Automatic data processing at the ESRF: An overview of new and old features

Max Nanao
EMBL Grenoble
Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP
Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP
Currently at ESRF

- A suite of data processing and analysis infrastructure for all datasets collected at MX undulator beamlines (and BM14)
Some elements of the ESRF online data analysis portfolio

• “Heads up Display” of data from the last crystal
• Automatic indexing and processing
• Presentation of results in ISPyB
Some elements of the ESRF online data analysis portfolio

- “Heads up Display” of data from the last crystal
- Automatic indexing and processing
- Presentation of results in ISPyB

Some elements of the ESRF online data analysis portfolio

- “Heads up Display” of data from the last crystal
- Automatic indexing and integration
- Presentation of results in ISPyB

Some elements of the ESRF online data analysis portfolio

- “Heads up Display” of data from the last crystal
- Automatic indexing and processing
- Presentation of results in ISPyB

Some elements of the ESRF online data analysis portfolio

- “Heads up Display” of data from the last crystal
- Automatic indexing and processing
- Presentation of results in ISPyB

Automatic processing is the starting point for more advanced analysis

- SAD
- Grouped processing
- Processed data
- RIP
- MR
Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP
Grouped data processing…

- Multi position datasets becoming more and more common (and easier in MxCube 2.0!)
- The palette of data collection schemes is beginning to outpace the “back end”: what can be automatically processed
- Need for a simple, rapid and robust system to automatically process data coming from these more complex schemes
A few challenges

• Consistent indexing between wedges
• Reliable indexing
• Timely results
• Minimal user intervention
• Multiple elements require coordinated modification
  • MXCuBE
  • autoprocessing server
  • ISPyB data model + display
Internals

- Pointless used on all data to determine the SG+Bravais lattice
- Individual wedges are re-indexed as required
- XSCALE used to scale wedges together
A few examples:

- 1 crystal, 5 x 15° wedges, 1 position, cryo
- 1 crystal, 4 x 15° wedges, 4 positions, cryo
- 6 crystals, one position per xtal, CrystalDirect tray, microdiffractometer with plate gripper, Room temperature
A few examples:

- 1 crystal, 5 x 15° wedges, 1 position, cryo
- 1 crystal, 4 x 15° wedges, 4 positions, cryo
- 6 crystals, one position per xtal, CrystalDirect tray, microdiffractometer with plate gripper, Room temperature
A few examples:

- 1 crystal, 5 x 15° wedges, 1 position, cryo
- 1 crystal, 4 x 15° wedges, 4 positions, cryo
- 6 crystals, one position per xtal, CrystalDirect tray, microdiffractometer with plate gripper, Room temperature
Today:

• Review of autoprocessing at ESRF
• Some new and/or less well known features of ESRF automatic MX data processing
  • Grouped data processing
  • Automatic SAD phasing
  • Automatic MR
    • Dimple
    • Unit cell based MR
  • Automatic RIP
Automatic Structure Solution

Pointless Lattice + all SGs in lattice

SHELXC

SHELXD

+ hand

SHELXE, Solv -10%

SHELXE, Solv +10%

Invert hand

SHELXE, Solv -10%

SHELXE, Solv

SHELXE, Solv +10%
Some statistics from visitor data

- **Datasets**: 3665
- **Enough signal**: 455
- **SHELXD success**: 254
- **Solved**: 93
Some statistics from visitor data

- Datasets: 3665
- Enough signal: 455
- SHELXD success: 254
- Solved: 93
Some structures solved automatically

Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP
Automatic MR

- MR and difference map peak identification possible through DIMPLE if the user specifies the PDB in MXCube.

DIMPLE - a pipeline for the rapid generation of difference maps from protein crystals with putatively bound ligands
M. Wojdyr, R. Keegan, G. Winter and A. Ashton
Dimple MR and difference peak identification

EDNA Autoprocessing

- Space Group
  - From ISPyB
  - From MxCuBE

- PDB
  - From ISPyB
  - From /PROCESSED_DATA/xds_

Dimple

Results (mtz + pdb)

For manual run: dimple.sh mypdb.pdb

If no pdb is given

dimple.sh
Automatic MR

- If the user does not specify the PDB?
  - The DIMPLE pipeline makes a script for running DIMPLE later
  - Search PDB for similar unit cell parameters

Automatic MR by unit cell

- Fast processing cell parsed
- Unit cell dimensions submitted to Nearest-Cell
- List of PDBs in different families returned
- Top $n$ downloaded from EBI
- Search model cleaned up (First protein chain, no waters)
- PHASER
- Evaluation by Z-scores. Failed directories are deleted
- Refmac restrained refinement
- Water updating in Phenix.refine

**nearest-cell**

*A fast and easy tool for locating crystal matches in the PDB*

Unit-cell:  
Space-group:  
Submit

Today:

- Review of autoprocessing at ESRF
- Some new and/or less well known features of ESRF automatic MX data processing
  - Grouped data processing
  - Automatic SAD phasing
  - Automatic MR
    - Dimple
    - Unit cell based MR
  - Automatic RIP
Automatic segmented Radiation Damage Induced Phasing

- Large oscillation dataset split into sub datasets for RIP
- Very well adapted to large datasets (i.e. PILATUS datasets)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic SAD phasing</td>
<td>Functional and solving structures Needs IspyB integration</td>
</tr>
<tr>
<td>Automatic Grouped data processing</td>
<td>Re-coded for new MxCube, still in testing but deployed Needs IspyB integration</td>
</tr>
<tr>
<td>Auto-MR</td>
<td>Installed January 2014, in testing</td>
</tr>
<tr>
<td>Dimple MR</td>
<td>Functional but needs MxCube 2.0 modification, Needs IspyB integration</td>
</tr>
<tr>
<td>Auto-RIP</td>
<td>Beta testing</td>
</tr>
</tbody>
</table>
People:

**ESRF**
- Matias Guijarro
- Sasha Popov
- Marjolaine Bodin
- Solange Delageniere
- Stephanie Monaco
- Daniele de Sanctis
- Thomas Boeglin
- Gordon Leonard
- Olof Svensson
- Marcus Oscarsson

**EMBL**
- Hassan Belrahili
- Florent Cipriani
- Josan Marquez
- Ulrich Zander
- Andrew McCarthy
List of autoprocessing directories

Confirm that autoprocessing results exist in each dir

No -- wait up to 500s

Run POINTLESS on all data to Determine SG and cell

Loop through all directories

compare POINTLESS lattice with the integrated lattice. Are the lattices the same?

Yes

Has a reference dataset been chosen?

No

set the current dataset to be the reference

run XDS, job = CORRECT, Specify SG and cell

Yes

specify the reference dataset with REFERENCE_DATA_SET

XSCALE together all datasets

Generate stats and graphs

convert to MTZ via XDSConv

No

Has a reference dataset been chosen?

No

run XDS, job = ALL, Specify SG and cell

Yes

specify the reference dataset with REFERENCE_DATA_SET