Introduction to Protein Powder Diffraction: Overview

R.B. Von Dreele, Advanced Photon Source
Argonne National Laboratory

Thanks: DOE/OS/BES, P. Stephens, I. Margiolaki, J. Wright
Proteins – polymers of amino acids - representations

Space filling (HEWL+NAG₃)
129 AA

Ribbon diagram – botox
~1100 AA

Random coil

α-helix

β-sheet

Schematic – insulin
102 AA in 4 chains
1/3 of shown
**Why proteins & powders?**

- ~6-10% of all identified DNA sequences yield PDB deposited structures (NB: only 2% of DNA give protein sequences!)
- ~30% lost growing suitable crystals
- Powders in general much easier to make
- Apart from structure – protein materials crystallography unexplored
- Proteins are polyphasic – powder discovery
- Etc., etc., etc.
- The challenge!
Protein powders – “ideal” (1μm & no μstrain)

T₆ Zn insulin; NSLS X3b1; \( \lambda = 1.401\text{Å} \)

- Sharp peaks! (better than NIST SRM’s!)
Instrumentation

Laboratory CuKα

12 Multidetector/analyzers
11BM - APS
Faster data collection - Image Plate Detector – MAR345
Avoid radiation damage

Beam focused to MAR345 surface & MAR345 offset 6cm up

Sample (spun, 1x1mm)

Beam stop

\[ \lambda = 0.6488\text{Å} \ (20\text{keV}) \]

\(~300\text{mm}\)

\(~725\text{mm}\)

1BM/APS June 2002; thx Peter Lee
Rings – protein pattern (HEWL) –
X-rays 30s @ 20kV on MAR345; <1mg HEWL

Texture free sample & no graininess – 1μm “perfect” powder
Resolution limit – 1.85Å
Residual solvent scattering – background

Inner most ring – d~55Å
(110) Reflection, lowest order for tetragonal lysozyme
2Θ ~ 0.67deg

Beam stop holder

~9000 F_{hkl} for HEWL >2Å

(Air, solvent & Kapton background subtracted)
Compare image plate with analyzer/detector

11BMB – 10min scan

1BM/MAR345 – 1sec exposure

3-4X broader peaks!
How do we get the most out of this data?

- Use multiple data sets to recover overlaps
  \[ \textit{more information} \]
- Well known molecular geometry – bonds, angles, planes & torsions (primary structure) in 20 amino acids
- Well known chain folding patterns (coupled torsion angles – secondary structure)
  \[ \textit{restraints & constraints as rigid bodies} \]
- Long chain polymers – can structure least squares to fit; band matrix (not full matrix)
**pH induced phase changes in HEWL (acetate buffer)**


Also see lattice parameter changes with pH – not isotropic
Solvent & radiation damage induced lattice changes for HEWL

Salt & pH effects

Radiation exposure ~8% loss in 4.5m

Compare: [NaCl] ~1-5% strain; pH similar; rad.dam. ~0.1% strain
cf. FWHM 0.035°2Θ ~ 0.02% strain @18°2Θ – easy to see
2.5-2.0 Å resolution range – [NaCl] sequence

Obsd & calcd powder patterns

0.25M NaCl

Peak shifts!

1.25M NaCl

.: Use multiple data sets
Stereochemical restraints – additional “data”

\[ M = f_y \sum w_i(Y_{oi} - Y_{ci})^2 \]

Powder profile (Rietveld)

\[ + f_a \sum w_i(a_{oi} - a_{ci})^2 \]

Bond angles

\[ + f_d \sum w_i(d_{oi} - d_{ci})^2 \]

Bond distances

\[ + f_t \sum w_i(-t_{ci})^4 \]

Torsion angle pseudopotentials

\[ + f_p \sum w_i(-p_{ci})^2 \]

Plane RMS displacements

\[ + f_v \sum w_i(v_{oi} - v_{ci})^4 \]

van der Waals distances (if \( v_{oi} < v_{ci} \))

\[ + f_h \sum w_i(h_{oi} - h_{ci})^2 \]

Hydrogen bonds

\[ + f_x \sum w_i(x_{oi} - x_{ci})^2 \]

Chiral volumes

\[ + f_R \sum w_i(-R_{ci})^4 \]

“φ/ψ” pseudopotential

\[ w_i = 1/\sigma^2 \] weighting factor

\[ f_x \] - weight multipliers (typically 0.1-3)
Ramachandran plot - main chain torsion angles $\Phi$ & $\Psi$

- **$\beta$-sheet region**
- **“most favored” regions**
- **$\alpha$-helix region**

**Bad – clashes between AA side chains**
\( \Phi/\psi \) surface restraint - Ramachandran plot

Surface described by 4 terms of

\[
R_c = A_o + \sum_{i=1}^{n} A_i \exp[-B_i (\phi_{oi} - \phi_c)^2 - C_i (\psi_{oi} - \psi_c)^2 - D_i (\phi_{oi} - \phi_c)(\psi_{oi} - \psi_c)]
\]
Constraints: A residue rigid body model for phenylalanine

\[ 3t_{xyz} + 3Q_{ijk} + \psi + \chi_1 + \chi_2 = 9 \text{ variables} \]

vs 33 unconstrained xyz coordinates
Quaternion - Quaternions? What the ......?

- Wikipedia: “In mathematics, quaternions are a non-commutative extension of complex numbers.”
- Most useful in computer games & aircraft controls (especially fighters) also in GSAS-II for structure drawing
- 4D vector; 1 real & 3 imaginary

\[ Q_{ijk} = r + ai + bj + ck \]

(can be reversed with real at end)
- Represent rotations of objects – quicker & easier than Eulerian matrix rotations – no “gimbel lock”
At completion of refinement
check quality of result - PROCHECK

Stereochemistry of resulting protein structure – subject to restraints
PROCHECK – report how well we did
1) Most amino acid $\phi/\psi$ torsion angles within “most preferred region”
2) Also compare distances & angles with “expected”
3) Other “scoring” programs as well
Multidata HEWL powder refinement - Structure quality?

Ramachandran plot – 90% most favored

Total OMIT map – protein & H₂O
HEWL – superposition of 3 determinations ([NaCl],pH5;[NaCl],pH4 & RD) & H$_2$O independently detn.

H$_2$O – many common positions (& some not)

Variations?

RMSD~0.4Å (all protein atoms)

**Protein structure solution 1** Example – new Zn-insulin phase

Grind $T_3R_3$ complex in agate mortar with mother liquor
High resolution synchrotron x-ray powder patterns (X3b1/NSLS)

Immediately after grinding
Additional samples showed transition over a day or two

Indexed – R3
$a=81.275\text{Å}, c=73.024\text{Å}$
New phase – $T_3R_3DC$

Indexed – R3
$a=81.084\text{Å}, c=37.537\text{Å}$
same as single xtal
High Resolution X-ray Powder Diffraction on Proteins

Zn insulin structure determined from powder diffraction data
• R3 unit cell $a=81.276\,\text{Å}, c=73.037\,\text{Å}$
• Indexed from pattern
• $V=418,000\,\text{Å}^3$!!
• $>1600$ atoms!!; $102$ AA
• Rietveld refinement (GSAS)
• $R_{wp}=3.74\%$

1st Molecular replacement solution!!
3 parameter problem
**Schematic of $T_3R_3DC$ Zn-insulin complex.**

**Powder RT structure** PDB=1FUB

**Same structure as --**

**Single crystal – Lo T phase** PDB=1G7A

*(Confirmed by single crystal study!)*

View down 3-fold axis - front $T_3R_3$ turned 9° wrt back $T_3R_3$


More molecular replacement – 2nd SH3 ponsin domain

- 67 residue protein; 554 atoms – microcrystals only from 1st xtalization trials
- High res powder diffraction data & radiation damage - lattice parm shifts
- ESRF – ID31; ~10keV & ~15keV → $d_{\text{min}} \sim 2.3-2.8 \AA$; P2$_{1}$2$_{1}$2$_{1}$;~24x36x72\AA
- Extract $F_0^2$ & solve – molrep; PDB 1W70 (38% homology) & 1OOT (40%) @ 3.7\AA & 36 HOH molecules
- Subsequent single xtal – same structure (0.52\AA & 1.20\AA rmsd)

Margiolaki, et al. (2007) JACS 129, 11865

SH3 2nd ponsin powder total OMIT map details
Ligand binding - HEWL/NAG$_n$ – powder data from complexes

Example: HEWL/NAG$_2$

$\Delta F$ map from HEWL & extracted $F_o$

Refinement – band matrix approximation


X3b1/NSLS in July 2001
HEWL at 100K with a mixture of PEG 400 and methanol

Find PEG 400 in OMIT maps from powder data: right mix – no broadening on cooling

What does GSAS do?

Mostly (in ~125,000 lines of Fortran)
NB: small memory code <1MB
Very widely used & trusted:
6-7000 citations in 25+ years

Thanks to Lynn McCusker for maze
GSAS & EXPGUI interfaces

EXPDTS data setup option (<?>,D,F,K,L,P,R,S,X)
EXPDST data setup options:

<?> - Type this help listing
D - Distance/angle calculation set up
F - Fourier calculation set up
K n - Delete all but the last n history records
L - Least squares refinement set up
P - Powder data preparation
R - Review data in the experiment file
S - Single crystal data preparation
X - Exit from EXPEDT

GSAS – EXPEDT (and everything else) – text based menus with help, macro building, etc. (1980’s user interface!)

EXPGUI: access to GSAS
Typical GUI – edit boxes, buttons, pull downs etc.
Liveplot – powder pattern display
(1990’s user interface Tcl/Tk)
**GSAS & GSAS-II: code models**

1980’s

- PC-GSAS – thin wrapper GUI
- Keyboard interface only
- .EXP file, etc.
- GSAS programs – each is a Fortran exe (common library of routines)

2010’s

- Slow GUI code – wxPython & common project file name.gpx
- Fast code – numpy array routines (& a few fortran routines)
- Python – ideal for this
GSAS-II: modern GUI – 3 frame layout + console

Main menu
Data tree
Submenu
Data tabs
Data window
Graphics window
Drawing tabs

NB: Dialog box windows will appear wanting a response
Advanced visualization in GSAS-II

Powder profile – easy zoom

Contour plot

Waterfall plot

Structure drawing
Advanced visualization in GSAS-II: numbers as pictures

μ-strain surface

Texture – sph. harmonics

v-cov matrix
Why python?

Code snippet – charge flipping all inside a “while” loop

NB: CEhkl is expanded Fhkl over full sphere & zero filled out to 1/resolution limit as an array

```python
CERho = np.real(fft.fftn(fft.fftshift(CEhkl)))*(1.+0j)          #fft Fhkl → ρ(xyz)
CESig = np.std(CERho)                                           #get σ(ρ)
CFrho = np.where(np.real(CERho) >= flipData['k-factor']*CESig,CERho,-CERho)  #CF ρ → ρ’
CFrho = np.where(np.real(CERho) <= flipData['k-Max']*CESig,CFrho,-CFrho)  #U atom CF!
CFhkl = fft.ifftshift(fft.ifftn(CFrho))                        #fft ρ(xyz) → F’(hkl)
CFhkl = np.where(CFhkl,CFhkl,1.0)                              #avoid divide by zero
phase = CFhkl/np.absolute(CFhkl)                              # get Φ(hkl) from F’
CEhkl = np.absolute(Ehkl)*phase                                #apply Φ to F
Ncyc += 1                                                      #count tries
sumCF = np.sum(ma.array(np.absolute(CFhkl),mask=Emask))        #∑ F
DEhkl = np.absolute(np.absolute(Ehkl)/sumE-np.absolute(CFhkl)/sumCF)  #∑DF
Rcf = min(100.,np.sum(ma.array(DEhkl,mask=Emask)*100.))         #R-value for CF
```

This stuff is fast! ~1s/cycle for 500K reflections/map points
New Hessian LSQ - modified Levenberg/Marquardt Algorithm

Steps:
1. Compute \( A_{ij} = \sum w \frac{\partial I_i}{\partial p_i} \frac{\partial I_j}{\partial p_j} \) \text{ SLOW step}
2. Normalize \( A'_{ij} = A_{ij} / \sqrt{A_{ii}A_{jj}} \)
3. Compute \( \chi^2(p) \)
4. Select \( \lambda (=0.001) \)
5. Modify \( A''_{ii} = A'_{ii} (1 + \lambda) \) \text{ FAST steps}
6. Solve for \( \delta p \) (unnormalized!) & compute \( \chi^2(p+\delta p) \)
7. If \( \chi^2(p+\delta p) > \chi^2(p) \) then \( \lambda \times 10 \) go to 5
8. Else apply \( \delta p \) to \( p \) & go to 1 (new cycle)
9. Quit when \( \chi^2(p) - \chi^2(p+\delta p) / \chi^2(p) < 0.0001 \)

NB: all in 36 lines of python; all double precision
NB^2: this thing is exceedingly robust
GSAS-II Future – fill in rest of the maze

- Restraints – done
- Rigid bodies - done
- Charge flipping - done
- Monte Carlo/Simulated Annealing – slow
- CIF Publication – in process
- Other scattering
  - Neutron TOF
  - Small angle
  - Reflectometry

NB: nonatomistic models for these 2

Distribution: via XOR subversion, Google “GSAS-II”; e-mail: >100 users so far
Obtaining GSAS & GSAS-II
Both from APS “subversion” sites – most up-to-date

- GSAS & EXPGUI - [https://subversion.xor.aps.anl.gov/trac(EXPGUI)] (Link)
- GSAS-II - [https://subversion.xor.aps.anl.gov/trac/pyGSAS](https://subversion.xor.aps.anl.gov/trac/pyGSAS)
- GSAS-II is easily found from Google, GSAS is a little harder as GSAS is a common acronym – e.g. Graduate School of Arts & Sciences at various universities, Greater Seattle Aquarium Society, Games and Simulation Arts and Sciences!, Great Sheffield Art Society, etc.
- GSAS-II requires Python – we suggest “EPDFree” from Enthought; use python 2.7; 32 bit or 64 bit version are supported. Protein work will require 64 bit & lots of machine memory
- Afternoon session – just some demonstrations.