[Return to Main Documents Page]

SASSIE-web: Quick Start

Introduction

SASSIE-web is an online simulation and analysis tool for the modelling of biomolecular structures using small angle scattering data. It is based on the original, standalone, program SASSIE (Curtis *et al.* 2012) and retains all of its core features. This guide is designed to get you familiar with the basic features of the program as quickly as possible. The features covered will be:

- Monomer Monte Carlo simulation to create an ensemble of varied trial protein structures
- Calculation of theoretical scattering from protein models
- Comparison of theoretical and experimental scattering data
- Minimization of those structures that best fit the experimental scattering data

Important Note: Before you start this tutorial you will need to register for an account for and login to SASSIE-web. Instructions on how to register can be found here.

The only data needed to work through this tutorial are:

- A starting structure in PDB format: gag_start.pdb
- A PSF topology file: gag_start.psf
- An experimental scattering curve: sans_data.sub

You should download these files to your computer now. A good idea is to create a directory called *SASSIE-web-tutorial* and save all downloads from the tutorial in this location.

You will also need to familiarize yourself with a molecular viewing program that can display PDB and DCD files. We recommend VMD and provide a quick tutorial here.

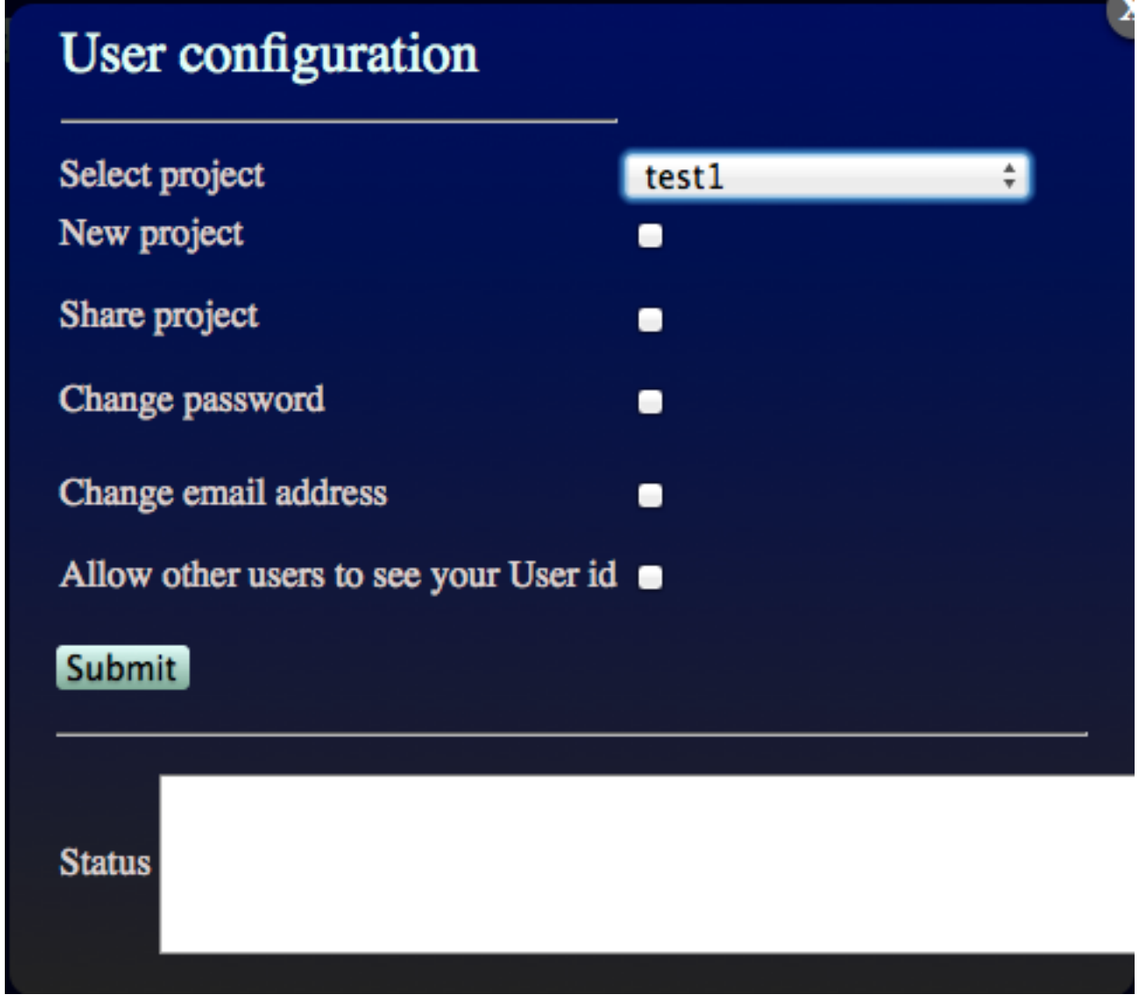
SASSIE-web Interface

Once logged in to SASSIE-web the page should look something like the figure below.

- **Session Management**: Section at top right containing the following icons:
  - Filing Cabinet: File browser
  - Cogs: Job management
  - Head: User configuration
- **Main Menu Toggle**: Lines at top left:
  - Hides or shows the main menu
- **Main Menu**: Provides links to the different categories of modules available in SASSIE

During this tutorial, when instructed to select something from the **Main Menu** but no menu is visible on the left hand side of the page you must click on the **Main Menu toggle** to reveal it.

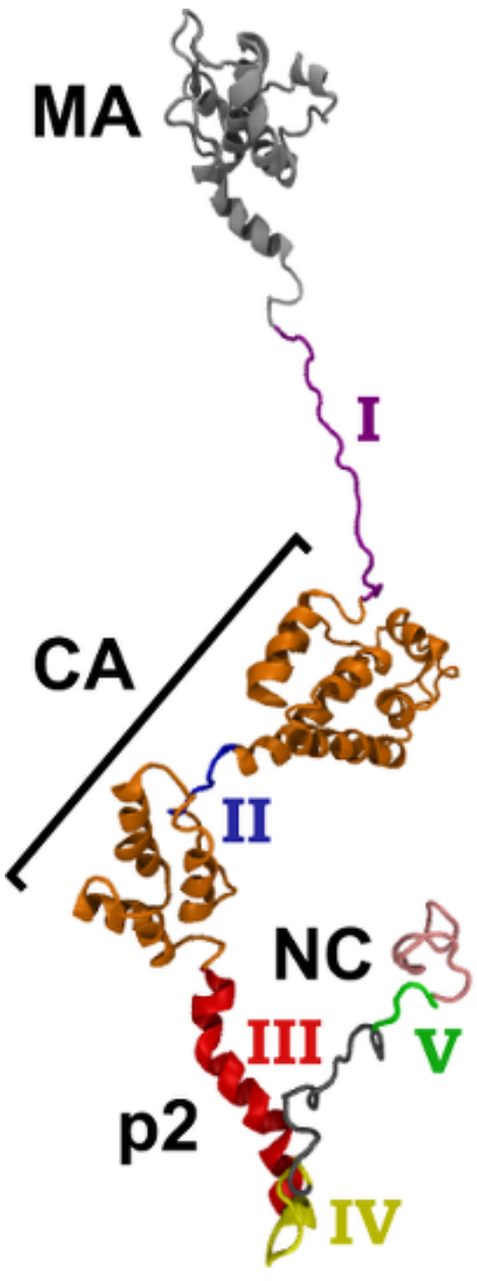
To choose a project name where your work will be stored, click on the Head icon. The User Configuration menu will appear.



Choose an existing project or create a new project name and click the 'Submit' button to connect to the project. The project name should now appear at the top of the web page. Here, we have chosen project name 'test1'.

Starting Structure

In this tutorial we will model the conformation of the HIV-1 Gag protein following the study of Datta *et al.* 2007. HIV Gag is a long polyprotein which is cleaved to form the functional proteins required by the virus. The viral proteins which form the domains in Gag are the matrix (MA), capsid (CA), p2 and nucleocapsid (NC). A structure stitched together from evidence based on crystal structures and models of the individual domains is shown below. A similar structure will be used as the starting structure for our simulations.



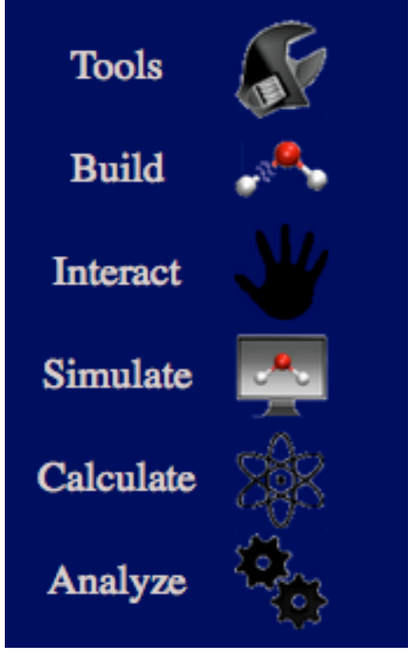
Domain	Flexible Region	Residues
MA		1 - 122
linker	I	123 - 144
CA		145 - 276
	II	277 - 282
		283 - 353
p2	III	354 - 377
linker	IV	378 - 389
NC		390 - 407
	V	408 - 412
		413 - 432

Datta *et al.* 2007 identified 5 flexible regions (labelled I-V), which are highlighted in the figure above. The table alongside the picture shows the residues which make up each region (we are going to need this information to select regions to be varied when we run Monte Carlo simulations).

Data Interpolation

Data interpolation is necessary to create a new data file that is spaced on a uniform grid from the experimental data file. More information on the module is available in the Data Interpolation documentation. Here we interpolate the SANS data from a HIV-1 Gag protein at a concentration of 1 mg/ml in a D2O buffer.

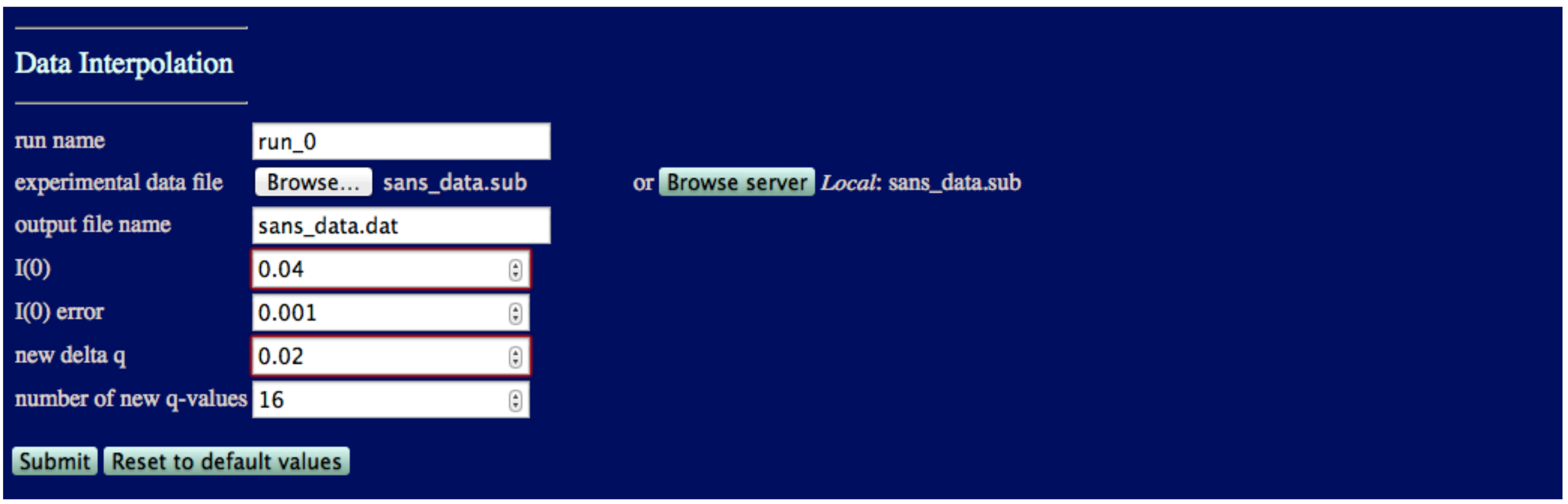
Select the 'Tools' button at the top of the **Main Menu**. Click on **Main Menu Toggle** if necessary.



This will reveal a list of buttons for each tool running across the top of the page (just below the top bar with the **Session Management** and **Main Menu Toggle** icons).

Select the 'Data Interpolation' button from this menu.

You should now see a page like the one below. This page is used to enter all of the information needed to do the data interpolation.



The figure shows the values for each field as required for data interpolation.

Edit the values on your screen to match the screenshot. An explanation of the field and how to edit it can be found below.

run name: user defined name of folder that will contain the results.

experimental data file: Name of input file with experimental data with at least three columns: q, I(q), and error in I(q). Here we use the *sans_data.sub* file.

output file name: Name of file that will contain the interpolated data. Here we choose the name *sans_data.dat*.

I(0): Experimentally determined value of scattering intensity at q = 0. Here we used the value of 0.04 that was derived from a Guinier fit to the data.

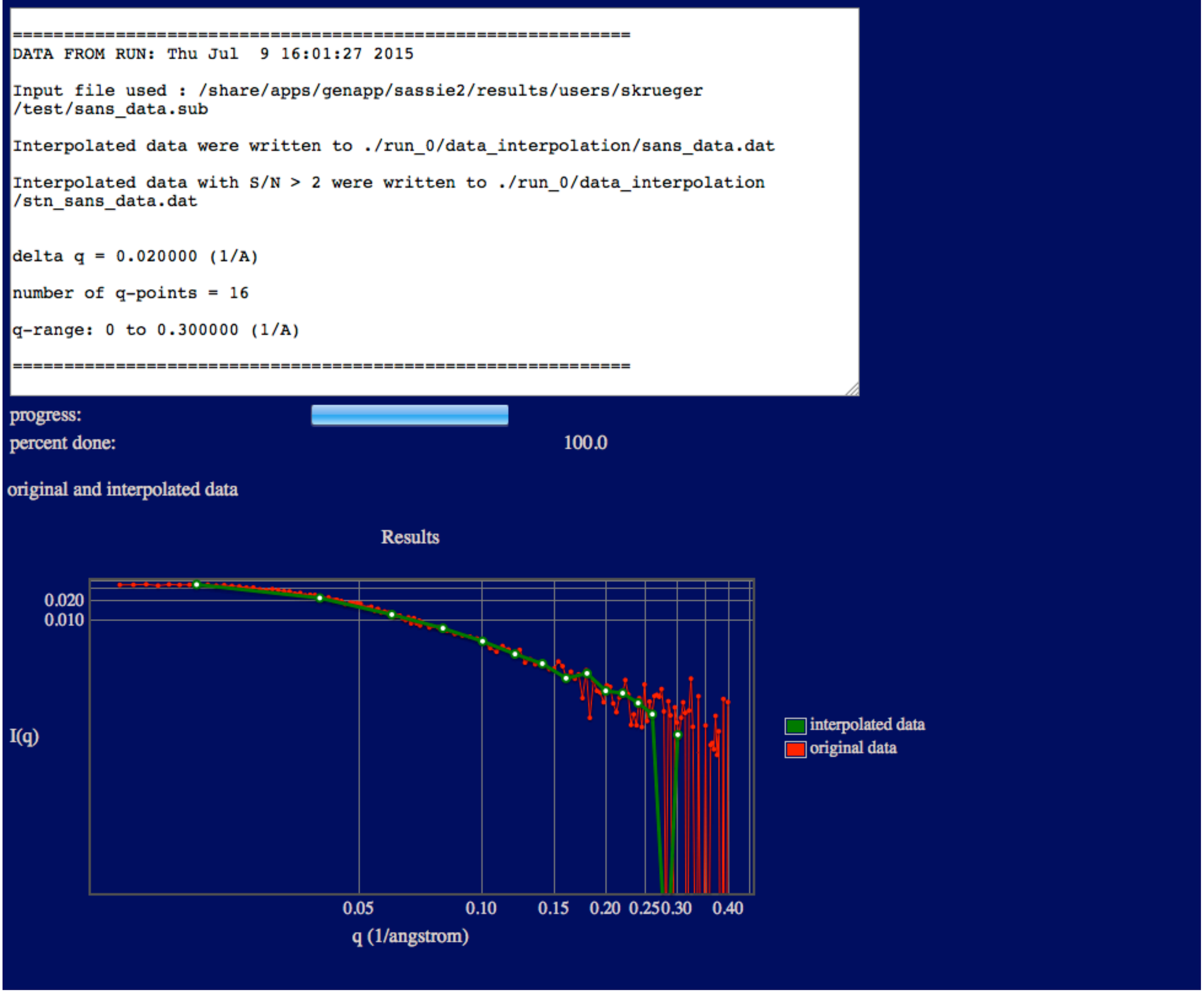
I(0) error: Experimentally determined value of the error of the scattering intensity at q = 0. Here we use the value of 0.001 that was obtained from the Guinier fit to the data.

new delta q: Desired spacing of q-values (1/Angstrom). This should be chosen so that your first interpolated data point falls within the q-range of the experimental data. For this tutorial, the value has been set to 0.02 since the first data point occurs at a value of ~0.013.

number of new q-values: Integer number of desired q-values. For this tutorial, the value has been set to 16 to that the maximum q value is 0.3.

Once you have understood the input fields and made sure that your values agree with the figure click on the 'Submit' button to start simulation.

As the run continues the progress bar beneath the submit button should update. A graph beneath this should will show the variation of the radius of gyration over the steps of the Monte Carlo simulation. Once complete the output should look similar to the figure below.



The output will show a plot of the original and interpolated data, the name of the input file and the name of the interpolated data file as well as the directory in which it is located.

Note that roll-over help will indicate options to resize, zoom and reset the view of the plot.

What have we generated:

test1/run_0/data_interpolation

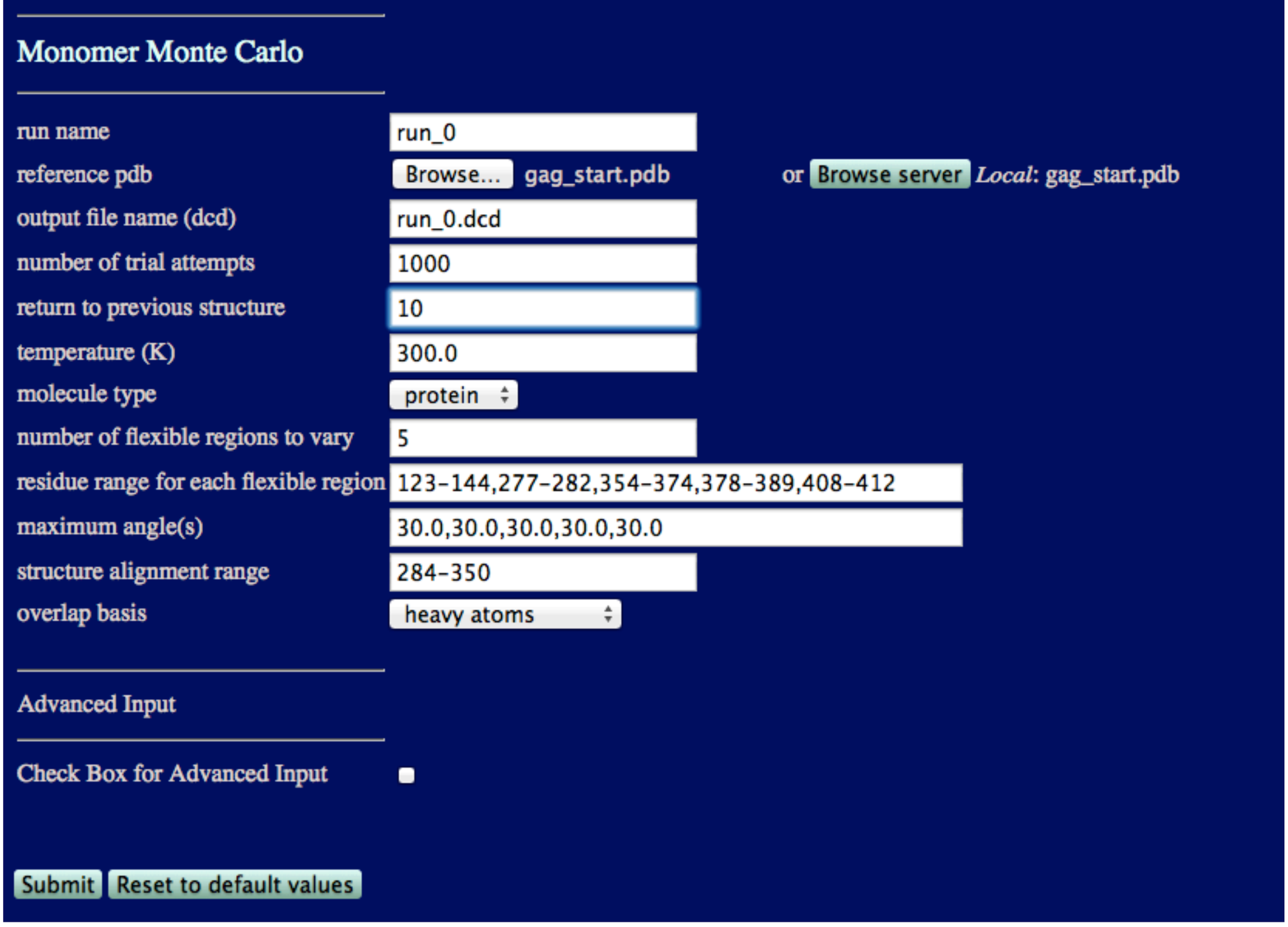
- *sans_data.dat*: text file containing the interpolated SANS data
- *stn_sans_data.dat*: text file containing a the interpolated SANS data truncated at the q-value where the signal-to-noise drops below a value of 2

Structure Variation - Monte Carlo (Monomer)

The primary way to vary structures in SASSIE is via Monte Carlo simulations which rotate the backbone dihedral angles of flexible regions within proteins to sample a wide range of structures. More information can be found in the Monomer Monte Carlo documentation. Here we setup and run such a simulation before visualizing the range of stuctures produced.

Select the 'Simulate' button from the Main Menu of SASSIE-web and then click on the 'Monomer Monte Carlo' button.

You should now see a page like the one below. This page is used to enter all of the information needed to run a Monte Carlo simulation.



The figure shows the values for each field as required for our simulation. An explanation of some of the fields can be found below.

reference pdb: The starting structure for the simulation. Here we use the *gag_start.pdb* file.

number of trial attempts: Number of times the simulation will try to vary the structure (some structures will be discarded by the Monte Carlo algorithm) For this tutorial set the value to 1000. For real studies tens of thousands of structures are needed.

return to previous structure: Number of discarded structures in a row that are considered before returning to a randomly-selected structure that was previously accepted

number of flexible regions to vary: single number

residue range for each flexible region: comma-separated list of the range of residues to vary for each flexible region

maximum angle(s): comma-separated list of the maximum angle sampled in a single Monte Carlo step for each flexible region

structure alignment region: a single range of residues for structural alignment of all the flexible segments. This makes it easy to make visual comparisons of each frame in the output trajectory.

overlap basis: Select either heavy atoms, all, backbone or enter atom name. The atom name option will spawn futher inputs:

- **overlap basis:** Enter an atom name to check for overlap.
- **overlap cutoff (angsgtroms):** Overlap basis atoms closer than this distance defines an overlap condition.

Once you have understood the input fields and made sure that your values agree with the figure click on the 'Submit' button to start simulation.

As the run continues the progress bar beneath the submit button should update. A graph beneath this should will show the variation of the radius of gyration over the steps of the Monte Carlo simulation. Once complete the output should look similar to the figure below.

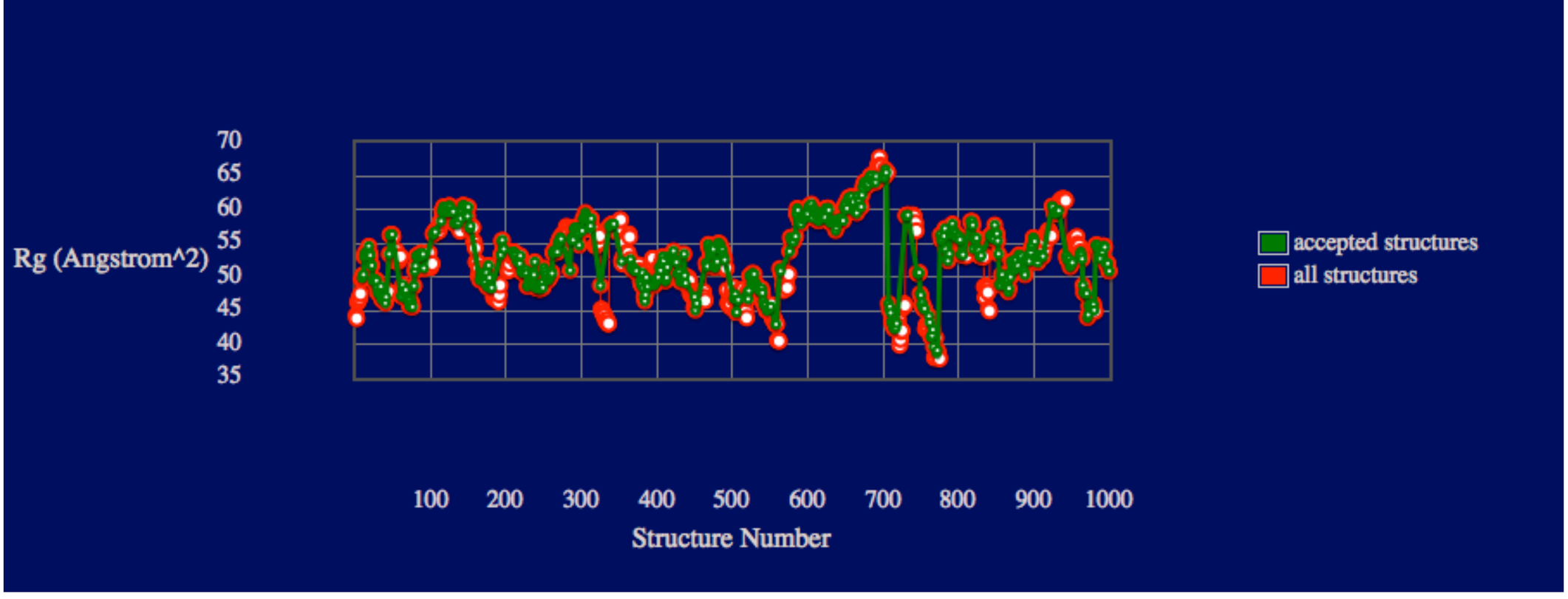
```
============================================================
DATA FROM RUN: Fri Jul 10 09:40:22 2015

Average accepted rg2 = 53.429966

Configurations and statistics saved in ./run_0/monomer_monte_carlo/ directory

lowest Rg = 37.832101    highest Rg = 67.812830
accepted 692 out of 1000 : 69.200000 percent
overlap check discarded 308 out of 1000 moves : 30.800000 percent
Rg cutoffs discarded 0 out of 1000 moves : 0.000000 percent

minimum x = -106.000069  maximum x = 144.011806 -> range: 250.011875 Angstroms
minimum y = -102.160418  maximum y = 95.319356 -> range: 197.479775 Angstroms
minimum z = -112.573038  maximum z = 78.222151 -> range: 190.795189 Angstroms

============================================================
```

progress:

percent done: 100.0

all rg and accepted rg data

Rg Results



What have we generated:

test1/run_0/monomer_monte_carlo

- *run_0.dcd*: DCD file containing all of the structures accepted by out Monte Carlo simulation
- *run_0.dcd.accepted_rg_results_data.txt*: text file containing radius of gyration for all structures that made it into the DCD file
- *run_0.dcd.all_rg_results_data.txt*: text file containing radius of gyration for all structures generated
- *run_0.dcd.stats*: text file containing statistics for our Monte Carlo run

Visualization

You should now download the output trajectory using the file browser.

- Click on the filing cabinet icon in the **Session Management** area.

Note: the 'Configurations and statistics saved in' line in the output gives a relative path under the project directory.

- Click on the triangle next to the 'test1' directory name to reveal the 'run_0' directory created by our simulation.
- Check the box next to 'run_0' to select it for download.
- Beneath the file tree is an option labelled 'Compression type'. Select an option suitable for your operating system from the list (for Windows select 'zipped' for Linux 'bzip2 tarball').
- Click the 'Download' button.

A progress bar will appear monitoring the upload of your files to the server. Once complete a link will appear beneath the download button.

- Click on this to download your data.

Once the download is complete uncompress the file in a location of your choice.

You should now load the PDB gag_start.pdb (you will find this in the run_0 directory you just downloaded) and DCD (run_0/monomer_monte_carlo/run_0.dcd) into VMD to observe the variation produced even in our very short Monte Carlo simulation. Remember the DCD file contains coordinates alone, you need to load the PDB first so that the visualization software knows about the atoms they represent and how they are connected. You can also download and visualize the DCD file run_0.dcd generated from this particular Quick Start run for comparison.

SAS Curve Calculation

Next we calculate a theoretical scattering curve for each of the trial structures we have generated. The SASSIE workflow operates by calculating the scattering intensities at evenly spaced Q values and matching these against interpolated experimental values.

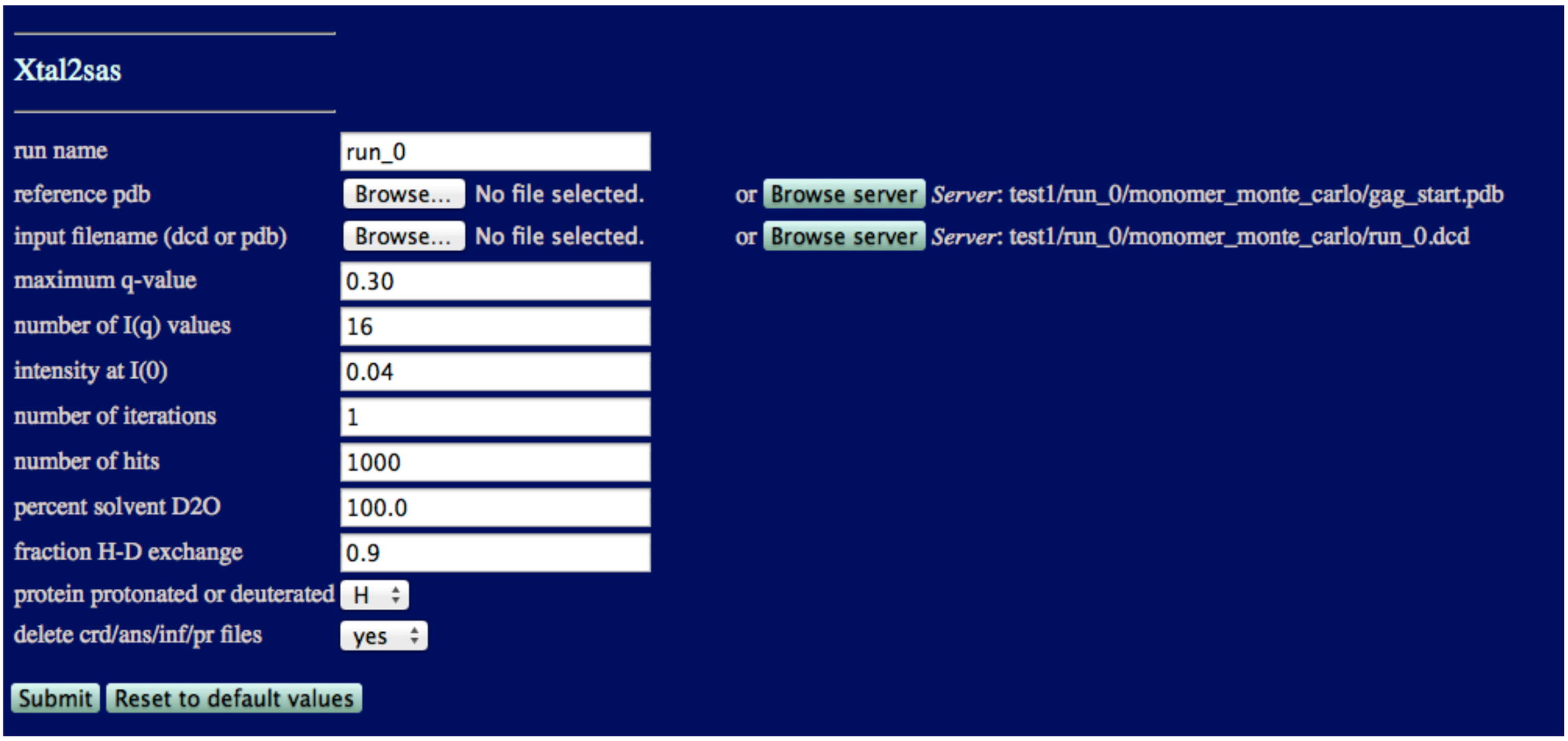
The file sans_data.dat contains our previously interpolated experimental data. In order to create the correct data points in our theoretical curves we need three pieces of information:

- Intensity and Q=0, I(0): 0.04
- Maximum value of Q: 0.3 (units are inverse Angstroms)
- Number of points in the curve: 16

A number of scattering calculators are available in SASSIE. Here we use Xtal2sas. More information can be found in the Xtal2sas documentation.

- Select 'Calculate' from the Main Menu.
- Click the 'Xtal2sas' button.

Now you need to enter the information to run the scattering calculator. Other than the values listed above you can keep the default values for this tutorial (see figure below).



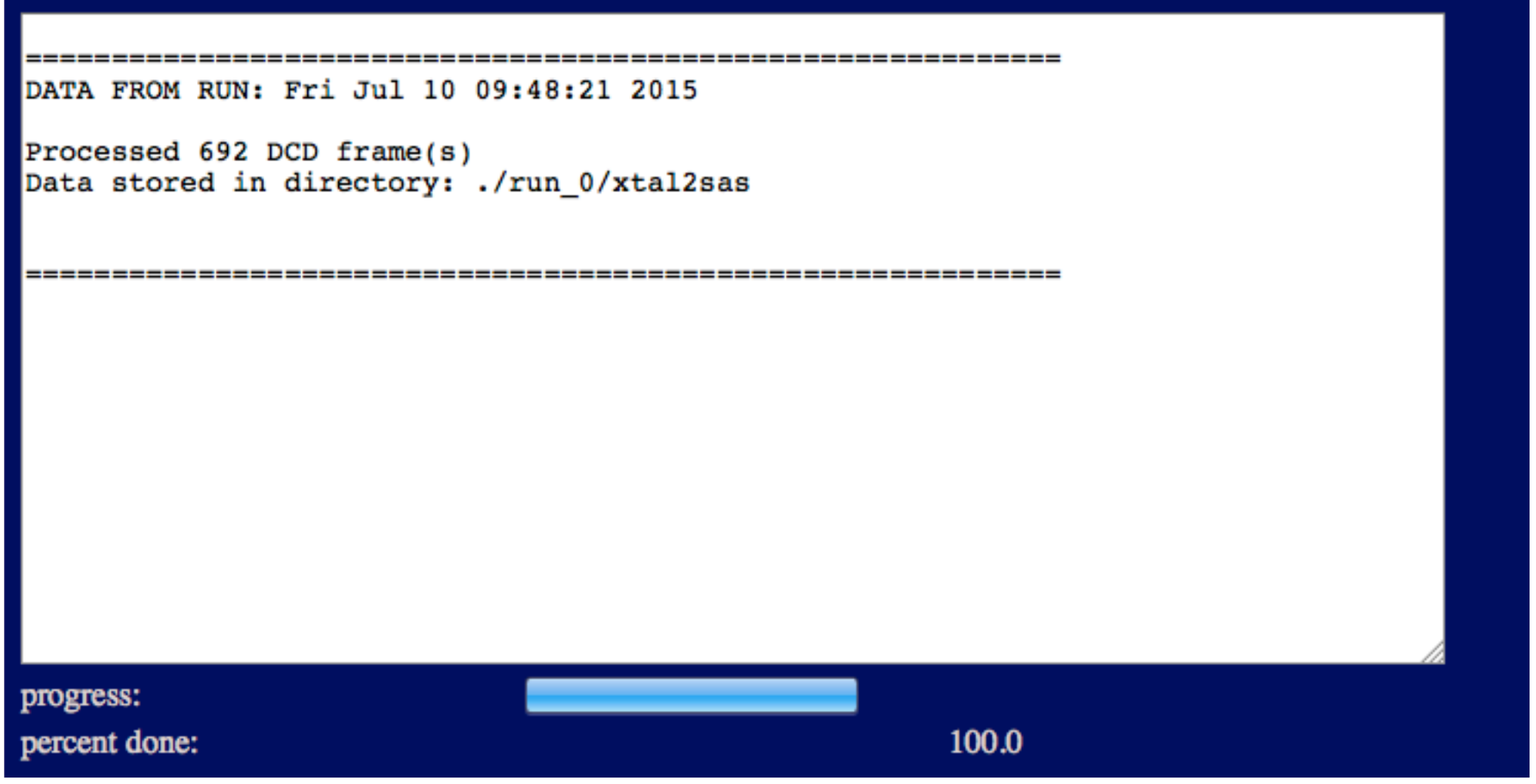
The **reference pdb** is the same one we used to start the simulation and the **input filename** (DCD) comes from the previous step, so both are already uploaded to the SASSIE-web server. To select both of them follow you the same procedure:

- Click on the 'Browse server' button next to the appropriate field.
- Navigate to the file requires
  - PDB: test1/run_0/monomer_monte_carlo/gag_start.pdb
  - DCD: test1/run_0/monomer_monte_carlo/run_0.dcd
- Click 'OK'

Now all input fields should be complete.

- Click 'Submit'

A scattering curve will be calculated for all of the structures generated by the Monte Carlo simulation (the progress bar should reach 100% and a message stating the run finished appear in the window beneath when the job has completed).



What have we generated:

test1/run_0/xtal2sas

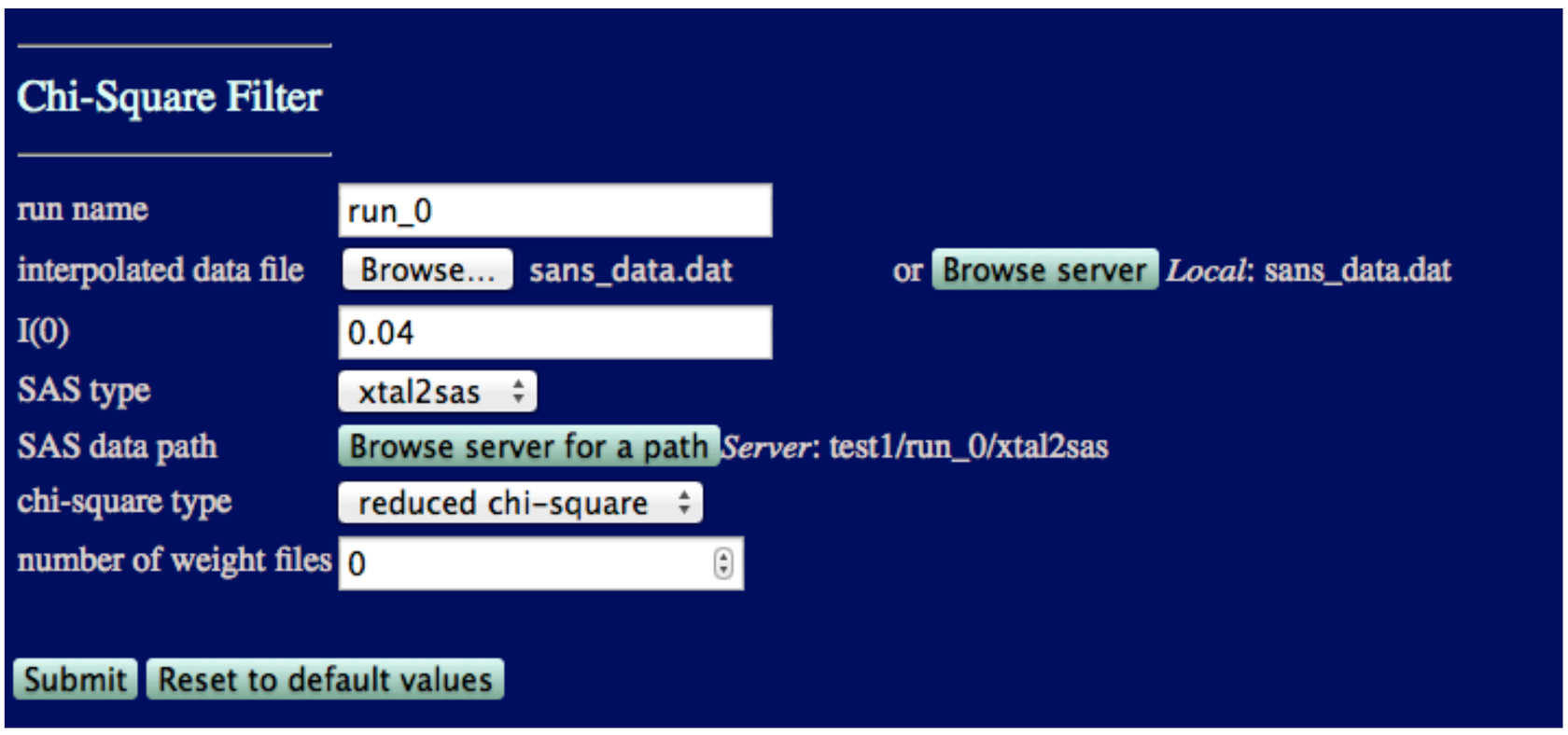
- **.iq*: files for each structure we generated containing theoretical scattering data
- **.log*: log files for each run of xtal2sas
- **.ans*: input file used for each xtal2sas run

Initial SAS Curve Comparison

Now we compare our theoretical curves to the experimental data to see which of our structures are plausible models of the real protein using Chi-Square Filter. More information can be found in the Chi-Square Filter documention.

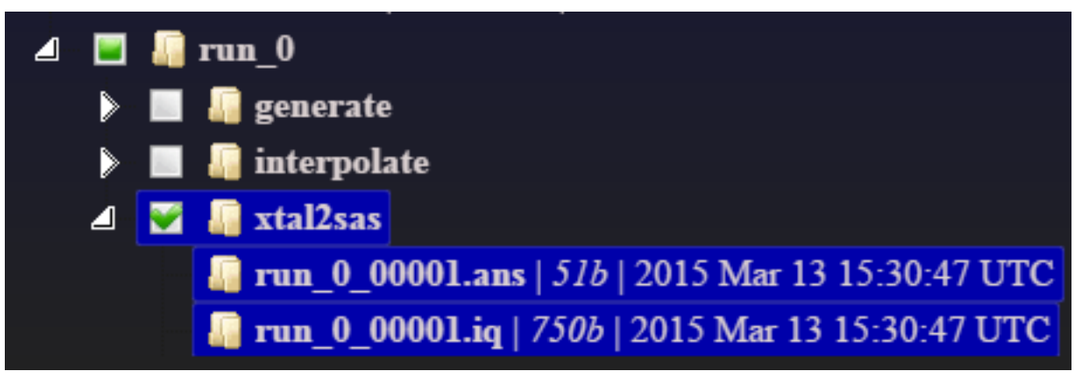
- Select 'Analyze' from the Main Menu.
- Click the 'Chi-Square Filter' button.

We now need to select the path containing the theoretical scattering curves and the file containing the experimental data. In addition we need to input the value of I(0) to enable comparison of the two curves (see the picture below).



To set the path to the scattering curves generated in the previus step:

- Click on the 'Browse server for a path' button
- Navigate to the test1 project folder and select the run_0/xtal2sas folder (see picture below)



- Click 'OK'

interpolated data file

- Click on the 'Choose File' button

- Navigate to and select the sans_data.dat file on your local computer
- Click 'OK'

I(0)

- Enter the value 0.04

We eventually may want to create 'weight files' that record which frames meet criteria that make them successful models of our data. This means those with low chi square values. However, we don't know the range of chi square values we have at this stage. So, we set the 'number of weight files' to 0 at this time.

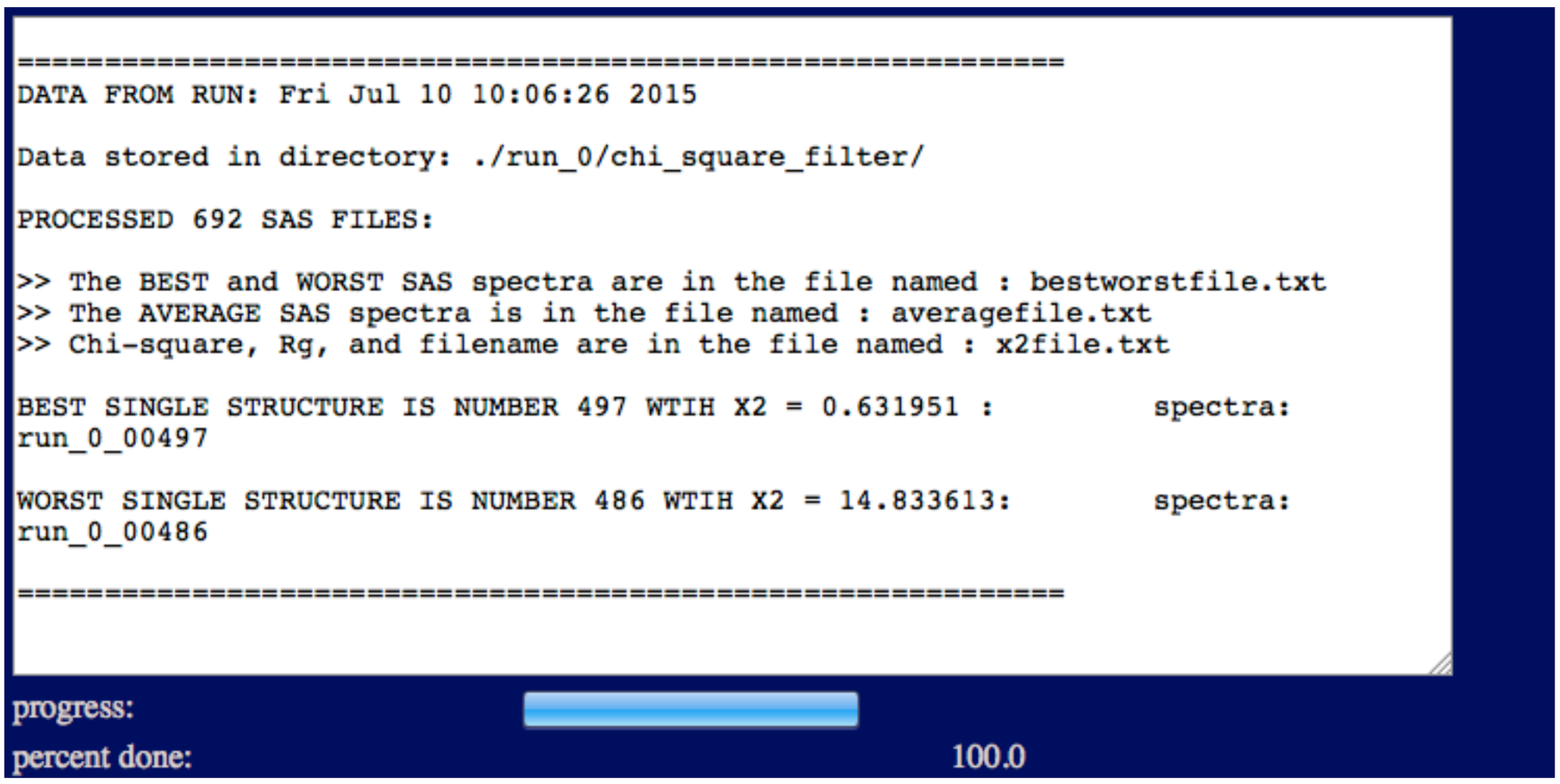
number of weight files

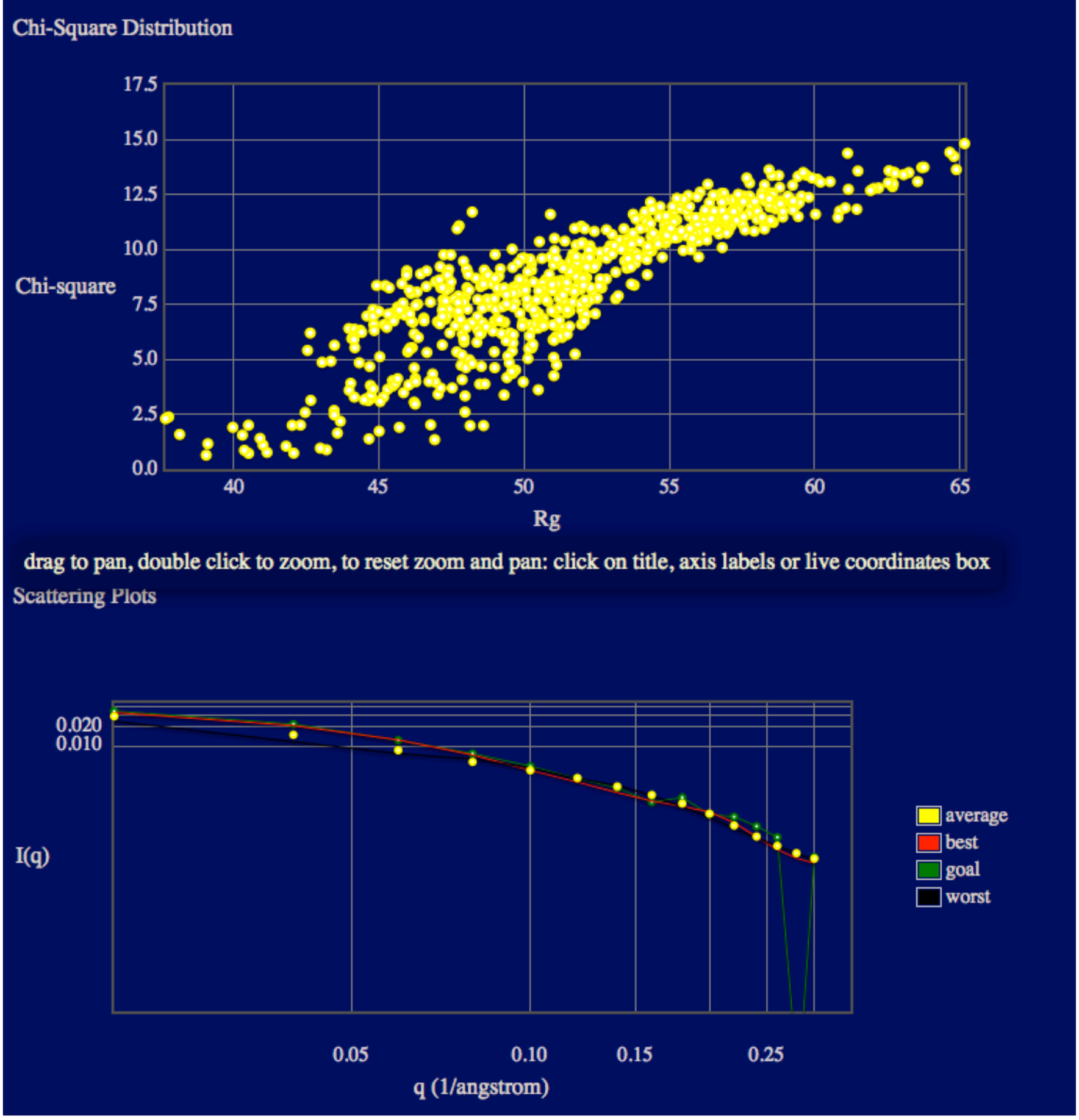
- Set this to 0 using the down arrow associated with the input box. If you type '0' in the box, press the TAB button to make sure this value is accepted.

Note: There are list boxes that allow the selection of the format of the input theoretical curves and the metric used to compare the curves. Here we wish to use the defaults of 'xtal2sas' and 'reduced chi-square'.

Click 'Submit'.

Once complete you the run you should see outputs similar to those below.





In the text output you will see the minimum chi square (X2) values is given.

The top plot shows the variation of chi squared (y-axis) with the radius of gyration (x-axis). Chi squared is a measure of the quality of fit of the theoretical curve to the experimental one. It is a percentage and the lower the value the better.

The bottom plot shows a direct comparison of the best, worst and average theoretical curves with experiment (goal).

What have we generated:

test1/run_0/chi_square_filter

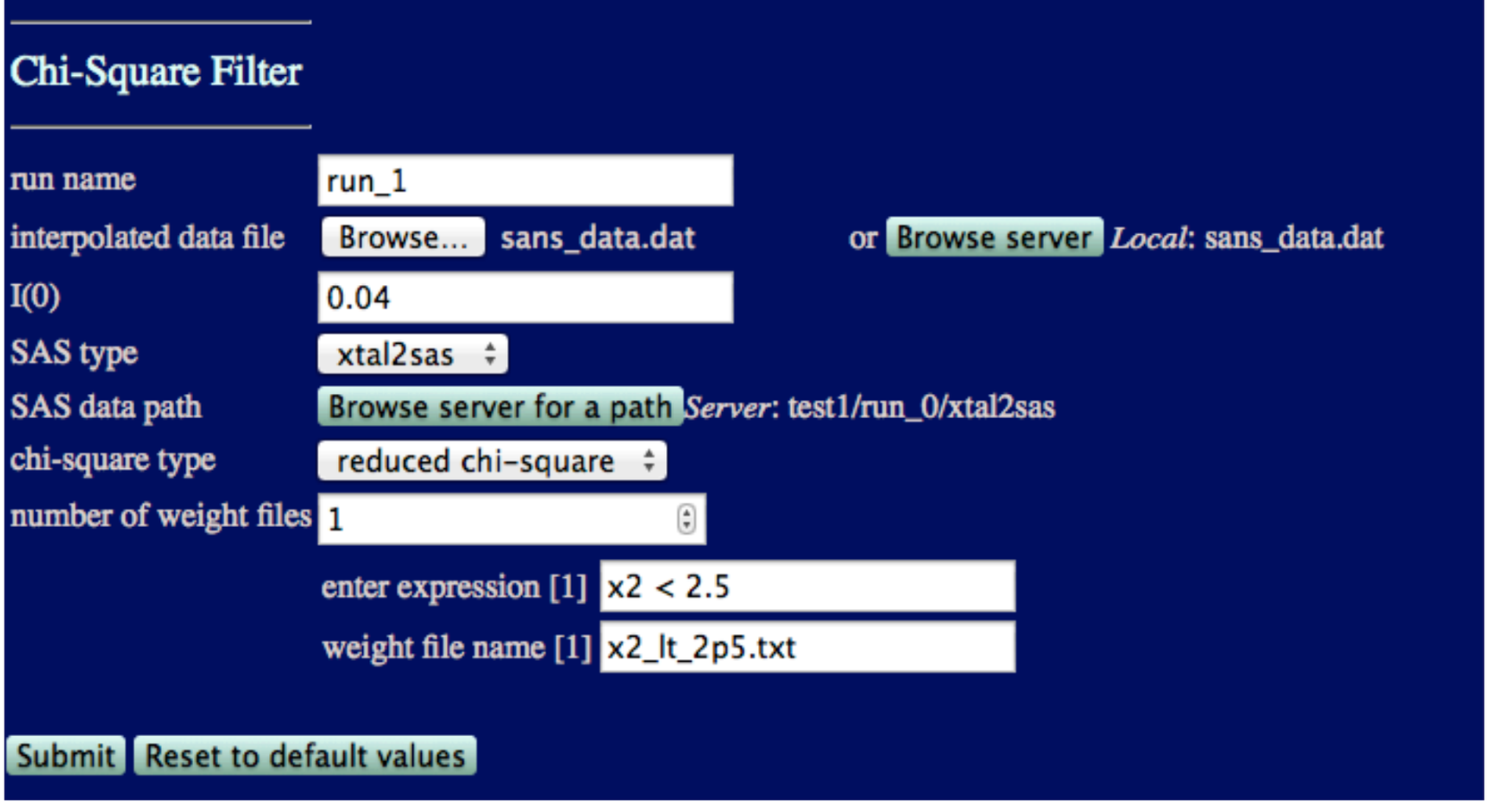
- *averagefile.txt*: average scattering curve for all structures
- *bestworstfile.txt*: best and worst scattering curves selected from all structures
- *sas_spectra_plot.txt*: goal, best, worst and average scattering curves
- *x2_vs_rg_plot.txt*: chi squared against radius of gyration for all structures
- *x2file.txt*: chi squared for all structures

spectra

- **.ciq*: scattering curves scalled to correct I(0) for each structure

Second SAS Curve Comparison

Now that we know the range of chi square values that we have, we can compare the theoretical curves to the data a second time and create a weight file that flags all structures with chi square values below a certain number. Now, we set the 'number of weight files' to 1.



run name:

- Since we already have a chi_square_filter folder in the run_0 directory, set the run name to run_1.

number of weight files

- Set this to 1 using the down arrow associated with the input box. If you type '1' in the box, press the TAB button to make sure this value is accepted.

Weight files contain information on which frames in our simulation meet specific criteria provided in the expression box.

enter expression

- Enter the following expression:

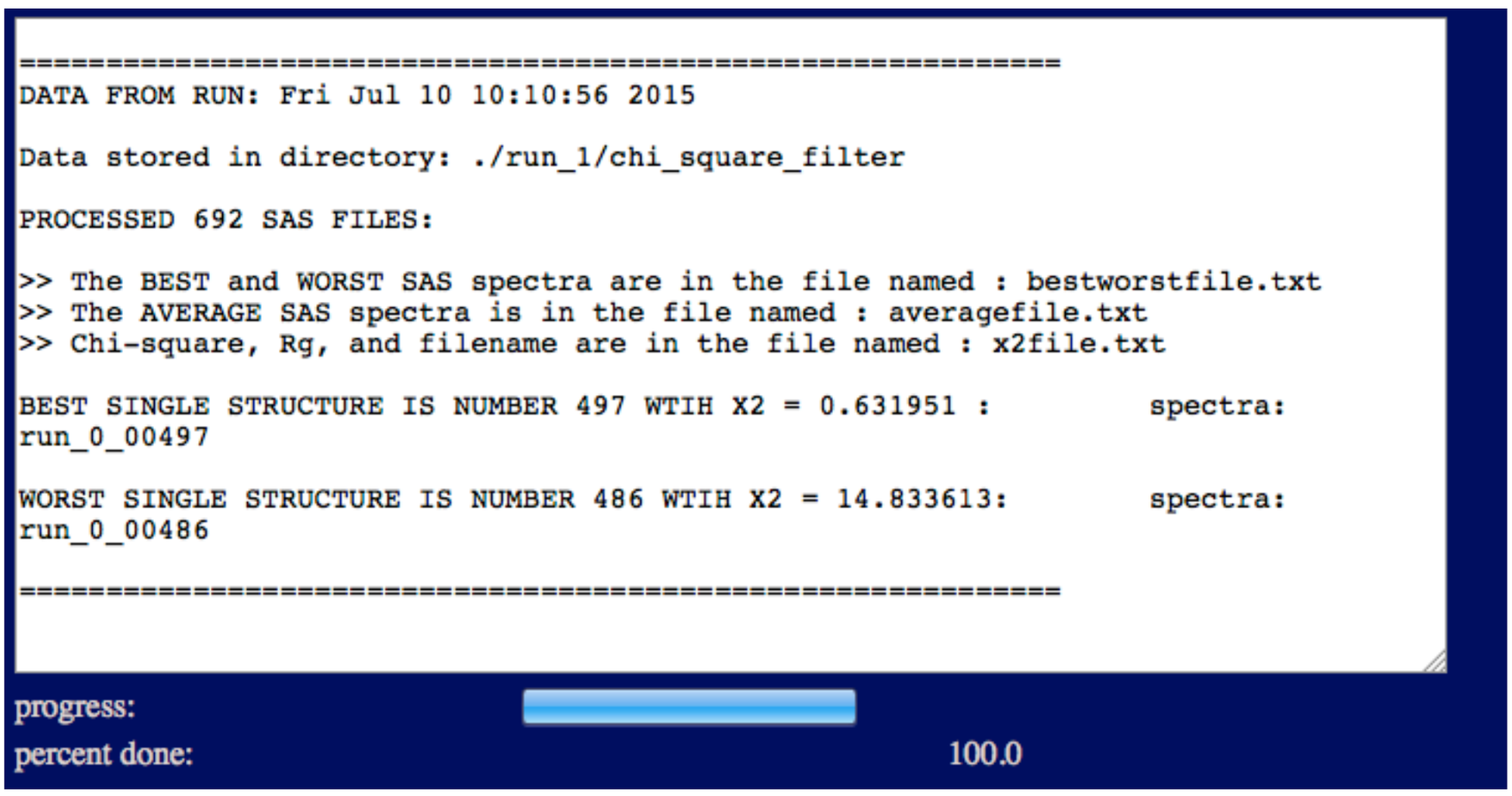
```
x2 < 2.5
```

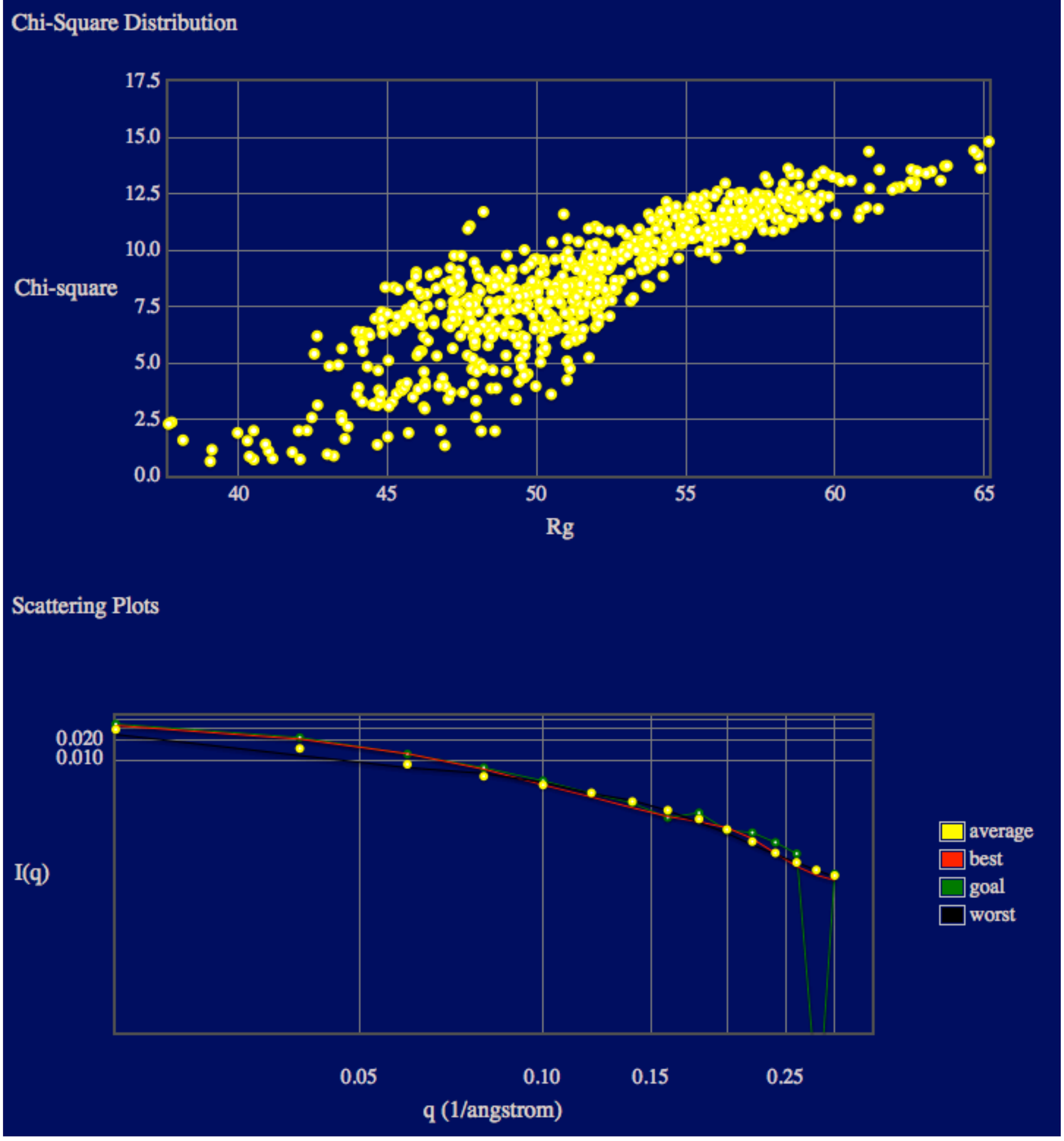
This selects all frames with a chi square less than 2.5. Adjust this value if necessary to suit the results from your simulation.

weight file name

- Enter x2_lt_2p5.txt
- Click 'Submit'.

Once complete you the run you should see outputs like those below.





These results are essentially the same as those from our first comparison above except that we have now generated a weight file.

What have we generated:

test1/run_1/chi_square_filter

- *averagefile.txt*: average scattering curve for all structures
- *bestworstfile.txt*: best and worst scattering curves selected from all structures
- *sas_spectra_plot.txt*: goal, best, worst and average scattering curves
- *x2_vs_rg_plot.txt*: chi squared against radius of gyration for all structures
- *x2file.txt*: chi squared for all structures
- *x2_lt_2p5.txt*: weights file selecting only frames with chi squared < 2.5

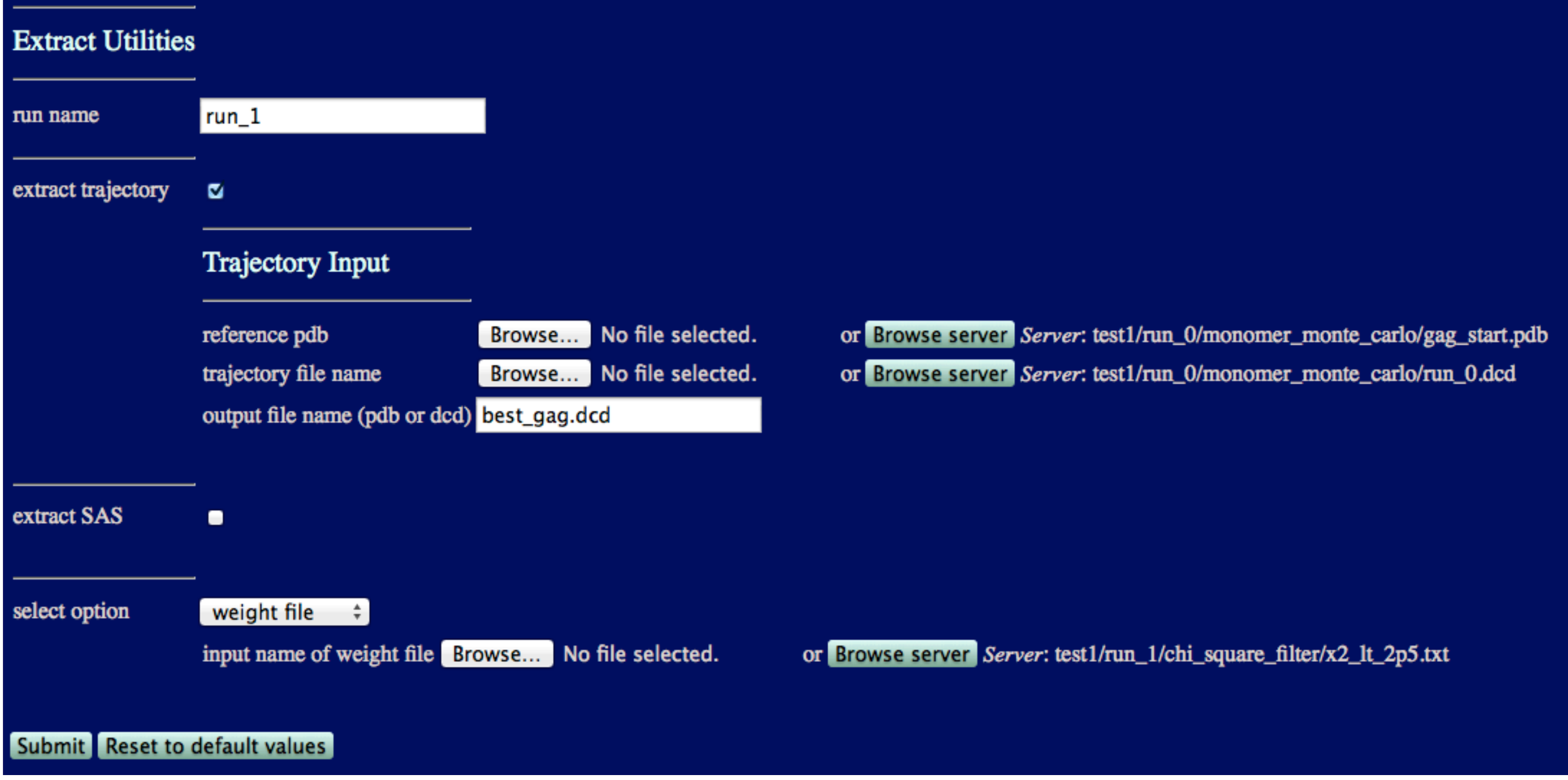
spectra

- **.ciq*: scattering curves scalled to correct I(0) for each structure

Trajectory Filtering

Now we can filter out the best fit structures and vizualize them using the Extract Utilities. More information can be found in the Extract Utilities documentation.

- Select 'Tools' from the Main Menu.
- Click the 'Extract Utilities' button.

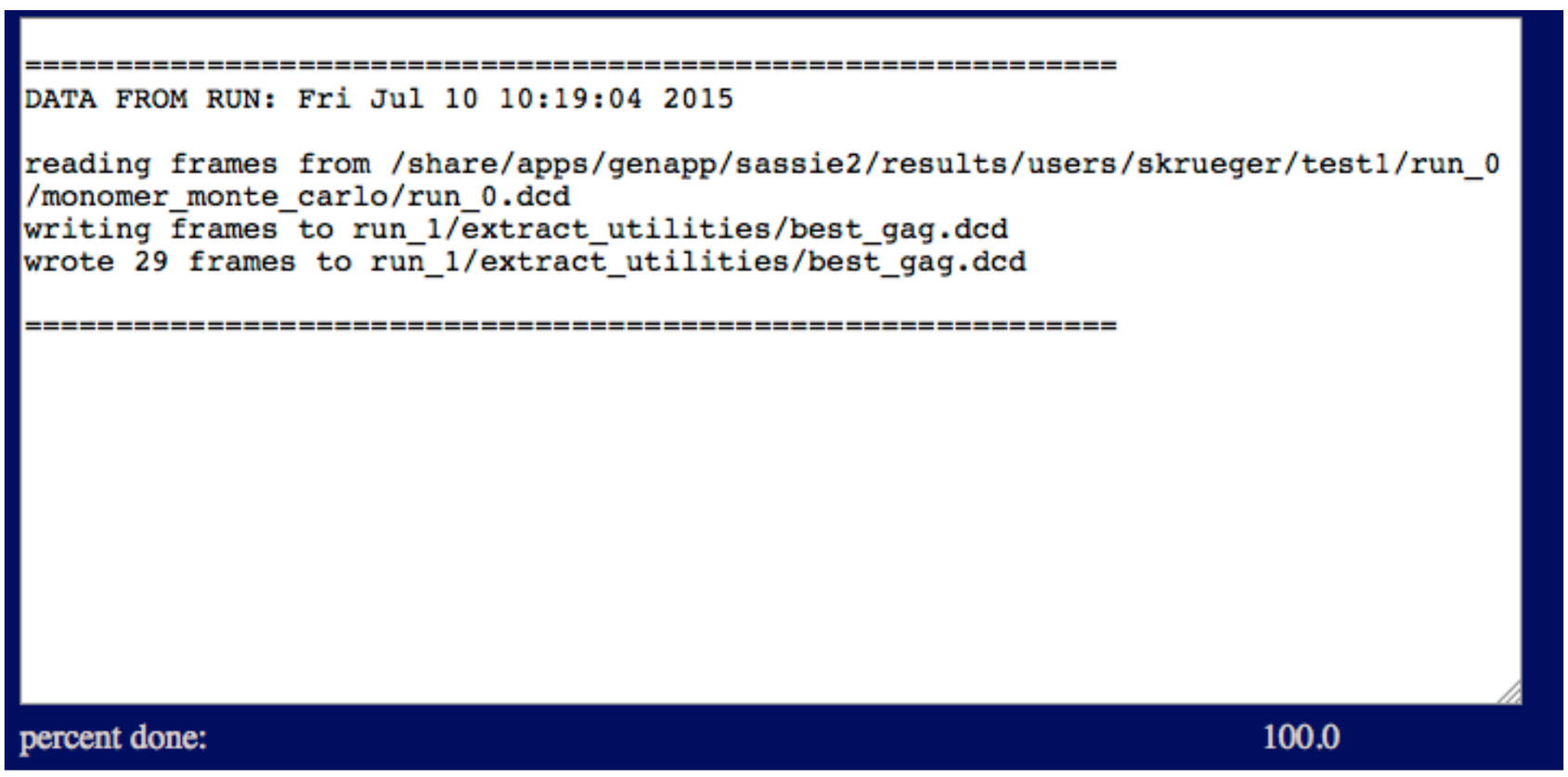


In this module we can select structures from the DCD we created from the Monte Carlo simulation using the weight files generated in the Chi-Square Filter module.

By now you should be familiar with how to select files from the server, so no detailed instructions are provided here.

- Check the tick box labelled 'extract trajectory' (this will reveal the options shown in the screenshot)
- Select the usual 'reference pdb' and the DCD output from the Monte Carlo simulation
- Input 'best_gag.dcd' as the 'output filename'
- Choose 'weight file' from the 'select option' listbox.
    - Where it says 'input name of weight file' select the 'x2_lt_2p5.txt' file generated in the last step that selects only the frames which had a chi squared value of less 2.5 (You may have chosen a different value and file name.)
- Click 'Submit'

When the process is finished your output should look like the one below.



In the event that none of your frames pass the filter then you can download these preprepared files and try the filtering process:

- DCD: run_0.dcd
- weights file: x2_lt_2p5.txt

What have we generated:

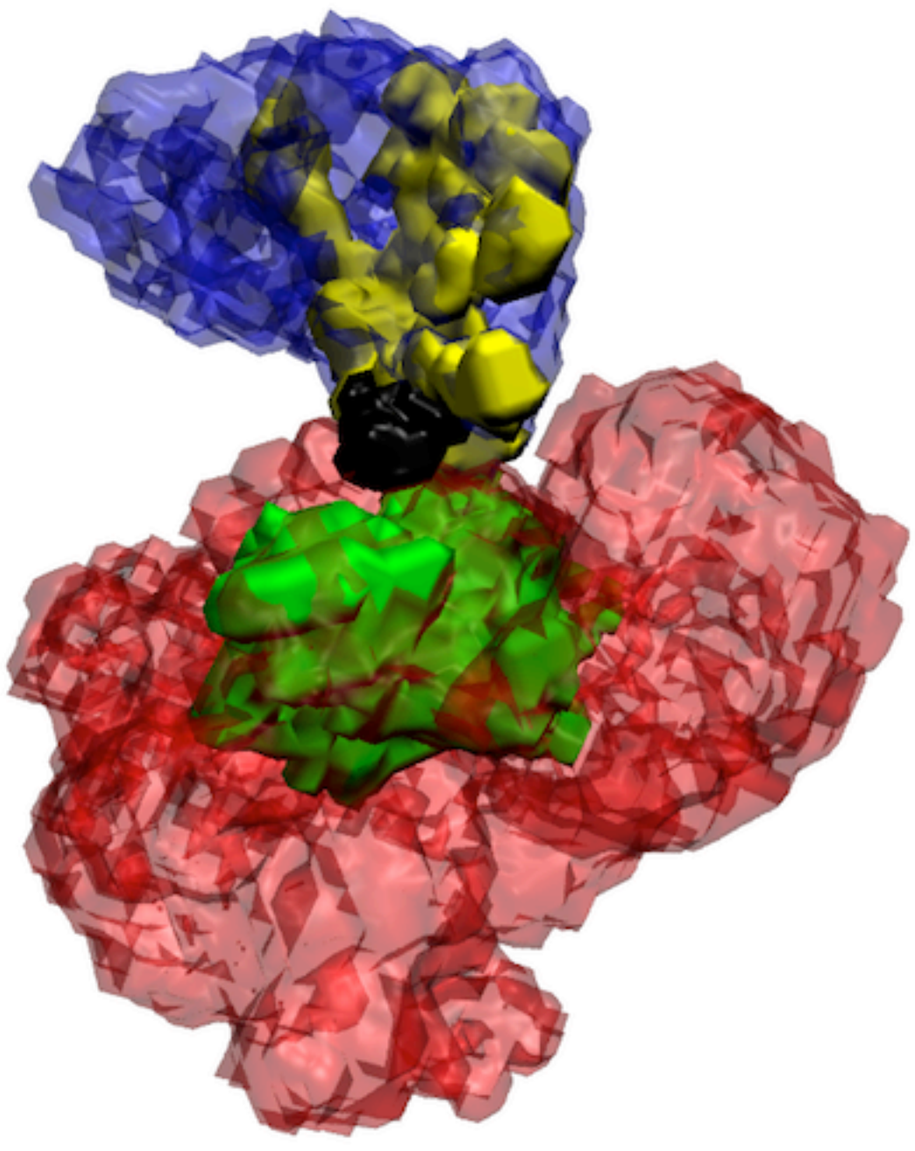
test1/run_1/extract_utilities

- *best_gag.dcd*: DCD containing only the frames for which the theoretical scattering curve is a good match to experiment.

Visualization

Download the 'best_gag.dcd' file as you did the unfiltered DCD and then vizualize the structure again in VMD (you will need to load a suitable PDB first as before). You should see that the filtered structures are all noticeably more compact that the starting structure and the majority of those in the unfiltered DCD.

Another way to visualize the structures sampled in the 'run_0.dcd' and 'best_gag.dcd' files for comparison is to use the Density Plot module. The density plot below shows the envelope sampled by all of the accepted structures as well as that sampled by only the best fit structures. The black region is the envelope represented by residues 283-353, which is approximately the alignment region that we defined in the Monomer Monte Carlo module. The blue and yellow regions represent the envelope sampled by residues 1-282 for all accepted structures (blue) and for the best fit structures (yellow). The red and green regions represent the envelope sampled by residues 354-431 for all accepted structures (red) and the best fit structures (green). This representation makes it easier to see that the envelope represented by the best fit structures is significantly smaller than that represented by all of the accepted structures. Remember that our sample has only 692 accepted structures and 29 best fit structures. For a real study, several thousand accepted structures would be needed to determine if this observation holds true.

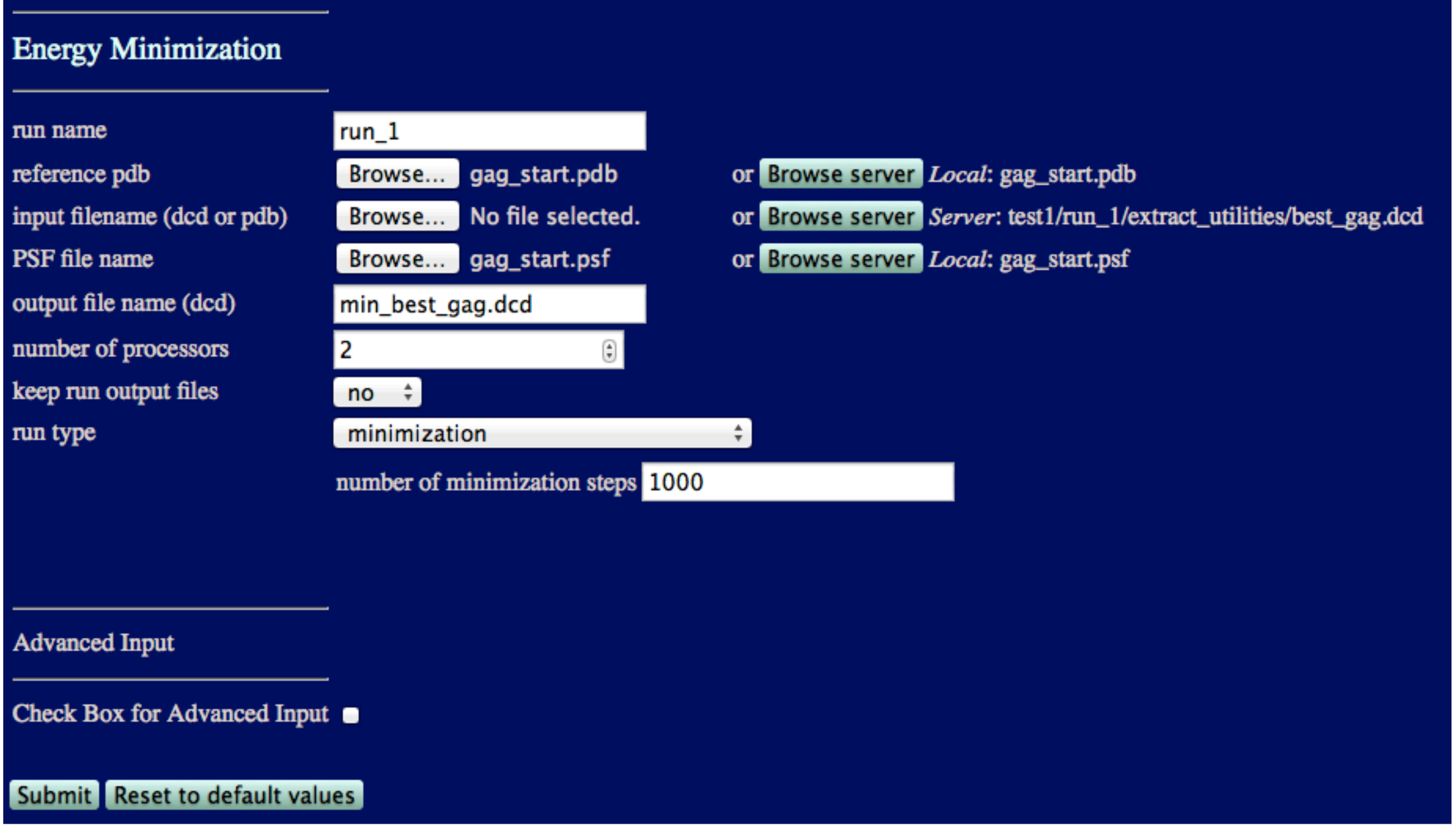


More information on how to use the Density Plot module can be found in the Density Plot documentation.

Minimization of Best Structures

Now we can minimize the best fit gag structures using the Energy Minimization module. More information can be found in the Energy Minimization documentation.

- Select 'Simulate' from the Main Menu.
- Click the 'Energy Minimization' button.



- **run name:** user defined name of folder that will contain the results.
- **reference pdb:** PDB file with naming information for coordinates that will be extracted. We are using the *gag_start.pdb* file.
- **input filename (dcd or pdb):** file containing starting conformation(s) for simulation. The number of atoms must match that in the reference pdb. For files with multiple frames each one will be simulated. We are using the *run_0.dcd* file.
- **PSF file name (dcd or pdb):** PSF file with topology information, must match the reference pdb and input dcd/pdb. Here we use the *gag_start.psf* file.
- **output file name (dcd):** filename for the output DCD contaiing the final frames resulting from simulation
- **number of minimization steps:** number of steps of the conjugate gradient minimization to apply to each structure. We are using 1000 steps for each structure in the DCD file.
- **number of processors:** number of processors used to run the simulation (1-4). We are using use 2 processors.
- **keep run output files:** choice of whether to retain NAMD log files and other output from each simulation
- **DCD write frequency:** user determined number of steps between the output of structures from each simulation
- **run type:** select which of the four combinations of minimization and molecular dynamics to run
- Click 'Submit'

When the process is finished your output should look like the one below.

```
============================================================
DATA FROM RUN: Fri Jul 10 10:57:09 2015

Total number of frames = 29

Minimized structures saved to : ./run_1/energy_minimization/

============================================================
```

What have we generated:

test1/run_1/energy_minimization

- *min_best_gag.dcd*: DCD file containing the minimized best fit gag structures.
- *min_best_gag_dcd.pdb*: reference PDB file that can be used to visualize the frames in the DCD file.

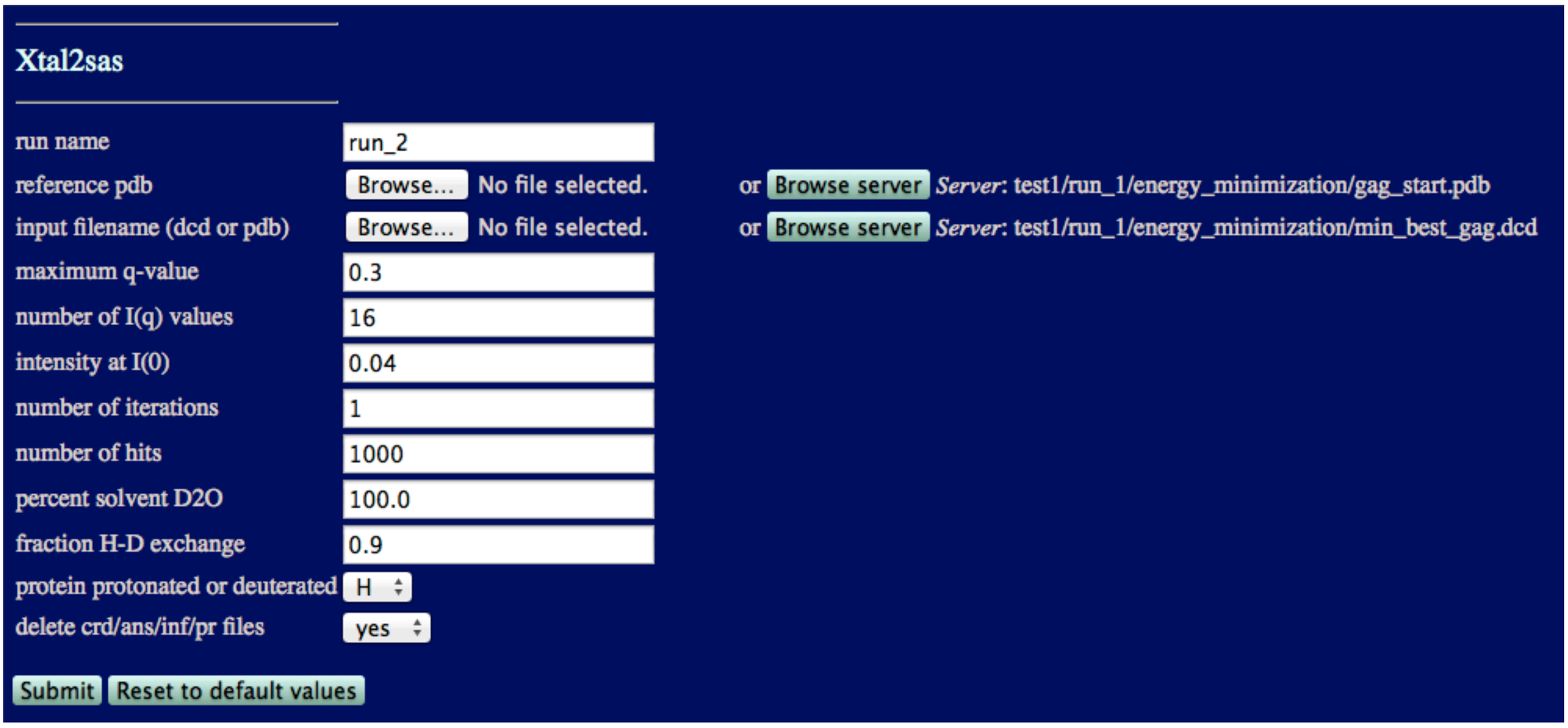
Visualization

Download the 'min_best_gag.dcd' and 'min_best_gag.dcd.pdb' files and then vizualize the structures in VMD. You can load 'best_gag.dcd' (along with a suitable PDB file) again as well for comparison. Go through the two structures frame by frame. You should notice very little difference in the structures.

Final SAS Curve Comparison

If desired, you can compare the minimized structures to the SANS data by calculating their theoretical SANS curves and comparing them to the SANS data again to see how different the best fit chi square values are after the minimization.

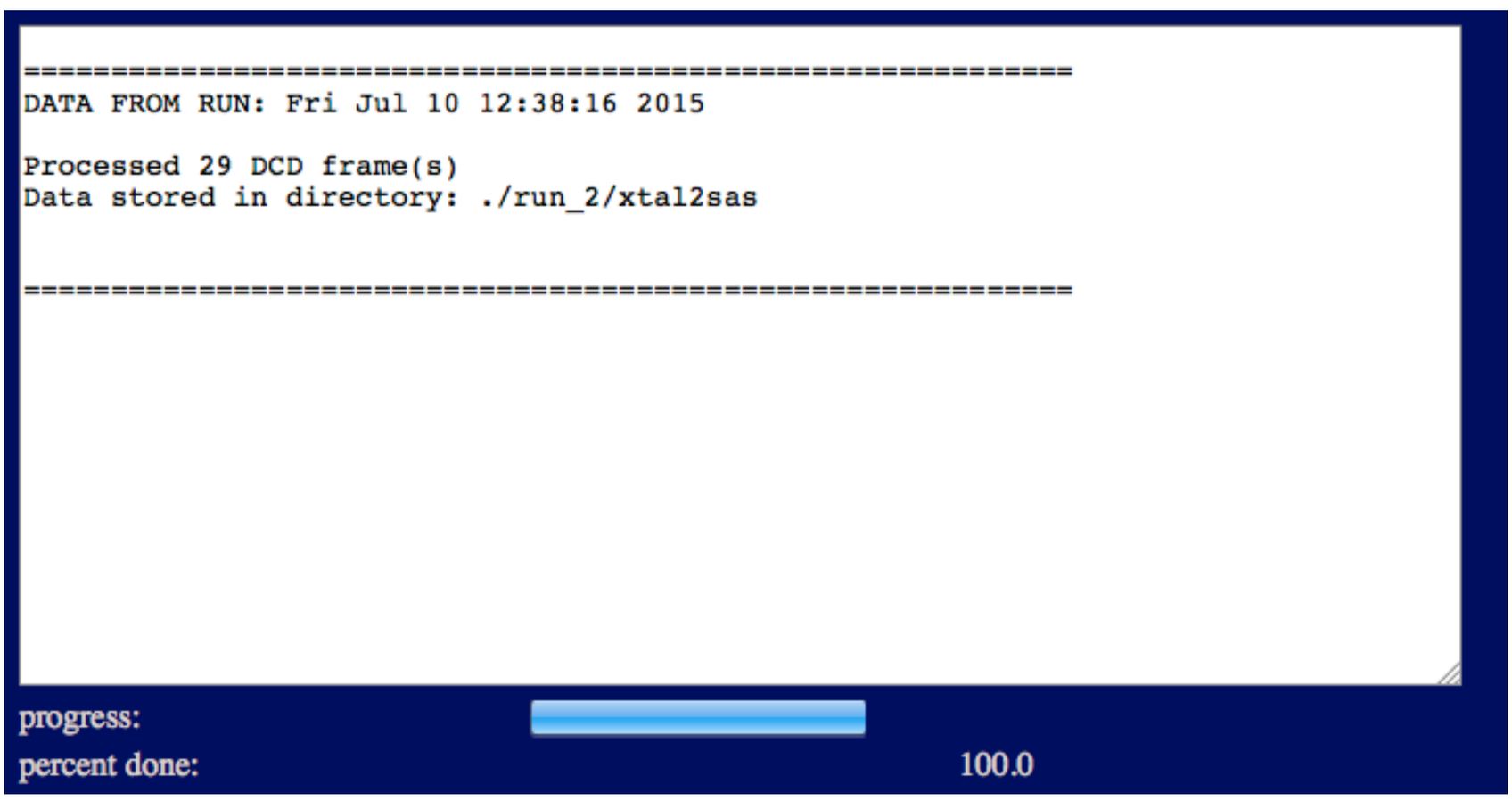
First, calculate the theoretical SANS curves using Xtal2sas.



run name:

Set the run name to run_2. Once the inputs have been entered, click 'Submit'.

Once the run is complete, you should see outputs like those below.

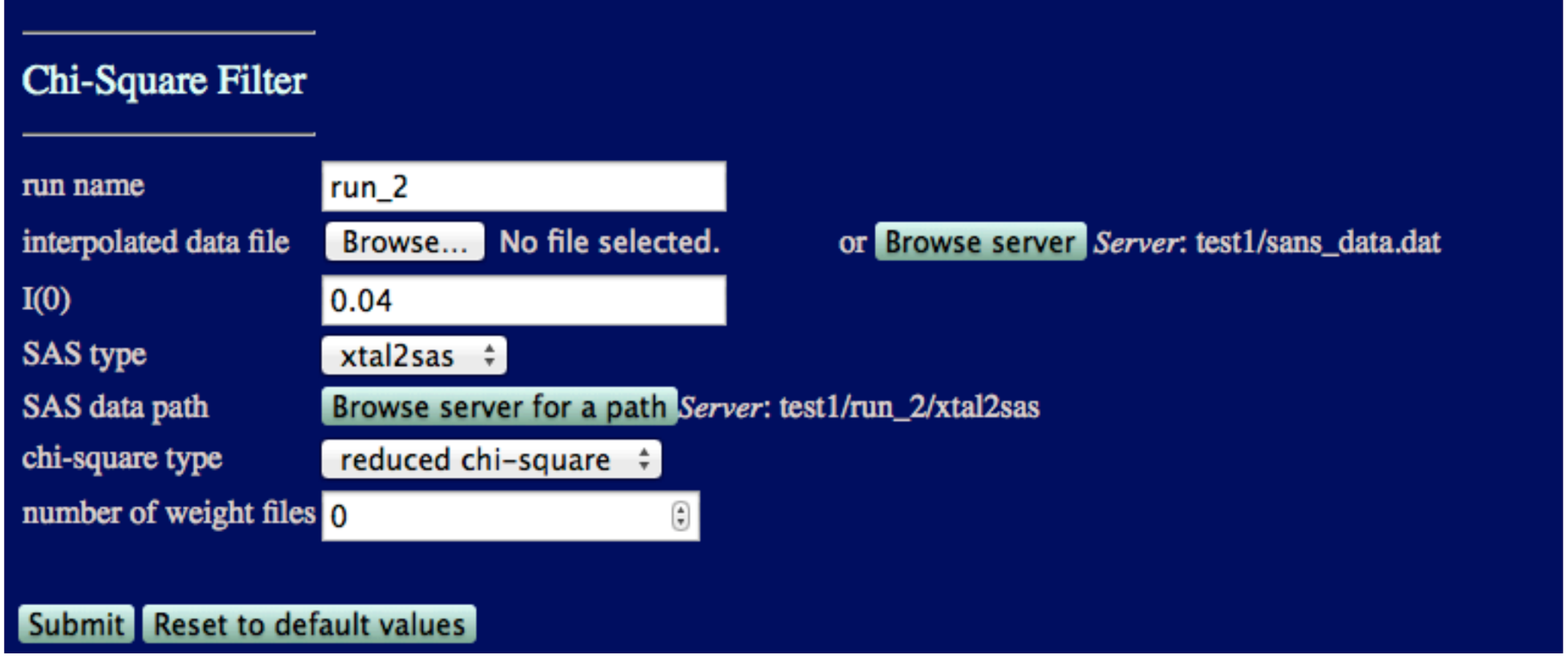


What have we generated:

test1/run_2/xtal2sas

- **.iq*: files for each structure we generated containing theoretical scattering data
- **.log*: log files for each run of xtal2sas
- **.ans*: input file used for each xtal2sas run

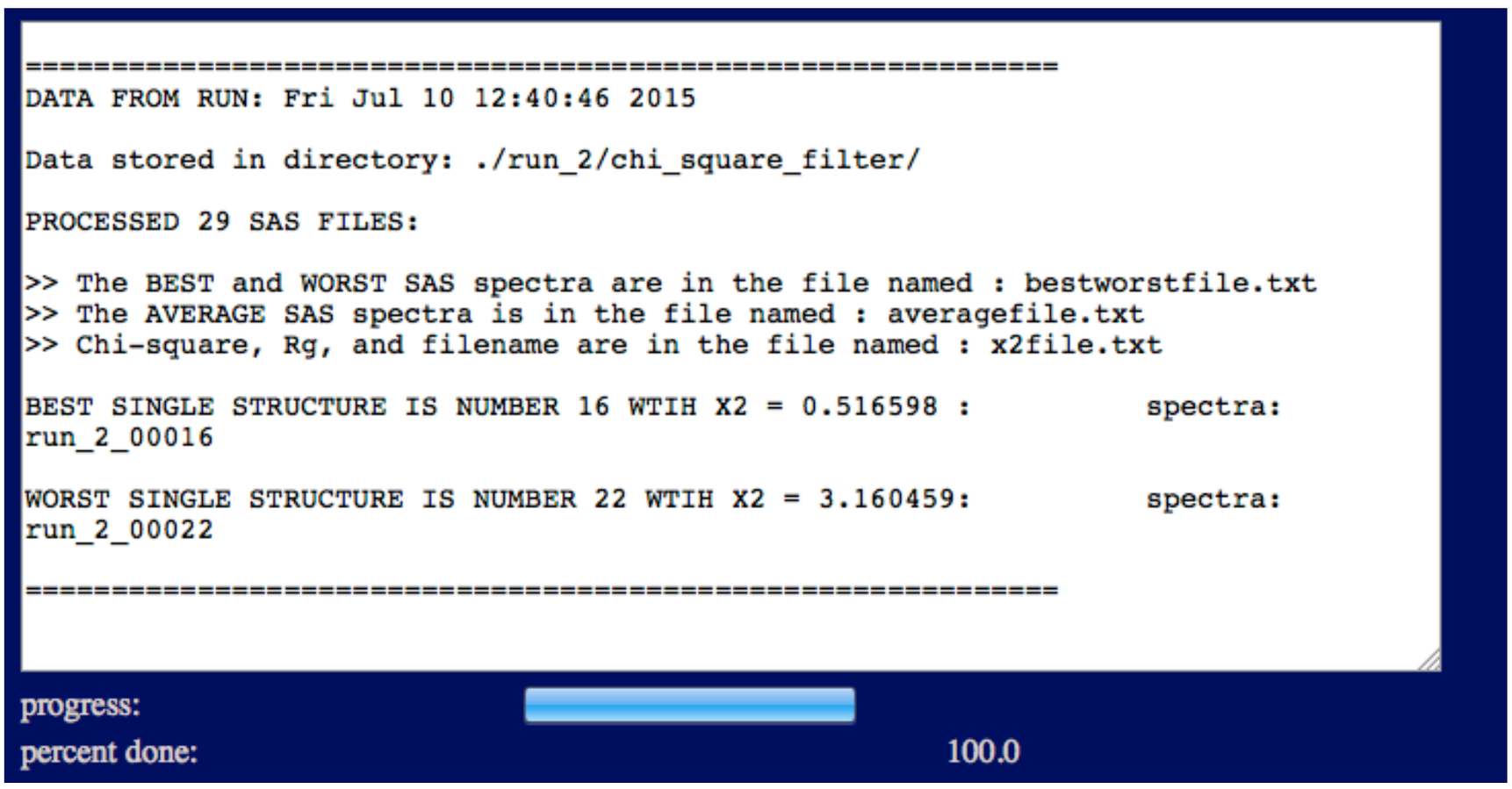
Then, compare the theoretical SANS curves to the SANS data using Chi-Square Filter.

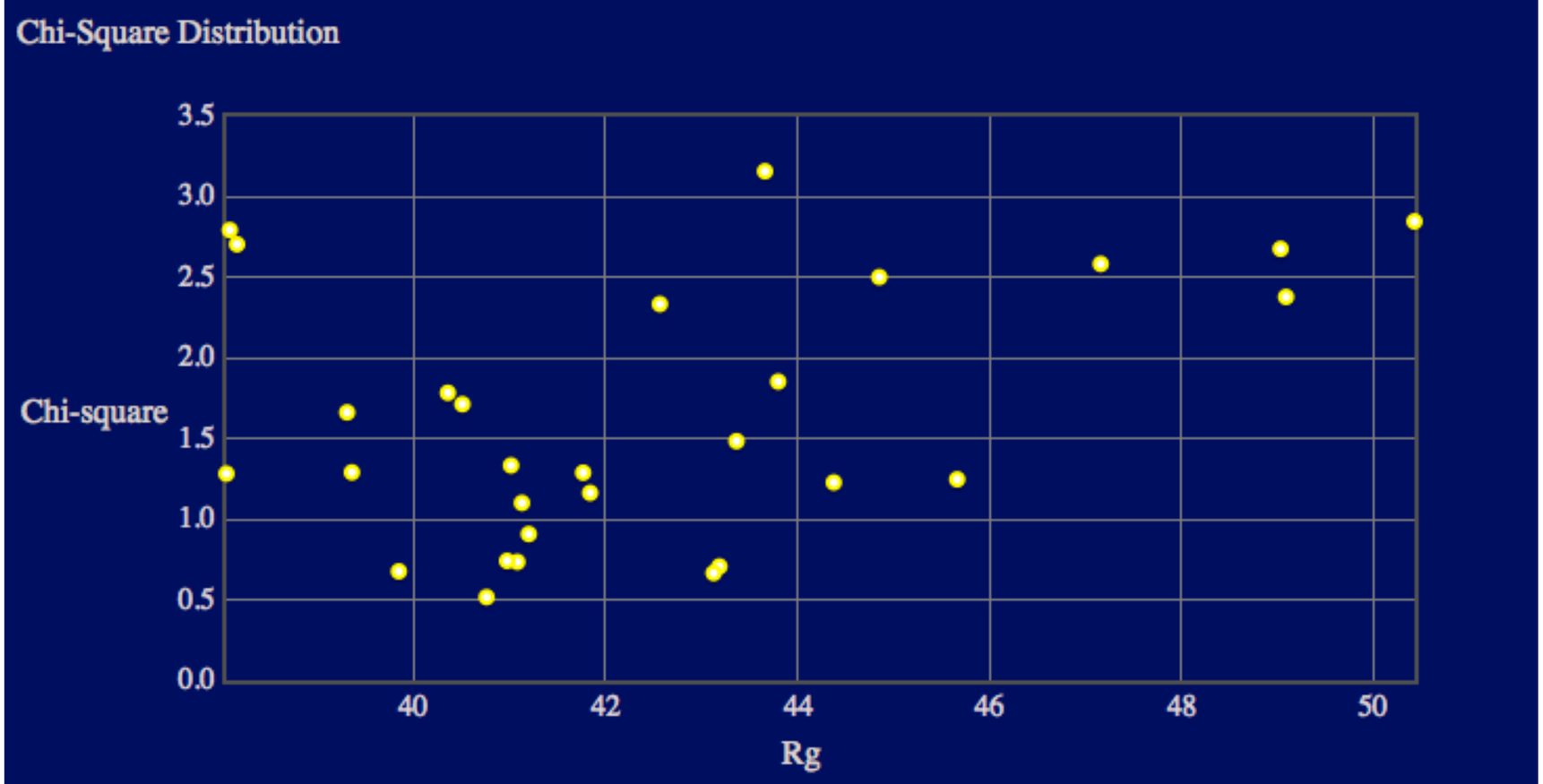


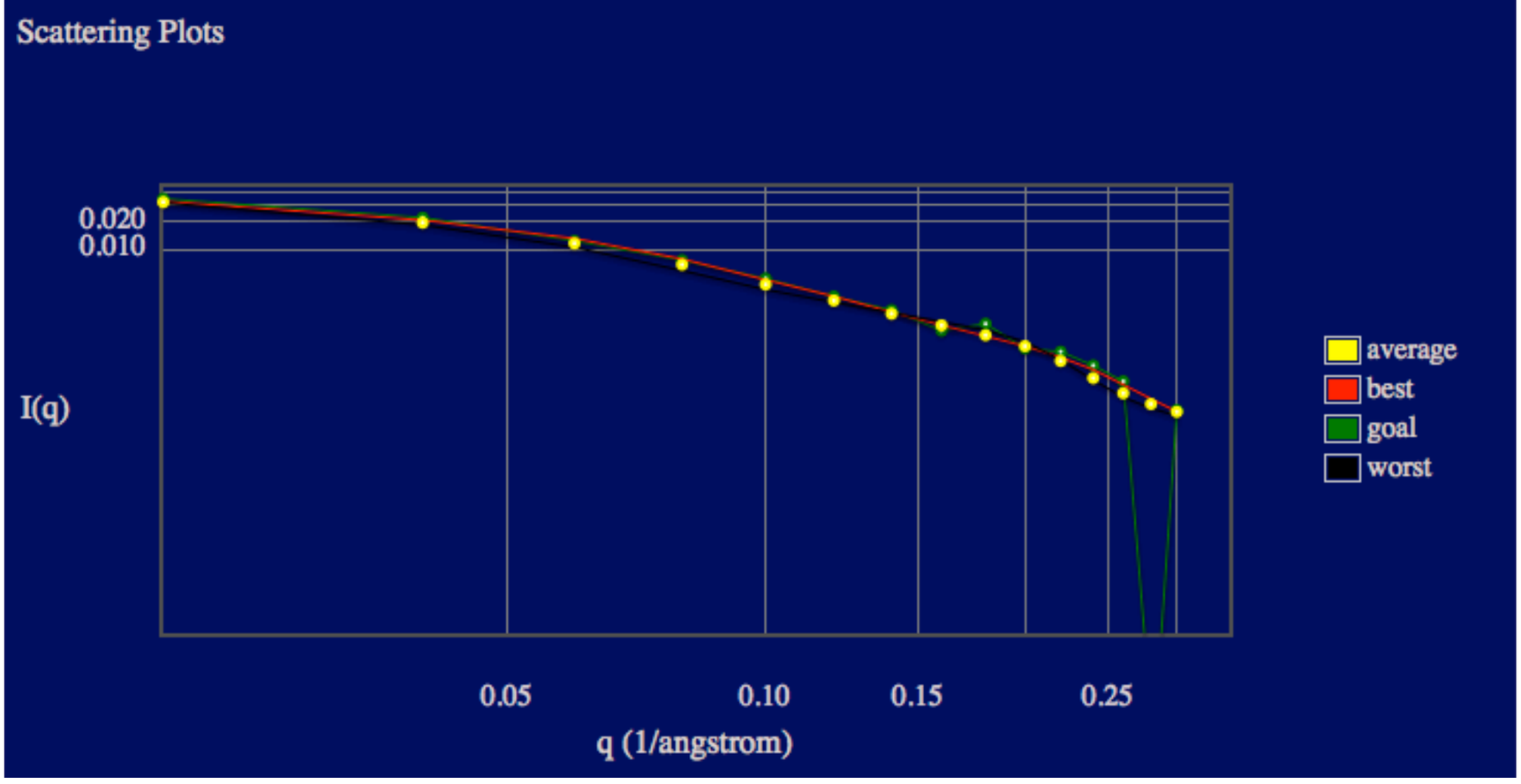
run name:

- Since we already have a chi_square_filter folder in the run_1 directory, set the run name to run_2.
- **number of weight files**
- Set the 'number of weight files' to 0 since we are already dealing with the best fit structures.
- Click 'Submit'.

Once the run is complete, you should see outputs like those below.







What have we generated:

test1/run_2/chi_square_filter

- *averagefile.txt*: average scattering curve for all structures
- *bestworstfile.txt*: best and worst scattering curves selected from all structures
- *sas_spectra_plot.txt*: goal, best, worst and average scattering curves
- *x2_vs_rg_plot.txt*: chi squared against radius of gyration for all structures
- *x2file.txt*: chi squared for all structures

spectra

- **.ciq*: scattering curves scalled to correct I(0) for each structure

References

1. Conformation of the HIV-1 Gag Protein in Solution S. A. K. Datta, J. E. Curtis, W. Ratcliff, P. K. Clark, R. M. Crist, J. Lebowitz, S. Krueger, A. Rein, J. Mol. Biol. 365, 812-824 (2007). BIBTex, Endnote, Plain Text
2. SASSIE: A program to study intrinsically disordered biological molecules and macromolecular ensembles using experimental scattering restraints J. E. Curtis, S. Raghunandan, H. Nanda, S. Krueger, Comp. Phys. Comm. 183, 382-389 (2012). BIBTeX, EndNote, Plain Text

Return to Main Documents Page

Go to top

Supported via CCP-SAS a joint EPSRC (EP/K039121/1) and NSF (CHE-1265821) grant