near 1 S that represents several percent of the total. But if you look later in the run, starting around scan 28 where the resolution of this species should actually be better, that peak is no longer seen. Similarly if you confine the *c(s)* analysis to different subsets of the run this peak varies considerably in both magnitude and position. So something a bit odd is going on here! But the fact that similar phenomenon are seen by fairly orthogonal analysis methods suggests this is more likely some sort of experimental problem or artifact rather than simply a data analysis artifact.

In my opinion to really be certain about whether this sample contains minor species at a level of a few percent or less one would need (1) data where the physical separation is much better, which one could easily obtain by using a full cell (column height of 12 mm) rather than the ~4 mm column employed here, (2) data from an instrument that is working optimally, and (3) at least one replicate sample.

Conclusions

1. The sedimentation coefficients of the main species obtained via *c(s)* or DCDT+ agree very well.
2. The new fitting algorithm in version 2 of DCDT+ gives correct results even when the choice of scans causes gross broadening and distortion of the peaks.
3. Both analysis approaches can easily determine the sedimentation coefficient and/or other properties of the major component. However when it comes to minor components, or the measuring properties to the last decimal place, both approaches can be sensitive to details of scan selection, which regions of the cell are included in the analysis, positioning of the meniscus and/or cell bottom, *etc.*