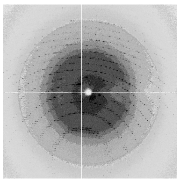
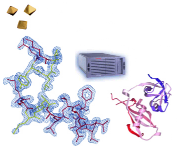
Brookhaven National Laboratory, 22 May 2008
Workshop on Raw Image Formats in Structural Biology

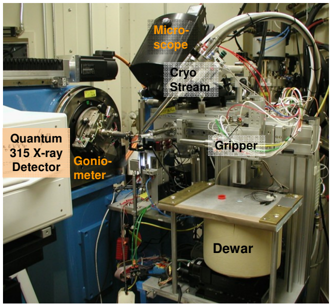
The Importance of Standard Image Formats for Scientific Progress

Nicholas Sauter
Lawrence Berkeley National Laboratory





Sector 5 ALS Automounter



- **ALS-style puck: 112 Crystal Samples**
- **Beamline Operating System (BOS) control**
- **Liquid Nitrogen Autofill**

High Throughput Screening:
- Screen for best crystal growth conditions
- Select the highest-quality samples from a batch
- Discovery of drug leads and protein-ligand complexes
- Enable multi-crystal dataset acquisition
- Perform initial characterization with minimal radiation dose

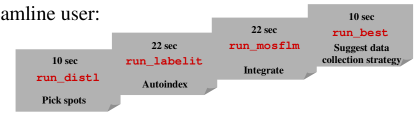
Real-time Analysis:
- Autoindex
- Measure the model fit (rmsd)
- Limiting resolution
- Mosaicity
- Ice rings & other artifacts

The challenge is to perform this analysis reliably in an automated setting!

Web-Ice: Images and screening results can be viewed both locally & over the Web

González (2008) **J Appl Cryst** 41:176

Basic idea: Command-line scripts are automatically run behalf of the beamline user:

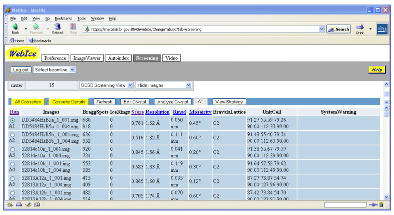


10 sec
`run_distl`
Pick spots

22 sec
`run_labelit`
Autoindex

22 sec
`run_mosflm`
Integrate

10 sec
`run_best`
Suggest data collection strategy


WebIce - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
Search
Print
Home Bookmarks
WebIce
Preference ImageViewer Autoindex Screening Video
Log out Select beamline
Help
All Cassettes Cassette Details Refresh Edit Crystal Analyze Crystal A1 View Strategy

https://sharptail.lbl.gov:8543/webice/ChangeTab.do?tab=screening

sauter 15 BCSB Screening View Hide Images

Run	Images	BraggSpots	IceRings	Score	Resolution	Rmsd	Mosaicity	BravaisLattice	UnitCell	SystemWarning
● A1	DD5484HsB5a_1_001.img DD5484HsB5a_1_004.img	680 918	0 0	0.761	1.62 Å	0.060 mm	0.45°	C2	91.27 55.59 79.26 90.00 112.33 90.00	
○ A2	DD5484HsB5b_1_001.img DD5484HsB5b_1_004.img	626 552	0 0	0.516	1.82 Å	0.111 mm	0.60°	C2	91.48 55.40 79.31 90.00 112.63 90.00	
○ A3	52814e10a_1_001.img 52814e10a_1_004.img	920 724	0 0	0.845	1.56 Å	0.041 mm	0.20°	C2	91.38 55.67 79.39 90.00 112.39 90.00	
○ A4	52814e10b_1_001.img 52814e10b_1_004.img	553 385	0 0	0.683	1.83 Å	0.119 mm	0.30°	C2	91.64 57.52 79.62 90.00 112.49 90.00	
○ A5	52813A12a_1_001.img 52813A12a_1_004.img	415 409	0 0	0.865	1.60 Å	0.035 mm	0.12°	C2	87.27 73.87 54.74 90.00 127.96 90.00	
○ A6	52813A12b_1_001.img 52813A12b_1_004.img	482 514	0 0	0.705	1.74 Å	0.070 mm	0.60°	C2	87.42 73.84 54.70 90.00 127.91 90.00	

- *Crystal screening table is visible on a Web page and downloadable as an Excel spreadsheet*
- *Web-based image viewer*
- *The same tabulated results are visible within the beamline GUI:*
  Blu-Ice at SSRL or BOS at ALS
- *System is under evaluation at APS (GMCA-CAT, IMCA-CAT, NE-CAT)*

Review of Last Year's Presentation: Increasing the Reliability of Automated Image Processing

DISTL Zhang *et al.*(2006) **J Appl Cryst** 39:112

LABELIT Sauter *et al.*(2004) **J Appl Cryst** 37:399

• Macromolecular diffraction patterns are very diverse. Basic well-known algorithms (*e.g.,* cell reduction & autoindexing) had to be rewritten to cover outlying cases. Legacy software (pre-2003) relied heavily on human input to recognize the challenging cases.
• Writing a new set of programs also involved extending support to all of the detector formats that the software users requested.

Thanks for providing detector information
Chris Nielson, ADSC
Michael Blum, Mar USA
Jim Pflugrath, Rigaku
Miroslav Kobas, Dectris

Detector Vendors
- Successful support
  - ADSC Quantum 4, 210, 315
  - Mar CCD
  - Mar Image Plate
  - Rigaku Raxis IV and HTC
  - Rigaku Raxis II (transformation rectangular pixels to square)
  - Rigaku Saturn 92 CCD
  - MacScience DIP 2030b
  - Pilatus-6M
- Very limited success
  - Bruker Proteus CCD (1K x 1K) proprietary spatial calibration
  - APS SBC 19BM / 19ID requires calibration file

Review of Last Year's Presentation: Increasing the Reliability of Automated Image Processing

DISTL Zhang *et al.*(2006) **J Appl Cryst** 39:112

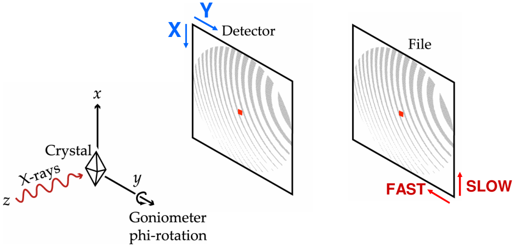
LABELIT Sauter *et al.*(2004) **J Appl Cryst** 37:399

• Macromolecular diffraction patterns are very diverse. Basic well-known algorithms (*e.g.,* cell reduction & autoindexing) had to be rewritten to cover outlying cases. Legacy software (pre-2003) relied heavily on human input to recognize the challenging cases.
• Writing a new set of programs also involved extending support to all of the detector formats that the software users requested.

Detector Vendors
- Successful support
  - ADSC Quantum 4, 210, 315
  - Mar CCD
  - Mar Image Plate
  - Rigaku Raxis IV and HTC
  - Rigaku Raxis II (transformat
  - Rigaku Saturn 92 CCD
  - MacScience DIP 2030b
  - Pilatus-6M
- Very limited success
  - Bruker Proteus CCD (1K x 1K) proprietary spatial calibration
  - APS SBC 19BM / 19ID requires calibration file

Gerd Rosenbaum: "Does the CBF standard support a data entry that records the incident intensity at various time points during the angular rotation (e.g. at 0.2 second intervals)? This would allow us to apply different scaling factors for reflections on the same image."

Inter-Related Coordinate Systems



• Without a standard file format, the coordinate relationships must be worked out individually for each detector type.

Other Difficulties with Diverse File Formats

- Local keyword dialects. The openness of the ADSC file format has allowed different facilities to utilize conflicting keywords.
  - Berkeley Center uses conflicting “DENZO_BEAM_CENTER” and “BEAM_CENTER” tags
- Coordinate system relationships are unspecified
  - There are 8 possible relationships between Detector and File coordinate systems. For ADSC detectors, two of them are in common use at different synchrotrons. LABELIT needs to maintain a list, keyed by DETECTOR serial number.
- Unit of measure is unspecified
  - ESRF writes MAR CCD beam center in mm instead of pixel units
- Redundant information
  - Start phi, end phi, and delta phi all defined.

New Results: Support for CBF in LABELIT

- LABELIT is built on top of a core library of C++ crystallography algorithms, the “Computational Crystallography Toolbox” or **`cctbx`**. Open source: http://cctbx.sf.net
- An adaptor module, **`cbflib_adaptbx`** (CBF library adaptor toolbox) has been added to **`cctbx`**. CBFlib can now be compiled in as an optional dependency.**
- The complexities of CBF function are encapsulated within wrapper C++ classes. We expose only the limited set of features that will actually be used for file reading and data processing, although this could be extended at any time. Currently 800 lines.
- Memory management is handled by constructors and destructors.
- The C++ wrapper classes are exposed at the Python scripting level with Boost.Python bindings.
- Error-handling macros are redefined so that C++ exceptions will be thrown and handled by the user code; this is propagated up to Python.
- Use Python scripts to rapidly prototype new approaches for data processing.

A Python Example

```
> from iotbx.detectors import ImageFactory
> C = ImageFactory("./MB_LP_1_001.CBF")
> C.show_header()
File: ./MB_LP_1_001.CBF
Number of pixels: slow=3072 fast=3072
Pixel size: 0.102588 mm
Saturation: 65000
Detector distance: 200.00 mm
Detector 2theta swing: 0.00 deg.
Rotation start: 85.00 deg.
Rotation width: 1.00 deg.
Beam center x=157.52 mm  y=157.52 mm
Wavelength: 0.979381 Ang.
```

• *LABELIT* can index the example data provided by Chris Nielson & *MOSFLM* integration works.
• But: I have yet to see a single CBF dataset that is either a complete dataset (for structure solution), or that is collected by a scientific user.

miniCBF support: The Pilatus-6M

- What file formats does the Pilatus detector produce? And why?
- miniCBF format is not CBF.
- CBF examples include convert_miniCBF.c ➔ does not quite work for the example cubic insulin dataset provided on the Pilatus web site.
- Header information is efficiently parsed by a 100-line python script.

```
###CBF: VERSION 1.5, CBFlib v0.7.8 - SLS/DECTRIS PILATUS detectors
data_run2_1_00001.cbf
_array_data.header_convention "SLS_1.0"
_array_data.header_contents
;
# Detector: PILATUS 6M, 60-0001, X06SA@SLS
# 2007/Oct/22 15:31:17.480
# Pixel_size 172e-6 m x 172e-6 m
# Silicon sensor, thickness 0.000320 m
# Exposure_time 0.395000 s
# Exposure_period 0.400000 s
# Tau = 124.7e-09 s
# Count_cutoff 1583782 counts
# Threshold_setting 6000 eV
# N_excluded_pixels = 949
# Excluded_pixels: BadPixelMap.tif
# Flat_field: (nil)
# Trim_directory: (nil)
# Wavelength 1.0332 A
# Energy_range (0, 0) eV
# Detector_distance 0.30000 m
# Detector_Voffset 0.00000 m
# Beam_xy (1230.10, 1310.50) pixels
# Flux 39.516 ph/s
# Filter_transmission 0.3560
# Start_angle 45.0000 deg.
# Angle_increment 0.2000 deg.
# Detector_2theta 0.0000 deg.
# Polarization 0.990
# Alpha 0.0000 deg.
# Kappa 0.0000 deg.
# Phi 0.0000 deg.
# Chi 0.0000 deg.
# Oscillation_axis  X, CW
# N_oscillations 1
;
_array_data.data
;
--CIF-BINARY-FORMAT-SECTION--
Content-Type: application/octet-stream;
    conversions="x-CBF_BYTE_OFFSET"
Content-Transfer-Encoding: BINARY
X-Binary-Size: 6225323
X-Binary-ID: 1
X-Binary-Element-Type: "signed 32-bit integer"
X-Binary-Element-Byte-Order: LITTLE_ENDIAN
Content-MD5: yga+qLsZ7ENzKf05endpKw==
X-Binary-Number-of-Elements: 6224001
X-Binary-Size-Fastest-Dimension: 2463
X-Binary-Size-Second-Dimension: 2527
```

miniCBF support: The Pilatus-6M

- The binary data is read using the crystallographic binary API

```
private_file = std::fopen(filename.c_str(),"rb");
cbf_read_widefile (cbf_h, private_file, MSG_DIGEST);

/* Find the binary data */
cbf_find_tag (cbf_h, "_array_data.data");
cbf_rewind_row (cbf_h);

/* Read data array parameters */
cbf_get_integerarrayparameters_wdims (
     cbf_h, &compression, &binary_id, &elsize, &elsigned, &elunsigned,
     &elements, &minelement, &maxelement,(const char **) &byteorder,
     &dim1, &dim2, &dim3, &padding);

ASSERT(elements == slow*fast); ASSERT(elsize == sizeof(int)); ASSERT(elsigned==1);

/* Read the binary data */
cbf_get_integerarray (cbf_h, //cbf handle
&id,                         //ptr to binary section identifier
begin,                       //array ptr
sizeof (int),                //element size
1,                           //flag of signed data type
sz,                          //elements requested
&nelem_read                  //elements read
                    );
```

Other issues with the Pilatus & its file format

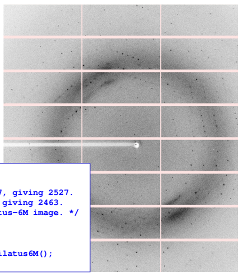
• The data are not spatially corrected. Dectris says the spot centroid σ is 0.3mm

```
                     Cubic Insulin Results:

                        Indexing Integration
Solution  Metric fit    rmsd         rmsd  crystal_system    unit_cell
:)  22      0.4819 dg   0.672       0.153        cubic cI   76.9  76.9  76.9
:)   1      0.0000 dg   0.755       0.158   triclinic aP    66.3  66.5  66.6
```

• The image has inactive pixels.
-1 = dead area
-2 = invalid pixel

• Pilatus-6M has 60 rectangular pixel arrays. Derive a class for identifying inactive area, which is then accessed in run time by a virtual function call:



```
class ActiveAreaPilatus6M: public ActiveAreaDefault {
  virtual bool is_active_area(int x,int y){
    /* x vertical: 12 blocks of 195, separated by 11 stripes of 17, giving 2527.
       y horizont:  5 blocks of 487, separated by 4 stripes of 7, giving 2463.
       Takes 0.02 seconds extra to apply this test to entire Pilatus-6M image. */
    return ( (x%212<195) && (y%494<487) ); }
};
...
if (vendortype=="Pilatus-6M") detector_location = new ActiveAreaPilatus6M();
if (!detector_location->is_active_area(x,y)) { /* make pink */ }
```

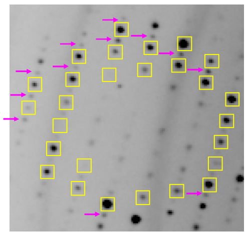
Advantages of Image Standardization

- The idea of permanently archiving raw data has recently regained currency [Baker, Dauter, Guss & Einspahr (2008). *Acta Cryst.* D**64**, 337].
- Ashley Deacon of the JCSG [Joint Center for Structural Genomics] has made over 100 datasets available (out of a total of ~600 PDB structures).
- The desire to make these datasets available for outside analysis also provides a rationale for requiring spatial correction information and beamline conventions to be recorded with the original data file.
- Image formatting questions naturally enter into this discussion.
- The JCSG experience provides a window into the types of discoveries possible when revisiting archived datasets.

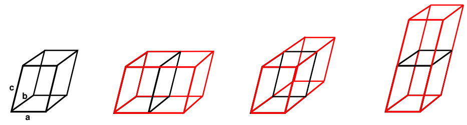
Findings:
- Sublattices
- Non-merohdral twinning
- Spot shape indicating phase transition

Pseudocentering: systematically weak Bragg spots

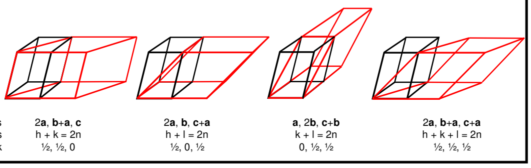
- The true symmetry is P2_1 with two protein molecules per asymmetric unit, related by a non-crystallographic translation.
- The NCS translation is ½ the cell length, approximating an additional symmetry operator, giving rise to alternating weak spots (Hauptman & Karle, 1953).
- If weak spots are ignored, the symmetry is *C*-centered orthorhombic with one protein molecule per asymmetric unit.
- Automatic indexing relies on picking the brightest spots, so it is easy to pick the oC cell by chance. [*Disclaimer: I don't know any example of this type of misindexing in published JCSG work.*]
- Lowering the spot-picking threshhold to find the weak spots is counterproductive.



Construction of the Sublattice: Cell Doubling

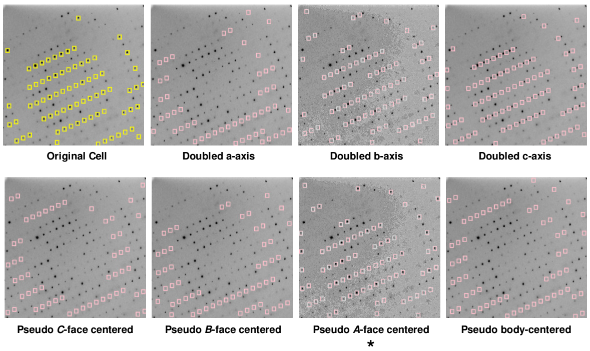


Basis vectors	**a**, **b**, **c**	2**a**, **b**, **c**	**a**, 2**b**, **c**	**a**, **b**, 2**c**
Strong reflections	hkl	h = 2n	k = 2n	l = 2n
Patterson peak	0, 0, 0	½, 0, 0	0, ½, 0	0, 0, ½



Basis vectors	2**a**, **b**+**a**, **c**	2**a**, **b**, **c**+**a**	**a**, 2**b**, **c**+**b**	2**a**, **b**+**a**, **c**+**a**
Strong reflections	h + k = 2n	h + l = 2n	k + l = 2n	h + k + l = 2n
Patterson peak	½, ½, 0	½, 0, ½	0, ½, ½	½, ½, ½

Evidence for Cell Doubling in the Raw Data



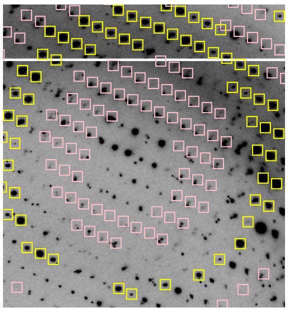
Original Cell | Doubled a-axis | Doubled b-axis | Doubled c-axis

Pseudo *C*-face centered | Pseudo *B*-face centered | Pseudo *A*-face centered * | Pseudo body-centered

Filtering out the decoy signals

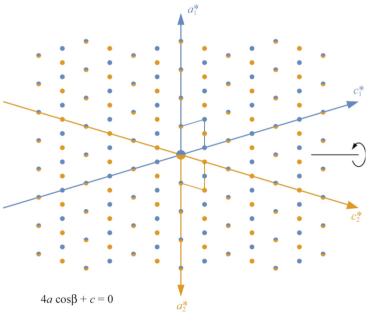
Weak spots do not match the profile of the main lattice

- Mismatched positions
- Split spots



Non-merohedral twinning

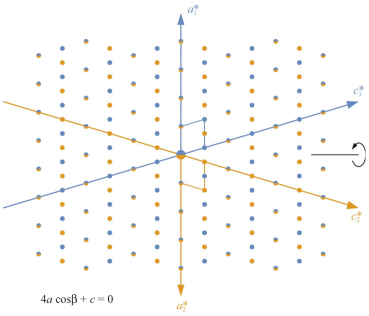
- Non-merohedral twin laws involve a symmetry operation belonging to a higher symmetry supercell, not to the point group of the diffraction pattern.



- In this example from Dauter (2003), twin domains (blue and orange) are related by the black two-fold rotation.
- A coincidental relationship in the unit cell measurements causes half of the blue spots to overlap with half of the orange spots.
- The apparent symmetry of the ensemble is higher than the intrinsic symmetry of each twin domain; but there is an unusual patten of absences—missing every other reflection on alternate layers.
- This is difficult to index; but having done so, easy to detwin.
- It is possible for the two domains to have unequal twin fractions, leading to a pattern of strong and weak reflections.

Non-merohedral twinning

• Non-merohedral twin laws involve a symmetry operation belonging to a higher symmetry supercell, not to the point group of the diffraction pattern.



• If the “blue” twin domain is predominant, the pattern can be correctly indexed and the structure correctly solved. However, the R factor will be degraded.
• The key to fixing it is to observe the weak “orange” reflections in the original data.
• An initial survey finds at least two such cases (of 100) in the JCSG database, so this phenomenon is fairly common.

In Summary

- There is still work to be done to demonstrate that real CBF-formatted data can be processed with *LABELIT* as part of an automated pipeline.
- There is information in the raw dataset that is not captured in the processed structure factor file deposited with the Protein Data Bank. Follow up analysis could potentially lead to re-refinement and an improved understanding of particular structures.
- The inclusion of spatial correction data with the CBF-formatted file is potentially critical for follow up analysis.

Brookhaven National Lab
Workshop on Raw Image Formats in Structural Biology
22 May 2008

Acknowledgements

Nicholas Sauter *Billy Poon* *Ralf Grosse-Kunstleve*	Computational Crystallography Initiative at Lawrence Berkeley National Lab
Paul Adams *Peter Zwart* *John Taylor* *Yun Zhou*	Berkeley Center for Structural Biology at the Advanced Light Source
Ana González *Mike Soltis* *Penjit (Boom) Moorhead* *Jinhu Song* *Ken Sharp* *Scott McPhillips*	Stanford Synchrotron Radiation Lab