ESS Neutron Protein Crystallography 2013
Aarhus, Denmark

shelxl: Refinement of Macromolecular Structures from Neutron Data

Tim Grüne
University of Göttingen
Dept. of Structural Chemistry

http://shelx.uni-ac.gwdg.de
tg@shelx.uni-ac.gwdg.de
Motivation for Neutron Macromolecular Crystallography

X-ray scattering \( f(2\theta = 0) = Z/\text{Å} \)

Neutron scattering length \( b_c \)

- Pharmaceutical application:
  - Binding interaction between drug and target usually hydrophobic
  - “Acting” interaction usually electrostatic / interchange of hydrogens
Benefit of neutron radiation

1. Detection of $H/D$ atoms (enzymatic reactions)

2. Detection of metals in their reduced state (e.g. $Fe^{2+}$ in rubredoxin, 4AR4 vs. 4AR3) — no production of free radicals, i.e. no radiation damage

Generation of Deuterated Crystals

1. Small Molecules: Storage in $D_2O$ (> 3 months)

$\Rightarrow$ only replaceable $H$, e.g. not methyl groups $CH_3$

2. Proteins: recombinant expression by E. coli with **fully deuterated media** ($D_2O$, amino acids, ...)
Problems of MX Neutron Measurements

- incomplete data, even from Laue data collection
- sample size
- costs for deuteration

Protein Data Bank (Nov 2012) 86,487 structures, 63 by neutron diffraction (29 unique by 90 % sequence cut-off)

Cambridge Crystallographic Database (Nov 2012) 624,927 entries, 1,565 neutron data
Joint X–ray + Neutron data Refinement


Stabilisation of refinement by X–ray data
Why Neutron Refinement Alone

- Main point of neutron structure solution: protonation states of carboxyl groups, histidine, ...  

- X-rays interact with electrons, i.e. hydrogen bonds from X-ray data appear shorter (0.9Å) than internucleic distances (1.1Å)  

- positions of hydrogen in water and $-\text{OH}$ groups cannot be inferred from X-ray data  

- While weak, hydrogen contribution is taken into account when calculating X-ray structure factors

⇒ X-ray and neutron data are different

This work: Feasibility study of refinement against neutron data alone

- Data: Rubredoxin (provided by Dr. Flora Meilleur)

- Program: SHELXL-2013
Neutron Refinement with Shelx1-97

- Previously (shelx1-97): Scattering bond length defined through SFAC card

  SFAC  C  0 0 0 0 0 0 0 0 6.6460 0 0 5.5500 0.76 12
  SFAC  H  0 0 0 0 0 0 0 0 -3.7390 0 0 1.7583 0.31 1
  SFAC  N  0 0 0 0 0 0 0 0 9.3600 0 0 11.010 0.71 14

- CHIV-restraint (chiral volume) does not work for $C_{\alpha}$ bound to 4 atoms:

  ** Bad CHIV: CA_1043 bonds to DA_1043 N_1043 CB_1043 C_1043 - ignored **

  ⇒ important restraints for protein structures missing
Shelx1-2012: NEUT card

1. Removes special treatment of hydrogen atoms

2. Short ("normal") SFAC syntax can now be used:

   SFAC  C  H  N

   Convenient consequence: compatibility with COOT (P. Emsley)

3. CHIV  CA works
The Rubredoxin Neutron Data


- PDB codes 3kyu, 3kyv, 3kyw, 3kyx, and 3kyy

- Neutron data for 3kyx kindly provided by Dr. Flora Meilleur (SNS Oak Ridge, USA)

- 3kyx: joint neutron + X-ray refinement, T=295 K, 1.6 Å, $R/R_{\text{work}} = 16.9\%, 19.7\%$

- $P2_12_12_1$, $a = 33.92$ $b = 34.93$ $c = 43.53$
Rubredoxin Neutron Data from LADI-III at ILL
Data and Parameters for Rubredoxin, $T = 295K$

$\lambda$ $3.3 - 4.2$ Å
$d$ $21.8 - 1.65$ Å (1.75 – 1.65 Å)
# reflections 16111
unique 4781 (237)
Completeness 72.1 % (30.3%)

817 atoms + 1Fe$^{2+}$ or $^{3+}$ ($C_{267}^{2}H_{396}N_{62}O_{87}S_{5}$) (excluding solvent molecules)
Refinement Results

- $R/R_{\text{free}} = 22.5\% / 29.5\%$
  
  (4256 reflections + 5759 for 2923 parameters)

- No disagreeable restraints

Confirmative:

- Dual Conformation of ILE 1011 not present in 3kyx
- Fragment of C-terminal GLU 1052 not present in 3kyx
- Conservative placement of water molecules: $11 \times D_2O$ vs. 28 in 3KYX
Restraints and Constraints

- Data to parameter ratio **very low** ($< 1.5$ without restraints)

- Neutron data: many more parameters than X-ray because of hydrogen atoms

- Refinement must be stabilised through **constraints** & **restraints**

**Restraints** mathematically like data, but cannot make up for lack of completeness

**Constraints** reduce number of parameters, *i.e.* real improvement of data:parameters
Standard Protein Restraints & Constraints

- **SHELXPRO**: Standard restraints for macro molecules (Engh & Huber Parameters, Acta Cryst. (1991), A47, 392–400) derived from CCSD: \texttt{DFIX}, \texttt{DANG}, \texttt{CHIV}, \texttt{FLAT}

- Treatment of hydrogens in X-ray crystallography: Riding hydrogen model

  = calculated \textit{H}-positions relative to bonded atom (\texttt{HFIX}/ \texttt{AFIX} in shelxl)

  = no extra parameters for X-ray data
New Restraints for $D/H$


- Only very few samples, e.g. searching for results in # 15 samples from CCSD

- *Guesstimate* distances (H-X distance for X-ray data in shelxl set 40 years ago ...)

Default effective X-H distances for $T = 20.0$ C

$\text{AFIX m} = 1 \quad 2 \quad 3 \quad 4 \quad 4[N] \quad 3[N] \quad 15[B] \quad 8[0] \quad 9 \quad 9[N] \quad 16$

$d(\text{X-H}) = 0.98 \quad 0.97 \quad 0.96 \quad 0.93 \quad 0.86 \quad 0.89 \quad 1.10 \quad 0.82 \quad 0.93 \quad 0.86 \quad 0.93$
Hydrogenation with \texttt{shelxl}

- Coordinate for non-H/ non-D usually known from X-ray experiment

- Addition of hydrogens with preset distance in \texttt{shelxl} with "HFIX" commands

\begin{verbatim}
HFIX_ASN 43 -1.5 1.02 N
HFIX_ASN 13 -1.5 1.09 CA
HFIX_ASN 23 -1.5 1.096 CB
HFIX_ASN 93 -1.5 1.02 ND2
\end{verbatim}

- Generation of Angular Distances restraints (\texttt{DANG}): measured from resulting .res-file after 0 cycles of refinement
More chemistry: CHIV vs. FLAT

CHIV Cζ : more flexibility for \( D \) atom to move out of aromatic ring

CHIV N available for neutron data
Constraining $U_{\text{iso}}(D/H)$

- SHELXL syntax: $U_{\text{iso}} = -5 \leq q \leq -0.5$: Isotropic ADP relative to previous “normal” atom

$$U_{\text{iso}}(D) = |q| * U_{\text{iso}}(X)$$

- avoid 1 parameter per $D$

- automatic with e.g. HFIX_ASN 43 -1.5 1.02 N

Heuristics values by bond type:

$N - D_{2/3} = -1.7$ || $C_a - D_a = -1.3$

$N - D = -1.3$ || $C - D_2 = -1.5$

$O - D = -1.7$ || $C - D_3 = -1.7$

- $R/R_{\text{free}} = 22.9\%/30.8\% \rightarrow 23.6\%/30.1\% = \text{Closure of gap between } R \text{ and } R_{\text{free}}$
Deuterium Saturation

- **Motivation:** Contamination of $^2H$ with $^1H$

- **shelxl:** Occupancy refinement for groups of atoms by free variables

<table>
<thead>
<tr>
<th>type</th>
<th>“exchange rate”</th>
<th>FVAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N - D &amp; O - D$</td>
<td>easily exchanged</td>
<td>2</td>
</tr>
<tr>
<td>$C_\alpha - D_\alpha$</td>
<td>medium exchange</td>
<td>3</td>
</tr>
<tr>
<td>$C - D_n$</td>
<td>slow to no exchange</td>
<td>4</td>
</tr>
</tbody>
</table>

- **Starting values for all three FVAR’s:** 0.9
Deuterium Saturation

1. Non-physical results for carbon bound $D$

<table>
<thead>
<tr>
<th>type</th>
<th>refined FVAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N - D_n$ &amp; $O - D$</td>
<td>0.937</td>
</tr>
<tr>
<td>$C_\alpha - D_\alpha$</td>
<td>1.113 ← non-sense</td>
</tr>
<tr>
<td>$C - D_n$</td>
<td>1.138 ← non-sense</td>
</tr>
</tbody>
</table>

2. Only Nitrogen and Oxygen bound $D$: FVAR=89.3%

3. Calculation of $D$ saturation $p$:

$$f \times b_c(D) = p \times b_c(D) + (1 - p) \times b_c(H)$$

$$0.89 \times 6.67 \text{ fm} = p \times 6.67 \text{ fm} - (1 - p) \times 3.74 \text{ fm}$$

$$p = 93\%$$

4. Easier than “dual conformation” with PARTs for $D$ and $H$
Data:Parameter: Replacing $R_{\text{free}}$ with 50-fold cross-validation

- Macromolecular Refinement without cross validation nearly impossible (over-fitting)


- 50-fold cross validation with 50 flagged reflections (conventional: 22.5%/29.5%):
  
  \[ R_1 : 22.9\% \pm 0.06\% \quad (R_{\text{free}}) : 29.6\% \pm 3.73\% \]

- Benefit: 4731 vs. 4256 reflections, ratio 1.62 vs. 1.46
Validation of Restraints


- data available at PDB contain $I^+$ and $I^-$, not only $F$

- rubredoxin ins-file refined against r4ar4sf.ent

- geometry checked with http://molprobity.biochem.duke.edu/
## Validation of Restraints

### Protein Geometry

<table>
<thead>
<tr>
<th>All-Atom Contacts</th>
<th>Clashscore, all atoms:</th>
<th>13.31</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB-ID 4AR4</td>
<td>21st percentile (^*) (N=456, 1.38Å ± 0.25Å)</td>
<td></td>
</tr>
</tbody>
</table>

Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.

### Protein Geometry

<table>
<thead>
<tr>
<th>Poor rotamers</th>
<th>0</th>
<th>0.0%</th>
<th>Goal: &lt;1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramachandran outliers</td>
<td>0</td>
<td>0.0%</td>
<td>Goal: &lt;0.05%</td>
</tr>
<tr>
<td>Ramachandran favored</td>
<td>51</td>
<td>96.23%</td>
<td>Goal: &gt;98%</td>
</tr>
</tbody>
</table>

**MolProbity score\(^*\)**

| 1.94 |
| 99th percentile \(^*\) (N=3847, 1.38Å ± 0.25Å) |

CB deviations >0.25Å

| 0 / 215 |
| 0.0% |
| Goal: >0% |

Bad backbone angles

| 0 / 267 |
| 0.0% |
| Goal: <0.1% |

---

### Protein Geometry

<table>
<thead>
<tr>
<th>All-Atom Contacts</th>
<th>Clashscore, all atoms:</th>
<th>11.86</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB-ID 4AR4</td>
<td>63rd percentile (^*) (N=1784, all resolutions)</td>
<td></td>
</tr>
</tbody>
</table>

Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.

### Protein Geometry

<table>
<thead>
<tr>
<th>Poor rotamers</th>
<th>1</th>
<th>2.50%</th>
<th>Goal: &lt;1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramachandran outliers</td>
<td>0</td>
<td>0.0%</td>
<td>Goal: &lt;0.05%</td>
</tr>
<tr>
<td>Ramachandran favored</td>
<td>51</td>
<td>98.08%</td>
<td>Goal: &gt;98%</td>
</tr>
</tbody>
</table>

**MolProbity score\(^*\)**

| 1.89 |
| 91st percentile \(^*\) (N=27675, 0Å - 99Å) |

CB deviations >0.25Å

| 0 / 217 |
| 0.0% |
| Goal: >0% |

Bad backbone angles

| 0 / 257 |
| 0.0% |
| Goal: <0.1% |

---

### Protein Geometry

<table>
<thead>
<tr>
<th>All-Atom Contacts</th>
<th>Clashscore, all atoms:</th>
<th>10.54</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB-ID 4AR4</td>
<td>68th percentile (^*) (N=1784, all resolutions)</td>
<td></td>
</tr>
</tbody>
</table>

Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.

### Protein Geometry

<table>
<thead>
<tr>
<th>Poor rotamers</th>
<th>0</th>
<th>0.0%</th>
<th>Goal: &lt;1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramachandran outliers</td>
<td>0</td>
<td>0.0%</td>
<td>Goal: &lt;0.05%</td>
</tr>
<tr>
<td>Ramachandran favored</td>
<td>51</td>
<td>98.98%</td>
<td>Goal: &gt;98%</td>
</tr>
</tbody>
</table>

**MolProbity score\(^*\)**

| 1.54 |
| 94th percentile \(^*\) (N=27675, 0Å - 99Å) |

CB deviations >0.25Å

| 0 / 207 |
| 0.0% |
| Goal: >0% |

Bad backbone angles

| 0 / 257 |
| 0.0% |
| Goal: <0.1% |

---

**PDB-ID 4AR4**

**4AR4** with shelxl restraints, no model building

**4AR4** with shelxl restraints, after 1\(^{st}\) round of model building
Conclusions

MX refinement of neutron data alone feasible

SHELXL provides powerful constraints ($U_{iso} = -1.7$)

Occupancy refinement for $D$ gives reasonable result

TODO: DNA restraints (PDB ID 1WQZ and 3QBA)
Acknowledgment

- Dr Flora Meilleur, Spallation Neutron Source, Oak Ridge, USA: Neutron data sets for Rubredoxin and Xylose Isomerase

- Prof George M. Sheldrick, University of Göttingen for discussion, advice and for SHELXL-2013

- Thomas Niklas (model building of 4AR4)