Implementation of R Code for Estimating the Radius of Gyration of Biological Macromolecules in Solution Using Small-Angle X-ray Scattering Data

1 Download R

In order to run this program you need to use the free, publicly available program “R”. R is a widely used programming language for statistical computing and graphics. To download R, visit the site “http://cran.us.r-project.org” and follow the instructions.

2 Set Up the Estimation Routines and Examples

1. Start the application R.
2. Within R, the working directory must first be changed in order to conveniently select the data. Click “File” > “Change dir...” and then select the folder containing the SAXS intensity curve data. Now it is simple to use any data sets in this folder.
3. Open the files file1.R, file2.R and file3.R from the R drop-down menu: click “File” > “Open script” and then navigate to the correct file location and click “Open.” Do this for each file.
4. If this is the first time you have run this program in R, you must install the R package “changepoint.” From the R drop-down menu:
   (a) Click “Packages” > “Install packages.”
   (b) Select a geographic location from the “CRAN mirror” menu that pops up; it is best to select a location near you for fast download speed. Click “OK.”
   (c) Select the “changepoint” package from the “packages” menu that pops up; click “OK.” The package will automatically download and install.

   Once the package has been installed, you do not need to repeat Step 4 upon subsequent runs.
5. Highlight everything in file1.R and run the code (Ctrl+A, Ctrl+R for Windows machines).

The program is now ready to analyze the example data sets or user-supplied SAXS data. Data in the input file must be organized into three columns, delimited by spaces or tabs if using a text file (.txt, .dat, etc.), or by commas for a .csv file. The columns must contain the following data in the following order:

angle (s)   intensity   standard deviation

The second column should NOT contain log intensity.

The R code in file2.R describes the analysis of a single replicate, using a sample SAXS data set for the molecule ovalbumin. The code also describes alternate file formats. See Section 3 Section 4 below. The R code in file3.R describes the analysis of multiple replicates, using 10 sample SAXS data sets for the molecule myoglobin. See Section 5 below.

3 Single Replicate Example: User-Specified Initial Angle

Included in this folder is a sample SAXS data set for the molecule ovalbumin. The following R code is included in file2.R. The code runs the analysis using the ovalbumin data set and should yield Figure 1, Figure 2, and Figure 3. The function `estimate_Rg` has three arguments: the first argument is the name of the data object read from the file, the second argument is the number of replicates, and the third (optional) argument indicates the index i of the initial angle s_i to be used in the analysis (that is, excluding the first $i-1$ data points near zero from the analysis). If the third argument is not included, then the program defaults to automatically determining any initial outlying data points.

```
data = read.table("oval_01C_S008_0_01.dat", header = FALSE)
estimate_Rg(data, 1, 5)
```

The program output is three plots (one may be concealed by the other) containing several pieces of information:

3.1 *Plot of log-intensity versus angle.*

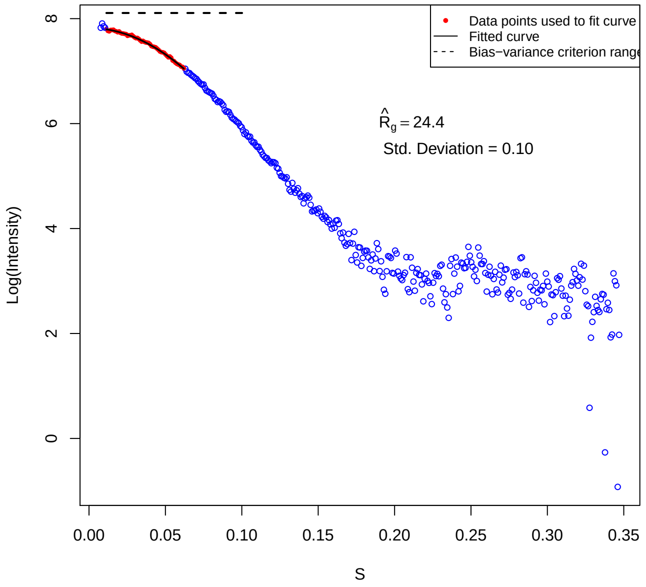

Figure 1: Plot of log intensity vs. s with the estimated R_g value and its standard deviation for a single replicate of ovalbumin.

- Data points (open blue dots and solid red dots) represent log intensity vs. angle s of the input data; this plot can be used to ensure the input data are correct.

- Specifically, the solid red data points are those that have been chosen for use in curve fitting by minimizing the bias-variance criterion.
- A quadratic fit of the solid red data points is indicated by the solid black curve. This curve is used to estimate R_g and its standard deviation. This curve does not need to fit the data perfectly; some bias is acceptable in return for smaller standard deviation.
- The resulting estimates of R_g and its standard deviation are given.
- A black horizontal dashed line indicates the range of possible values over which the bias-variance criterion is optimized.

3.2 Plot of log-intensity versus squared angle.

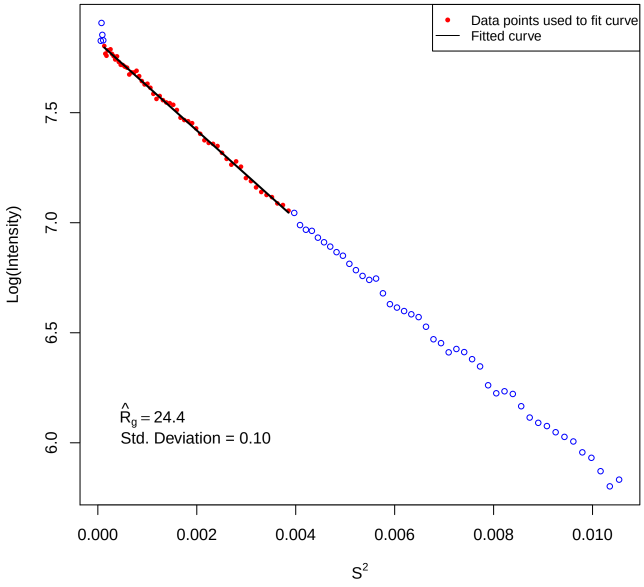

Figure 2: Plot of log intensity vs. s^2 with the estimated R_g value and its standard deviation for a single replicate of ovalbumin.

- Data points (open blue dots and solid red dots) represent log intensity vs. angle squared s^2 of the input data over which the bias-variance criterion is optimized.
- Specifically, the solid red data points are those that have been chosen for use in curve fitting by minimizing the bias-variance criterion.
- A fit of the solid red data points is indicated by the solid black line. This line is used to estimate R_g and its standard deviation. This line does not need to fit the data perfectly; some bias is acceptable in return for smaller standard deviation.
- The resulting estimates of R_g and its standard deviation are given.

3.3 Plot of residuals from quadratic fit.

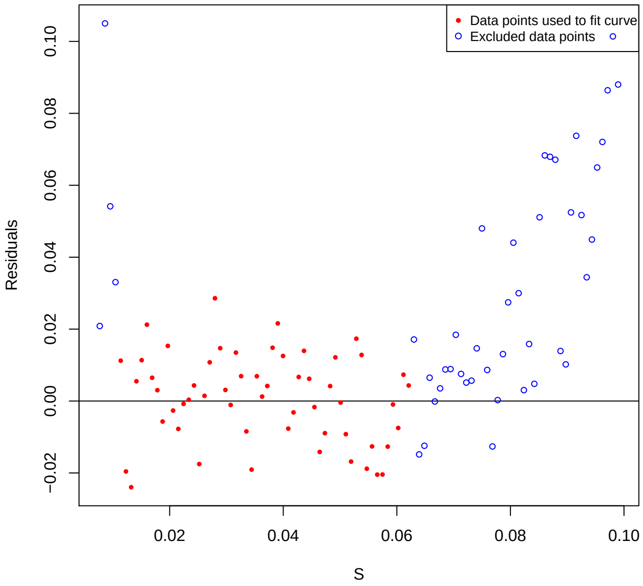

Figure 3: Plot of residuals vs. s for a single replicate of ovalbumin.

- The points (open blue dots and solid red dots) represent residuals vs. angle s of the input data; this residual plot can be used to ensure that the data window is a reasonable fit.
- Specifically, the solid red data points are those that have been chosen for use in curve fitting by minimizing the bias-variance criterion, and the open blue dots are not used in the fit.

To save a plot as a PDF file, first select the plot by clicking on it. Then, from the R drop-down menu, click "File" > "Save as" > "PDF..." Then select the save location, enter a name for the file, and click "Save."

4 Single Replicate Example: Automatic Selection of Initial Angle

By default, the program will automatically determine any initial outlying data points using a modified DFBETAS procedure. If the user does not enter any values for the initial angle, then the program will determine these points automatically and output the number of points removed from the curve. The following R code is included in file2.R. It runs the program using the data set for the molecule ovalbumin and should yield Figure 4, Figure 5, and Figure 6.

```
data = read.table("oval_01C_S008_0_01.dat", header = FALSE)
estimate_Rg(data, 1)
```

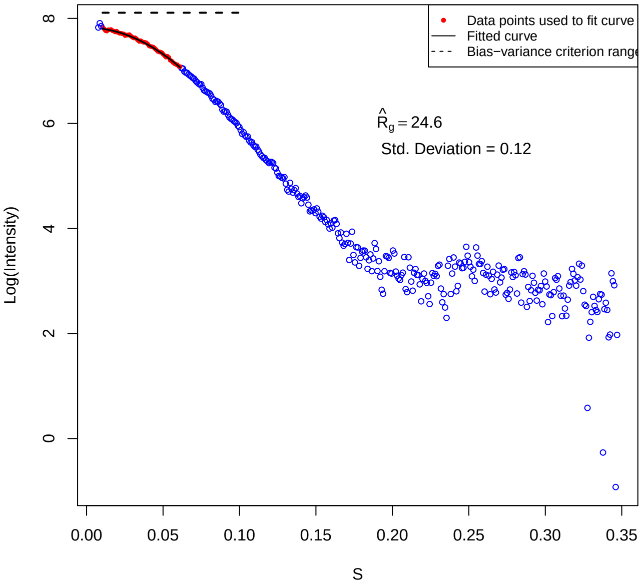

Figure 4: Plot of log intensity vs. s with the estimated R_g value and its standard deviation for a single replicate of ovalbumin with automatic outlier detection.

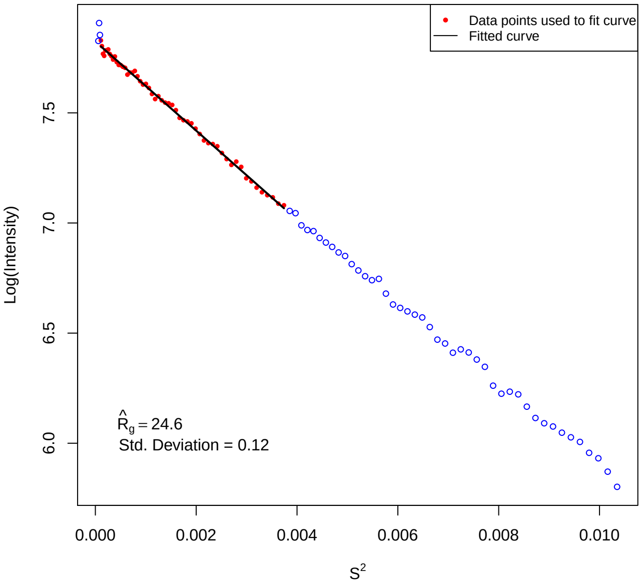

Figure 5: Plot of log intensity vs. s^2 with the estimated R_g value and its standard deviation for a single replicate of ovalbumin with automatic outlier detection.

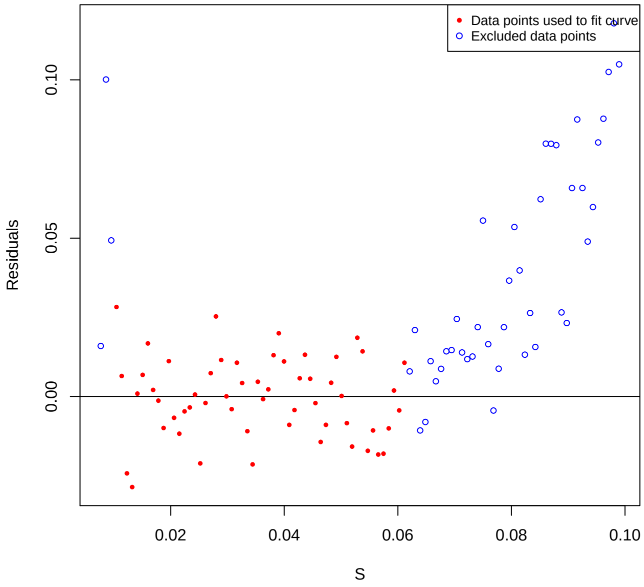

Figure 6: Plot of residuals vs. s for a single replicate of ovalbumin with automatic outlier detection.

5 Multiple Replicates Example

An important innovation of this new procedure is the ability to incorporate replicate SAXS intensity curves to determine a more accurate and precise estimate of R_g and its variance. The code below is included in file3.R and demonstrates how to apply this program with replicate data for the molecule myoglobin. First, the replicate data are read in from 10 different files:

```
data1 = read.table("myo2_07D_S215_0_01.dat", header = FALSE)
data2 = read.table("myo2_07D_S215_0_02.dat", header = FALSE)
data3 = read.table("myo2_07D_S215_0_03.dat", header = FALSE)
data4 = read.table("myo2_07D_S215_0_04.dat", header = FALSE)
data5 = read.table("myo2_07D_S215_0_05.dat", header = FALSE)
data6 = read.table("myo2_07D_S215_0_06.dat", header = FALSE)
data7 = read.table("myo2_07D_S215_0_07.dat", header = FALSE)
data8 = read.table("myo2_07D_S215_0_08.dat", header = FALSE)
data9 = read.table("myo2_07D_S215_0_09.dat", header = FALSE)
data10= read.table("myo2_07D_S215_0_10.dat", header = FALSE)
```

See file2.R for alternate file formats.

Next, the data are combined into a matrix with the following columns in the following order:

angle (s), intensity for first replicate, . . . , intensity for last replicate

In this example, we first combine all data into a matrix with all ten replicates, then use subsets of the data to illustrate estimation with one, three, and ten replicates:

```
# For illustration, look at one replicate, three replicates, and ten replicates.
# First combine the data into one big ten-replicate matrix.
# Keep angle and intensity from replicate 1 (columns 1 and 2 but not 3),
# intensity from replicate 2 (column 2 only),
# intensity from replicate 3 (column 2 only),...,
# intensity from replicate 10 (column 2 only).
#
combined_data = cbind(data1[1:400,-3],data2[1:400,2],data3[1:400,2],
data4[1:400,2],data5[1:400,2],data6[1:400,2],
data7[1:400,2],data8[1:400,2],data9[1:400,2],
data10[1:400,2])
```

It remains to specify the initial angle, or let it be selected via automatic outlier detection. In this example, we specify in each case (one, three or ten replicates) that no points are to be deleted:

```
# Run the estimation code with one replicate
# (only the first two columns of the combined data), with no points deleted:
estimate_Rg(combined_data[,1:2], 1, 1)
# Run the estimation code with three replicates
# (only the first four columns of the combined data), with no points deleted:
estimate_Rg(combined_data[,1:4], 3, rep(1,3))
# Run the estimation code with all ten replicates
# (all eleven columns of the combined data), with no points deleted:
estimate_Rg(combined_data, 10, rep(1,10))
```

Table 1 summarizes the results. In each case, the estimate of R_g is similar but using more replicates increases the precision of the estimate.

As with the single replicates case, the program output is three plots (one may be concealed by the other). The only difference is that all the replicate data are plotted. See Figure 7, Figure 8, and Figure 9 for the case of ten replicates.

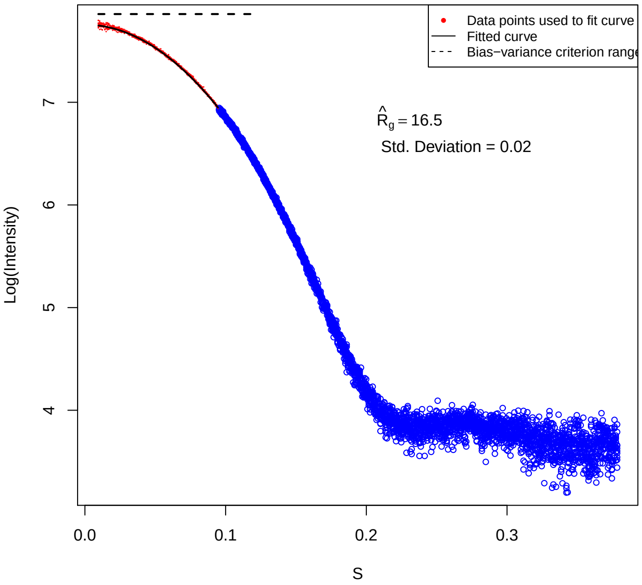

Figure 7: Plot of log intensity vs. s with the estimated R_g value and its standard deviation for ten replicates of myoglobin.

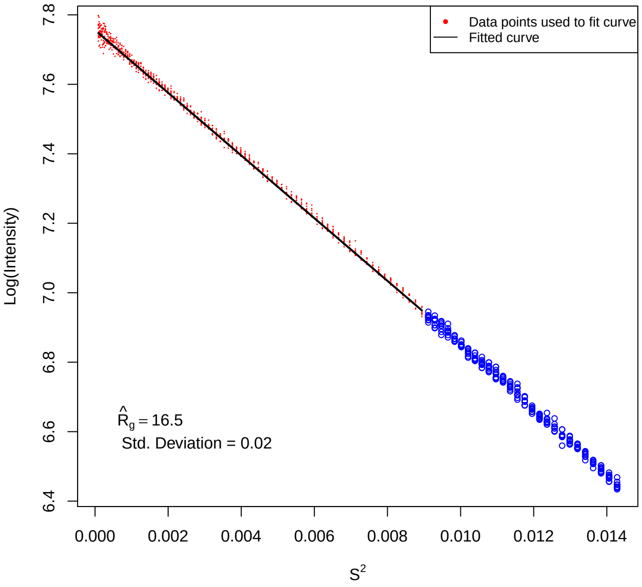

Figure 8: Plot of log intensity vs. s^2 with the estimated R_g value and its standard deviation for ten replicates of myoglobin.

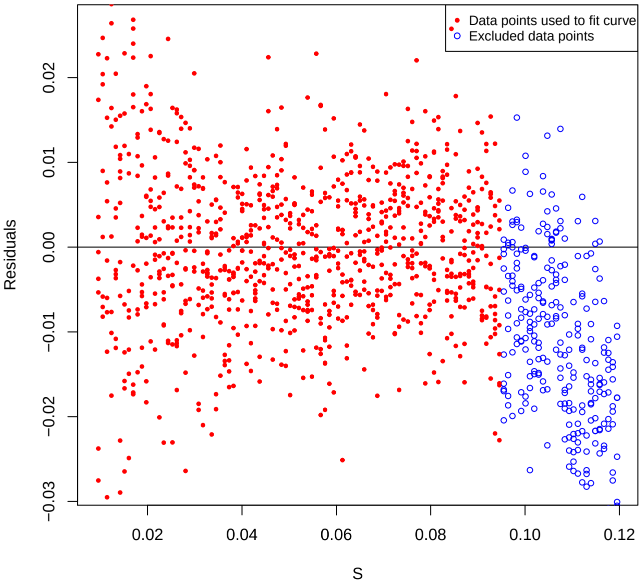

Figure 9: Plot of residuals vs. s for ten replicates of myoglobin.

Table 1: Results for estimating R_g using the new procedure for the molecule myoglobin with one, three, and ten replicate SAXS intensity curves. In each case, $\widehat{R}_g$ and its standard deviation are given.

Replicates	$\widehat{R}_g$	SD($\widehat{R}_g$)
1	16.469	0.030
3	16.452	0.029
10	16.452	0.019

The third argument can be altered to select different initial points from each replicate. For example, if you want to eliminate the first three points of the fourth replicate while deleting no points from the other nine replicates, you would use the following code:

```
estimate_Rg(combined_data,10,c(1,1,1,4,1,1,1,1,1,1))
```

Alternatively, you can delete the third argument, in which case the program uses the modified DFBETAS criterion as an outlier detection algorithm to determine one, common initial point for all of of the replicate curves. For automatic selection of the common initial point in the example with three replicates, use

```
estimate_Rg(combined_data[,1:4],3)
```

Similarly, for automatic selection of the common initial point in the example with ten replicates, use

```
estimate_Rg(combined_data, 10)
```

6 Problems and possible solutions

Problem: The following error message appears when executing the program. “Error in [.data.frame(M, , 2) : undefined columns selected”
Solution: Make sure there is only a one-line header in the data file.

Problem: The following warning message appears when reading in a file. “Incomplete final line found by readTableHeader on ‘filename’”
Solution: The final line of your text or CSV doesn’t have a line feed or carriage return.