Data reduction and processing tutorial

Petr V. Konarev

European Molecular Biology Laboratory, Hamburg Outstation

BioSAXS group
Vacuum cell

25-31 October 2010 EMBO Course

EMBL BioSAXS beamline X33, 2010

Optics

Secondary Slit systems (tungsten blades)

SS3    SS2    SS1

Rho coated Zerodur mirror
radius 2.2 km to infinity

White beam from bending magnet

cooled, bendable Si111 triangular monochromator
asymmetric cut \( \alpha = 8.5^\circ \)

Beam monitors

Primary Slit system
water cooled

Vacuum cell

Hutch

Completely redesigned 2005-2010
Efficiency gain: about 5-10 times

Nice User
PILATUS Pixel X-ray Detector at X33

PILATUS 1M (10*100K modules)
Active area 70*420 mm², pixel size: 170µm
Readout time: 3.6ms, framing rate: 50Hz
In user operation on X33 from 22.11.07

AgBeh pattern
Ag-Behenate 2D Pattern (FIT2D View)

intensity decrease owing to polarisation = 0.7662
we is applied to the data.
rm active data region (this may take some time)

intensity decrease owing to polarisation = 0.9151
we is applied to the data.
rm active data region (this may take some time)

intensity decrease owing to polarisation = 0.9819
we is applied to the data.

0

2300 Number of overloaded pixels : 429

motion has been extracted from the file header:

an = Mon Jun 6 11:05:32 2005
distance = 675.38 mm
1.5400 Angstroms
art of the oscillation = 357.576 degrees
d of the oscillation = 357.576 degrees
changed using the "GEOMETRY" command.

instead of 6900 requested
rm active data region (this may take some time)
ted = 300 (13%)
ted = 600 (26%)
ted = 900 (39%)
ted = 1200 (52%)
ted = 1500 (65%)
ted = 1800 (78%)
ted = 2100 (91%)

intensity decrease owing to polarisation = 0.9819
we is applied to the data.

a = -0.87800686E-01 6.283624
a = 0.887858 0.6356816
1: 2250
Ag-Behenate 2D Pattern (with MASK)

25-31 October 2010 EMBO Course

2300 Number of overloaded pixels = 429

Information has been extracted from the file header:

- Mon Jun 6 11:05:32 2005
distance = 675.38 mm

1.3400 Angstroms

- of the oscillation = 357.576 degrees
d of the oscillation = 357.576 degrees

- changed using the "GEOMETRY" command.)

- instead of 6960 requested
Ag-Behenate 1D Pattern (FIT2D View)

```
intensity decrease owing to polarisation = 0.7662
ue is applied to the data.)
rm active data region (this may take some time)

intensity decrease owing to polarisation = 0.9151
ue is applied to the data.)
rm active data region (this may take some time)

intensity decrease owing to polarisation = 0.9819
ue is applied to the data.)
0

2300 Number of overloaded pixels = 429
mation has been extracted from the file header:
an = Mon Jun 6 11:05:32 2005
distance = 675.38 mm
1.5400 Angstrom
art of the oscillation = 357.576 degrees
D of the oscillation = 357.576 degrees
changed using the "GEOMETRY" command.)
instead of 6900 requested
rm active data region (this may take some time)
ted = 300 ( 13%)
ted = 600 ( 26%)
ted = 900 ( 39%)
ted = 1200 ( 52%)
ted = 1500 ( 65%)
ted = 1800 ( 78%)
ted = 2100 ( 91%)

intensity decrease owing to polarisation = 0.9819
ue is applied to the data.)
E = -0.878068E-01 6.283624
E = 6.087858 0.6356916
1: 2250
```
Creating angular axis

Angular axis file

Spacing: 2.145e-2
Width: 11.43

Spacing: 7.138e-3
Width: 6.923

Spacing: 5.349e-3
Width: 6.340

Spacing: 4.270e-3
Width: 5.365
Raw data reduction steps

- Radial integration of 2D image into 1D curve
- Check for radiation damage, averaging of different time frames (total of 8 frames, each for 15 seconds)
- Associated errors in the data points are computed from the numbers of counts using Poisson statistics
- Mask file is used to eliminate beamstop and inactive detector area
- Exact coordinates of the beam center are required for integration (determined from AgBeh data)
- Data are normalized to the pindiode value (intensity of the transmitted beam)
- Data are transferred into ASCII format containing 3 columns: \( s \quad I(s) \quad Er(s) \)
An automated SAXS pipeline at X33

- Data normalization
- 2D-1D reduction
- Data processing
- Check for radiation damage
- Computation of overall parameters
- Database search
- *Ab initio* modelling
- XML-summary file generation

Hardware-independent analysis block

### Summary XML

<table>
<thead>
<tr>
<th>Run #</th>
<th>File</th>
<th>Description</th>
<th>Conc. mg/ml</th>
<th>(R_g) nm</th>
<th>Guinier points</th>
<th>(D_{max}) nm</th>
<th>(M_r) kDa</th>
<th>(V_{total}) mm³</th>
<th>Quality %</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>02w003.dat</td>
<td>bsa</td>
<td>4.0</td>
<td>3.2±0.2</td>
<td>61-230</td>
<td>12.6</td>
<td>68</td>
<td>190</td>
<td>84</td>
<td>13±20</td>
</tr>
<tr>
<td>6</td>
<td>Fe3_006.dat</td>
<td>FeO</td>
<td>2.2</td>
<td>5.2±0.3</td>
<td>47-77</td>
<td>12.0</td>
<td>219</td>
<td>526</td>
<td>77</td>
<td>13±20</td>
</tr>
<tr>
<td>7</td>
<td>Fe3_007.dat</td>
<td>FeO</td>
<td>2.2</td>
<td>5.2±0.3</td>
<td>26-75</td>
<td>12.7</td>
<td>229</td>
<td>592</td>
<td>89</td>
<td>13±20</td>
</tr>
<tr>
<td>9</td>
<td>Fe3_008.dat</td>
<td>FeO</td>
<td>1.2</td>
<td>5.1±0.1</td>
<td>27-61</td>
<td>18.3</td>
<td>206</td>
<td>346</td>
<td>89</td>
<td>13±51</td>
</tr>
<tr>
<td>11</td>
<td>Fe3_011.dat</td>
<td>FeO</td>
<td>1.5</td>
<td>5.0±0.0</td>
<td>24-60</td>
<td>16.0</td>
<td>189</td>
<td>473</td>
<td>90</td>
<td>13±02</td>
</tr>
<tr>
<td>10</td>
<td>Fe3_013.dat</td>
<td>FeO</td>
<td>4.0</td>
<td>5.5±0.3</td>
<td>33-72</td>
<td>22.0</td>
<td>235</td>
<td>496</td>
<td>86</td>
<td>13±12</td>
</tr>
<tr>
<td>14</td>
<td>Fe3_014.dat</td>
<td>FeO</td>
<td>4.0</td>
<td>5.6±0.3</td>
<td>20-24</td>
<td>21.4</td>
<td>209</td>
<td>467</td>
<td>88</td>
<td>13±22</td>
</tr>
</tbody>
</table>
Program PRIMUS- graphical package for data manipulations and analysis

♦ statistical analysis, 2D integration and scaling of raw data in MAR and OTOKO format, transforming MAR image plate/gas detector data into ASCII format (modules MAR-PRIMUS, Sapoko and Binasc)

♦ data manipulations (averaging, background subtraction, merging of data in different angular ranges, extrapolation to infinite dilution)

♦ evaluation of radius of gyration and forward intensity (Guinier plot, module AUTORG), estimation of Porod volume

♦ calculation of distance/size distribution function $p(r)/V(r)$ (module GNOM)

♦ data fitting using the parameters of simple geometrical bodies (ellipsoid, elliptic/hollow cylinder, rectangular prism) (module BODIES)

♦ data analysis for polydisperse and interacting systems, mixtures and partially ordered systems (modules OLIGOMER, SVDPLOT, MIXTURE and PEAK)

P.V. Konarev, V.V. Volkov, A.V. Sokolova, M.H.J. Koch, D.I. Svergun
PRIMUS: graphical user interface
PRIMUS: data processing

Buffer before sample: BSA 5.4 mg/ml
Buffer after
PRIMUS: buffer averaging

25-31 October 2010 EMBO Course

Average buffer before
Average buffer after
buffer averaged

Data processing

File to be opened: A30000.dat
Plot: view experimental data
Working directory: F:\data\hsa
Plot: view experimental data

Running | Input pending in Graphics Window
Plot data

sample: BSA 5.4 mg/ml
buffer averaged

PRIMUS: data processing

File: A05000.dat
Plot: view experimental data
Working directory: F:\data\bsa
File to be opened: A05000.dat
Plot: view experimental data

Running
Input pending in Graphics Window
**PRIMUS: buffer subtraction**

**Sample:** BSA 5.4 mg/ml

**Buffer Averaged Subtracted Curve**

---

### Data Processing

<table>
<thead>
<tr>
<th>#</th>
<th>File Name</th>
<th>Range</th>
<th>Units</th>
<th>nDeg</th>
<th>Sync</th>
<th>nEnd</th>
<th>Conc</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A29000.dat</td>
<td>Select</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.000</td>
</tr>
<tr>
<td>2</td>
<td>A23000.dat</td>
<td>Select</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>A39000.dat</td>
<td>Select</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>Avglog06.dat</td>
<td>Select</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.000</td>
</tr>
</tbody>
</table>

---

**File to be opened:** Averag06.dat

**Plot:** view experimental data

**Working directory:** c:data\bsa

**Plot:** view experimental data

---

**Running** | **Input pending in Graphics Window**
PRIMUS: data at different angular ranges

25-31 October 2010 EMBO Course
PRIMUS: merging data

25-31 October 2010 EMBO Course

PRIMUS

Graphics Window

Plot data

saxs range
waxs range

Data processing

Active|Toggle| File Name  | Range  | Units | nEnd | Sync | nEnd | Conc | Multiplier
---|-------|-----------|-------|-------|------|------|------|------|----------
✓ | #1 | subc_saxs.data | Select | 1 | 1 | 1 | 962 | 1.000 | 1.000
✓ | #2 | subc_waxs.d | Select | 1 | 1 | 1 | 776 | 1.000 | 0.3300
#3 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000
#4 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000
#5 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000
#6 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000
#7 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000
#8 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000
#9 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000
#10 | Select | 1 | 1 | 1 | 9699 | 1.000 | 1.000

Plot
Average
Subtract
Divide
Subset
Divist
ZarConC
Adjust
Guinier
Flat
Red
Finish

Running | Input pending in Graphics Window

04-Nov-2000 17:51:48

File: data\Subc

Merge: manipulation with data
Plot: view experimental data
Merge: manipulation with data
Plot: view experimental data
PRIMUS: merging data

Plot data

saxs range
waxs range
merged curve

25-31 October 2010 EMBO Course
**Guinier Plot**

- **$R_g = 2.68 \pm 1.11 \times 10^{-2}$**
- **$I_0 = 271.07 \pm 0.605$**

\[ I(s) = I(0) \cdot \exp(-s^2 \cdot R_g^2 / 3) \]

**$R_g$ – radius of gyration**

\[ M \approx M_{BSA} \cdot \left( \frac{I(0)}{I_{BSA}(0)} \right) \]
PRIMUS: Porod volume

\[ V = 2 \pi^2 I(0) / Q = 2 \pi^2 I(0) / \int_0^\infty s^2 (I(s) - K) \, ds \]

\[ V = 92.37 \]

\( V \) – excluded volume of particle
Extrapolation to zero concentration

Human kinesin \( c = 1 \text{ g/l}, 3 \text{ g/l}, 6 \text{ g/l}, 9 \text{ g/l}, c \to 0 \)
Merging for making final composite curve

25-31 October 2010 EMBO Course
Checking Guinier plot

Guinier Plot: globular particle
File name: sk_023.dat

- Particle type = 0
- Points 20 to 144
- $d_R$ limits = 0.460 to 1.42
- $R_g$ = 3.55 ± 2.86e-3
- $I_0$ = 14644. ± 10.4

Average protein concentration:
- 9.8 mg/ml
- 9.8 mg/ml
- 1.4 mg/ml
- 1.4 mg/ml
- 3.8 mg/ml
- 3.8 mg/ml
- 5.7 mg/ml
- 5.7 mg/ml
- 8.5 mg/ml

Date: 25-31 October 2010 EMBO Course

C:\PKONAREV\Conferences\EMBL
Merging for making final composite curve

25-31 October 2010 EMBO Course

log I(s)

9.3 mg/ml
1.4 mg/ml
3.8 mg/ml
5.7 mg/ml
8.5 mg/ml
merged

sasplot1-1
sasplot2-1
sasplot3-1
sasplot4-1
sasplot5-1
sasplot6-1
Real/reciprocal space transformation

\[ p(r) = r^2 \gamma(r) \quad \text{distance distribution function} \]

\[ \gamma_0(r) = \frac{\gamma(r)}{\gamma(0)} \]

Probability to find a point at distance \( r \) from a given point inside the particle.
Distance distribution function from simple shapes

SAXS patterns

Pair distribution functions
Distance distribution function of helix

\[ I, \text{ rel. units} \]

\[ p(r), \text{ rel. units} \]

\[ S, \text{ nm}^{-1} \]

\[ \text{distance, nm} \]
Indirect Fourier Transform

Input file(s): bsa.a29.dat  JOB = 0
Real space: Ny = 2.05 , I(y) = 0.2886E+03

GNOM Parameters

- Input1: bsa.a29.dat
- Input2: none
- Output: gnoma.out
- Job: 0
- Rmin: 0.0
- Rmax: 8.000

Gnomer fit using 100 points:
- Radius of gyration: 3.798
- I(0) estimate: 275.0
- Estimated Pore Volume: 98.66

Running  Input pending in Graphics Window
In the original version of GNOM the maximum particle size $D_{\text{max}}$ is a user-defined parameter and successive calculations with different $D_{\text{max}}$ are required to select its optimum value.

This optimum $D_{\text{max}}$ should provide a smooth real space distance distribution function $p(r)$ such that $p(D_{\text{max}})$ and its first derivative $p'(D_{\text{max}})$ are approaching zero, and the back-transformed intensity from the $p(r)$ fits the experimental data.

Estimation of $D_{\text{max}}$ with GNOM (under-estimation)

Poor fit to experimental data

Distance distribution function $p(r)$ goes to zero too abruptly
Estimation of $D_{\text{max}}$ with GNOM (over-estimation)

Good fit to experimental data

BUT: Distance distribution function $p(r)$ becomes negative
Estimation of $D_{\text{max}}$ with GNOM (correct case)

Good fit to experimental data

Distance distribution function $p(r)$ goes smoothly to zero
The maximum size is determined from automated comparison of the $p(r)$ functions calculated at different $D_{\text{max}}$ values ranging from $2R_g$ to $4R_g$, where $R_g$ is the radius of gyration provided by AUTORG.

The calculated $p(r)$ functions and corresponding fits to the experimental curves are compared using the perceptual criteria of GNOM (Svergun, 1992) together with the analysis of the behavior of $p(r)$ function near $D_{\text{max}}$ and the best $p(r)$ function is chosen for the final output.

Examples for data processing

Go to directory /Examples/Primus/

../AgBeh - data from saxes and waxes range as well as the merged curve

../BSA - standard sample for calibration of molecular mass

../Lysozyme – example of p(r) from globular particle

../Lymazine – example of p(r) from hollow globular (virus-like) particle

../ISWI_Cterm – example of p(r) from elongated particle