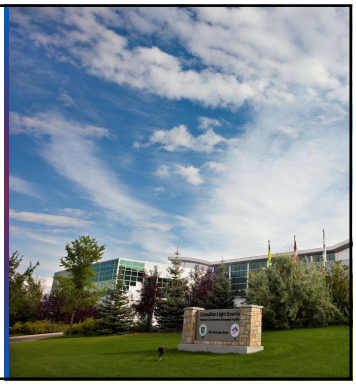
10th Annual
CLS MX Data Collection School
VIRTUAL EDITION

Principles of Data Processing

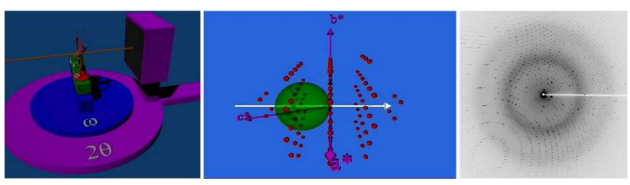
Michel Fodje



Outline

- What is crystallographic data
- Overview of data processing
- Modeling the diffraction experiment
- Data processing steps
- Was processing successful?
- Are the results good enough?
- Data Processing Software:
  - Mosflm, HKL2000, DIALS, XDS

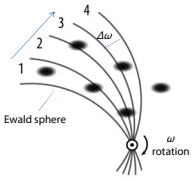
What is crystallographic data?

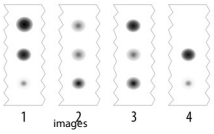


The rotation method

- A set of successive images
- Collected while rotating crystal by small angle ($\Delta\omega$) about fixed axis (ω)
- Diffraction condition is satisfied when reciprocal lattice intersects the Ewald's sphere
- Diffraction pattern recorded on a detector (camera).

What is crystallographic data?





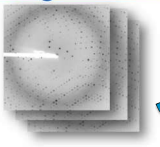
- Crystallographic data contains
  - information about the structure of the crystal in the cell dimensions and indices
  - Information about the contents of each unit cell (the protein/macromolecule) in the intensities.

Canadian Light Source | Centre canadien de rayonnement synchrotron

4 | THE BRIGHTEST LIGHT IN CANADA | lightsource.ca

The Goal of Data Processing

- Extraction intensities and standard deviations of crystallographic reflections from diffraction images
- Steps
  - Model reflection positions
  - Integrate intensities
  - Correct, Scale and Refine Intensities
  - Assign Space-group
- All data reduction packages perform these steps
  - only differ in implementation



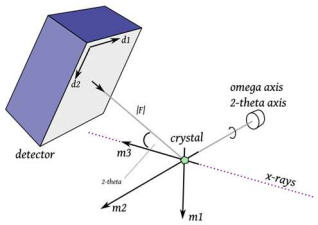
H	K	L	I	SIGMA(I)	XD	YD	ZD
0	0	-4	8.587E+07	1.819E+06	2183.2	2084.7	149.1
0	0	-6	7.899E+05	1.998E+04	2258.6	2101.2	144.4
0	0	-8	4.535E+05	1.627E+04	2334.3	2117.5	139.7
0	0	-10	4.341E+05	1.819E+04	2410.3	2133.4	135.0
0	0	-12	3.582E+05	2.051E+04	2486.6	2149.0	130.2
0	0	14	2.036E+06	4.926E+04	1502.7	1938.0	191.3
0	0	-14	2.126E+06	5.314E+04	2563.5	2164.2	125.3
0	0	16	1.369E+07	2.947E+05	1425.7	1923.2	196.2
0	0	-16	1.301E+07	2.805E+05	2640.9	2178.9	120.5
0	0	18	6.215E+05	2.687E+04	1348.0	1909.1	201.2
0	0	-18	6.160E+05	3.147E+04	2718.8	2192.9	115.5
0	0	20	2.451E+04	2.451E+04	1269.7	1895.6	206.2
0	0	-20	0.871E+03	2.730E+04	2797.1	2206.1	110.6

Canadian Light Source | Centre canadien de rayonnement synchrotron

5 | THE BRIGHTEST LIGHT IN CANADA | lightsource.ca

Modelling the diffraction experiment

- Coordinate systems
  - Detector coordinate system, Goniostat coordinate system, Crystal coordinate system
- Experimental Parameters:
  - Direction of incident beam, detector origin, detector pixel size, photon energy/wavelength, starting ω and Δω angle per oscillation, crystal to detector distance, beam divergence
- Note: different programs use different conventions



Canadian Light Source | Centre canadien de rayonnement synchrotron

11/12/2020

6 | THE BRIGHTEST LIGHT IN CANADA | lightsource.ca

Modelling spot positions - XDS

- Detector coordinate system
  - Right-handed orthonormal $\{\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}\}$
  - Origin closest point between detector and crystal (X_0, Y_0) found at distance |F| from crystal
  - Pixel (X,Y) on detector surface represented by vector

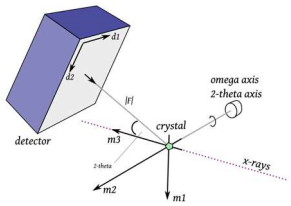
$$(X - X_0)\mathbf{d_1} + (Y - Y_0)\mathbf{d_2} + F\mathbf{d_3}$$

  in lab coordinate system
- Goniostat coordinate system
  - Right-handed orthonormal $\{\mathbf{m_1}, \mathbf{m_2}, \mathbf{m_3}\}$
  - Origin at crystal
  - Determined from incident beam vector S0 and rotation axis m2 such that

$$\mathbf{m_1} = \frac{(\mathbf{m2} \times \mathbf{S0})}{|\mathbf{m2} \times \mathbf{S0}|}$$

  and

$$\mathbf{m_3} = \mathbf{m1} \times \mathbf{m2}$$



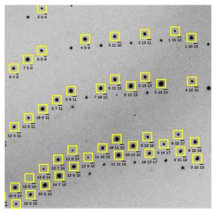
XDS – Modelling spot positions and shape

- Crystal coordinate system
  - Right-handed $\{\mathbf{b_1}, \mathbf{b_2}, \mathbf{b_3}\}$
  - Reciprocal basis $\{\mathbf{b^*_1}, \mathbf{b^*_2}, \mathbf{b^*_3}\}$
  - Represents un-rotated crystal (ie $\boldsymbol{\omega} = 0°$)
  - Any reciprocal-lattice vector can be expressed as

$$\mathbf{p}^*_0 = h\mathbf{b}^*_1 + k\mathbf{b}^*_2 + l\mathbf{b}^*_3$$

  where *h, k, l* are integers
  - Location of all diffraction peaks can be calculated from the parameters
    $\mathbf{S_0}$, $\mathbf{m_2}$, $\mathbf{b_1}$, $\mathbf{b_2}$, $\mathbf{b_3}$, X_0, Y_0, F, $\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$, $\boldsymbol{\omega_0}$ and $\boldsymbol{\Delta\omega}$.
  - Only remaining parameters are spot shape and extent
- Spot shape and extent parameters
  - Gaussian model
  - σ_R: Standard deviation of reflecting range (Mosaicity)
  - σ_B: Standard deviation of beam divergence
  - Spot coordinate system $\{\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}\}$ which transforms all spots to the same shape.

Auto-Indexing



H	K	L	I	SIGMA(I)
0	0	-4	8.587E+07	1.819E+06
0	0	-6	7.899E+05	1.998E+04
0	0	-8	4.535E+05	1.627E+04
0	0	-10	4.341E+05	1.819E+04
0	0	-12	3.582E+05	2.051E+04
0	0	14	2.036E+06	4.926E+04
0	0	-14	2.126E+06	5.314E+04
0	0	16	1.369E+07	2.947E+05
0	0	-16	1.301E+07	2.805E+05
0	0	18	6.215E+05	2.687E+04

Determines the

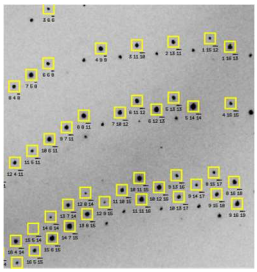
- *Crystal orientation*
- *Approximate cell dimensions*
- *Possible crystal symmetry*

Auto-Indexing

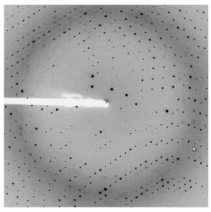
- A list of strong spots occurring in images is found
- A strong spot is one which exceeds the mean by a given number in units of σ, typically 3-5σ.
- Reciprocal basis vectors are found from the spots using various methods:
  - Mosflm, 1-D FFT
  - HKL2000, 3-D FFT
  - XDS, 3D difference vector cluster analysis
- A quality score is calculated
- Parameters are refined using more reflections.

Auto-Indexing - XDS

- The process during which XDS determines the parameters
  - A list of strong spots occurring in images is found
  - A strong spot is one which exceeds the mean by >3-5 σ
  - Reciprocal basis vectors $\{\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*\}$ are found from the spots using local indexing
    - Three independent vectors $\mathbf{b}^*_1, \mathbf{b}^*_2, \mathbf{b}^*_3$ are selected which maximize a quality function
  - Ideally each spot corresponds to a reciprocal-lattice vector
    - $\mathbf{p}_0^* = h\mathbf{b}_1^* + k\mathbf{b}_2^* + l\mathbf{b}_3^*$
    - Unfortunately "alien" spots are often present



Integration



H	K	L	I	SIGMA(I)
0	0	-4	8.587E+07	1.819E+06
0	0	-6	7.899E+05	1.998E+04
0	0	-8	4.535E+05	1.627E+04
0	0	-10	4.341E+05	1.819E+04
0	0	-12	3.582E+05	2.051E+04
0	0	14	2.036E+06	4.926E+04
0	0	-14	2.126E+06	5.314E+04
0	0	16	1.369E+07	2.947E+05
0	0	-16	1.301E+07	2.805E+05
0	0	18	6.215E+05	2.687E+04

Determines

- *The intensities of each reflection, i.e. the total number of photons scattered by the family of planes represented by each index.*
- *And their standard deviations*

Integration

- ***Successful integration is only possible if the model of the experiment is accurate***
  - Requires accurate experimental parameters
  - Requires successful auto-indexing

Integration

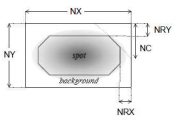
- In each region where a reflection is expected, pixels which belong to a reflection and those which belong to the background are identified
- The intensities of the non-background pixels are added up
  - Not possible if any pixels are overloaded.
- The total intensity of the background pixels is then subtracted
- Using counting statistics, an estimate of the error in the intensity (sigma) is calculated.

Integration

- Two types of integration algorithms:
  - 2D integration: Mosflm, HKL2000
    - Reflections are integrated using a 2D profile for each each image
    - Partial reflections are only added together during scaling
    - Work better for images collected with "thick" slicing.
  - 3D integration: XDS, d*Trek, DIALS
    - Reflection are integrated using a 3D profile which includes multiple adjacent images.
    - Produce better modelling of spot shape and more accurate spot centroids.
    - Works well with "thin"-sliced data which generally produce more accurate reflections.

Integration

- Background determination differences
  - Mosflm/HKL2000
  - XDS/DIALS
    - Strong pixels are removed until remaining pixels represent samples from a random distribution
    - Background is the mean intensity of remaining pixels

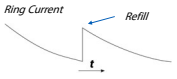


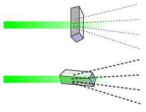
Integration - XDS

- For each frame
  - Identify strong pixels in the image
  - Label each pixel with the indices of nearest reflection
  - 3D profile is determined for all pixels labelled with the same index
  - A reflection is rejected if its centroid deviates too much from expected position
  - Mosaicity (σ_R) and Beam divergence (σ_B) are estimated
    - 3D Gaussian fitting
  - Background is determined
    - Strong pixels are removed until remaining pixels represent samples from a random distribution
    - Background is the mean intensity of remaining pixels
  - Integrated intensity is estimated by profile fitting

Correction & Scaling

- Integrated intensities need to be corrected for
  - Changes in beam intensity and variations of illuminated crystal volume
  - Absorption of incident and diffraction beams
  - Radiation damage
  - Variations in detector sensitivity on its surface
  - If multiple crystals were used
    - Differences in crystal sizes
    - Differences in crystalline order
  - Combined effect:
    - Correlations between reflection and frame number/angle
    - Correlations between reflection and spatial location on detector surface





Scaling

- Assumptions:
  - Statistically independent observations of intensities of symmetry related reflections recorded at different angles, frame numbers (time), spatial location on the detector surface, or different crystals, must be equal
  - If more than one crystal was used, intensities of equivalent reflections from different crystals must be equal.
- Scaling involves adjusting the integrated intensities to satisfy as many of those assumptions as reasonably possible.
- For MAD/MIR experiments relative intensity differences are more important than absolute intensities
  - multiple datasets should be scaled relative to each other.

Space-group assignment

- Identification of the correct space-group should be postponed for as long as possible until all intensities have been integrated.
- The correct space-group can be determined manually by examining systematic absences and unit cell parameters.
- Software can automate the process, e.g. POINTLESS, XDS, DIALS
- Re-integrating the data after space-group assignment may improve the data quality.

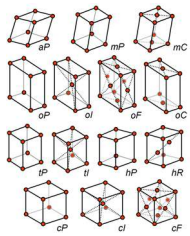


Image credits: © Bernhard Rupp

Was data processing successful?

- Always carefully inspect output files/reports to make sure each step was successful.
- Carefully inspect reflection profiles. Symptoms of problems include:
  - Off-centered profiles, Incorrectly predicted spots, (Mis-indexing, Crystal slippage, or change in incident beam direction)
  - Reflections extending to edge of profile
  - Many unintegrated reflections

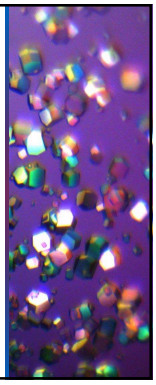
Are the results good enough?

- Important indicators include,
  - completeness,
  - # of outliers, systematic absences if present,
  - I/sigma(I), R-factors, CC½, high resolution limit, anomalous signal, etc.
- R-merge is a poor guide to data quality
  - Does not take multiplicity into account
  - Not a good indicator of the high resolution limit
  - Variants such as R-meas and R-pim are only slightly better
- Better alternatives for checking internal consistency:
  - I/sigma(I) ≥ 2.0
  - Chi-squared ≈ 1
  - Rejecting data may improve the R-values but not the model. *See Acta Cryst. (2013) D69, 1215-1222*

Are the results good enough?

- Pearson correlation coefficient (CC) is a better a better statistic.
- CC ≥ 0.3 for anomalous signal is a good high resolution cut-off for phasing
- CC½, (related to CC) is a good statistic for assessing the information content in the reflections.
  - For CC½ around 0.2-0.4, I/sigma(I) will be around 0.5-1.5 and there is little information left in the reflections.

Data Processing Software



XDS – X-ray Data Streamer

- Software for processing single-crystal monochromatic diffraction data recorded by the rotation method.
- XDS can process small-molecule and macromolecular diffraction data from a wide variety of image formats
- Developed by Wolfgang Kabsch at Max Planck Institute : http://xds.mpimf-heidelberg.mpg.de/
- Free for academic and non-commercial use
- Primarily command line
  - Third Party GUI's available
- References:
  - *J. Appl. Cryst.* (1988). 21, 67-72.
  - *J. Appl. Cryst.* (1988). 21, 916-924.
  - *J. Appl. Cryst.* (1993). 26, 795-800.
  - *Acta Cryst.* (2010). **D66**, 133-144.
  - *Acta Cryst.* (2010). **D66**, 125-132.

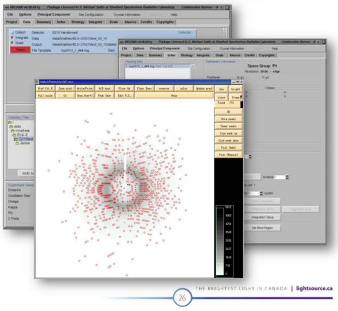
Canadian Light Source Centre canadien de rayonnement synchrotron

25 THE BRIGHTEST LIGHT IN CANADA | lightsource.ca

HKL2000

- Software from HKL Research Inc
- GUI
  - Command line programs (denzo, scalepack)
- License required to use
- Not available at CMCF

https://www.hkl-xray.com/hkl-2000



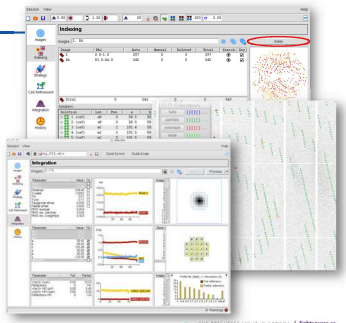
Canadian Light Source Centre canadien de rayonnement synchrotron

26 THE BRIGHTEST LIGHT IN CANADA | lightsource.ca

iMOSFLM

- GUI for MOSFLM
- Included with CCP4
  - Can be installed separately
- Available for Windows

https://www.mrc-lmb.cam.ac.uk/harry/imosflm/



Canadian Light Source Centre canadien de rayonnement synchrotron

27 THE BRIGHTEST LIGHT IN CANADA | lightsource.ca

DIALS - Diffraction Integration for Advanced Light Sources

- Open Source Collaborative initiative
- Newest Data Processing Package
  - Uses knowledge acquired from history of crystallographic data processing
  - Written in Python and C++
- Command line interface
  - GUI available in CCP4

https://dials.github.io/