Introduction

Structural modelling of bio-macromolecules from small-angle scattering is a highly demanding technique. Interpretation of scattering data from a sample in terms of molecular structure requires not only is careful data correction (i.e. subtraction of background and solvent scattering), samples must be homogeneous (i.e. pure, monodisperse, not significantly flexible), and must also be dilute (i.e. no significant attractive or repulsive interactions between molecules). Under these circumstances, the form factor for each molecule is the same, and the structure factor can be neglected, such that:

$$I(q) = N.(\Delta\rho.V)^2.P(q) \quad (1)$$

Where N is the number density of molecules in solution, $\Delta\rho$ is the contrast of the molecules and V is the volume of each molecule. These 3 quantities are constants, which means that the measured $I(q)$ profile is directly proportional to the molecular form factor $P(q)$, and can be modelled directly to determine the structure of molecule in solution. An important takeaway for the equation above is that particles with larger sizes (larger volumes), or larger contrasts, contribute to the total scattering in such a way that that an object double the size of the molecule of interest, will scatter four times more strongly than the molecule of interest. Similarly, a contaminant with double the contrast will scatter 4 times more strongly than the molecule of interest.

Size Exclusion Chromatography with Small-angle X-ray Scattering (SEC-SAXS)

As outlined above, sample purity is of the highest importance for collecting data for structural modelling. In many instances it can be difficult to establish solution conditions for a protein that allow it to be kept for long periods of time without seeing adverse effects (such as irreversible association/aggregation, or molecular degradation). Concentrating a protein (e.g. by centrifugal concentration), storage, defrosting (if the protein has been stored frozen), and general sample handling, can cause damage to the protein in solution. Each of these actions

will be encountered by a scientist preparing samples, transporting them to the Synchrotron, and handling them before measurement. These issues are molecule and buffer dependent, and cannot be predicted in advance. To assist in the collection of high quality data for the purposes of structural modelling, the use of an in-line size exclusion chromatography (SEC) set-up can be advantageous.

SEC involves passing a solution through porous media. The size range of molecules resolved by a SEC column is dependent on the characteristics of the porous media, but smaller molecules have a higher probability of entering the pores and elute more slowly. Large molecules how a lower probability of entering the pores and elute more quickly. Thus, the use of SEC allows the separation of large aggregates (which scatter strongly) and smaller degradation products from you molecule of interest.

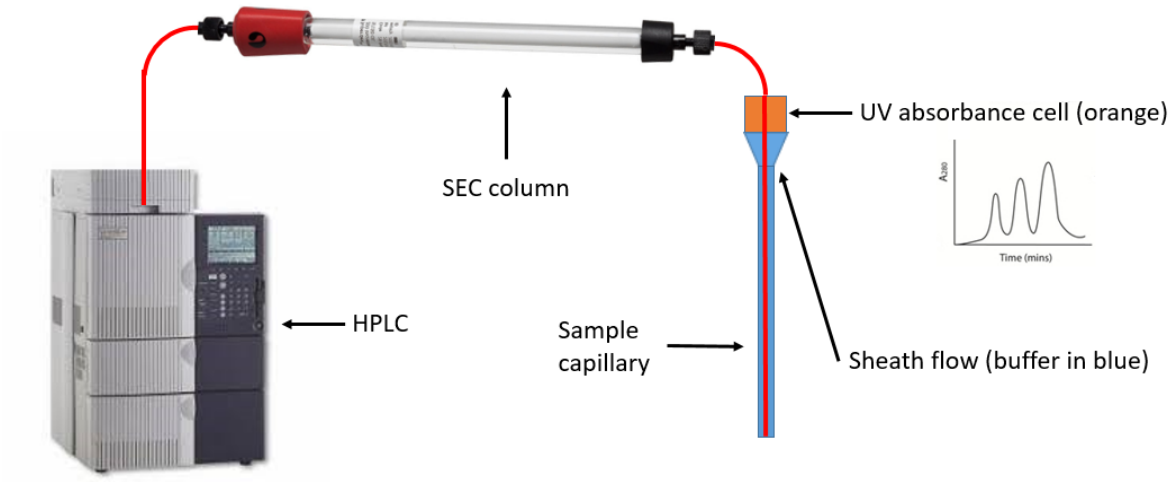

Figure 1: SEC-SAXS setup. Sample is pumped through the SEC column using a HPLC unit. As the sample leaves the column it passes through a UV absorbance cell (the chromatogram is recorded, from which the concentration of protein at each time point can be calculated). After leaving the UV absorbance cell, the solution passes into a sheath flow setup (ref). This setup creates a laminar stream of solution that is not in contact with the capillary wall, which prevents fouling of the capillary. As the solution is passed through the capillary, X-ray scattering data is measured (typically at 1 second intervals).

Experimental

Here, a 10 mg/mL solution of Glucose Isomerase from *Streptomyces rubiginosus* (Hampton Research HR7-102) was dialysed against 50mM Tris pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$. A 60 µL aliquot was injected onto a Superdex S200 5/150GL column equilibrated with 50mM Tris pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$. The SAXS instrument setup is summarised in Table 1.

Table 1 SAXS data collection details

Data Collection Parameters	**Glucose Isomerase (*Streptomyces rubiginosus*)**
Instrument	SAXS-WAXS (Australian Synchrotron)
Beam geometry	250 µm (h) × 450 µm (v)
Wavelength (Å)	1.078
Sample to detector distance (m)	2.791
q-range ($Å^{-1}$)	0.005 – 0.36
Exposure time for each frame (s) *	1
Configuation	In-line SEC-SAXS (S200 5/150 GL) with sheath flow
Injection concentration (mg/mL)	1.0
Injection volume (µL)	60
Flow rate (mL/min)	0.45
Temperature (K)	283
Absolute intensity calibration	Water

Data Reduction and preliminary analysis

In regular circumstances, you would normally work through the data reduction. However, this is a potential challenge for this virtual practical. Instead, the data reduction steps will be outlined here, and you will be provided with radially averaged data to work with.

The first step of reducing SEC-SAXS data is to identify at which point the protein has eluted. The A280 trace can give an indication of this, but it is also useful to create a contour plot of the data (Figure 2). For this contour plot I identified that frames 190-220 appeared to be representative of the solvent scattering (scattering from the buffer solution immediately before protein started eluting). These frames were average and subtracted away from all data to highlight the protein scattering, which can be observed between frames 220-350. Each

frame between 220 and 250 has been corrected for solvent scattering and is included in your data folder for further analysis.

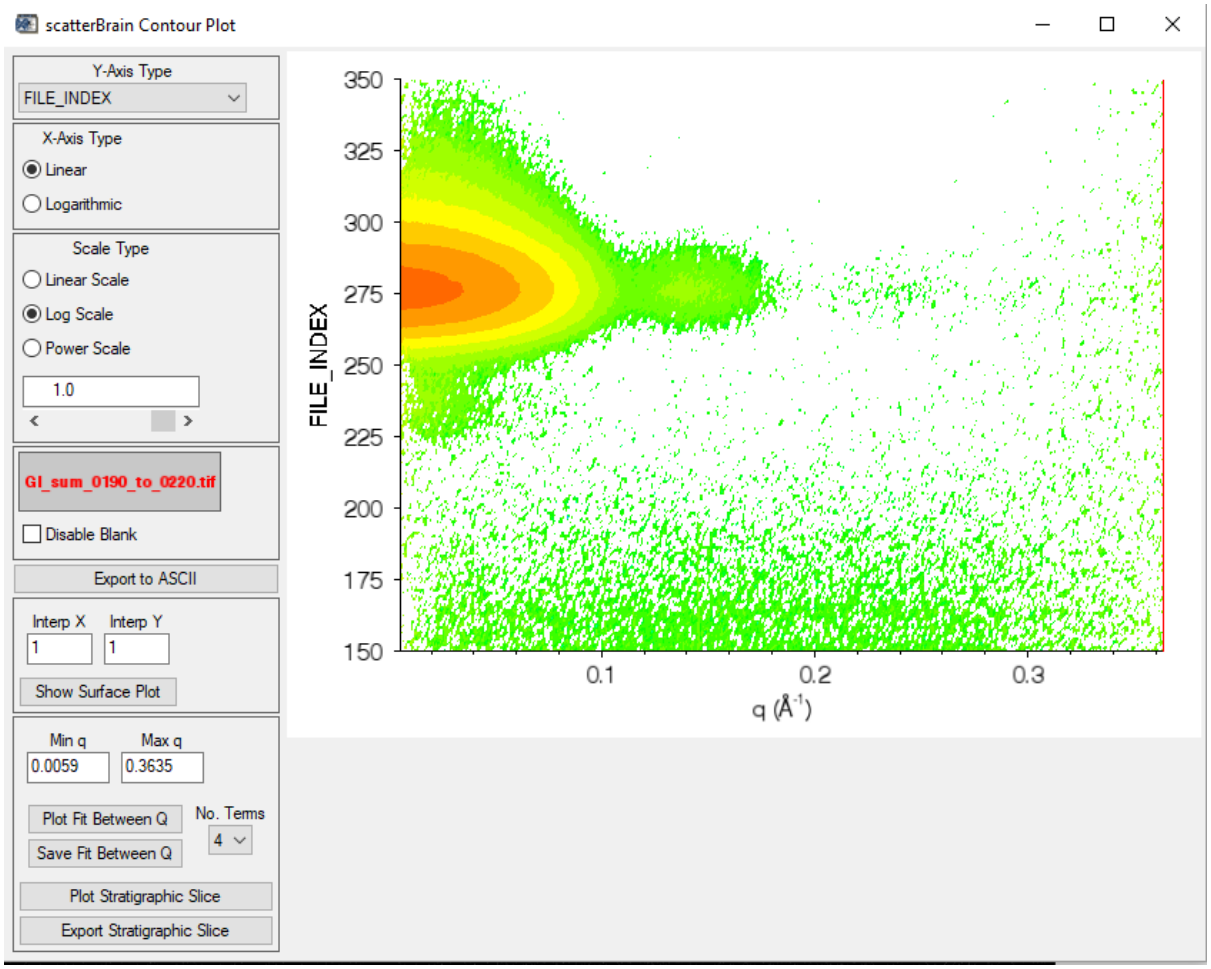

Figure 2: Contour plot from Scatterbrain. The x-axis shows the q-range, while the y-axis represents the file number. The intensity is represented by the colouring (red is more intense, while white is less intense). Here a background has been subtracted to highlight the protein scattering

Due to the large number of files, you will first use a program from the ATSAS package called AutoRg to automatically perform a Guinier analysis on each dataset. On a PC, open the command prompt (type `cmd` at the Start Menu). On a Mac, open a terminal window. In the window, navigate to the directory containing the subtracted data frames. Assuming a typical installation of the ATSAS software, type the following:

```
"C:\Program Files (x86)\ATSAS 3.0.3\bin\autorg.exe" -o GI_autoRg.csv -f csv *.dat
```

or

```
/Applications/ATSAS/autorg -o GI_autoRg.csv -f csv *.dat
```

These commands will automatically perform a Guinier analysis on each individual file on a PC and a Mac respectively. Load GI_autoRg.csv file into a spreadsheet and plot *I*(0) and *R*g as a function of frame number (you can also add custom error bars if you wish). Use a secondary y-axis to show *I*(0) and *R*g on different scales. You will notice that autoRg did not perform Guinier analysis on all data, only for data collected as the protein eluted as the remaining data are too noisy to reliably extract the parameters.

Q1: Qualitatively describe the relationship between Rg and I(0)? Why do I(0) and Rg change in this way? HINT: Is there evidence of an attractive or repulsive interaction between molecules in solution at higher concentrations?

Data in the ranges of frames 262-268 (the leading edge of the elution peak), frames 274-280 (at the elution peak), and 286-292 (the trailing edge of the elution peak) have been averaged for you. Load each of these scattering curves into PrimusQt and determine *R*g for each curve (Radius of Gyration button under the analysis tab). This will determine *R*g automatically (like the autoRg program did), but try changing the data range and see what it does to the radius of gyration and to the weighted difference plot (ΔI/σ) – Note: the upper *s.R*g limit should not exceed 1.3, and the difference plot should show a random distribution.

Q2: Are the Rg values significantly different for each of the data ranges (within estimated uncertainty)? Which range would you choose for further analysis (Frames 262-268, 274-280, 286-292)? Why?

Distance distribution

Inverse Fourier transformation of the scattering data yields the pair-distance distribution function (PDDF) - A histogram of pair-distances in the molecule of interest (it starts at 0, and is 0 past the maximum dimension of the molecule, D_{max}). Interpreting this real space distribution is more intuitive than interpreting the scattering data, but unfortunately as only a small section of Fourier space is measured, and the experimental data can be noisy, the data cannot be directly transformed. Instead, an indirect transformation procedure is used, whereby, a maximum dimension is chosen and a smooth PDDF is optimised such that when it is transformed, it is consistent with the scattering data. The smoothness of the curve can be influenced by the "Alpha" parameter – increasing this parameter will increase curve smoothness, and past a certain value, will increase systematic deviations between the indirect transformation and the experimental scattering curve. This relationship is known as a bias-variance trade-off – if alpha is too low, then the bias of the model is high (fitting noise), and the variance is high (the difference between the model and real $p(r)$).

In turn, highlight each of the scattering curves in PrimusQT and calculate the distance distribution function under the Analysis tab. An estimate of the optimal maximum dimension and smoothness parameter will be determined by Primus. These estimates are usually good, but may not always yield the best value. Try changing first the maximum dimension and then the smoothness parameter and look at the impact each has. In ideal circumstances, the maximum dimension should be chosen such that the $p(r)$ approaches the r-axis asymptotically, and the smoothness should be chosen to be the largest value that does not introduce systematic features in the difference plot

Q3: Letting PrimusQt determine the smoothness parameter, what Dmax value yields a p(r) function that approaches the r-axis asymptotically for each data range? Does the p(r) remain positive at all distances?

Q4: Looking at the Rg and Dmax values, would you change your opinion which dataset you would choose for further analysis?

Estimating molecular mass

If data is not placed on an absolute scale, mass estimates can be made in a relative manner by comparison of *I*(0) values of standard protein samples. However, this data has been placed on a relative scale against scattering from water, thus, mass estimates can be made directly. By definition, at $q = 0$, $P(q) = 1$, thus substituting into equation (1)

$$I(0) = N.(\Delta\rho.V)^2 = N.(\Delta\rho.V)^2 = \frac{c.N_A}{M}.(\Delta\rho.V)^2$$

Where c is the concentration, M is the molecular mass, and N_A is Avagadro's number. Additionally, the volume of the molecule can be recast in terms of the partial specific volume ($\bar{v}$) and molecular mass as (M)

$$V = \frac{M.\bar{v}}{N_A}$$

Which yields

$$I(0) = \frac{c.N_A}{M}.(\Delta\rho.\frac{M.\bar{v}}{N_A})^2 = \frac{c.M}{N_A}.(\Delta\rho.\bar{v})^2$$

On rearrangement

$$M = \frac{I(0).N_A}{c.(\Delta\rho.\bar{v})^2} \qquad (2)$$

Thus to estimate the mass we need to know, *I*(0), concentration, contrast and partial specific volume. The value for *I*(0) can be taken from either the Guinier or p(r) analysis (both should be similar), and the contrast and partial specific volume can be calculated from the primary sequence of the protein.

The first step is to get the sequence of the protein (https://www.uniprot.org/uniprot/P24300 - Click on the FASTA button in the Sequence section (also take note of the mass on the right, it is a number around 40000) and paste it into both formula boxes on http://smb-research.smb.usyd.edu.au/NCVWeb/input.jsp. Enter a title at the top of the form and press submit and take the X-ray contrast values from the "*Tabulated scattering length densities and contrasts"* table. Also note the partial specific volume from just below last table (You may notice that this website gives a lot of extra information. This is because it is designed for determining contrasts of 2 component systems for neutron scattering). Both these numbers will be needed to estimate the particle mass.

Q5: From the calculated sequence: What is the partial specific volume of Glucose Isomerase? What is the contrast of Glucose Isomerase (don't forget the 10^{10})? What is the mass of glucose isomerase protein based on the sequence?

The other quantity required to estimate the mass is the concentration. Concentration is usually estimated from the extinction coefficient of the protein (estimated from the primary sequence), and the absorbance at 280nm. Take the protein sequence and paste it into https://web.expasy.org/protparam/ and press "Compute parameters". Look for the Abs 0.1% values (these are the calculated absorbance values for a 1 g/L or 0.1% w/v solution with a path length of 1 cm). You may note that there are 2 values here (with Cys residues reduced, and with Cys residues oxidised); this is because disulphide bonds absorb at 280 nm. You don't need to worry about that here because there is only 1 Cys residue in the sequence, so no disulphide bonds.

Q6: Why would we want to use the Abs 0.1% instead of the molar extinction coefficient?

The absorbance at 280 nm is monitored by a detector placed immediately above the sample capillary. It is placed close to the X-ray beam so that the absorbance measured is practically the same that would be measured at the sample position (offset by a few seconds as the sample flows through the absorbance cell first). If doing a SEC-SAXS experiment on the SAXS-WAXS beamline at the Australian Synchrotron be sure to check the path length of the UV cell. At the time these data were collected, the thickness was 0.247 cm, and this needs to be taken into account to calculate concentration.

In Excel, take the A280 data provided and add it to the spreadsheet containing the I(0) and Rg data from earlier. First determine offset between the I(0) data and the A280 data – you can do this graphically, or estimate it based on the highest 5 A280 values correspond to the 5 highest I(0) values (the correction should only 1-2 seconds). Create a new column for the corrected time. In an adjacent column, calculate the protein concentration, by dividing by both the cell path length and Abs 0.1% value. Finally, we need to average the concentration over the range measured for each dataset. A simple approximation is to just take the numerical average of the concentration over the measured frames. A better way is to integrate A280 curve (by the trapezoidal rule is sufficient) and then divide this area by the measurement time (NOTE: the measurement time is 7 seconds in each case, but the A280 values do not exactly line up with the SAXS data measurement times. Here, just average the concentration over the best approximation to the actual range the data was measured over). The concentration will be in g/L, but to use equation (2), you will need to convert to g/cm^3.

Q7: What is the average concentration (in g/cm^3) of protein for each of the datasets (Frames 262-268, Frames 274-280, and Frames 286-292)?

The final parameter required to determine the mass from the scattering data is $I(0)$. You will have determined the $I(0)$ value during either Guinier or $p(r)$ analysis. However, in the sheath flow setup at the SAXS/WAXS beamline at the Australian Synchrotron, your protein solution

only fills about half the diameter of the capillary. You will need to correct for this as the absolute scaling assumes the sample fills the whole capillary. The correction factor for this data is 2.05 (be sure to get this number from your instrument scientist collecting scattering data in this configuration).

Multiply your *I*(0) values by 2.05.

Q8: Using equation (2), what is the apparent molecular mass of the Glucose Isomerase in solution for each dataset? Does this value differ from the value calculated from the primary sequence? If so why? HINT: look at the ratio between the mass calculated using equation (X) and that calculated from the primary sequence.

Q9: What are possible sources of error when determining the molecular mass of the protein?

Prediction of scattering profiles from crystal structures

Many high resolution structures exist for Glucose Isomerase. Programs such as CRYSOL can predict the scattering profile of given a crystal structure. The predicted scattering profile can be compared to experimental data to determine the relative agreement between the structure of the protein in solution and the crystal structure.

Go to http://www.rcsb.org/structure/4XIS and take a look at the structure "Biological assembly 1". For those who haven't seen a protein crystal structure before, they are usually represented by these ribbons that trace the path of the amino acid chain. This simplifies the structure, and allows easy identification of the structural motifs (alpha helix, beta sheet, coil).

Q10: Each protein that forms the molecule is coloured differently. How many individual proteins make up the functional glucose isomerase molecule? Is this consistent with molecular mass calculated from I(0)?

Download the crystal structure 4XIS. Under the analysis tab in PrimusQt, press the "Crysol and Sreflex" and compare the experimental scattering data with profiles calculated from the predicted crystal structure for 4XIS.

Q11: What is the χ^2 statistic for comparison of the scattering curve calculated from 4XIS and the experimental data? Try with and without a background correction – does this improve the fit? Which dataset(s) agree best with the crystal structure? Is this the dataset you chose earlier for further analysis?

Q12: What does the background correction tell you about the reduced data? Is the background scattering high, low, correct? Were the frames chosen for solvent scattering representative of the solvent present as the protein eluted? If not, what may have been the issue?

We will not attempt it here, but a more advanced application involving the calculation of scattering profiles from crystal structures is rigid body modelling. In rigid-body modelling, you define rigid subunits, assemble them, then calculate a scattering profile and compare it to the experimental data. These modelling programs will then move these rigid subunits (penalising steric clashed) in such a way as to improve the fit between the predicted scattering profile and the experimental data.

Ab initio modelling

Quite often, the protein structure of interest may not have ever been crystallised, nor has a homologous protein. In such cases, it is not possible to compare the experimental scattering data with a scattering profile generated from a crystal structure. In cases like these, a common approach is to optimise the distribution of "beads" or "dummy atoms" that represent scattering density of the molecule of interest and optimise their position to improve the agreement between the experimental data and predicted scattering curve. This is known as *ab initio* or dummy atom modelling. There are a few important considerations for this approach. The dummy atoms need to be compact, and they need to be connected. Further there may be symmetry considerations. Symmetry is a difficult to determine using small-angle

scattering. It can reduce the number of degrees of freedom in your model, but it can also bias your model. Typically, you may try low symmetry point groups to see if the model naturally possesses features of a higher symmetry point group before modelling with a higher symmetry point group. In other cases, you may have evidence for a certain point group from a similar protein. Here, we already have a crystal structure to guide us.

Take a look at the crystal structure, taking note of symmetry axes. With 4 subunits, there are a number of possible arrangements (Schoenflies and Hermann-Mauguin notation respectively): C_1 or 1 (no symmetry), C_2 or 2 (possessing a single 2-fold axis), C_4 or 4 (possessing a single 4-fold axis), and D_2 or 222 (possessing a 2-fold axis, and a 2-fold axis perpendicular to the original axis)

Q13: What point group does glucose isomerase possess? Check your answer against the "Global Symmetry" on the PDB page: http://www.rcsb.org/structure/4XIS

Using the point group symmetry you have determined, and your chosen dataset, click on Dammif under the Analysis tab in PrimusQt. Use manual selection of parameters, and feel free to tweak the Guinier and $p(r)$ analysis performed (but don't feel obliged to). Enter your symmetry (this program places a P in from of the Hermann-Mauguin point group (i.e. P1, P2, P4, P222). By default the program should average 10 models, however, here, we will just run the program once.

Q14: What is the χ^2 value for the fit? Take a look at the averaged dummy atom model. Is the shape and dimensions consistent with the crystal structure?