Validation of macromolecular structures

- Integral part of structure determination and modelling
- A critical step to ensure the integrity of structural biology data
- Evaluating the reliability / accuracy of three-dimensional models of biologic macromolecules

Three main points:
- Validity of experimental data
- Consistency of the generated model with experimental data
- Consistency of the model with known biological, physical and chemical facts
Validation = Critical assessment

- How good is my model?
- Does it explain all data that was used?
- Does it explain all prior knowledge that was available?
- Does the model explain all the data that was not used (= cross-validation)?
- Is the model the best possible, most parsimonious explanation for the data?
- Are the testable predictions on the model correct?

Fyffe et al. Cell 2001
Validation is essential for data archiving

Aspects to consider with respect to archiving:

- Is the model ready for publication and archiving?
- How my model compares to other models?
- How much other people can rely their science on the model?
- Basis for high-throughput analysis (selecting suitable targets)
Why do mistakes happen? SAS-specific problems

Wrong data range:

Example: Lysozyme data up to 0.1 Å⁻¹  
Lysozyme data up to 0.3 Å⁻¹

Increasing accuracy / resolution

Increasing data range
Why do mistakes happen? SAS-specific problems

DAMMIF reconstruction, no constraints

DAMMIF reconstruction in P2

DAMMIF reconstruction in P2, Prolate anisometry constraint

Increasing accuracy / resolution

Increasing number of constraints
Why do mistakes happen? SAS-specific problems

Wrong constraints: 
DAMMIF reconstruction, no constraints

All models fit equally well the SAXS data!

DAMMIF reconstruction in P2, Prolate anisometry constraint

BUT: 
DAMMIF reconstruction in P2, Oblate anisometry constraint
Why do mistakes happen? SAS-specific problems

- Limitations in data
  - Incomplete data:
    - Data range not suitable for protein size
    - Data range not suitable modelling approach (SAXS vs. WAXS)
  - Low data quality:
    - Noisy data (detector problems, low concentration…)
    - Aggregated sample

- The human factor
  - Bias in the interpretation of the data / model
  - Inexperience
  - No time for validation
  - Incorrect background knowledge:
    Wrong sequence / MW information, incorrect atomic models for rigid-body modelling / hybrid approach, wrong symmetry constraints
VALIDATION OF SAS DATA
SAS data quality control

1. Initial checkup for aggregation/interparticle interaction (Guinier plot)

\[ \log[I(s)] \text{ vs. } s \]

\[ \log[I(s)] \text{ vs. } s^2 \]
SAS data quality control

1. Initial checkup for aggregation/interparticle interaction (Guinier plot)

2. Sanity check of model-free parameters ($D_{\text{max}}$, $I(0)$, $R_g$, MW)

   - Do the obtained values match with the expected ones? (if known from previous work)
SAS data quality control

1. Initial checkup for aggregation/interparticle interaction (Guinier plot)

2. Sanity check of model-free parameters ($D_{\text{max}}$, MW, $R_g$)

3. Concentration / Time-dependence of SAS profiles
   - Time-dependence of $R_g$ and $I(0)$
     $\rightarrow$ Radiation-induced aggregation
   - $I(0)/c$ and $R_g$ not constant over concentration series
     $\rightarrow$ Oligomerization process
The importance of reporting

- SAXS ‘Table 1’ of experimental settings and model free parameters ($D_{\text{max}}$, MW, $R_g$, $I(0)$)
- Reporting either values for each sample at every point in a concentration series or data interpolated to zero concentration
- Details how the scattering data were scaled and programs employed for data analysis/modelling

<table>
<thead>
<tr>
<th>Data collection parameters</th>
<th>CHI</th>
<th>naringenin-CHI</th>
<th>CHI Δlid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument</td>
<td>P12 at EMBL/DESY, storage ring PETRA III, Germany</td>
<td>0.2 × 0.12 mm²</td>
<td></td>
</tr>
<tr>
<td>Beam geometry</td>
<td>1.24</td>
<td>0.008 – 0.47</td>
<td>20 × 50</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.1 – 18.2</td>
<td>0.8 – 2.6</td>
<td>1.3 – 9.8</td>
</tr>
<tr>
<td>q-range (Å⁻¹)</td>
<td>283</td>
<td>283</td>
<td>283</td>
</tr>
<tr>
<td>Exposure time (ms)</td>
<td>20 × 50</td>
<td>20 × 50</td>
<td>20 × 50</td>
</tr>
<tr>
<td>Concentration range (mg ml⁻¹)</td>
<td>0.1 – 18.2</td>
<td>0.8 – 2.6</td>
<td>1.3 – 9.8</td>
</tr>
<tr>
<td>Temperature (K)</td>
<td>283</td>
<td>283</td>
<td>283</td>
</tr>
</tbody>
</table>

| Structural parameters*    | | | |
| $l(0)$ (arbitrary units)  | 25450 ± 10 | 24640 ± 10 | 26710 ± 10 |
| $R_g$ (from $P(r)$) (Å)   | 37 ± 2 | 36 ± 2 | 35 ± 2 |
| $l(0)$ (arbitrary units)  | 25450 ± 30 | 25100 ± 20 | 27220 ± 20 |
| $R_g$ (from Guinier)      | 36 ± 3 | 37 ± 3 | 36 ± 3 |
| $D_{\text{max}}$ (Å)      | 130 ± 5 | 110 ± 5 | 110 ± 5 |
| Porod volume (10^3 Å³)    | 320 ± 20 | 320 ± 20 | 270 ± 20 |

| Molecular mass determination* | | | |
| $M_{\text{Porod}}$ (from Porod volume) (kDa) | 190 ± 10 | 190 ± 10 | 160 ± 10 |
| Contrast ($\delta \rho \times 10^{10}$ cm⁻¹) | 3.047 | 3.047 | 3.047 |
| $M_{\text{max}}$ (from $I(0)$, (kDa)) | 190 ± 20 | 190 ± 20 | 150 ± 10 |
| Calculated hexameric MM from sequence (kDa) | 194.4 | 194.4 | 181.2 |
| SASDB entry code            | SASDAL6 | SASDAM6 | SASDAN6 |

| Software employed          | Automated radial averaging (Petoukhov et al., 2007) | PRIMUS (Konarev et al., 2003) | DAMMIN (Svergun, 1999) |
| Primary data reduction     | | | |
| Data processing            | | | |
| Ab initio analysis         | | | |
| Validation and averaging   | | | |
| Rigid body modelling       | | | |
| Computation of model intensities | | | |

*Reported for infinite dilution of concentration series measurements

Thomsen et al. 2015 Acta Cryst. D
Validation and quality estimates of SAS models
SAS-based *ab initio* modeling

- No prior structural knowledge needed

- Molecules presented as densely packed assemblies of beads (DAMMIN/F) OR as dummy residues (GASBOR)

- Monte-Carlo approaches employed to construct assemblies whose theoretical scattering profiles fit optimally the experimental data

- Typically 10 to 20 independent models generated

Multiple ab initio models and post-processing

- Multiple independent modeling runs required to reduce ambiguity

With multiple models:
- Find those that are most similar (uniqueness of reconstruction is not guaranteed)
- Superimpose and average them
- Restart fitting process using the averaged model

20 ab initio bead models of myoglobin (DAMMIF)
All structures fit equally good the measured SAXS data
Comparing SAS-models from an ensemble

- Superimpose models pairwise (principle axis alignment, gradient minimization, local grid search)

- Compute the similarities between the models:
  Similarity metric - Normalized Spatial Discrepancy (NSD)
  NSD < 1 implies similar models

<table>
<thead>
<tr>
<th>File</th>
<th>Aver</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.05</td>
<td>0.00</td>
<td>0.98</td>
<td>0.92</td>
<td>1.02</td>
<td>1.11</td>
<td>1.02</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>1.04</td>
<td>0.98</td>
<td>0.00</td>
<td>0.98</td>
<td>0.96</td>
<td>0.99</td>
<td>1.11</td>
<td>1.02</td>
</tr>
<tr>
<td>3</td>
<td>1.02</td>
<td>0.92</td>
<td>0.98</td>
<td>0.00</td>
<td>0.96</td>
<td>1.03</td>
<td>1.08</td>
<td>1.05</td>
</tr>
<tr>
<td>4</td>
<td>1.06</td>
<td>1.02</td>
<td>0.96</td>
<td>0.96</td>
<td>0.00</td>
<td>1.01</td>
<td>1.10</td>
<td>1.07</td>
</tr>
<tr>
<td>5</td>
<td>1.07</td>
<td>1.11</td>
<td>0.99</td>
<td>1.03</td>
<td>1.01</td>
<td>0.00</td>
<td>1.13</td>
<td>0.92</td>
</tr>
<tr>
<td>6</td>
<td>1.08</td>
<td>1.02</td>
<td>1.11</td>
<td>1.08</td>
<td>1.10</td>
<td>1.13</td>
<td>0.00</td>
<td>1.08</td>
</tr>
<tr>
<td>7</td>
<td>1.05</td>
<td>0.97</td>
<td>1.02</td>
<td>1.05</td>
<td>1.07</td>
<td>0.92</td>
<td>1.08</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>1.05</td>
<td>0.95</td>
<td>1.00</td>
<td>0.98</td>
<td>0.97</td>
<td>1.03</td>
<td>1.13</td>
<td>1.06</td>
</tr>
<tr>
<td>9</td>
<td>1.14</td>
<td>1.15</td>
<td>1.21</td>
<td>1.07</td>
<td>1.16</td>
<td>1.23</td>
<td>1.20</td>
<td>1.04</td>
</tr>
<tr>
<td>10</td>
<td>1.06</td>
<td>1.09</td>
<td>1.01</td>
<td>1.03</td>
<td>1.03</td>
<td>1.07</td>
<td>1.12</td>
<td>1.01</td>
</tr>
<tr>
<td>11</td>
<td>1.11</td>
<td>1.13</td>
<td>1.16</td>
<td>1.07</td>
<td>1.06</td>
<td>1.14</td>
<td>1.03</td>
<td>1.10</td>
</tr>
<tr>
<td>12</td>
<td>1.07</td>
<td>1.12</td>
<td>1.02</td>
<td>1.03</td>
<td>1.11</td>
<td>1.08</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>13</td>
<td>1.09</td>
<td>1.09</td>
<td>0.98</td>
<td>1.00</td>
<td>1.06</td>
<td>1.06</td>
<td>1.10</td>
<td>1.06</td>
</tr>
<tr>
<td>14</td>
<td>1.10</td>
<td>1.11</td>
<td>1.12</td>
<td>1.02</td>
<td>1.20</td>
<td>1.10</td>
<td>1.08</td>
<td>1.11</td>
</tr>
<tr>
<td>15</td>
<td>1.16</td>
<td>1.15</td>
<td>1.21</td>
<td>1.09</td>
<td>1.22</td>
<td>1.10</td>
<td>1.16</td>
<td>1.20</td>
</tr>
<tr>
<td>16</td>
<td>1.02</td>
<td>1.00</td>
<td>0.96</td>
<td>0.94</td>
<td>0.94</td>
<td>0.99</td>
<td>1.02</td>
<td>1.02</td>
</tr>
<tr>
<td>17</td>
<td>1.07</td>
<td>1.10</td>
<td>0.96</td>
<td>1.00</td>
<td>1.05</td>
<td>1.02</td>
<td>1.10</td>
<td>1.03</td>
</tr>
<tr>
<td>18</td>
<td>1.05</td>
<td>1.03</td>
<td>1.01</td>
<td>1.09</td>
<td>0.96</td>
<td>1.03</td>
<td>1.07</td>
<td>1.06</td>
</tr>
<tr>
<td>19</td>
<td>1.05</td>
<td>1.00</td>
<td>1.00</td>
<td>1.06</td>
<td>1.06</td>
<td>1.08</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>20</td>
<td>1.08</td>
<td>1.07</td>
<td>1.02</td>
<td>1.06</td>
<td>1.13</td>
<td>1.17</td>
<td>0.94</td>
<td>1.11</td>
</tr>
<tr>
<td>Aver</td>
<td>1.07</td>
<td>1.05</td>
<td>1.04</td>
<td>1.02</td>
<td>1.06</td>
<td>1.07</td>
<td>1.08</td>
<td>1.05</td>
</tr>
</tbody>
</table>

The myoglobin example

Mean value of NSD : 1.071
Standard deviation of NSD : 0.036

Refinement of SAS-models

- A bead probability density map can be generated within the search volume
- Take the averaged model – but this will not fit the data
- Take the model that has the least NSD to all others – this fits the data
- Use averaged model and restart DAMMIN/DAMMIF to fit the experimental data

Resolution of SAS models?

Xtallographic structure 2.25 Å

5 Å resolution

10 Å resolution

15 Å resolution  20 Å resolution

SAS-based ab initio models?
For MX and other diffraction methods, resolution is typically derived using Bragg’s law:

$$\text{Resolution} = \frac{2\pi}{s_{\text{max}}}$$

- A nominal theoretical resolution limit based on data range

$$s_{\text{max}} = \frac{5}{R_g}$$

$$s_{\text{max}} = \frac{7}{R_g}$$

$$s_{\text{max}} = \frac{9}{R_g}$$
Quality assessment and validation approaches

- Resolution limitations in SAS-based modeling
  - Signal-to-Noise Ratio (SNR) in the data
  - Data range
  - Spherically averaged data → Ambiguity problem
  - Search model used for reconstruction (Bead models vs. Dummy residue models)

There is no external objective standard by which the resolution of SAS-models could be evaluated such as the real-space distance criteria

**THUS:** The “crystallographic” resolution $2\pi/s_{\text{max}}$ does not work.
Quality assessment and validation approaches

- MX, NMR and atomic-resolution EM models can be quality assessed using stereo-chemical criteria
  - Knowledge-based scores which evaluate how models fit with the known features of proteins
    (e.g. Molprobity, CING, PROCHECK or ResProx)

Distribution of $\phi$, $\psi$ angles in PROCHECK
Reid et al. Structure(2011) 19, 1395-1412
Quality assessment and validation approaches

- MX, NMR and atomic-resolution EM models quality assessed with stereo-chemical criteria
  - Knowledge-based scores which evaluate how models fit with the known features of proteins
  (e.g. programs like Molprobity, CING, PROCHECK or ResProx)

**PROBLEM for SAS**:  *Ab initio* SAS models do not reveal atomic detail
  → A statistics based approach is not applicable

<table>
<thead>
<tr>
<th>Metric</th>
<th>Whole archive (#Entries)</th>
<th>Similar resolution (#Entries, resolution range(Å))</th>
</tr>
</thead>
<tbody>
<tr>
<td>R$_{free}$</td>
<td>91344</td>
<td>4533 (1.80-1.80)</td>
</tr>
<tr>
<td>Clashscore</td>
<td>102346</td>
<td>5383 (1.80-1.80)</td>
</tr>
<tr>
<td>Ramachandran outliers</td>
<td>190387</td>
<td>5320 (1.80-1.80)</td>
</tr>
<tr>
<td>Sidechain outliers</td>
<td>100360</td>
<td>5319 (1.80-1.80)</td>
</tr>
<tr>
<td>RSRZ outliers</td>
<td>91569</td>
<td>4547 (1.80-1.80)</td>
</tr>
</tbody>
</table>
Quality assessment and validation approaches

- For MX and other diffraction methods, resolution is typically defined using Bragg’s law

- MX, NMR and atomic-resolution EM models quality assessed with stereo-chemical criteria
  - Knowledge-based scores which evaluate how models fit with the known features of proteins
    (e.g. programs like Molprobity, CING, PROCHECK or ResProx)

- MX cross-validation using $R_{\text{free}}$

**PROBLEM for SAS**: The low information content of SAS data prevents computing of a ‘SAS $R$-free’ equivalent
Quality assessment and validation approaches

- For NMR models, quality assessed based on the RMSD of the reconstructed atomic ensembles compatible with the data

**PROBLEM for SAS:**
- The ensemble RMSD value tends to overestimate the true variance & the true resolution
- SAS models do not have a one-to-one point correspondence

PDBe NMR validation report: 2KNR
Quality assessment and validation approaches

- The resolution of EM model estimated by Fourier Shell Correlation (FSC) method

  \[ FSC = \text{Normalized cross-correlation coefficient between two 3-dimensional volumes over corresponding shells in Fourier space} \]  
  \[ (= \text{as a function of spatial frequency}) \]

- An analogous approach can be employed for SAS-based models

Resolution assessment: Can the variability of a model ensemble be used to estimate its resolution?
Fourier shell correlation (FSC) approach

\[ FSC(s) = \frac{\sum_{[s,\Delta s]} A_{lm}(s_i) \cdot B_{lm}^*(s_i)}{\sqrt{\sum_{[s,\Delta s]} |A_{lm}(s_i)|^2 \cdot \sum_{[s,\Delta s]} |B_{lm}(s_i)|^2}} \]

- **A** and **B** are two models and \( A(s) \) and \( B(s) \) their scattering amplitudes using spherical harmonics presentation

\[ FSC = \text{Normalized Fourier shell correlation coefficient between the scattering amplitudes of two molecules as a function of spatial frequency } s \]

Similar approach routinely used in EM studies

**Model A**

\[ A(s) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{lm}(s) Y_{lm}(s) \]

**Model B**

\[ B(s) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} B_{lm}(s) Y_{lm}(s) \]

Tuukkanen et al. IUCrJ 2016, In press
FSC approach – Ensemble variability measure

- Evaluates the consistency of models in reciprocal space
- Variability definition: The spatial frequency $s$ at which FSC equals 0.5
- The optimal cut-off value tested by model calculations on randomized atomic structures

Structural alignment of model B against model A
FSC Workflow

Several independent *ab initio* models

Pairwise structural alignment of models

Pairwise FSC calculations

- Structural alignments using SUPCOMB, NSD metric
- Ensemble of *N* structures $\Rightarrow N \times (N - 1) / 2$ comparisons

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>14.71</td>
<td>0.00</td>
<td>14.29</td>
<td>13.82</td>
<td>13.23</td>
<td>14.54</td>
<td>14.13</td>
<td>0.00</td>
<td>13.37</td>
</tr>
<tr>
<td>2</td>
<td>14.71</td>
<td>0.00</td>
<td>14.21</td>
<td>14.62</td>
<td>14.29</td>
<td>14.13</td>
<td>13.37</td>
<td>14.30</td>
<td>14.80</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>14.21</td>
<td>0.00</td>
<td>14.45</td>
<td>13.74</td>
<td>14.89</td>
<td>13.82</td>
<td>14.79</td>
<td>14.21</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>14.29</td>
<td>14.62</td>
<td>14.45</td>
<td>0.00</td>
<td>13.89</td>
<td>12.82</td>
<td>13.59</td>
<td>14.53</td>
<td>0.00</td>
<td>13.37</td>
</tr>
<tr>
<td>5</td>
<td>13.82</td>
<td>14.29</td>
<td>13.74</td>
<td>13.89</td>
<td>0.00</td>
<td>13.37</td>
<td>13.82</td>
<td>13.44</td>
<td>13.97</td>
<td>14.05</td>
</tr>
<tr>
<td>7</td>
<td>14.54</td>
<td>13.37</td>
<td>13.82</td>
<td>13.59</td>
<td>13.82</td>
<td>13.97</td>
<td>0.00</td>
<td>13.97</td>
<td>13.44</td>
<td>13.97</td>
</tr>
<tr>
<td>8</td>
<td>14.13</td>
<td>14.30</td>
<td>14.79</td>
<td>14.53</td>
<td>13.44</td>
<td>14.13</td>
<td>13.97</td>
<td>0.00</td>
<td>13.52</td>
<td>14.21</td>
</tr>
<tr>
<td>9</td>
<td>0.00</td>
<td>14.80</td>
<td>14.21</td>
<td>0.00</td>
<td>13.97</td>
<td>13.67</td>
<td>13.44</td>
<td>13.52</td>
<td>0.00</td>
<td>14.29</td>
</tr>
<tr>
<td>10</td>
<td>13.37</td>
<td>0.00</td>
<td>0.00</td>
<td>13.37</td>
<td>14.05</td>
<td>14.45</td>
<td>13.97</td>
<td>14.21</td>
<td>14.29</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Example: Myoglobin

- 20 DAMMIF models
  - Data range up to $s = 0.5 \, \text{Å}^{-1}$
- 190 pairwise FSC computations

Ensemble statistics:
The variability range = 12.2 – 20.1 Å
The standard deviation = 2.8 Å
Accurate variability assessment with average FSC

- Final variability estimate based on the average FSC over all pairwise correlation curves
- No need to smooth data by increasing the shell width $\Delta s$ in reciprocal space as in EM

Variability based on the ensemble, $\Delta_{\text{ensemble}} = 17.2$ Å
FSC Workflow

- Several independent \textit{ab initio} models
- Pairwise structural alignment of models
- Pairwise FSC calculations

<table>
<thead>
<tr>
<th>FSC</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variability estimate =
The spatial frequency $s$ at which the average FSC based on the pairwise internal cross-correlations equals 0.5

Average FSC curve
Benchmarking of the FSC approach using synthetic data

High-resolution xtal structures

1ATT
1FA2
1FS3
1IGD
1WC2
3V03
3VGZ
4Z0T

Synthetic data using CRYSOL

DAMMIF/GASBOR modeling runs

Pairwise structural alignments & FSC computations

FSC
1/d, Å\(^{-1}\)
Benchmarking of the FSC approach using synthetic data

- 107 proteins of various MWs and shapes
- Synthetic SAXS data generated using CRYSOL
- 20 independent \textit{ab initio} models for each protein using 5 different data ranges
  \rightarrow \text{Variability estimates}

<table>
<thead>
<tr>
<th>Protein</th>
<th>PDB id</th>
<th>MW, kDa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antithrombin III</td>
<td>1ATT</td>
<td>97.4</td>
</tr>
<tr>
<td>Beta-Amylase</td>
<td>1FA2</td>
<td>226.1</td>
</tr>
<tr>
<td>Ribonuclease A</td>
<td>1FS3</td>
<td>13.7</td>
</tr>
<tr>
<td>Protein G IgG-binding domain</td>
<td>1IGD</td>
<td>6.7</td>
</tr>
<tr>
<td>Glucose isomerase</td>
<td>1OAD</td>
<td>349.9</td>
</tr>
<tr>
<td>Subtilisin</td>
<td>1SCA</td>
<td>27.4</td>
</tr>
<tr>
<td>Ubiquitin</td>
<td>1UBQ</td>
<td>8.6</td>
</tr>
<tr>
<td>Carbonic Anhydrase</td>
<td>1V9E</td>
<td>58.2</td>
</tr>
<tr>
<td>Beta-Endoglucanase</td>
<td>1WC2</td>
<td>20.0</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>1WLA</td>
<td>17.7</td>
</tr>
<tr>
<td>Amine Oxidase</td>
<td>2C10</td>
<td>673.0</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>3LZT</td>
<td>14.9</td>
</tr>
<tr>
<td>BSA</td>
<td>3V03</td>
<td>66.0</td>
</tr>
<tr>
<td>Beta-propeller YncE</td>
<td>3VGZ</td>
<td>155.4</td>
</tr>
<tr>
<td>Oxoacyl reductase</td>
<td>4Z0T</td>
<td>28.2</td>
</tr>
</tbody>
</table>
Benchmarking of the FSC approach using synthetic data

- The selection of data ranges for DAMMIF modeling based on the $R_g$ of the proteins
- For GASBOR modeling two fixed $s_{\text{max}}$ values were used

<table>
<thead>
<tr>
<th>Protein</th>
<th>Modeling</th>
<th>s-range</th>
<th>$\Delta_{\text{ensemble}}$, Å</th>
</tr>
</thead>
<tbody>
<tr>
<td>3LZT</td>
<td>DAMMIF</td>
<td>$sR_g = 5$</td>
<td>16.71</td>
</tr>
<tr>
<td></td>
<td>DAMMIF</td>
<td>$sR_g = 7$</td>
<td>15.92</td>
</tr>
<tr>
<td></td>
<td>DAMMIF</td>
<td>$sR_g = 9$</td>
<td>16.07</td>
</tr>
<tr>
<td></td>
<td>GASBOR</td>
<td>0.5 Å⁻¹</td>
<td>14.63</td>
</tr>
<tr>
<td></td>
<td>GASBOR</td>
<td>1.0 Å⁻¹</td>
<td>14.53</td>
</tr>
<tr>
<td>1FS3</td>
<td>DAMMIF</td>
<td>$sR_g = 5$</td>
<td>18.78</td>
</tr>
<tr>
<td></td>
<td>DAMMIF</td>
<td>$sR_g = 7$</td>
<td>16.22</td>
</tr>
<tr>
<td></td>
<td>DAMMIF</td>
<td>$sR_g = 9$</td>
<td>11.53</td>
</tr>
<tr>
<td></td>
<td>GASBOR</td>
<td>0.5 Å⁻¹</td>
<td>11.07</td>
</tr>
<tr>
<td></td>
<td>GASBOR</td>
<td>1.0 Å⁻¹</td>
<td>12.38</td>
</tr>
</tbody>
</table>

The selection of data ranges for DAMMIF modeling based on the $R_g$ of the proteins.

For GASBOR modeling two fixed $s_{\text{max}}$ values were used.

The selection of data ranges for DAMMIF modeling based on the $R_g$ of the proteins.

For GASBOR modeling two fixed $s_{\text{max}}$ values were used.
How SAS-model ensemble variability is related to model resolution?

Cross-validation against X-ray crystallographic structures
Cross-validation against x-ray crystallographic structures

- FSC comparisons between ab initio models and the corresponding high-resolution structures - Xtal structures assumed to be error- and noise-free

- Cross-validated resolution $\Delta_{cc}$ = The spatial frequency $s$ at which the FSC between a model and the corresponding xtal structure equals 0.5

$$\text{FSC} = \frac{1}{d, \text{Å}^{-1}}$$
Cross-validation against x-ray structures – Example: Myoglobin

Average cross-validated resolution: \( \Delta_{cc} = 19.6 \ \text{Å} \)

Ensemble statistics:
- The resolution range = 16.1 - 31.5 Å
- The standard deviation = 4.2 Å
Cross-validation against x-ray structures – Myoglobin

- Average resolution based on internal comparison: $\Delta_{\text{ensemble}} = 17.2 \text{ Å}$ (blue)
- Average cross-correlated resolution: $\Delta_{cc} = 19.6 \text{ Å}$ (orange)
The linear relationship between Resolution and $\Delta_{ens}$

Linear correlation observed for both bead (Pearson correlation coefficient $r = 0.80$) and dummy-residue (Pearson correlation coefficient $r = 0.86$) models. The 95% confidence intervals are shown by red dotted lines and the 95% prediction intervals by blue dotted lines.
For all benchmark proteins, $\Delta_{cc}$ was found to be somewhat higher than $\Delta_{ens}$

A linear correlation between model resolution and ensemble variability established

Discrepancy can be explained by the presence of constraints (such as interconnectivity & compactness) in *ab initio* modeling

The use of the linear models provides a conservative estimate of the resolution of ensembles of unknown protein
Quality assessment and cross-validation of rigid-body models
Rigid-body modeling against SAS data

The goodness-of-the-fit, $\chi^2$

$$\chi^2 = \frac{1}{N-1} \sum_{i=1}^{N-1} \left( \frac{I_{\text{exp}}(s_i) - cI(s_i)}{\sigma(s_i)} \right)^2$$

- Rigid-body modelling is typically also repeated several times

  ➜ Estimating variability & RMSD based clustering
Problem: Ambiguity of rigid body models

How to distinguish the correct quaternary structure of a protein complex among several SAXS models?

Additional information is ALWAYS required to resolve or reduce ambiguity of interpretation
Data integration approach to reduce ambiguity

- All possible constraints & restraints should be collected
- Finding a structure based on SAS data with constraints included in the modelling target function
- Finding a structure which satisfies all constraints in an ensemble of structures consistent with SAS data

**Experimental constraints**
- Structural Interaction Templates
- Site-Directed Mutagenesis
- NMR restraints
- *In vivo* crosslinking
- FRET

**Computational constraints**
- Physics-based scoring functions
- Knowledge-based scoring functions
- Binding site predictions
- Surface residue conservation
- Surface shape complementarity
Structural basis for antigen recognition by TG2-specific autoantibodies in celiac disease

- Question: How celiac disease autoantibodies recognize transglutaminase 2 (TG2)
- The interaction between TG2 and a celiac disease epitope anti-TG2 antibody (Fab fragment) was studied by SAXS and combination of biochemical techniques

The scattering profiles and theoretical fits of the complex (pink), TG2 (gray), and the Fab fragment (green)

Collaboration with Melissa Graewert
Xi Chen et al. J. Biol. Chem. 2015 290: 21365-21375
Rigid-body models of TGA2 in complex with the Fab fragment

- 17 complex models generated with SASREF without any constraining
- Six different clusters based on NSD
- Hydrogen/deuterium exchange experiments and prior biological knowledge indicate to group $f$
TGA2 in complex with the Fab fragment –
Rigid-body model representative of group f

- Residues of TG2 within 5 Å distance to residues of the Fab fragment in yellow
- Residues selected for mutagenesis analysis are colored in red

Binding of antibody Fab 679-14-E06 to mutants of TG2 as assessed by ELISA
Atomistic MD simulations to refine and analyse interfaces

- Refinement of SAXS rigid-body complex models
  - Replica-exchange simulations for fast conformational sampling
  - Study of binding interface region
    → structural arrangements upon binding
    → improvement of SAXS data fitting in an iterative process

- Usage in validating rigid-body modeled protein complexes
  - Stability analysis of complexes (RMDS, RMSF,...)
  - Time-averaged interaction energies between protein subunits
Structure models derived from MD simulations of the interaction between TG2 and the Fab fragment

- MD simulations using NAMD and CHARMM36 all-atom force field
- The rigid body model representative of group $f$ as a starting model
- After 1.1 ns an equilibration state was reached (the backbone RMSD $\leq 1.0$ Å)
- The total mean binding = 475 kcal/mol (The electrostatic contribution = 447 kcal/mol
 Van der Waals interactions contribution =28 kcal/mol)
Structure models derived from MD simulations of the interaction between TG2 and the Fab fragment

- MD simulation reveals the involvement of the water network around His-134 in interacting with the heavy chain of Fab fragments.

- This water network is disrupted by replacing histidine with alanine -> disease relevant mutation.

- Conclusions on atomic detail possible when SAXS data used with atomic structures of subunits/domains and other (biochemical / computational) methods.
Joint use of SAS data with other methods

- Additional information (structural or biochemical) is ALWAYS required to resolve or reduce ambiguity of SAS data interpretation
- SAS provides complementary information to other structural methods like MX, NMR, EM, etc.
  - Cross-validation of SAS models against other structural models of the same system
  - Cross-validation of atomic models from other structural methods against SAS data
- Topics covered in several excellent talks during this course
  SAXS & AUC - Olwyn Byron
  SAXS & biochemical methods - Maria Vanoni
  SAXS & NMR - Annalisa Pastore
  SAXS & crystallography - Rob Meijers
Possible use of SAS in combination with Electron microscopy

- Use a solution scattering *ab initio* structure as a starting reference for EM reconstruction

  Hsp90 heat-shock protein


- Comparison of SAS models and independent EM reconstructions

  Tumour suppressor p53 and its complex with DNA

  Tidow, H et. al. (2007) *Proc Natl Acad Sci USA*, **104**, 12324
EM2DAM: Cross-validation of SAS models and EM maps

- Tool for computing SAXS profiles from EM maps of proteins
- EM2DAM fills the EM density with dummy residues located at the pixel size distance from each other
- The user should only provide a contour level value defining the particle density
- Output file has a PDB-like format

- Can be used, e.g. to compute theoretical scattering profiles
- Validation and comparison of EM maps with SAXS data
EM2DAM: Cross-validation of SAS models and EM maps

DENSITY MAP (MRC format) from EMDB

GroEL: EMD-1080

EM2DAM

- Contour level

BEAD MODEL

Theoretical SAXS profile
EM2DAM: Cross-validation of SAS models and EM maps

DENSITY MAP (MRC format) from EMDB

GroEL: EMD-1080

EM2DAM

Contour level

BEAD MODEL

Experimental SAXS data of GroEL (Cy Jeffries)
EM2DAM: Cross-validation of SAS models and EM maps

DENSITY MAP (MRC format) from EMDB

EM2DAM
- Relaxed Contour level

BEAD MODEL
STARTING SEARCH VOLUME FOR DAMMIN REFINEMENT

DAMMIN
- Damstart Search volume

GroEL: EMD-1080

Fit against the experimental SAXS data
EM2DAM: Cross-validation of SAS models and EM maps

- Basic validation information available for all EMDB entries

- Volume graphs: Map-density distribution, Volume estimate, Radially averaged power spectrum (RAPS)

- Comparison of RAPS and experimental/theoretical SAXS data provides means for validation

Ardan Patwardhan, EMBL-EBI
EM2DAM: Cross-validation of SAS models and EM maps

Example: GroEL EMD-1080

- RAPS curve of the EM map can be compared to experimental SAXS data
- Possibility to use for searching structurally similar macromolecules
Macromolecular Validation - Summary

- High SAS data quality and sanity checks basic requirement for any structural analysis

- Ambiguity problem reduced/solved by using additional biochemical / structural data

- Validation of SAS models using quality measures:
  - NSD and clustering approach
  - Resolution estimate of ab initio models

- Cross-validation of SAS models against structural models obtained by other methods
Acknowledgements

Dmitri Svergun (EMBL-HH)
Gerard Kleywegt (EMBL-EBI)
Ardan Patwardhan (EMBL-HH)

The EMBL-HH BioSAXS group

„The human understanding is not composed of dry light, but is subject to influence from the will and the emotions, a fact that creates fanciful knowledge; man prefers to believe what he wants to be true... for what man had rather were true he more readily believes.“ Sir Francis Bacon Novum Organum Scientiarum

Funding: The EMBL Interdisciplinary Postdoc Programme (EIPOD) under Marie Curie COFUND actions
BMBF research grant BioSCAT