AXDPRO/KYLIN Data Reduction and Data Collection Monitoring
(Linux Version)

§1. Introduction

This Linux version of KYLIN package is intended to be used at synchrotron beamlines, home sources or productive groups for monitoring data collection and/or processing single-crystal X-ray diffraction data, which includes:

1. KYLIN (a fully-functional data procession program, **§2, §4**)

2. AXDPRO (Automatic X-ray Diffraction Data Processing, **§2, §3**)
A single-line-command UI automatically processes the single crystal diffraction data by distributing the following multiple modular tasks parallel onto clusters and/or PC/Workstations or local clouds.

RCMDDENZO	{DEZNO/SCALEPACK*, SPGR4D, 3DSCALE}
RCMDDIALS	{DIALS*, CCP4*}
RCMDKYLIN	{KYLIN, SPGR4D}
RCMDXDS	{XDS*, CCP4, SPGR4D}

(*Has to be installed separately before using the module).

3. Tools (**§5)**
Some handy tools are also included in the package, such as monitoring data collection and backup data on-the-fly, constructing ssh-based local cloud for distributed computing, measuring beam center, interfaces for structure prediction (**§5.9)** and ligand-bindings screening (**§5.10)** by exploiting AlphaFold?, etc.

FYI: Part of the package has been used as a tool to perform on-the-fly quality control of data acquisitions at SER-CAT (APS Sector#22) since 2008.

§2. Installation

1. Download and save "sgxpro.tar" in a directory where you want to put the package. If installation is performed by 'root', the package will be available to all users. Otherwise, it will be limited to the current user. Go to the directory, then type the following command:

tar xvf sgxpro.tar; ./sgxpro_install

2. Type command sslistalias to create an alias of your data collection site configuration, which will be used where a siteconf is needed.

The following commands are helpful to start depending on what to do:

1). Type command sstestrtd to see how to setup and test using 'kid' to monitor data collection on-the-fly with color-coded warning system. Or you may bypass the setup testing, type command kid and directly go to 'Example 4' at the bottom to see how.

2). Type command axdpro to see how to auto-process data with multiple programs parallel.

3). Type command ssaxdpro to see how to search and auto-process with axdpro all the data sets having been collected and saved in a folder.

… … (see §5 for more tools)

FYI: The GUI of KYLIN data processing can be launched by command sgxpro.

*(If the commands not found, you may need to logout and re-login after the installation).

*ReleaseUpdate: https://github.com/axdpro

*Suggestions? Bugs? Please send to FUZQ@UGA.EDU, Thanks!

§3. Automatic Single-Crystal Diffraction Data Processing with AXDPRO

Type command 'axdpro', 'rcmddenzo*', 'rcmddials', 'rcmdkylin', or 'rcmdxds' will show a brief instruction on how to use these single-command-line UIs, which can always run locally, such as:

rcmdxds @localhost siteconf …

1. AXDPRO Default

Default on which modules to use and how to distribute multiple parallel tasks is defined in the header of helper script file 'axdpro'. With a text editor, you may change the default by modifying the header according to the computing facility available. Once the default is defined, you may simplify the commands, such as:

axdpro siteconf …

2. To fully exploit the automation capability of the package, a desirable computing system would be either a Linux cluster, or an ssh-based local KYLIN cloud consisting of multiple stand-alone Linux PCs and/or workstations and/or clusters that can be defined either by

1). Define a temporary KYLIN cloud as the default in the header of helper script file 'axdpro'.

or

2). Use command **sskylincloudconfig** to construct a KYLIN cloud for use by axdpro.

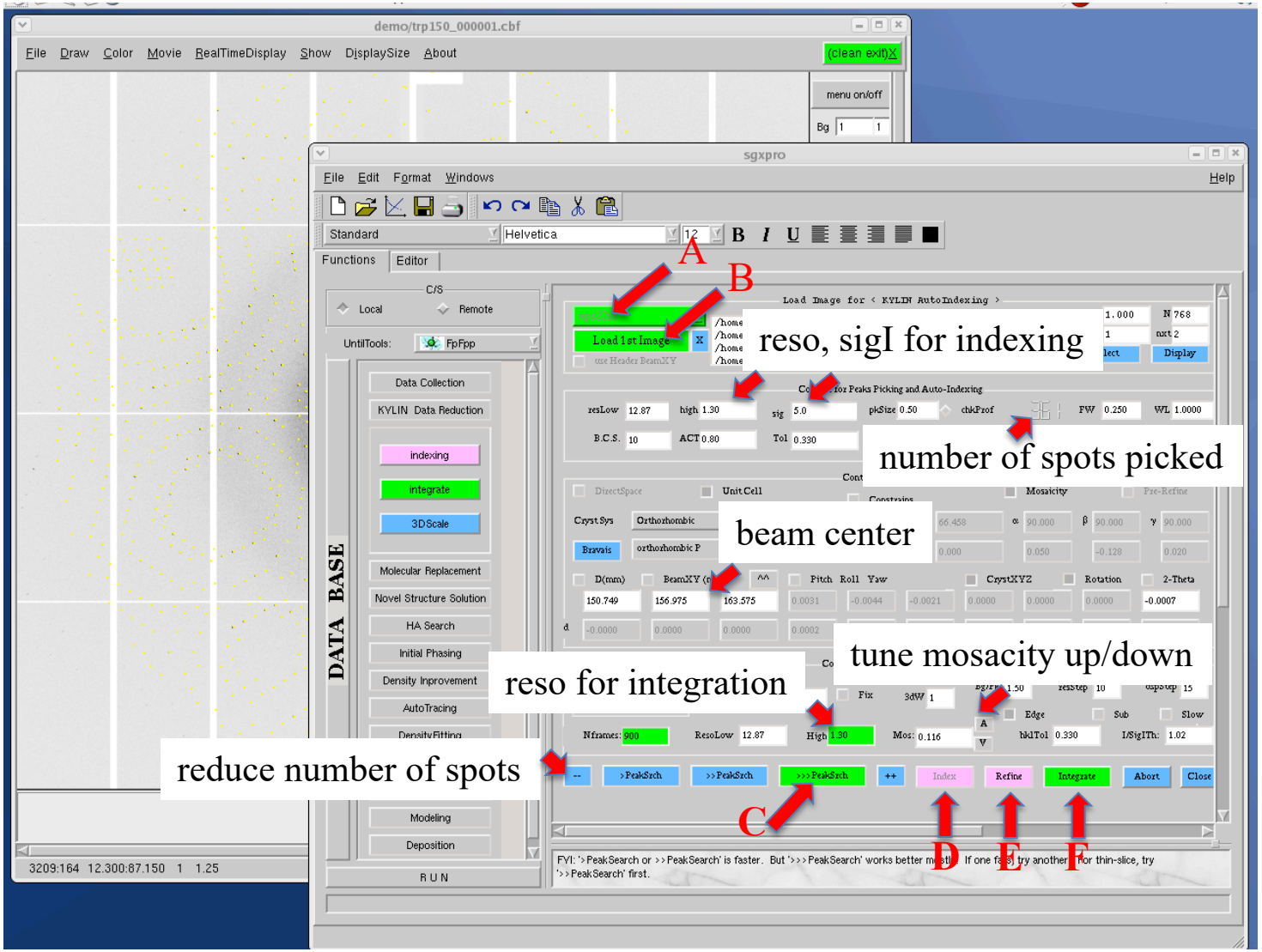
3. ZSUM Quality Briefing to Find the Best Data Set Quickly: zsum –b, zsum –b folder

*As HKL?000 blocked the external use of DENZO since 2014, the module RCMDDENZO may not work.

§4. Use KYLIN Data Reduction from GUI

1. Start KYLIN, Peaks Search, Index, Refine, and goto Integration

Type command sgxpro to start KYLIN GUI (currently attached to the "Functions" window of SGXPRO as shown below). The two windows "Functions" and "Editor" can be switched by click on it.

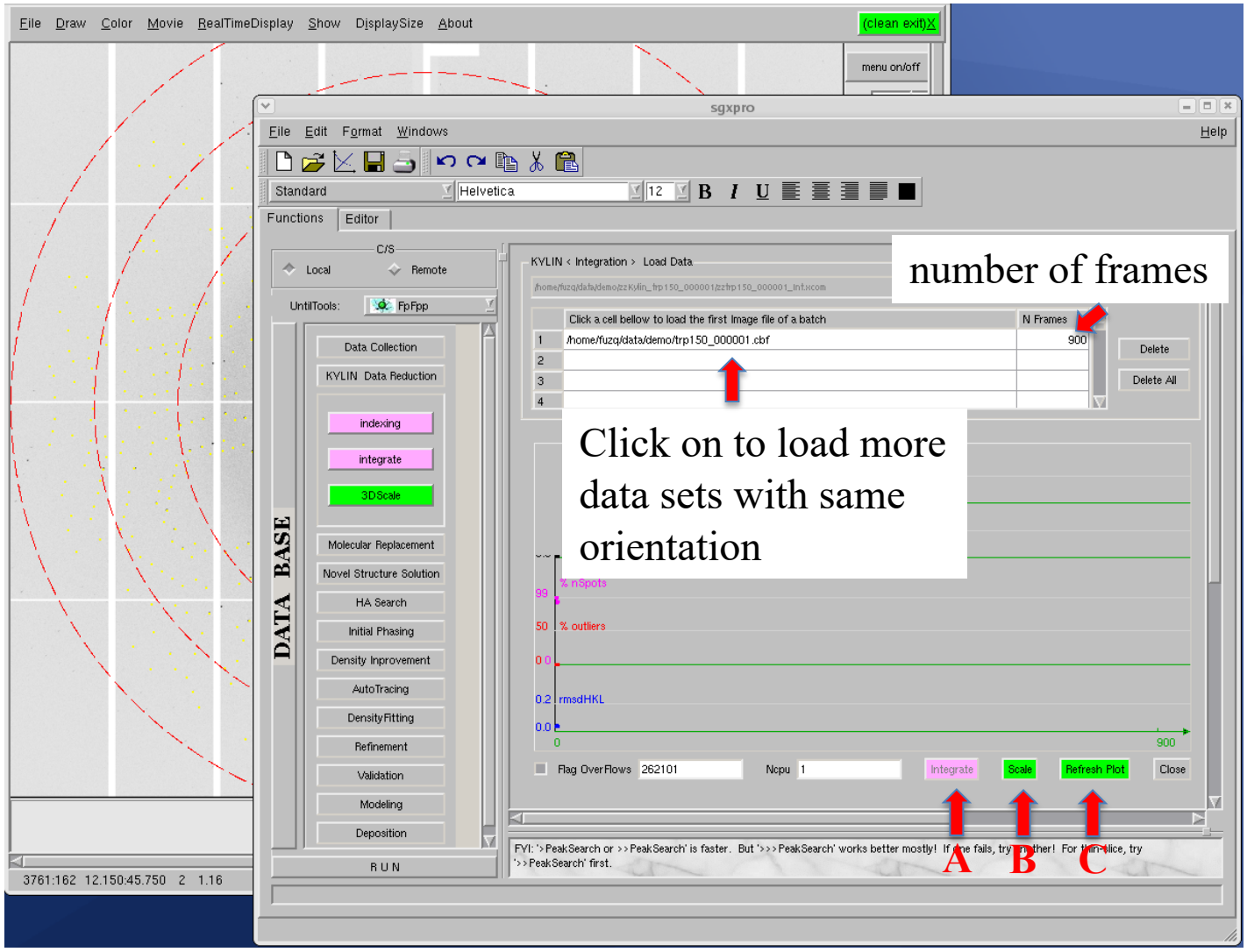


A: Select a siteconf
B: Load 1st Image or *_master.h5
C: >>>PeakSearch
D: Index
E: Refine
F: (goto) Integration
For most cases, only the above default clicks are needed.

*Otherwise, 'reducing number of spots', 'reso for indexing', give a smaller sigI i.g. 3 for very weak data before peak search, or correct beam center, etc. may help improve.

*If Refine failed to cover the spots by predicted (yellow dots on the display), you can manually tune mosaicity up/down.

2. Integration



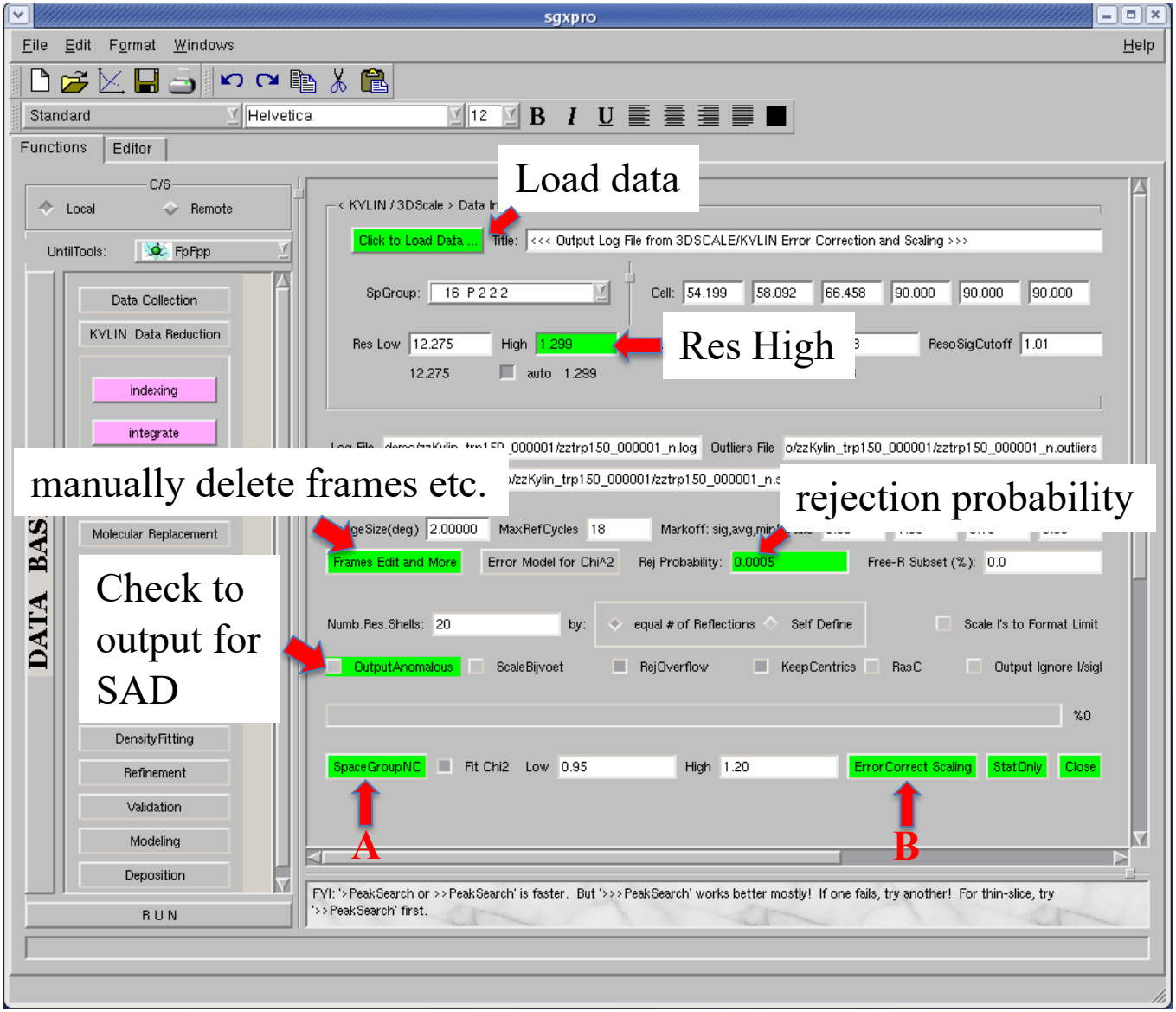
A: Integration
B: (goto) Scale
For most cases, only the above default clicks are needed.

*If you have multiple data sets with same orientation (such as those from helical segments collection), you may load and integrated together. However, data with different orientations have to be integrated separately, and merged at scaling.

*Ncpu is the number of processor(s) to be used by integration. Parallel integration will start automatically if more than 1 is given.

*The statistics plots will progress on-the-fly only if 1 processor is used, which can be refreshed by the "Refresh Plot" button.

3. Experimental errors correction and Scaling



*Bad frames will be automatically rejected as outliers during the scaling, which can be manually deleted if known.

A: SpaceGroup
B: EorrorCorrectionScaling

If scaling is started after integration, all the data have been loaded automatically. Otherwise, integrated data in '.kint' can be loaded manually.

*The 'SpaceGroup' in the current version is for none-centric symmetries only.

*For very noisy data, increase the rejection probability may improve the statistics, which is strongly not recommended for common data.

*'Res High' is the highest resolution of the data, which will be used as the starting cutoff. A more reasonable cutoff will be estimated during scaling. Changing this to a higher value manually may affect the final cutoff estimated.

4. Space Group Determination at Data Processing

1). KYLIN/RCMDKYLIN and RCMDDENZO use **SPGR4D** to determine the space group, which will summarize the results in '*_s.log' file (always check this file carefully). Some times, the would-be extinct reflections may not be harvested during the integration, leading to the uncertainty of space group determined from data reduction. SPGR4D will produce a warning when this happens. If the space group can't be determined, the point group will be used. In this case, you may just change the space group in the reduced data '*.sca' file, and test it in phasing or structural refinement steps, which won't affect the structural solution. The truth is, only Laue groups have major impacts during data reductions!

2). Data integrated with triclinic (P1)

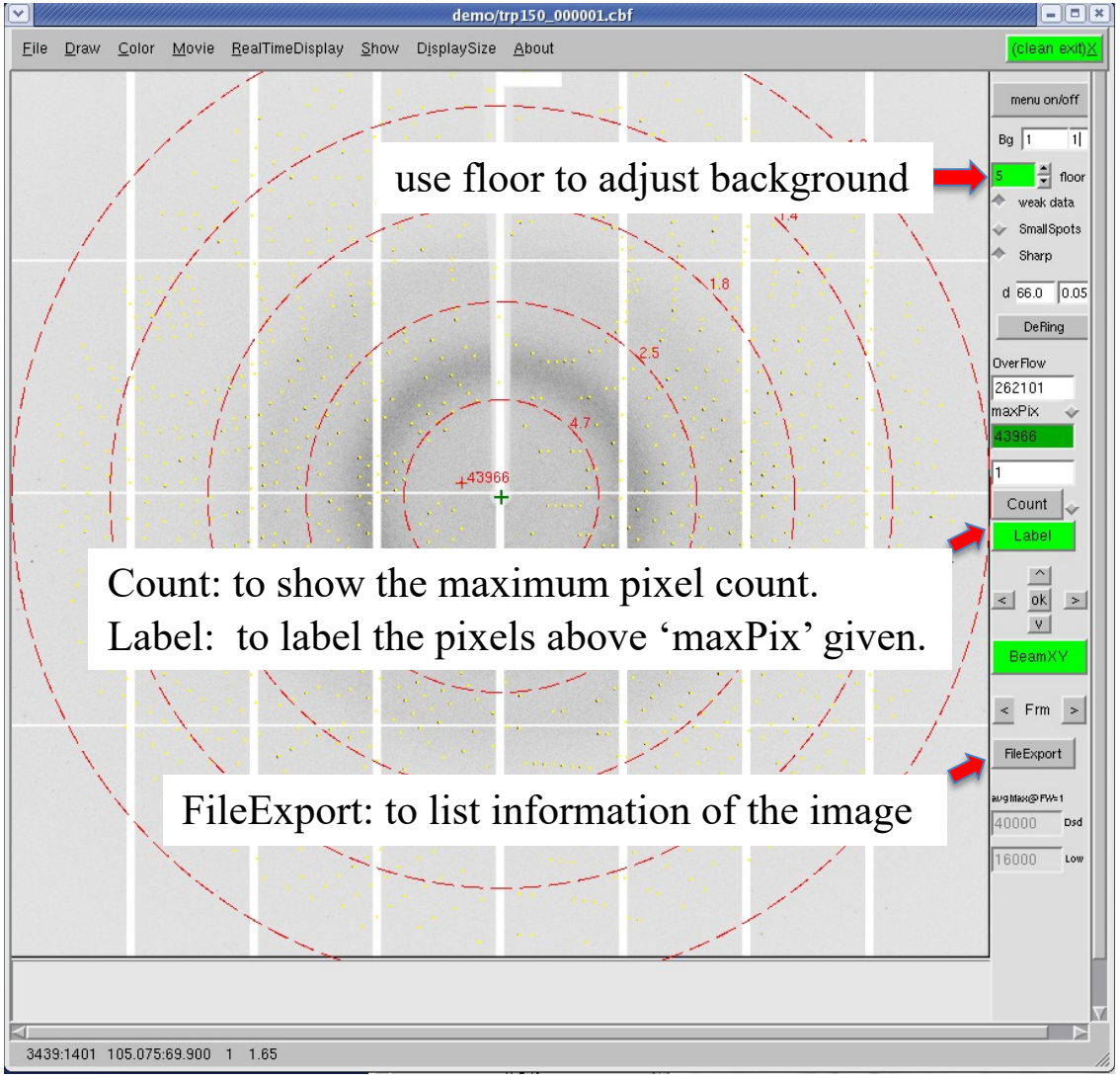
If not sure about the crystal system, you can always choose triclinic at indexing and integrating the data. The "SpaceGroup" function of KYLIN scaling will survey the data to see if a higher symmetry is reasonable, and automatically re-index and scale the data accordingly. However, choosing the correct crystal system at indexing may slightly improve the final results in some cases.

FYI: When a favorite program doesn't work, twinning or other hypothetical corruptions are more likely blamed than the program's weakness. AXDPRO uses most of the programs available to make sure that the data are processed by the best program for every crystal collected, as different programs behave differently, one may outperform others for some crystals and *vice versa.* Its parallel mutiple-module design will allow to add new module based on a program if proved to be unique. Each module can also be used independently by a single-line-command.

§5. Tools

Listed above is just the basic for routine data processing with AXDPRO/KYLIN, which would work for most cases. There are many parameters that can be adjusted for challenging cases, or for experienced users who like to fine-tune. In addition, there are some handy tools, such as listed in the following:

1. KID (Kylin Image Display, type command 'kid | more' to see details on usage and examples)



KID is a simple script using KYLIN display:

1). Monitor data collection on-the-fly:

kid siteconf any -rtd

Type command sstestrtd to see how to set up.

2). Movie and survey images:

kid siteconf imgFile nFrm –step 1

kid siteconf img_000001.hdf nFrm –step 1 (for data saved in img_master.h5)

3). Display one image:

kid siteconf imgFile

kid siteconf img_000101.hdf (for data saved in img_master.h5)

4). Search & display images by pattern name:

kids siteconf "/data/*_s_000001.cbf"

kids siteconf "/data/*_master.h5"

FYI:

*Right-click on the image to zoom-in.

*Try the functions in the drop-down menu to see the image display basics.

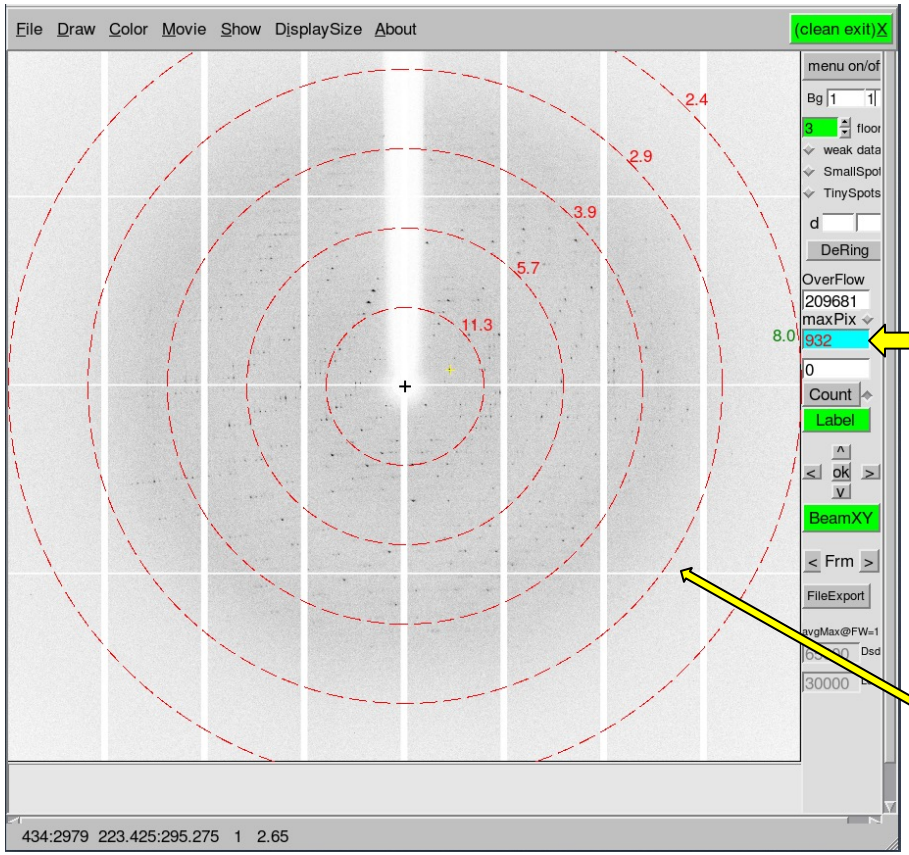
*Click on 'About' to see color codes assessing the exposure etc.

*Use '(clean exitX)' to exit when done.

… …

Use KID to monitor data collection on-the-fly

may help choose appropriate **Transmission/Exposure, Detector Distance etc.**



KID assess the diffraction image, and uses 'maxPix' background colors to flag exposure levels by:

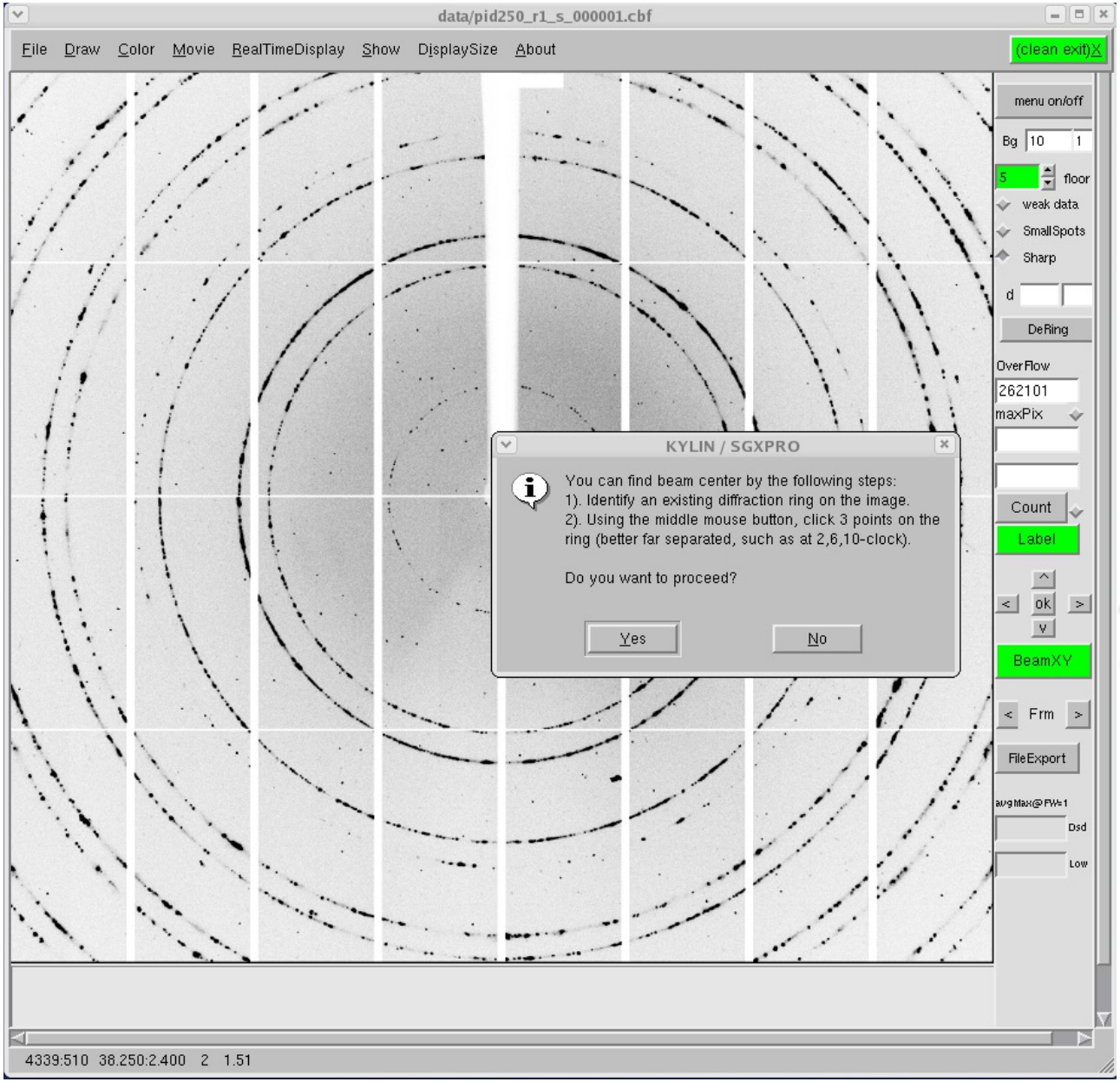
- Cyan: possible under-exposure.
- White: ok.
- Green: desirable.
- Purple: royal, or high?.
- Red: some pixels(cross-labeled) overflow or count over-range.

background color of maxPix

the second highest resolution ring

A Rule-of-Thumb about Detector Distance: Computer is more sensitive than human eyes to recognize the weakest spots on an image. Therefore, while moving back the detector to maximize the spot-separations (spatial resolution) for better data quality, to avoid losing resolution of reduced data, it's better to make sure no visible diffraction spots go beyond the **2nd highest resolution ring**.

2. Measuring Beam Center; Checking the Detector Distance



1). Data processing programs rely on accurate beam centers to function properly. KYLIN display can estimate beam center with any diffraction ring(s) (either from phosphor powder ball, alignment pin, or ice in sample loop, etc.).

*Due to the visual and/or inaccuracy of these rings, it is strongly recommended to further refine the beam center thus measured with single crystal diffraction data at processing.

⬅ Manual adjust

⬅ Click on BeamXY to start

2). Once beam center is calibrated, from drop-down menu Draw ➡ choose the right *CalDetD to check if the energy or the detector distance needs calibration, by clicking on the powder diffraction ring at the saidResolution from Si, Se, Ice, or Pball.

FYI: 'DeRing': If PeakSearch picked up many suspicious spots around a ring, you may click on the ring, then click 'DeRing' to remove these spots before index.

3. ZSUM quality briefing (type command 'zsum -help' to see detailed usage and examples)

Nowadays with high-speed detectors and powerful auto-processing software, many users adopt the 'collect-more-and-cut-later' strategy to collect data on every sample that looks collectable. This strategy works well only if you can process and monitor quality of the reduced data on-the-fly. ZSUM harvests and provides a brief summary of all data sets automatically processed, which would allow you monitor the **quality of processed data**, help adjust data collection parameters (i.e. Transmission and/or Exposure Time and/or Detector Distance and/or Frame Width etc.) to get the best data from your crystals before too late. Shown below is the zsum output from a users' data.

Reso	Rmerg(low high)	Rdn	Comp	I/sig	CC1/2	Mos	UnitCell	Symm	Location
		dir: /data/ID_xxx.raw/puck8/p8/					**{0.25/0.25/350.0 1 1546,}**		
5.72	0.110(0.079 0.846)	7.4	98.4	0.9	0.52	xxxxx	C2	zzCmdxds_puck8_Pn8_000001/*sum.log	
5.98	0.086(0.067 0.455)	7.2	98.5	0.8	0.48	xxxxx	C2	zzKylin_puck8_Pn8_000001/*sum.log	
		dir: /data/ID_xxx.raw/puck8/p9/ {0.5/0.25/ 350.0 1 2775,}							
4.04	0.104(0.048 1.225)	7.4	98.8	1.2	0.57	xxxxx	C2	zzCmdxds_puck8_Pn9_000001/*sum.log	
5.06	0.149(0.084 0.593)	7.6	99.3	1.0	0.54	xxxxx	C121	zzCmddials_puck8-09_Pn9_000001/*_sum.log	
4.26	0.095(0.081 0.435)	7.3	97.3	1.1	0.56	xxxxx	C2	zzKylin_puck8_Pn9_000001/*sum.log	
		dir: /data/ID_xxx.raw/puck8/p10/ {0.5/0.25/350.0 1 17943,}							
3.53	0.113(0.031 1.410)	7.6	98.7	1.1	0.61	xxxxx	C2	zzCmdxds_puck8_Pn10_000001/*sum.log	
3.39	0.098(0.037 0.482)	7.5	98.4	1.2	0.67	xxxxx	C2	zzKylin_puck8_Pn10_000001/*sum.log	
… …									

a). **dir: /data/ID_xxx.raw/puck8/p8/**
Folder containing the data (a good way to organize is to collect each data set into a separate folder).

b). **{0.25/0.25/350.0 1 1546,}** = {FW/ExposureTime/Distance Size maxPix}

In a few seconds by look, you can quickly find:

c). The quality of processed data from different programs can be significantly different. Nothing wrong! They are ready for you to choose the real best (best crystal, best collected, from best program for the crysal processed). They can also be ranked by a simple script program to present on a control program's GUI.

d). Some data sets may be under-exposed if not due to weak-diffraction samples.

4. spgr4d for space group determination

SPGR4D is part of the KYLIN package for space group handling. It was created for generating (3+1)-dimensional super space groups, and was adopted here for studying the 3-dimensional space groups. The current version is customized for none-centered symmetries only.

*If you want to check on an integrated data from DENZO (including HKL?000), in the directory with *.x files, type the following command:

spgr4d mydata_000001.x or spgr4d *000001.x (if only one data set in the directory):

*In addition to space groups of standard settings, SPGR4D also supports the none-standard settings, such as P21221 etc. ('mhklabc' can change *.sca file from such a none-standard to the standard setting).

*If the data from integration is not enough, a point group may be suggested.

*Same as others, it can't separate pairs of P3121 and P3221 etc. at the data reduction stage.

5. ssbeamxy siteconf (to show beam centers for different programs).

6. sshstbsdauthconf (to set host-based authentication among Linux computers)

*It has to be run by superuser, and may not work as expected for your system due to the differences of Linux distributions.

7. ssbackup (to backup data on-the-fly, 'ssbackup –help' for usage details).

ssbackup –a –k (for common users without superuser priviledge).

8. sstoolh5 (to search and/or convert diffraction data H5 files into *.cbf image files).

9. ssafsub

Submit jobs from outside to SER-CAT's server running AlphaFold to predict structure from sequence.

10. ssafscreen

Submit jobs from outside to SER-CAT's server for screening to find binding ligands with AlphaFold-Multimer.

*FYI: Other ssafsub/ssafscreen related helper commands:

1). sstoolaf --- Management tool for ssafsub/ssafscreen etc.

2). ssgetslog --- Check status or get results from AlphaFold jobs on SER-CAT's server.

3). rssalphafold, ssalphafold --- Interfaces for submit and manage batch jobs through a load-balanced scheduler onto the server running AlphaFold-Multimer.

4). sspdbbfplot --- Analyze and create pLDDT vs Residue plot of a predicted model, and derive AISIDscore for ligand-binding evaluation. It is called inside rssalphafold/ssalphafold by default.

5). ssafchks --- Harvest data from outputs of a ligand-screening project and present the final list of AISIDscores for all the bait-ligand pairs.

6). ssafchk, ssafchkstime, ssseqaac, sstrsc etc. --- Miscellaneous scripts for check, survey data from ligand screening project.

7). ssseqsplt --- Read and split a file containing protein sequences (such as that from UniProt database) into single-sequence files with fasta format.

8). ssseqmlt, ssseqmltsort --- Create composite (bait-ligand dimer) sequence files and the script to run ligand-screening batch jobs through 'rssalphafold', 'ssalphafold' interfaces.

* The tools listed under 9 and 10 above can also be installed separately as a subset without SGXPRO by:

- 1. Download 'serafclient.tar': scp username@164.54.208.23:/usr/local/sgxpro/serafclient.tar .
- 2. Save serafclient.tar file into a folder in the executable-paths, and extract: tar -xvf serafclient.tar

References:

1). SPGR4D
Fu,Z.-Q. and Fan,H.-F: A Computer Program to Derive (3+1)-Dimensional Symmetry Operations from Two-line symbols. **J.Appl.Cryst.** 30:73-78 (1997).

2). KYLIN/3DSCALE
Fu,Z.-Q.: Three-dimensional Model-free Experimental Error Correction of Protein Crystal Diffraction Data with Free-R Test. **Acta Cryst. D**61:1643-1648 (2005)

3). SGXPRO
Fu, Z.-Q., Rose, J. & Wang, B.-C.: 'SGXPro: A Parallel Workflow Engine Enabling Optimization of Program Performance and Automation of Structure Determination'. Acta Cryst. D61:951-959 (2005).

4). KYLIN
Fu,Z.-Q.: KYLIN Data Reduction of Single Crystal Diffraction (to be published).

5). AXDPRO
Fu,Z.-Q.: Automatic Processing of Single Crystal Diffraction Data with Multiple Programs Parallel (to be published).

Eof AXDPRO/KYLIN

§6. SGXPRO for Macro-molecular Structural Solution (obsolete)

Caution: SGXPRO for macro-molecular structural solution hasn't been updated since 2007. Many of its functional modules for partial structural solution, phasing, electron density map tracing, auto building, refinement, validation etc. may need updated. Archived below are just those from old releases. Therefore, it strongly recommended not to use this part before updates are checked and released.

1. On-Line Help

From sgxpro GUI click on 'Help' -> 'OnLine Quick Start', or web link http://www2.ser.aps.anl.gov/sgxpro/sgxpro_readme_openoffice.html, a brief instruction on how to use the SGXPRO suite will be displayed by FireFox web browser. If you are a new user, it would be helpful to take a few minutes reading through these quick start sections.

2. Manual

The 'sgxpro_readme.doc' file inside the sgxpro/UserGuide directory contains a brief manual of SGXPRO suite. This document is in MS Word format. A color print would be more readable.

Please send your suggestions and/or bugs report if any to fuzq@.
Thanks!

3. Notices

<<< NOTICE 1 >>>

SGXPRO supports many popular programs which maybe have already been installed on your computing system. Based on the feedbacks from users, the following program/program suites are needed to take advantages of SGXPRO workflow search engine to quickly find the better solutions expected for the given data sets:

CCP4, COOT (for density&model display, MTZ formatting).
SHELXD (one of the major paths for heavy-atom sites searching).
SOLVE/RESOLVE (one of the major paths for novel structure solution).
EPMR_LINUX (one of the major paths for MR structure solution).

If you do not have these programs installed yet, you have to acquire them seperately from the authors. SGXPRO does not require special installation or configuration of these programs. You just follow the instructions from these packages to install them either before or after the installation of SGXPRO.

If for any reason, an above program is available on the system, but with a different name, for example 'epmr' instead of 'epmr_linux', you may change the name inside "sgxproaddon_linux.config" file for this program. Then, SGXPRO will take it as your preferred later on.

<<< NOTICE 2 >>>
Inside the ‘sgxpro’ directory, there are several executable in the packages:
sgxpro_b32 --- 32-bit Linux PC/Workstation.
sgxpro_b64 --- 64-bit Linux PC/Workstation.
sgxpro_b64_beowulfcluster --- 64-bit Beowulf Linux Cluster.
isascom_b64 --- for 64-bit Linux
isascom_b32 --- for 32-bit Linux
At installation, the appropriate executable will be automatically selected based on the system setup. However, if SGXPRO doesn't properly, you may need to try a different executable for your Linux computing system by:
1). Go to the ‘sgxpro’ directory.

2). Copy the selected executable, for example:
cp sgxpro_beowulfcluster sgxpro
cp isascom_b64 isascom

See section "2. Platforms Supported" in the "sgxpro_readme.doc" file for more details. If your system is a cluster, please see section "4.4 SGXPRO Configuration".

4. History of Updates

<<< Version 1.02 2010-03-18 >>>

1). Updated to work with work with the new PDB sites for template structure searching.

<<< Version 1.01 >>>

1). Upon suggestions, a soft lock is implemented to avoid interruption of the running jobs for users who share a computer in public area.

<<< Version 1.00 >>>

1). Some users reported that few of the tasks may take very long because of bad data in high resolution shells if data have been processed too aggressively.
An internal function module is added to automatically detect and cleanup potential bad tasks on-the-fly.

2). A stand-alone 'sgxjobcheck' is also added to allow user manually check and stop multiple jobs. Type 'sgxjobcheck help' to see how to use it.

3). 'x3d' is renamed as 'spgr4d' to avoid a naming conflict.
4). The 'ssbackup' script is released for superusers/sudoers to backup data on Linux.

<<< Version 0.90 >>>
1). On-Line help is implemented.
2). A brief quick start instruction is compiled.

<<< Version 0.80 >>>
1). PHASER is added as one of the major modules in searching molecular replacement solutions. The other two modules are EPMR and AMORE.
2). More information is added to the '*.summary' file.

<<< Version 0.70 >>>
1). Parallel ShelxD is implemented.
2). Simplified graphical user interface for 'Novel Structure Solution" module is implemented, that allows user to setup parallel workflow engine to search the program and parameter space without going through individual GUIs. When MAD data are loaded, both MAD and SAD phasing are tried.
3). Space group determination tool 'spgr4d' is added.

<<< Version 0.60 >>>
1). Simplified graphical user interface for 'Molecular Replacement Solution" module is implemented, that allows user to do the BLAST search for sequence-base homologue structures and setup molecular replacement jobs from a single page.

2). Tool for molecular weight calculation from sequence.

<<< Version 0.50 >>>
1). Support for Beowulf Linux cluster is implemented.
2). Tool for solvent content estimation.
3). Tool for f'&f" plot.

<<< Version Beta >>>
Fu, Z.-Q., Rose, J. & Wang, B.-C. (2005). 'SGXPro: A Parallel Workflow Engine Enabling Optimization of Program Performance and Automation of Structure Determination'. Acta Cryst. D61:951-959.

5. **Bench Mark as Measured in March 2005**

The current SGXPRO suite runs on both Linux PC and Beowulf Linux cluster. The following are times taken on different computers to solve a novel protein structure with 250AA to resolution of 2.0 angstroms:

6'30" 8-CPU 2.6GHz PC
16'20" 4-CPU 2.6GHz PC
60'00" 2-CPU 2.6GHz PC
12'10" 30-CPU 1.0GHz Linux Cluster

EOF