XDS package and related tools for X-ray data processing

Keitaro Yamashita
RIKEN SPring-8 Center

2014-05-26 ACA
Blackboard Sessions: Data Processing With The Pros

SPring-8 campus

Contents

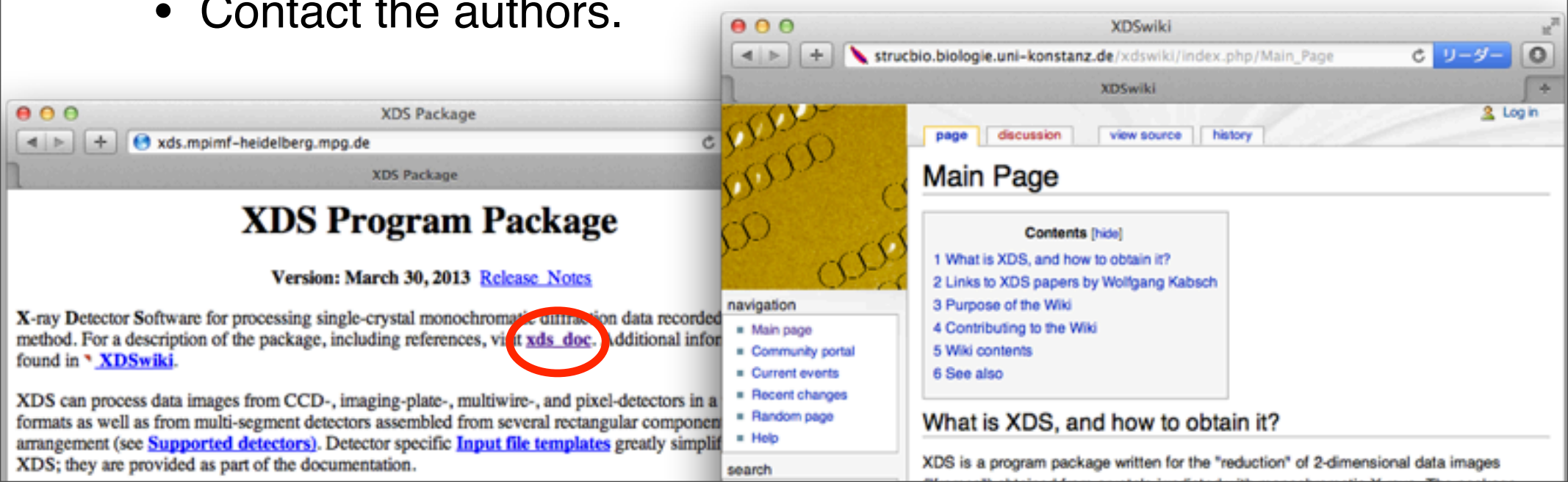
- Introduction of *XDS*
  - Features, algorithms, source of information, and concepts
- Data processing example
  - Thaumatin dataset by ‘shutterless’ data collection

XDS

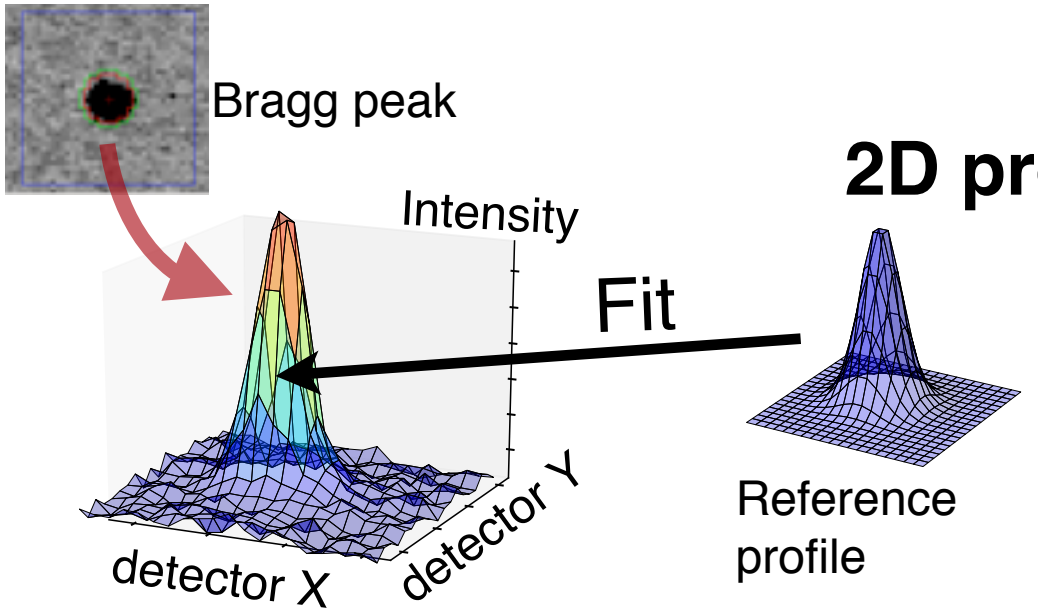
- Developed by Wolfgang Kabsch (MPI Heidelberg) in 1986.
- Kay Diederichs (University of Konstanz) joined in 2007.
- I (Keitaro Yamashita) am just a heavy user.
- Updates every 6 months.
- Closed source, but well documented.
- Integration by 3D profile fitting.
- Parallel processing with OpenMP or multiprocess.
- Recently, official GUI has been developed (XDSGUI).

Source of information

- Official website http://xds.mpimf-heidelberg.mpg.de/
  - download, documentation
- XDSwiki http://strucbio.biologie.uni-konstanz.de/xdswiki/
  - tips and tricks
  - tools (generate_XDS.INP, XDSGUI, etc.)
- CCP4BB
- Contact the authors.

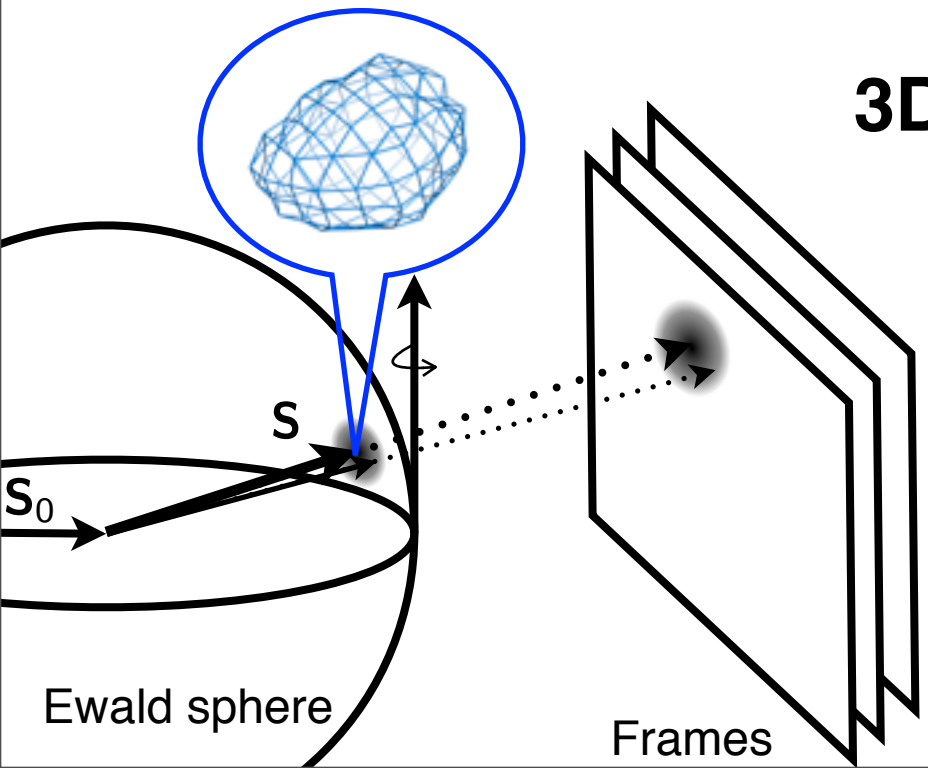


Profile fitting: 2D vs 3D



2D profile (*MOSFLM*, *DENZO*)

- Construct reflection profile on detector surface (2D)
- Each partial reflection is integrated and summed during scaling.



3D profile (*XDS*, *d*TREK*)

- Consider angular direction as well as detector surface (1+2D)
- Partial reflections are summed during integration.

How to use *XDS*?

- Prepare input file manually or by script and run xds
- Use GUI
  - XDSi, iXDS, XDSAPP, etc. (third-party)
  - XDSGUI (official)
- Use pipelines
  - *xia2* (CCP4)
  - *autoPROC* (GlobalPhasing Ltd.)
  - *xdsme* (Soleil)
  - *autoxds* (SSRL)
  - *autoprocess* (CMCF)
  - ..and so on.

skipped in the talk

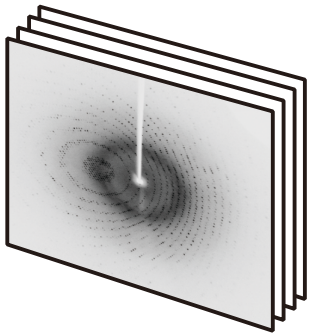
XDS package and related tools

- *XDS*
  the main program (indexing, integrating, scaling of single dataset).
- *XSCALE*
  scaling multiple xds outputs, zero-dose extrapolation, statistics with arbitrary resolution binning.
- *XDSCONV*
  converting xds output to other programs' format.

- *XDSSTAT*
  additional statistics based on scaling result of xds (XDS_ASCII.HKL)
- *XDSGUI*
  Graphical User Interface for XDS. Plots during/after processing.

Data processing with *XDS* package

Single dataset



XDS →

XDS_ASCII.
HKL

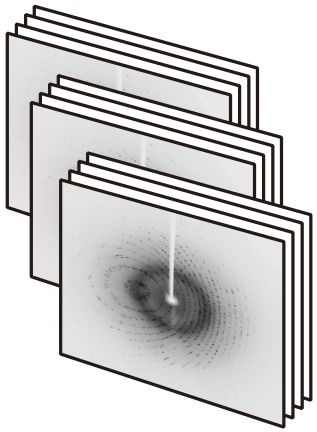
(Scaled and unmerged intensity)

XDSCONV →

MTZ file

or other format (CNS, SHELX)

Multiple datasets (Multi-wavelengths, Multiple crystals)



XDS →

XDS_ASCII.
XDS_ASCII.
XDS_ASCII.
HKL

XSCALE →

scaled.hkl

XDSCONV →

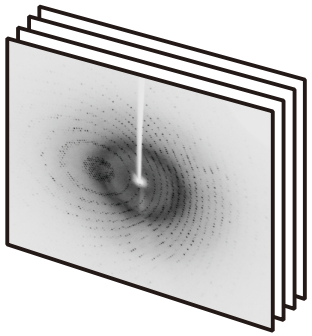
or

scaled.hkl

XDSCONV →

skipped in the talk

Possible routes of *XDS* + *Aimless*



XDS

(INTEGRATE)

INTEGRATE.
HKL

XDS

(CORRECT)

XDS_ASCII.
HKL

Pointless

CORRECTION Options:

- only MODULATION
- No correction (only correct polarization and absorption by air and detector sensor)

Pointless

(Polarization correction is applied)

MTZ file

(multi-batch)

Aimless
(& *ctruncate*)

MTZ file

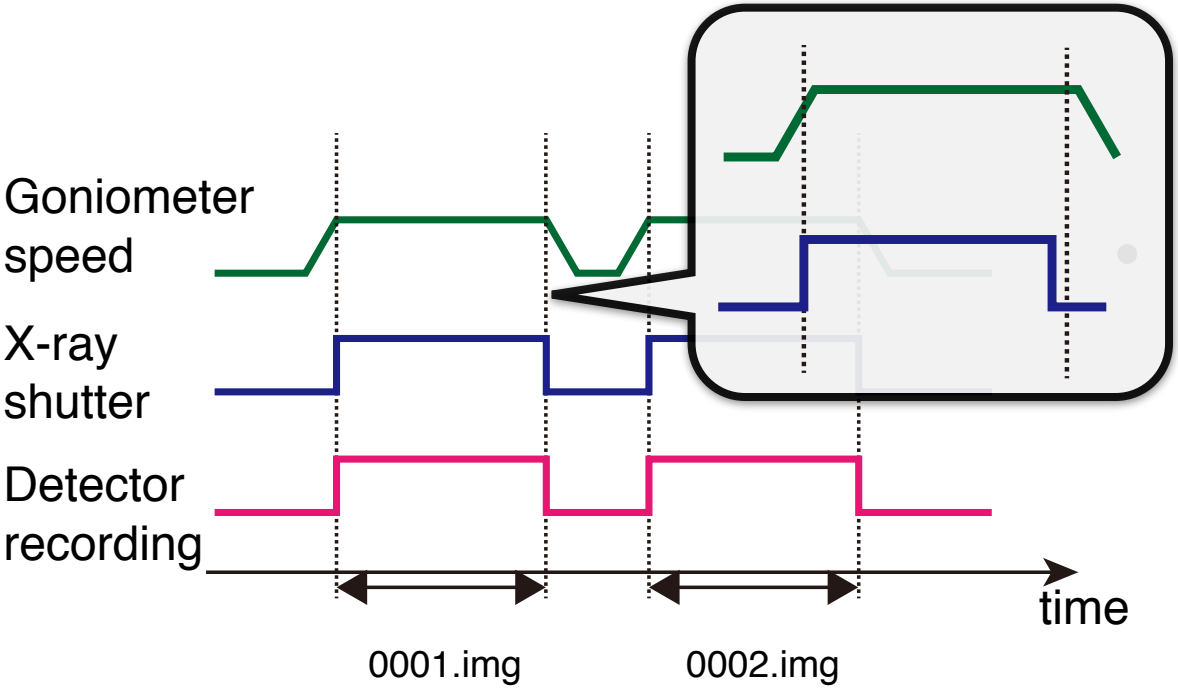
(Scaled)

Data processing example

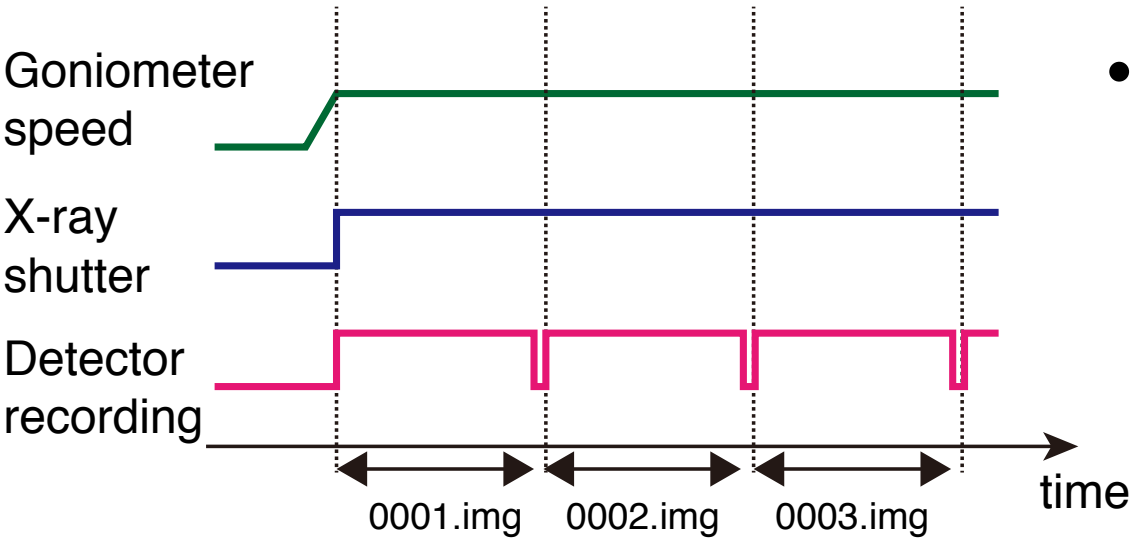
Thaumatin ‘shutterless’ dataset

- Kindly provided by Dr. K. Rajashankar (NE-CAT 24)
- Beamline: Advanced Photon Source 24-ID-C
- Detector: PILATUS 6M
- Oscillation ($\Delta\phi$): 0.2°/frame, continuous rotation
- Exposure: 0.2 s/frame
- Number of frames: 1,800 (360°)
- Wavelength: 1.2398 Å (10 keV)
- Camera distance: 250 mm

What is ‘shutterless’?



- With-shutter data collection
  - Slow.
  - Goniometer/X-ray shutter synchronization error and shutter jitter may exist.



- Shutterless data collection
  - Fast.
  - No mechanical error.
  - Requires detector with ignorable readout time.

skipped in the talk

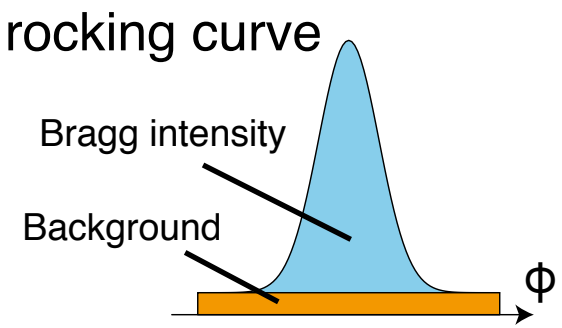
Fast detector enabled it!

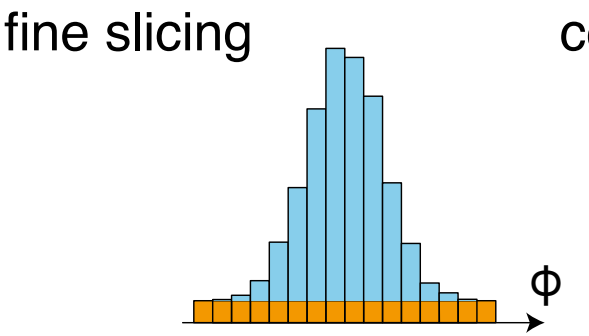
PAD (pixel array detector) is one of the fast detectors.

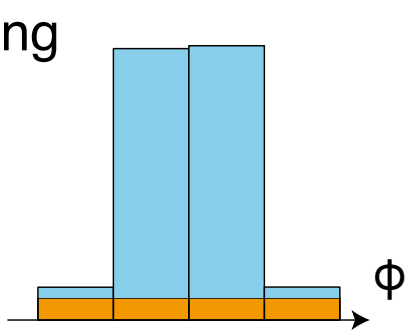
- Fast readout (PILATUS 6M: 3 ms)
- Free from intrinsic/readout noise

Data collection with PAD

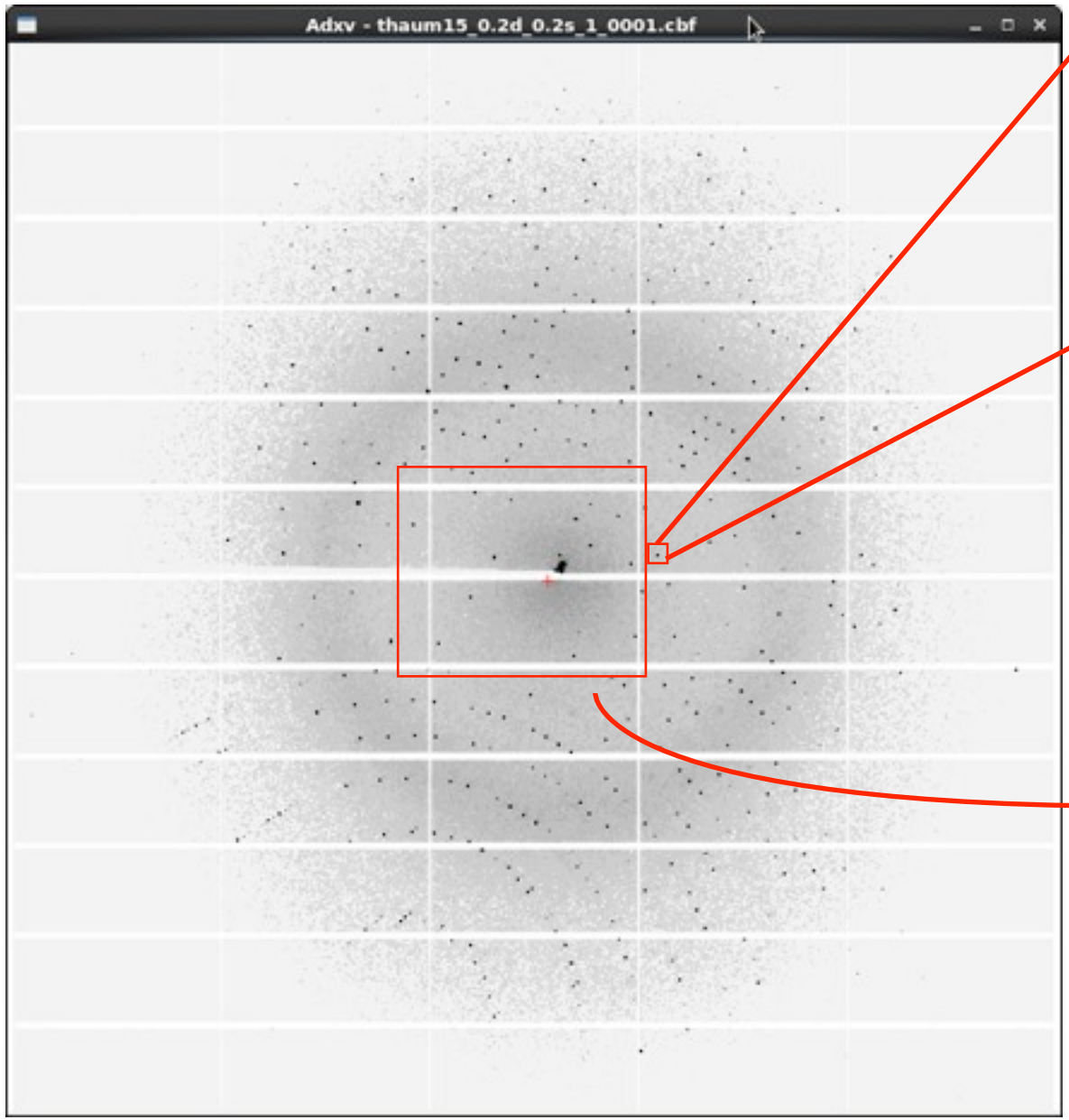
- Ideal for fine ϕ-slicing with shutterless continuous rotation
  - Better sampling of 3D reflection profile
  - Smaller background and better spot separation
  - Optimal Δϕ is mosaicity/2 (Mueller, 2012)
- For the same S/N, one can expose less; therefore less radiation damage, higher multiplicity



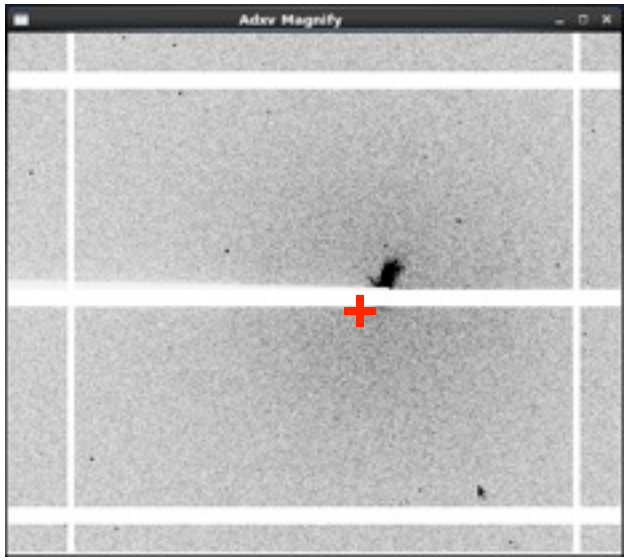




First of all - inspect frames



Tiny spots
(some are only one pixel)



Possibly wrong beam center coordinate

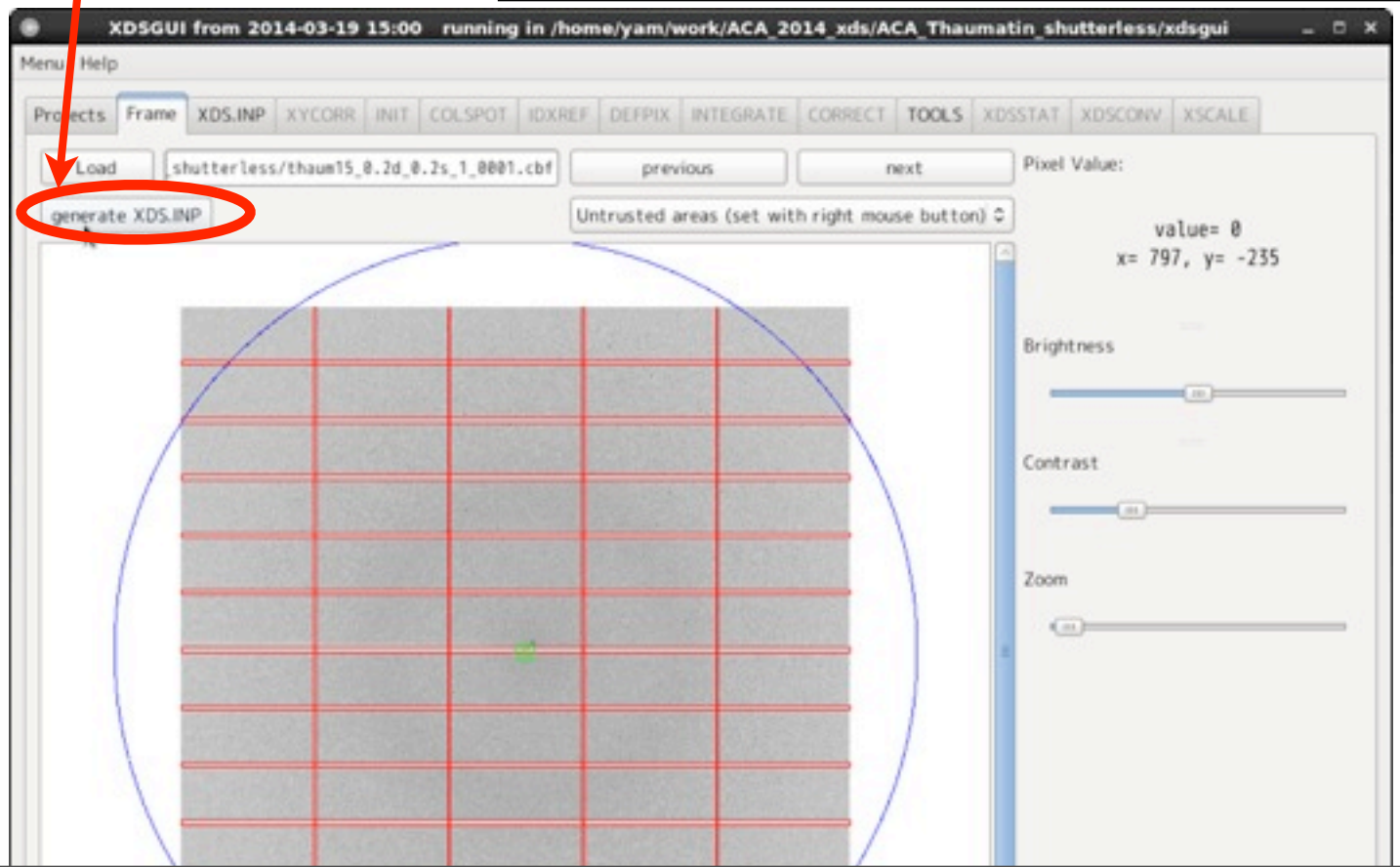
Let’s begin with XDSGUI

Click this button to run generate_XDS.INP script.

Then header of frame is read and
ready-to-use XDS.INP is given in next “XDS.INP” tab.

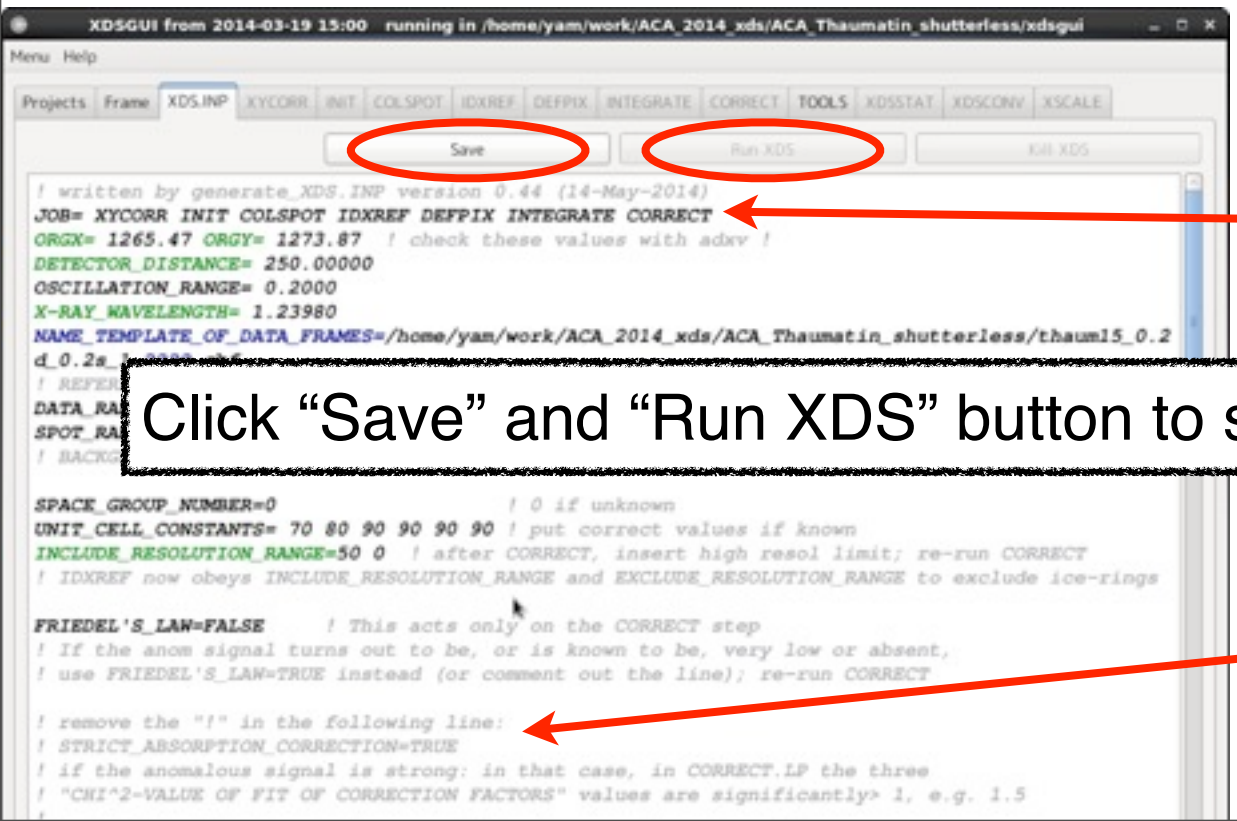
In command line, one should do:

```
$ generate_XDS.INP "/data/frms/mydata_1_???.img"
```



XDS.INP

- the input file for XDS run (xds & xds_par).
- As XDS does not see most image header, all needed information for data processing should be written in XDS.INP.
- Different name is not allowed (XDS automatically reads XDS.INP in the working directory).



JOB= card controls which job *XDS* will do.

Click “Save” and “Run XDS” button to start!

After ‘!’ is ignored (comment)

XDS steps (JOB)

XYCORR	Correction of detector surface distortion and parallax
INIT	Initial background estimation
COLSPOT	Peak search for indexing
IDXREF	Indexing (determine unit cell dimension and orientation)
DEFPIX	Determination of the detector region for integration
INTEGRATE	Integration by 3D profile fitting Polarization correction (Assuming unpolarized beam)
CORRECT	Empirical corrections (DECAY, MODULATION, ABSORP) Non-empirical corrections (polarization, air/sensor absorption) Rejection, Merging statistics.

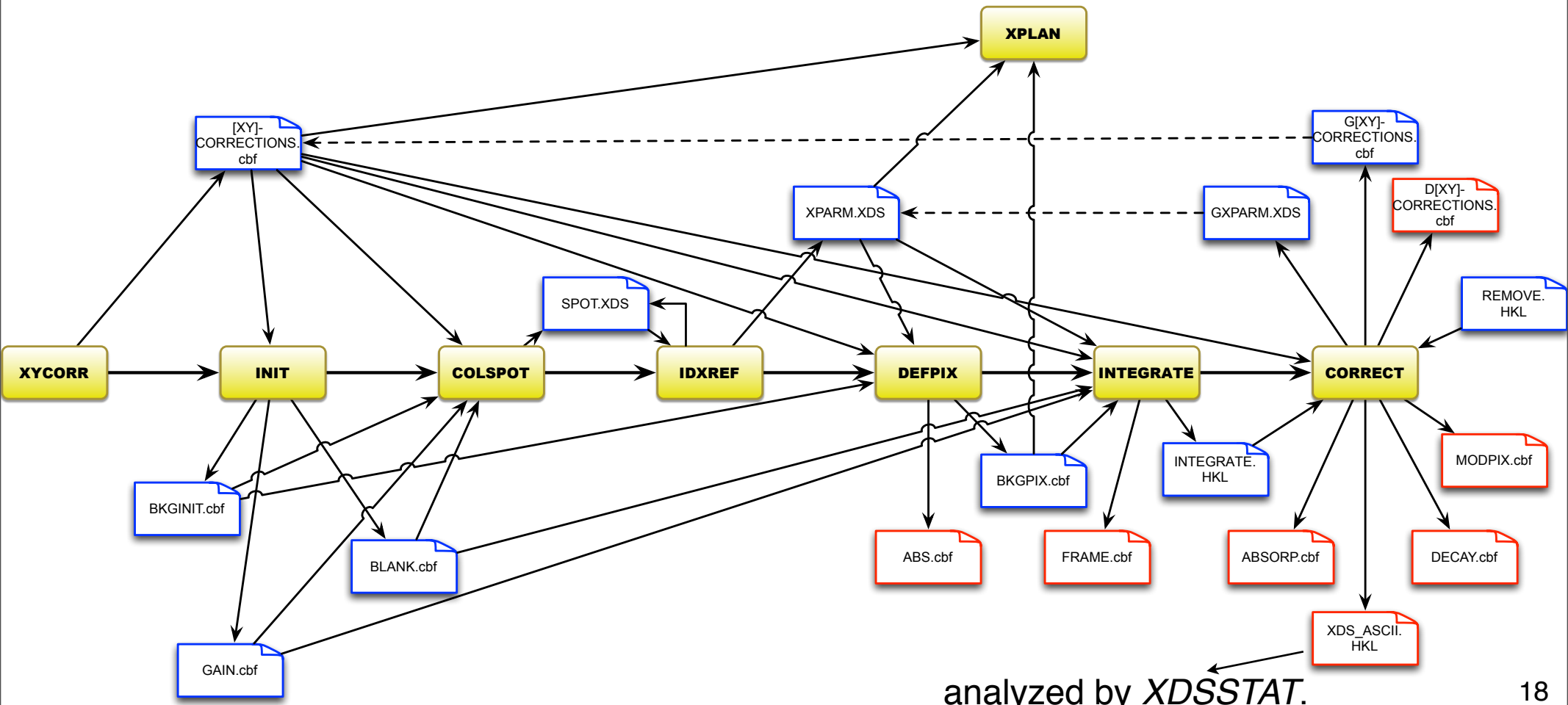
XDS JOBs and related files

JOB name

File needed by each JOB.

File supplied for inspection by user.

Job.LP (logfile) is not shown here, but written for every job.



analyzed by *XDSSTAT*.

skipped in the talk

‘Checkpoints’ in data processing

COLSPOT	Spot positions? (FRAME.cbf)
IDXREF	How many lattices? rmsd between observed and calculated positions? If not correctly indexed, check beam center, refined parameters, ...
DEFPIX	Masked region (e.g. beamstop shadow) looks ok? (BKGPIX.cbf)
INTEGRATE	Predictions (FRAME.cbf) encircle the actual spots? Refinement is stable? Mosaicity and beam divergence looks reasonable? (INTEGRATE.LP)
CORRECT	ISa? Statistics? (CORRECT.LP) R_{meas} by frame? R_d by framediff? (by XDSSTAT) R_{meas} on detector surface? (by XDSSTAT)

skipped in the talk

IDXREF

- Read spot coordinates (x, y, frame number) and determine unit cell dimension and orientation by vector-clustering method.
- All necessary geometric information for INTEGRATE is written as XPARM.XDS.
- Based on my experience, the most important and difficult step (for non-good data).
- If failed, try:
  - change SPOT_RANGE= to change frames for indexing.
  - change REFINE(IDXREF)=.
  - check if beam center (ORGX=,ORGY=) is correct.
  - specify space group and unit cell if known.
  - change peak search parameter and re-run COLSPOT.

skipped in the talk

XPARM.XDS (March 30, 2013~)

Starting frame

Starting angle

Oscillation range

generated by IDXREF
required by INTEGRATE

```
XPARM.XDS    VERSION March 30, 2013
    1        30.0000     1.0000 -0.000558  1.000000  0.000479
 λ  1.541800          0.000031       -0.000074          0.648592
    1      38.0516      77.4951      77.6363  89.961  90.050  90.018
   12.321515          2.109153       35.939613   = a
  -15.062855         76.014046        0.677926   = b
  -71.920357        -14.423973       25.432476   = c
       1        3000        3000     0.100000     0.100000
 1502.497925     1499.867798      100.000000
    1.000000          0.000000         0.000000
    0.000000          1.000000         0.000000
    0.000000          0.000000         1.000000
      1           1        3000           1        3000
  0.00     0.00     0.00  1.00000  0.00000  0.00000  0.00000  1.00000  0.00000
```

Rotation axis

Incident beam direction

Space group and unit cell

NX=, NY=, QX=, QY=

Camera distance

ORGX=, ORGY=

segment information

INTEGRATE

INTEGRATE.LP

```
*****************************************************************************
                     PROCESSING OF IMAGES       1 ...      25
*****************************************************************************

USING    8 PROCESSORS

*** DEFINITION OF SYMBOLS ***
IER      = ERROR CODE AFTER ACCESSING DATA IMAGE
             0: NO ERROR
            -1: CANNOT OPEN OR READ IMAGE FILE
            -3: WRONG DATA FORMAT
SCALE    = SCALING FACTOR FOR THIS DATA IMAGE
NBKG     = NUMBER OF BACKGROUND PIXELS ON DATA IMAGE
NOVL     = NUMBER OF OVERLOADED REFLECTIONS ON DATA IMAGE
NEWALD   = NUMBER OF REFLECTIONS CLOSE TO THE EWALD SPHERE
NSTRONG  = NUMBER OF STRONG REFLECTIONS ON DATA IMAGE
NREJ     = NUMBER OF UNEXPECTED REFLECTIONS
SIGMAB   = BEAM_DIVERGENCE_E.S.D.=SIGMAB
SIGMAR   = REFLECTING_RANGE_E.S.D.=SIGMAR (MOSAICITY)

IMAGE IER  SCALE     NBKG NOVL NEWALD NSTRONG  NREJ  SIGMAB  SIGMAR
    1   0  1.203  5305465    0   2651     390     1  0.0247  0.0933
    2   0  1.204  5306652    0   2647     409     0  0.0256  0.1101
    3   0  1.209  5304717    0   2707     400     2  0.0252  0.1009
    4   0  1.208  5303456    0   2728     416     1  0.0257  0.0929
    5   0  1.208  5301649    0   2677     407     3  0.0261  0.0877
    6   0  1.212  5306367    0   2633     371     5  0.0254  0.1013
    7   0  1.212  5308069    0   2632     366     1  0.0260  0.0925
    8   0  1.212  5306555    0   2558     388     1  0.0250  0.0923
    9   0  1.215  5303601    0   2618     419     4  0.0247  0.0933
   10   0  1.217  5301791    0   2637     400     1  0.0257  0.1058
```

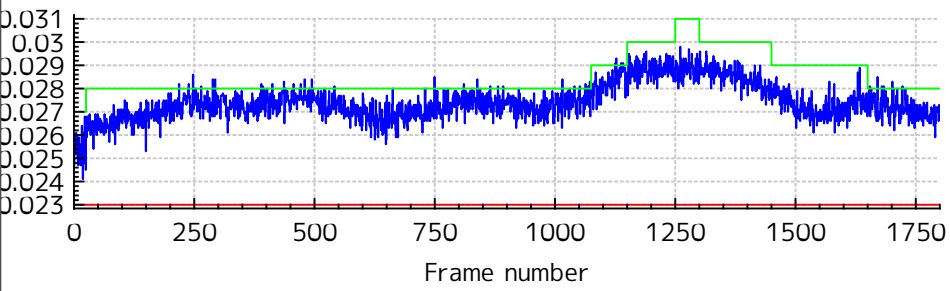
Process block of frames defined by DELPHI= (5° by default) at a time.

If too large (~2 or more), usually integration is going wrong way. Check if experimental geometry is correct and retry indexing.

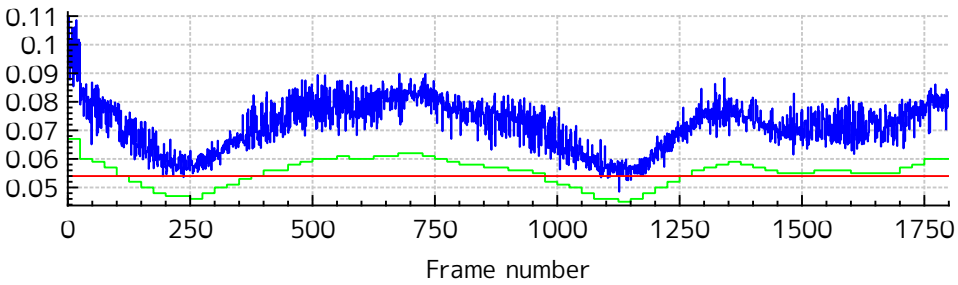
SIGMAB (spot spread) and SIGMAR (mosaicity) estimated for each frame.

INTEGRATE - check plots!

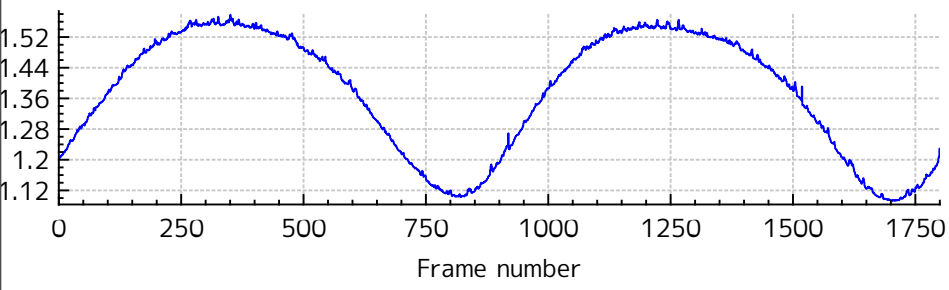
Beam divergence

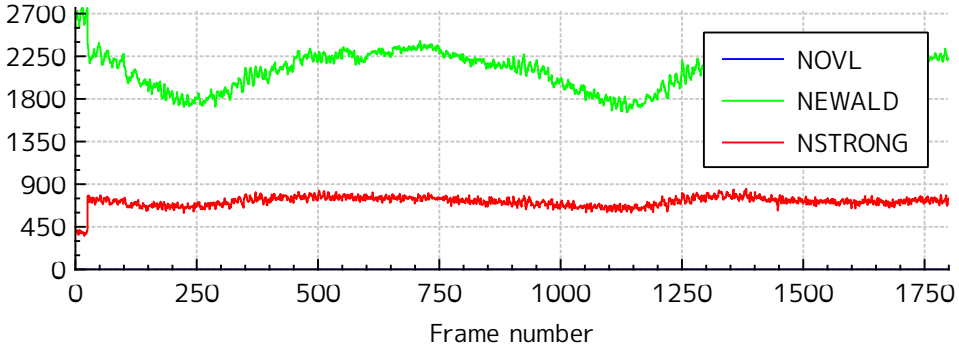


Mosaicity

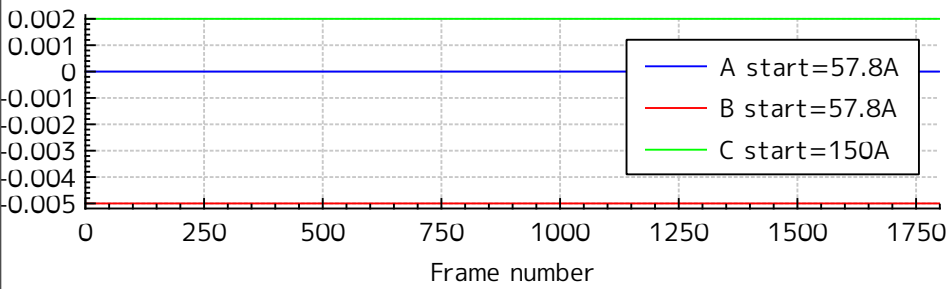


SCALE

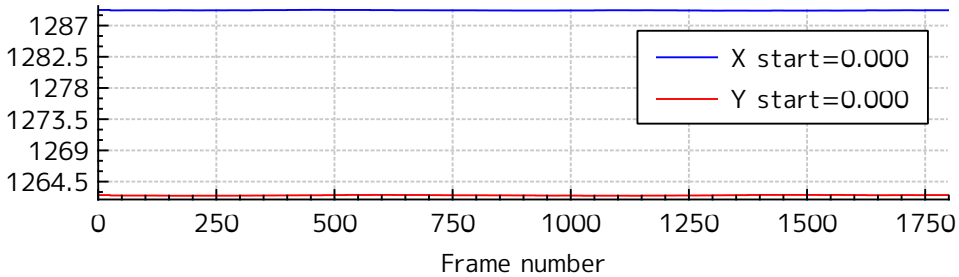




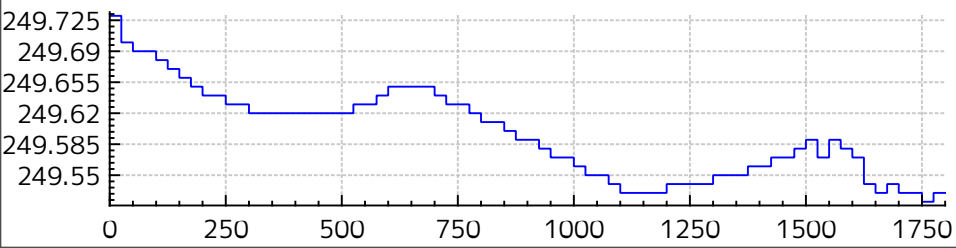
Change of Cell Axes [A]



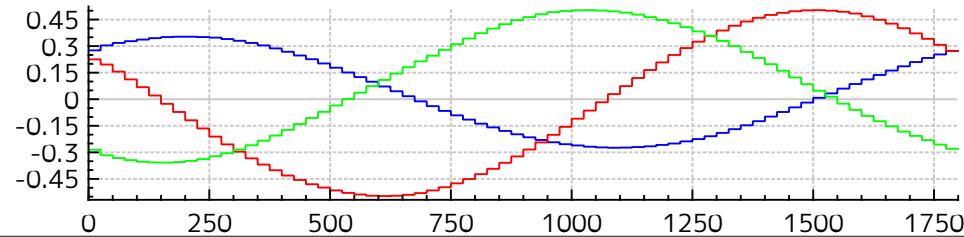
ence of DETECTOR COORDINATES (PIXELS) OF DIRECT



CRYSTAL TO DETECTOR DISTANCE (mm)



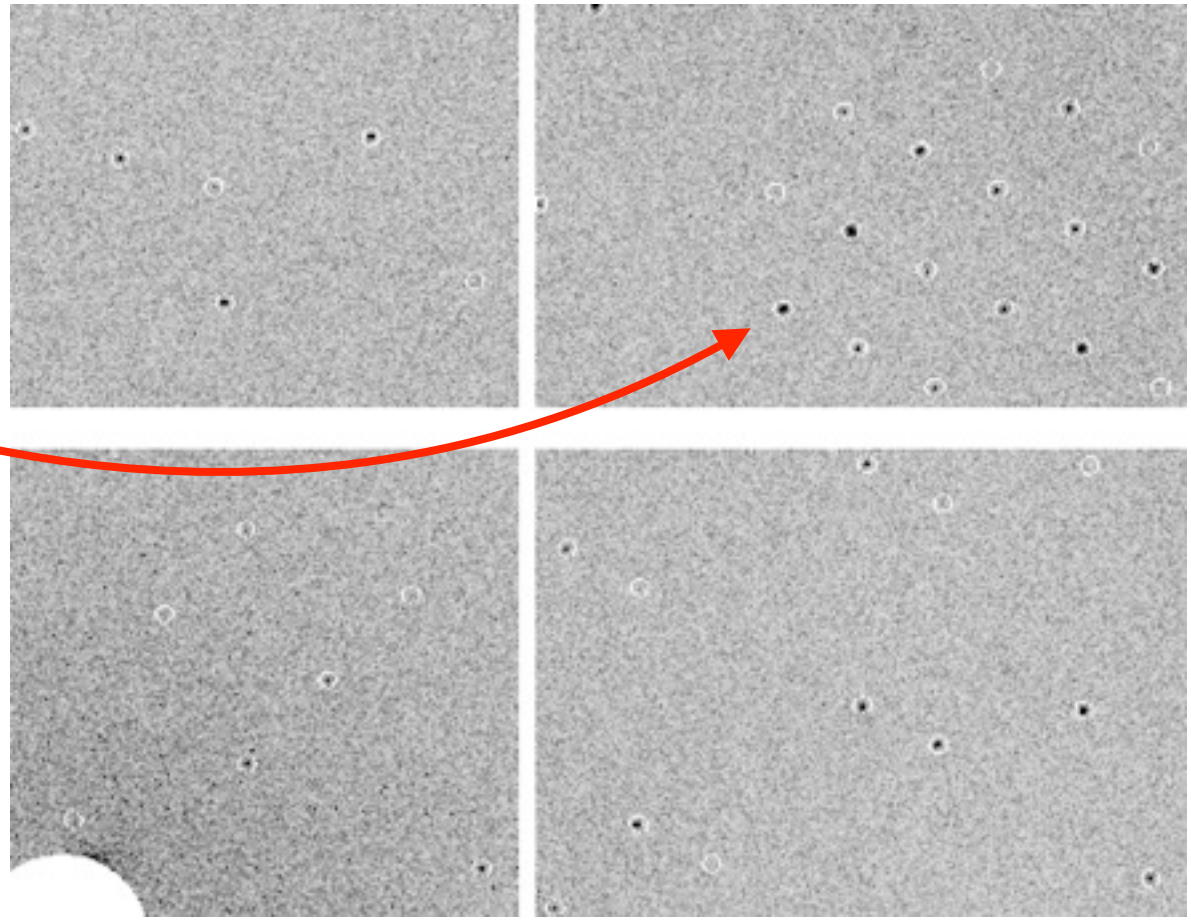
Crystal rotation [deg]



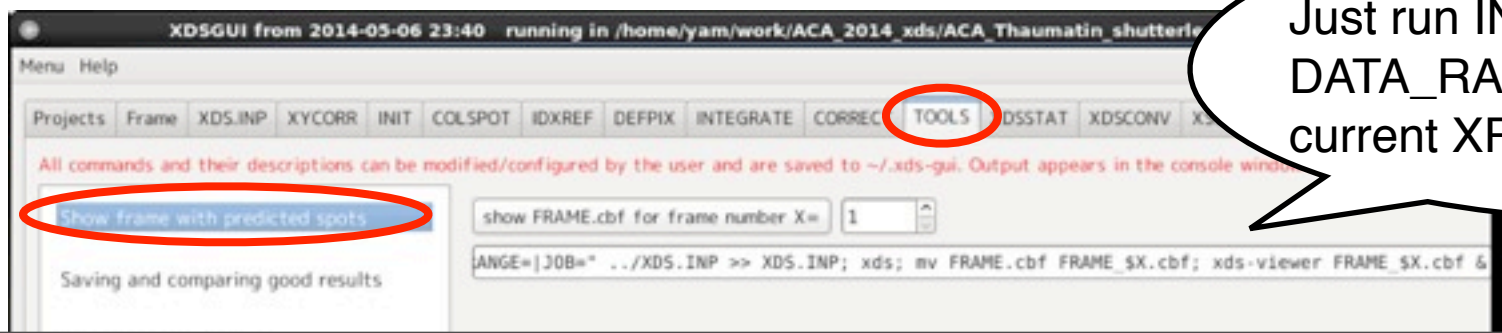
INTEGRATE - check predictions!

To see predictions of the last frame: check FRAME.cbf

Spots must be enclosed!



To see predictions of arbitrary frame: use TOOLS tab.



Just run INTEGRATE with DATA_RANGE= x x based on the current XPARM.XDS.

CORRECT - check symmetry and cell

In CORRECT.LP

```
 THE DATA COLLECTION STATISTICS REPORTED BELOW ASSUMES:
 SPACE_GROUP_NUMBER=   21
 UNIT_CELL_CONSTANTS=    81.75    81.79   150.02  90.000  90.000  90.000
```

Let pointless determine the symmetry..

```
$ pointless INTEGRATE.HKL
```

```
Nelmt  Lklhd  Z-cc    CC        N  Rmeas    Symmetry & operator (in Lattice Cell)

  1   0.955   9.82   0.98  577843  0.062     identity
  2   0.951   9.69   0.97  506629  0.087 *** 2-fold l ( 0 0 1) {-h,-k,l}
  3   0.949   9.65   0.96  514180  0.112 *** 2-fold k ( 0 1 0) {-h,k,-l}
  4   0.951   9.69   0.97  502669  0.091 *** 2-fold h ( 1 0 0) {h,-k,-l}
  5   0.953   9.75   0.98  503575  0.081 *** 2-fold   ( 1-1 0) {-k,-h,-l}
  6   0.955   9.84   0.98  505544  0.068 *** 2-fold   ( 1 1 0) {k,h,-l}
  7   0.949   9.65   0.96 1012994  0.109 *** 4-fold l ( 0 0 1) {-k,h,l}{k,-h,l}
....
Best Solution:    space group P 41 21 2
   Unit cell:   57.82  57.83    150     90.02     90  90.01
```

(with very clear systematic absences)

Edit XDS.INP and run CORRECT again..

```
JOB= CORRECT
SPACE_GROUP_NUMBER= 92
UNIT_CELL_CONSTANTS= 57.82  57.83    150     90.02     90  90.01
```

skipped in the talk

CORRECT - check the table.

Computed for reflections with merged $\langle I/\sigma(I)\rangle \geq -3$, which are used in downstream analysis.

completeness (= UNIQUE / POSSIBLE)

so-called R-merge (R-sym)

redundancy-independent R-merge (More important than R-merge..)

Important if anomalous

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr	SigAno	Nano
4.66	37203	2618	2619	100.0%	2.1%	2.0%	37203	116.39	2.1%	100.0*	79*	2.349	1025
3.30	64510	4697	4697	100.0%	2.1%	2.1%	64510	114.29	2.2%	100.0*	52*	1.470	2059
2.70	83826	6100	6100	100.0%	2.6%	2.5%	83826	90.39	2.7%	100.0*	47*	1.323	2758
2.34	99370	7147	7147	100.0%	3.2%	3.2%	99370	70.82	3.4%	100.0*	36*	1.196	3285
2.09	112368	8176	8176	100.0%	3.7%	3.7%	112368	61.32	3.8%	100.0*	32*	1.112	3797
1.91	121901	9032	9032	100.0%	4.5%	4.6%	121900	48.31	4.7%	99.9*	26*	1.059	4219
1.77	132580	9786	9786	100.0%	6.0%	6.2%	132580	35.62	6.2%	99.9*	20*	0.962	4605
1.65	86622	10516	10520	100.0%	7.5%	7.9%	86603	20.78	8.0%	99.7*	8	0.846	4953
1.56	35105	9335	11233	83.1%	8.4%	9.0%	34016	10.93	9.7%	99.1*	4	0.777	3005
total	773485	67407	69310	97.3%	3.1%	3.1%	772376	52.01	3.2%	100.0*	23*	1.099	29706

$\langle I/\sigma(I)\rangle$ for merged intensity

NUMBER OF REFLECTIONS IN SELECTED SUBSET OF IMAGES 779713
NUMBER OF REJECTED MISFITS 5877
NUMBER OF SYSTEMATIC ABSENT REFLECTIONS 351
NUMBER OF ACCEPTED OBSERVATIONS 773485
NUMBER OF UNIQUE ACCEPTED REFLECTIONS 67407

The reason why this number is different from that in table is because this number includes reflections with merged $\langle I/\sigma(I)\rangle < -3$. We should look at the table.

* If FRIEDEL'S_LAW=FALSE, Friedel pair is treated as different reflections. Otherwise, treated as unique reflections.

skipped in the talk

CORRECT - check ISa

Error model of XDS

$$\sigma^2(I_{hl}) = a\left[\sigma_0^2(I_{hl}) + b \cdot {I_{hl}}^2\right]$$

σ(I_{hl}) of each observation is adjusted with two global parameters *a*, *b* - such that the observed spread of intensities of symmetry-related reflections on average matches their variances.

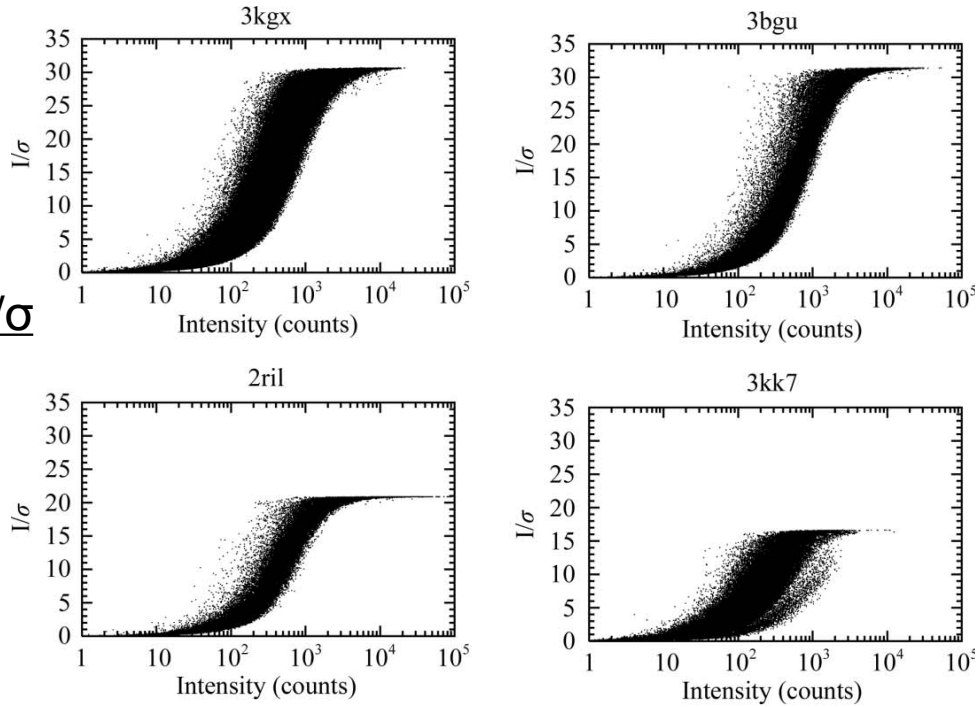
By the error model of XDS,

$$I^2/\sigma^2 = \left[a(\sigma_0^2/I^2 + b)\right]^{-1} \to (ab)^{-1} \quad (I \to \infty)$$

ISa is defined as the maximum value of unmerged *I*/σ

$$\mathrm{ISa} = (I/\sigma)^{\mathrm{asymptotic}} = \frac{1}{\sqrt{ab}}$$

However you increase the flux to reduce counting error, I/σ(I) cannot exceed ISa due to the systematic error.



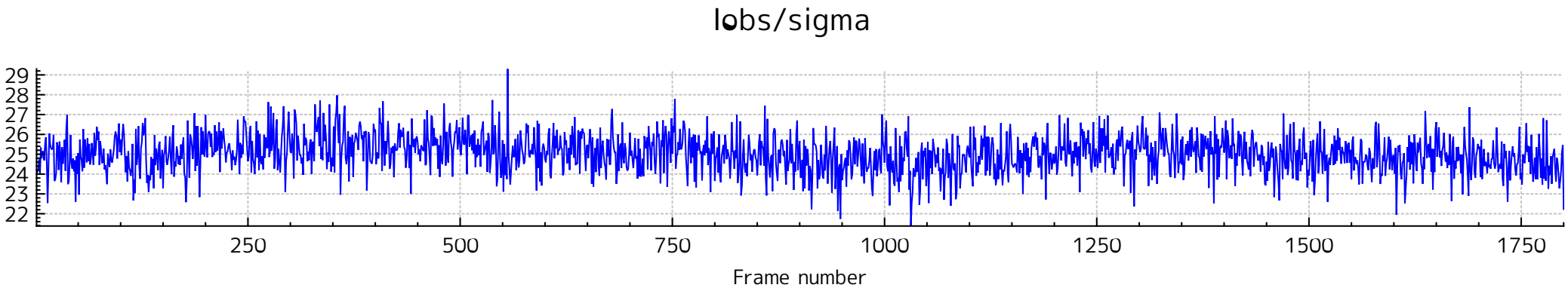
Diederichs (2010)

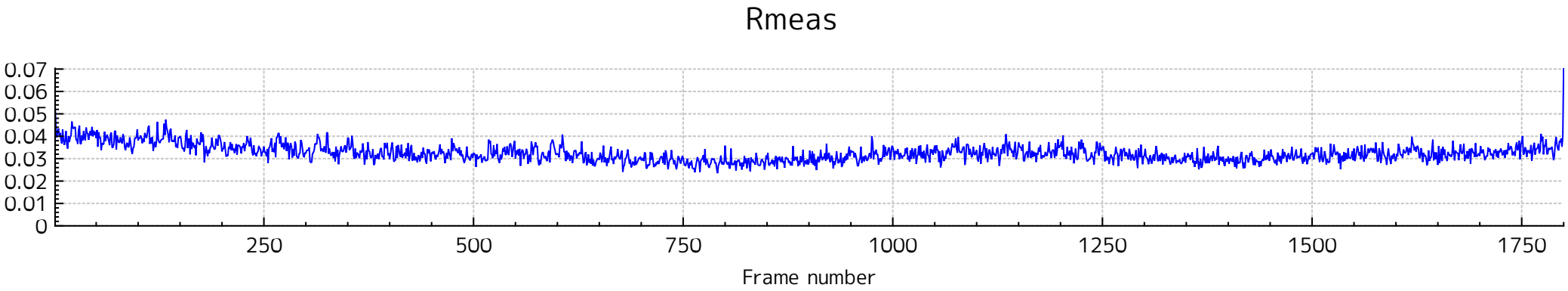
a	b	ISa
1.421E+00	3.619E-04	44.09

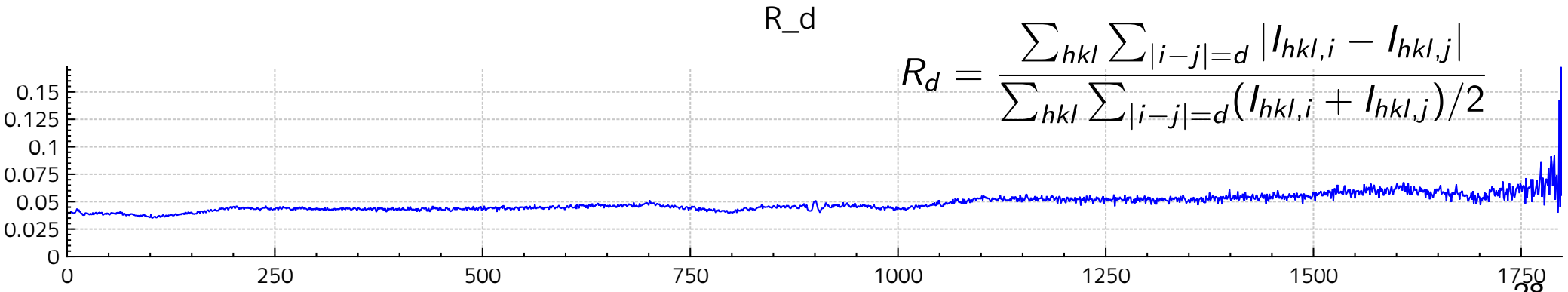
good value! (> 30)

XDSSTAT gives further information

for example,



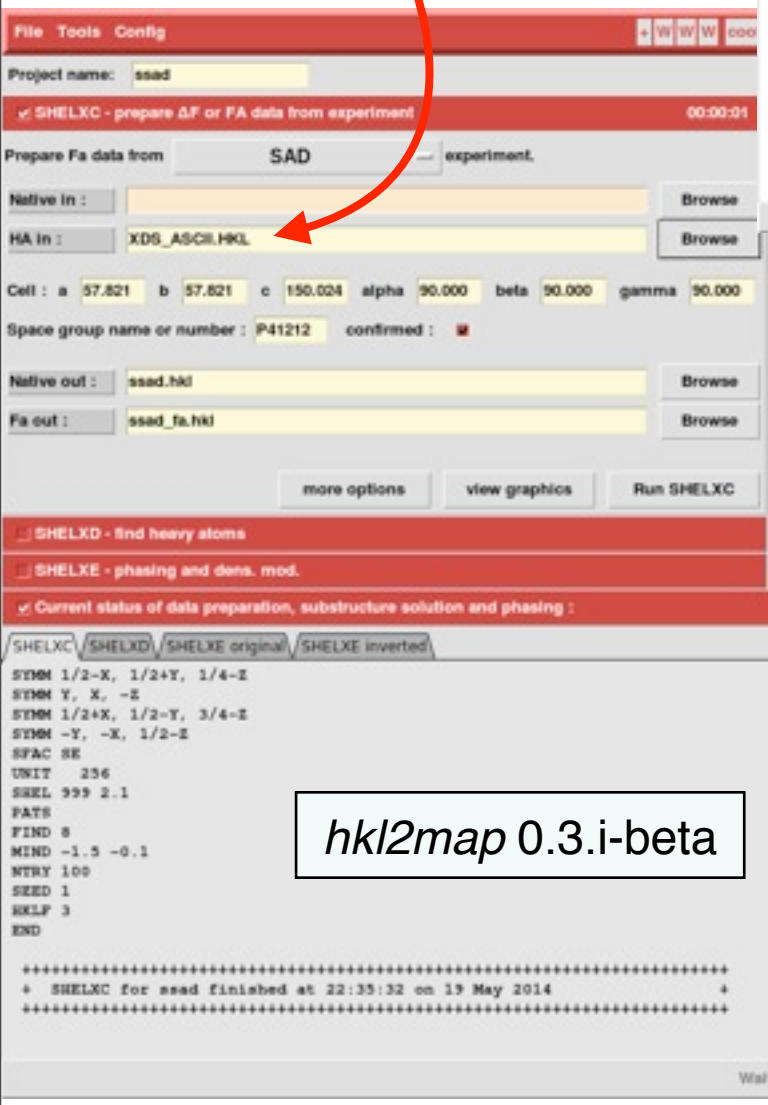




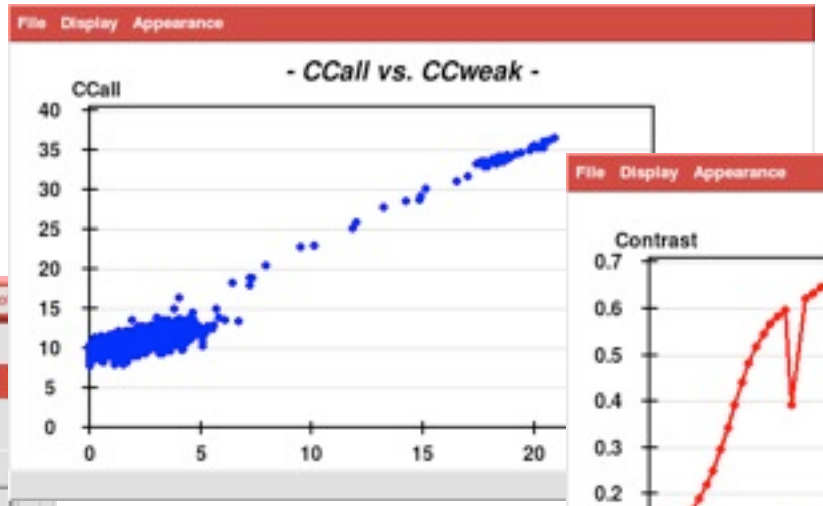
$$R_d = \frac{\sum_{hkl} \sum_{|i-j|=d} |I_{hkl,i} - I_{hkl,j}|}{\sum_{hkl} \sum_{|i-j|=d} (I_{hkl,i} + I_{hkl,j})/2}$$

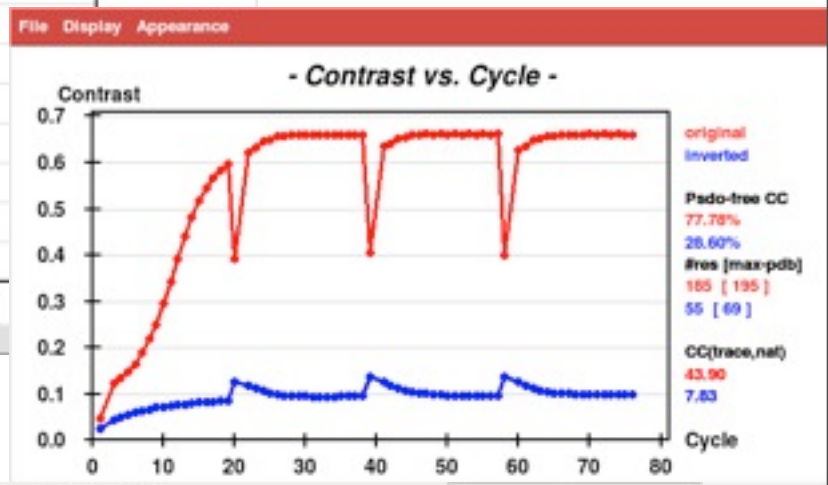
S-SAD phasing trial

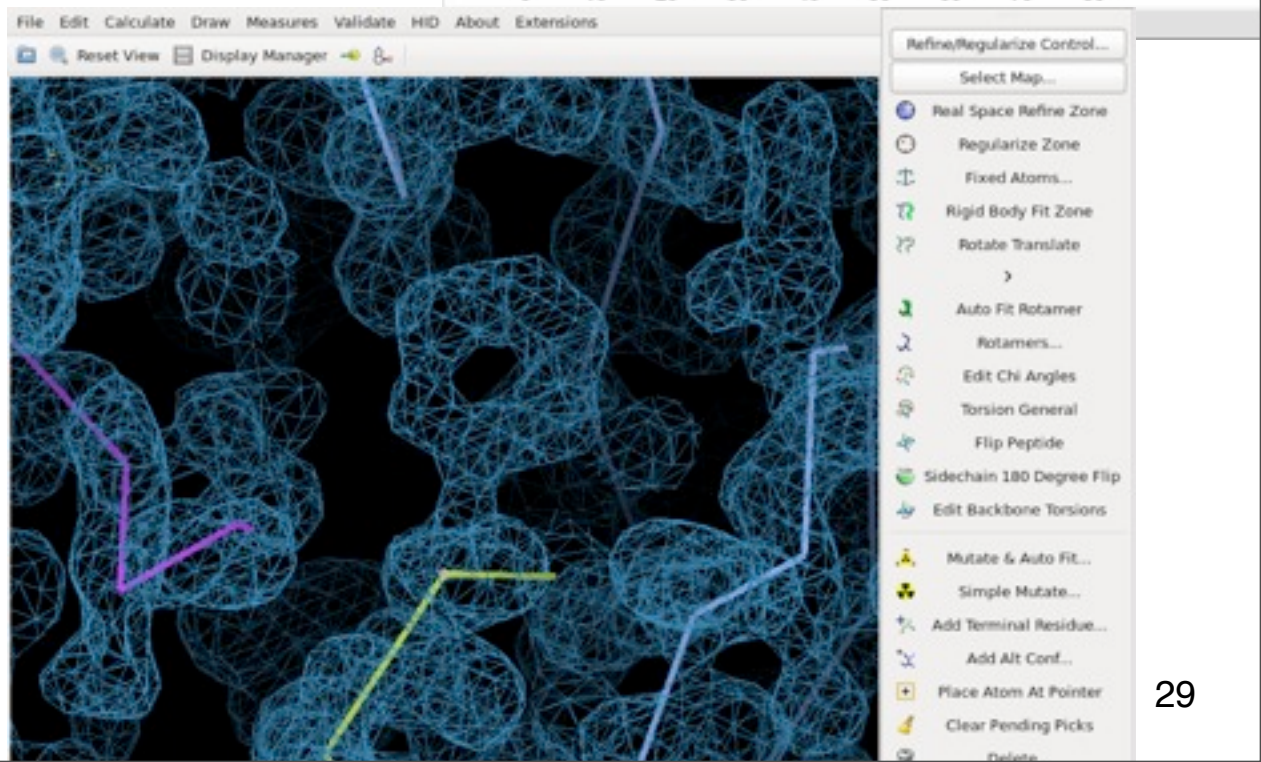
XDS_ASCII.HKL can be directly passed to shelxc.



hkl2map 0.3.i-beta







More information

- See XDSwiki articles about
  - Thaumatin (this talk)
  - Trypsin S-SAD
  - Trypsin inhibitor (Hard-to-identify Space Group)
  - Tough Spots
  - CagL (Possible Twinning)
- The slides of this talk can also be downloaded from ACA2014 article

Thank you for your interest!

If you had questions or comments on this talk,
please contact me: k.yamashita [at] spring8.or.jp