Ensemble Optimization Method on SAXS

EOM 2.0 – tutorial

Giancarlo Tria
giancarlo.tria@embl-hamburg.de
BioSAXS group @ EMBL Hamburg

EMBO Practical Course on Solution Scattering from Biological Macromolecular
October 21st, 2012
Kratky Plots to Detect Disorder

Kratky plot establishes an approximate relationship between $I(s)$ vs $s$ for folded and unfolded proteins.
Indications (not Proofs!!!) of Flexibility

► Smooth Scattering profiles and featureless Kratky Plots

► Large $R_g$ and $D_{max}$

► Absence of correlation peaks in the $p(r)$ function

► Low correlation densities in *ab initio* reconstructions

► Isolated domains in rigid body modelling

► Prediction of disorder using bioinformatics tools

http://www.idpbynmr.eu/home/science/research-tools.html
Detection of Flexibility: A Crucial Issue

SAXS curves

Analysis of the overall size descriptors ($R_g$, $p(r)$, Kratky)

Modelling: \textit{ab initio} (DAMMIN/DAMMIF) and Rigid body (BUNCH/CORAL)

Analysis of the differences

Go for flexibility!
Flexibility as mixture of different conformations

\[ I(s) = 4\pi \int_0^D p(r) \frac{\sin sr}{sr} dr \]

For monodisperse systems the scattering is proportional to that of a single particle averaged over all orientations.

\[ I(s) = \sum_k v_k I_k(s) \]

\( v_k \) = volume fraction
\( I_k(s) \) = scattering intensity from the \( k \)-th component
Ensemble methods in SAXS

The Ensemble Optimization Method (EOM)
Genetic Algorithm *(optimized ensemble size)*

Chromosome → Mutation → Crossing → Elitism → Generation 1

Generation 2

Chromosomes

\[ I(s) = \frac{1}{N} \sum_{n=1}^{N} I_n(s) \]
Modelling: Native vs. Random

Quasi $C_\alpha$-$C_\alpha$ Ramachandran plot

Bond angles vs. Dihedral angles

G. Kleywegt, Validation of protein models from $C_\alpha$ coordinates alone, JMB, 1997, 273, 371-376

Theoretical distribution of the bond and dihedral angles for random chains

$$R_g = R_0 \cdot N^\nu$$

$R_g$ Persistence Length
$\nu$ Solvent ‘quality’

Several experimental and theoretical studies establish $\nu \approx 0.588$ as an indication of the ‘random coil’ in chemically denatured (Urea or GuHCl) proteins.

Kohn et al. PNAS, 2004, 101, 12491
Unfolded protein …

- TAU protein isoform (124AA)

**Inputs for using EOM:**

- `sequence.seq`
- `curve.dat`

```
> CYLSRKMLDARENLKLDRMNRSLPHSCLQDKDFGLPQEMVEGDQLQKDPFVLYEMLQQSFNLFYTEHSSAAWDTTLLEQLCTGLQQQLDHLDTCRGQVMGEEDSELGNMDPIVTVKKYF
```
Ensemble Optimization Method on SAXS – EOM 2.0

<1min per repetition
... unfolded protein: results

<table>
<thead>
<tr>
<th>Filename</th>
<th>Rg</th>
<th>Dmax</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>00700eom.pdb</td>
<td>44.68</td>
<td>143.34</td>
<td>7</td>
</tr>
<tr>
<td>01193eom.pdb</td>
<td>36.62</td>
<td>115.47</td>
<td>2</td>
</tr>
<tr>
<td>02217eom.pdb</td>
<td>37.78</td>
<td>123.66</td>
<td>1</td>
</tr>
<tr>
<td>04413eom.pdb</td>
<td>51.98</td>
<td>167.94</td>
<td>14</td>
</tr>
<tr>
<td>05371eom.pdb</td>
<td>38.33</td>
<td>105.55</td>
<td>1</td>
</tr>
<tr>
<td>05721eom.pdb</td>
<td>40.61</td>
<td>127.53</td>
<td>6</td>
</tr>
<tr>
<td>09292eom.pdb</td>
<td>36.95</td>
<td>125.05</td>
<td>1</td>
</tr>
<tr>
<td><strong>Average values:</strong></td>
<td><strong>45.93</strong></td>
<td><strong>147.03</strong></td>
<td><strong>32</strong></td>
</tr>
<tr>
<td><strong>Chi-value:</strong></td>
<td><strong>0.883</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{R}_g = 45.05 \]
\[ R_g = 32.96 \]

\[ \bar{D}_{max} = 140.38 \]
\[ D_{max} = 101.02 \]
Missing loops (i.e. flat electron density map) ...

MRIGMV.........GGVQSHVLQ.....VLRDA

Kratky Plot

seq.seq

curve.dat

Nter.pdb

Cter.pdb

apo ferritin

pool

missing loop
30 AA
... Missing loops: results

<table>
<thead>
<tr>
<th>Rg</th>
<th>Dmax</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.51</td>
<td>81.35</td>
<td>44</td>
</tr>
</tbody>
</table>

Chi-value: 0.357

\[
\bar{R}_g = 24.46 \\
R_g = 24.28
\]
Flexible pentamer in solution ...
(full length protein measured in two buffers, with low and high ionic strength respectively)
Flexible pentamer in solution: results
(full length protein measured in two buffers with low and high ionic strength respectively)
Case extra: dodecamer (P62, 2 domains) + tRNA

158 AA N-terminal tail
high resolution (MX) N-terminal monomer domain (141 AA)
9 AA inter-domains linker
high resolution (MX) C-terminal monomer domain (270 AA)
30 N single strand tRNA

max distance in Å
subUnit
contact residues range
EOM Tests: Size of Pool

Number of chains: 10

Number of chains: 100

Number of chains: 1000

Number of chains: 5000

Number of chains: 10000

Number of chains: 64790
Resolution of Subpopulations by EOM ...

Generate a pool, select two subpopulations from it and calculate scattering curve for their union

Wide subpopulations

Narrow subpopulations
...resolution of Subpopulations by EOM

Genetic Algorithm

Scattering curves

Ensemble Optimization Method on SAXS – EOM 2.0

10% < Rg < 15%, 85% < Rg < 90%
5% < Rg < 10%, 90% < Rg < 95%
10% < Rg < 15%, 85% < Rg < 90%
15% < Rg < 20%, 80% < Rg < 85%
20% < Rg < 25%, 75% < Rg < 80%
25% < Rg < 30%, 70% < Rg < 75%
30% < Rg < 35%, 65% < Rg < 70%
35% < Rg < 40%, 60% < Rg < 65%
40% < Rg < 45%, 55% < Rg < 60%
45% < Rg < 50%, 50% < Rg < 55%
Pool

15-20%
Undistinguishable
ΔRg = 11.33

10-15%
Well distinguishable
ΔRg = 13.02
Take home messages (EOM 2.0)

• EOM allows one to quantitatively characterize the flexibility of a particle (what the conformations that the protein prefers in solution)

• Intrinsically unfolded protein (IDP) can be easily modelled with EOM (no Size limitations)

• SAXS in solution can be used as complementary technique to model flexible systems, disordered regions, etc.