SIMPLE 2017 Development Document

Responsibilities: Hans, Cyril, Dominika&Hans, Alex&Hans, Susan&Hans

Urgent 2dos (before workshop)

* should be possible to start refine=shc without oritab (exhaustive projection matching), needed for heterogeneity
* Restart mode for workflows, i.e. keep a status file that can replace the command line when given startit, should be possible to start from any previous iteration without having to give command line directives
* Need to deal with state=0 in heterogeneous processing, autoshrink number of states & bookkeep (evaluate whether to use native numbering or update to sequential)
* Consistency in output naming/directory structure in the distributed workflows and neat and tidy cwd following convergence
* add rejected dir to simple\_select and give possibility to input desired select/reject dirs.
* Autogenerate simple\_distr\_config.env file based on environment variables if it is not in cwd
* mailx -s "mail from prime2D" hans.elmlund@monash.edu < from\_prime2D
* Postprocessing of volumes in the distributed workflows
* Not only eo option in postproc vols, also lp
* Exception when partial stacks of wrong ldim
* Add cenlp option to relevant command lines
* Make unblur commander stack the powerspectra (SP mode)
* Docgen for distr\_exec

After workshop

2dos

* every time we execute a simple command line (by itself or from within the distributed environment) the command line should be appended to a simple.log file with a time-stamp
* higher level automodes for workflows (stitching workflows together, for example, chain together extract, 2D, 3D)
* It should be possible to set maxp (maximum number of particles in a class) and autosplit when pop becomes > maxp
* Clever command line: any set of parameters should be possible to input in the environment file and not asked for upon execution (but command line should be able to override)
* when automsk is turned on in 2D we must force the next round to search all refs, or it will go to shit
* class for memory allocation simple\_alloc. Keep track of mem exceptions and profile memory (hash table).
* Go over all random number generations and see if we can use matrices instead
* Replace integer rnaodm number generator with instrinsic one (Guide to Fortran 2008)

GPU

1. UNBLUR
2. PRIME2D

3) PRIME3D

Need support for the CTF formalism on GPU.

How do we deal with memorize sqsums with CTF on GPU?

How do we deal with leftover in chunk-based parallelisation?

Is chunk-based GPU execution the way forward?

If so, how does this splitter deal with nstates?

Tests

* Juha's code (icosahedrality of individual particle images)
* test that the Wiener filter works for initial model generation from particles
* test so that we can run PRIME2D/PRIME3D in all CTF/refine/eo modes
* evaluate stochastic momentum on the corr level
* evaluate deterministic momentum on the rec level. This is basically what Relion does with the slidebar where you select how much weight you want give to the reference vs. the data. This can be useful when you have “problematic data” (for example strong view preference)
* Need to test the knn structure 4 2D (million ribosomes)

Streaming

1. UNBLUR
2. asynchronous cavg/vol assemble (to reduce latency between distributed runs)
3. PRIME2D

Ideas

* Create a class scoring function based on the sum of the integrated weights of the class they belong to (minimum population threshold must apply). Correlate with the visual assessment of the cavg quality.
* COMPICK: code the common lines based idea for particle picking
* COMSEL: code the common lines based idea for cavg selection for ab initio rec

New developments

* Probabilistic SO(3) scatter search for high-resolution refinement
* Quasi-continuous sorting of conformational states (affinity propagation, shell-weighting or whatever works)
* implement tilt test

Exception handling

* exception handling class that makes more sensible outputs (especially when running the code in distributed mode). We need to create a database of exceptions and then have ONE control point where we check the stack for errors and report what has failed (similar to the command line dictionary)
* refs should be included in the mixed formats check
* when nr of commands are not sufficient the error message should tell what is missing
* when eo=yes is set the program should ask for fsc-file if missing & instruct better

Considerations

Using the strategy pattern for the different modes of refinement?

Known bugs

* simple\_qsys\_ctrl :: submit\_scripts; command execution failed with error Termination status of the command-language interpreter cannot be obtained. This means that when we execute a command line instruction from within Fortran we probably need to check that the exit status is healthy or iterate the command until it becomes successful and break out after a fixed number of iterations.
* merge\_rmat\_from\_parts sometimes gives EOF error when reading shellweights. I first thought that there might be a race condition, i.e. that the merging started despite all shellweights hadn’t been written, but this is impossible in the 3D case as the shellweight generation is decoupled from the search and in 2D you always merge the shellweights from the previous iteration—so it may turn out to be just flaky disks on the cluster.

Compilation

* port to ifort (Intel compiler)
* port to PGI (Portland group, with CUDA-FORTAN)

Documentation

Need to get a html code doc generator in place and fix the doc of every class. Will FORD source code documentation provide the solution?

<http://fortranwiki.org/fortran/show/FORD>

<https://github.com/cmacmackin/ford>

<http://jacobwilliams.github.io/json-fortran/index.html>

Books/webpages

* Structured parallel programming
* Structure and Interpretation of Computer Programs
* J-P Morrison. Flow-Based Programming: A New Approach to Application Developments. CreateSpace, 2nd ed.
* Mathematical Foundations of Imaging, Tomography and Wavefield Inversion
* Geometric Algebra for Computer Science (Revised Edition): An Object-Oriented Approach to Geometry (The Morgan Kaufmann Series in Computer Graphics) 1st Edition
* The Princeton Companion to Applied Mathematics
* Practical Machine Learning: http://www.computervisionmodels.com/
* Applied Stochastic Modelling, Second Edition (Chapman & Hall/CRC Texts in Statistical Science) 2nd Edition http://szeliski.org/Book/

Considerations

Should we implement a resolution update scheme for prime2D similar to that in the Unblur code?

Should we declare the params variables as protected?