EMBO Practical Course on Image Processing for Cryo EM
1 – 11 September 2015

Practical 6: Heterogeneity analysis, part 2

Introduction

If the structure of the initial complex is known, it can be used to generate reprojections that can be used in analysis of the data.

2D statistical analysis can also be used to reveal heterogeneity and separate mixed populations with conformational changes triggered by ligand binding. This approach needs preliminary alignment to reveal the differences in images caused by various orientations from differences produced by variable binding of ligands. Multi-Reference-alignment allows minimisation of orientation variations between complexes composing classes and increases the weight of localized changes induced by partial ligand occupancy. Multivariate Statistical analysis requires two successive classifications, the first based on eigenimages showing global variance of structural features due to different orientations, that we call “orientational” classes. This is followed by a second classification based on eigenimages showing localized variance arising from changes induced by partial ligand occupancy, that are termed “structural” classes. One the other hand having two models form the same data set can be separated by competitive alignment. In this method, each 2D EM projection data is aligned with all 2D reference projections of initial maps (whose number can be more as two). In addition to orientation parameters, the values of similarity measures are analyzed and the data is assigned to the group of the most similar 3D model. The classification can check the quality of separation and the orientation of the classes can be defined as it has been described previously.

As soon the reconstructions are obtained the measurement of resolution should quantify the level of reliable detail detectable in final maps. In practice, the detectability of features at a given resolution is determined by the SNR in that frequency range of the data. To quantify resolution, the SNR must be estimated as a function of spatial frequency. In single-particle and tomography data, both signal and noise are distributed over the whole spectrum, and there is no simple way to estimate the resolution. The most widely used method for determining the resolution of a single-particle reconstruction is the Fourier ring (in 2D) or Fourier shell correlation (FSC). The data set is split into two equivalent halves, usually by separating odd- and even-numbered images from the data stack. Separate reconstructions are calculated from the two halves, and their 3DFTs (F1, F2) are compared by cross-correlation in spatial frequency shells (f, f+Δf). The average correlation for each shell is plotted and typically shows a falloff from a correlation of 1 at low resolution down to 0 at high resolution. The spatial frequency at the 0.5 correlation threshold is commonly taken as the resolution estimate, but other criteria, for example, comparison to the noise level, or 0.143 have been proposed on the basis of SNR estimates. Another method, first proposed for 2D averages and subsequently extended to 3D structures, is the spectral signal-to-noise ratio (SSNR). In this case, the signal is estimated to be the reprojections of the map, and the noise is estimated by taking the difference between input images and the corresponding reprojections. This approach does not require the data set to be divided into halves. Like the FSC, the SSNR requires the aligned input images. It should also be noted that a good resolution value does not guarantee that the map is correct.
A further method has been proposed, the R-measure. It does not require the input data, but uses the final map itself, along with the surrounding region of the reconstruction outside the particle, for the resolution estimation. This method examines the correlations between adjacent

pixels in the FT of the reconstruction. For a map containing pure noise, adjacent transform pixels are uncorrelated.

Further reading:

1. Elad N, Clare DK, Saibil HR, Orlova EV. Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. J Struct Biol. 2008;162:108–120.
2. Henderson R. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. Proc Natl Acad Sci U S A. 2013 Nov 5;110(45):18037-41.
3. Henderson, R. Realizing the potential of electron cryo-microscopy. Q. Rev. Biophys. 2004, 37, 3.
4. Herman GT, Kalinowski M. Classification of heterogeneous electron microscopic projections into homogeneous subsets. Ultramicroscopy. 2008
5. Penczek PA. Variance in three-dimensional reconstructions from projections. In: Unser M, Liang ZP, editors. Proceedings of the IEEE International Symposium on Biomedical Imaging; July 7-10; Washington, DC. 2002. pp. 749–752.
6. Rosenthal, P. B.; Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J. Mol. Biol. 2003, 333, 721.
7. Saxton, W. O.; Baumeister, W. The correlation averaging of a regularly arranged bacterial cell envelope protein J. Microsc. 1982, 127, 127.
8. Scheres SH, Gao H, Valle M, Herman GT, Eggermont PP, Frank J, Carazo JM. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nature Methods. 2007;4:27–29
9. Sigworth FJ. A maximum-likelihood approach to single-particle image refinement. Journal of Structural Biology. 1998;122:328–339.
10. Sousa, D.; Grigorieff, N. Ab initio resolution measurement for single particle structures. J. Struct. Biol. 2007, 157, 201.
11. Spahn C.M., Penczek P.A. Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM. Curr Opin Struct Biol. 2009 Oct;19(5):623-31
12. Unser, M.; Sorzano, C. O.; Thevenaz, P.; Jonic, S.; El-Bez, C.; De Carlo, S.; Conway, J. F.; Trus, B. L. Spectral signal-to-noise ratio and resolution assessment of 3D reconstructions. J. Struct. Biol. 2005, 149, 243.
13. Unser, M.; Trus, B. L.; Steven, A. C. A new resolution criterion based on spectral signal-to-noise ratios. Ultramicroscopy 1987, 23, 39.
14. van Heel, M.; Schatz, M. Fourier shell correlation threshold criteria J. Struct. Biol. 2005, 151, 250.
15. White HE, Saibil HR, Ignatiou A, Orlova EV. Recognition and separation of single particles with size variation by statistical analysis of their images. Journal of Molecular Biology. 2004;336:453–460.

Login to the server using the instructions provided previously. Open a terminal window in your directory:

> cd PRAC-5-6

In this practical you will do the refinement of the two structures and will assess the resolution of them.

Please compare the masked reconstructions **3d_extra_den_a2_300_50m and 3d_no_den_a2_300_100m,** sections 110-160 using **DISPLAY**. These masked structures will be used to improve the separation of the mixed population. The masked reconstructions from the best 50 and 100 projections are used to make two **separate** anchor sets that will be used to refine angles of each set of classes. For that we will use two job batches: one for the structure with **an extra blob** and the other **without** it:

First of all processing of the data with the extra blob:

job6_r2_3d_ref_ext_den.b > & job6_r2_3d_ref_ext_den.log &

Time is ~ 25 min

These jobs do the following:

1. Calculation projections of **3d_extra_den_a2_300_50m** to obtain the new anchor set **anch_set_extra_den_0**
2. Run **EULER** for **clsum_extra_den_a2_300** using **anch_set_extra_den_0**
3. Calculate a new reconstruction **3d_extra_den_a2_300_n2** using the command **TRUE-THREE**
4. Sort classes according to the 3D errors and to extract the best first 100 into **clsum_extra_den_a2_300_n2_100** using **EXCLUSIVE-COPY** command.
5. Calculate a new reconstruction **3d_extra_den_a2_300_n2_100** using the command **TRUE-THREE** and selected classes **clsum_extra_den_a2_300_n2_100.** The output will be **3d_extra_den_a2_300_n2_100**

While you will look at this reconstruction using **DISPLAY** start the other job:
job6_r2_3d_ref_no_den.b > & job6_r2_3d_ref_no_den.log &

It should take ~ 25 min as well.

It will do the same processing as the previous job but will produce the structure from the data set without the blob. It will be

The steps of procedure are the same:

1. Calculation projections of **3d_no_den_a2_300_100m** to obtain the new anchor set **anch_set_no_den_1**
2. Run **Euler** for **clsum_no_den_a2_300** using **anch_set_no_den_1**
3. calculate a new reconstruction **3d_no_den_a2_300_n2** using the command **TRUE-THREE**
4. Sort classes according to the 3D errors and to extract the best first 100 using **EXCO-IM** command into file **clsum_no_den_a2_300_n2_100**
5. Calculate a new reconstruction **3d_no_den_a2_300_n2_100** using the command **TRUE-THREE** and selected classes **clsum_no_den_a2_300_n2_100.** The output will be **3d_no_den_a2_300_n2_100**

The structures have to be masked **using THREE-AUTO-MASKING** as was done in the previous practical with the same parameters to get masked structures for both complexes
.

Competitive alignment using re-projections of the previous 3Ds as references

You need to run a new **MRA (multi-reference alignment**) using re-projections of the two obtained models as references. The projections from the model with the ExtBlob have to be appended to the projections from the model with the NoBlob. Then the **M-R-A** should be run. The images that correspond to the first set of references (the highest correlation coefficients) will be selected in one subset (NoBlob) and the other that have a better correlation with the second part will be extracted to the second subset (ExtBlob)

To do this you have to run the batch **job_7_all_sort_r3.b** that does the folowing:

Input -> *output*

1. **THREE-FORWARD**: **3d_no_den_a2_300_n2_100m** -> **ref_no_den_a2_300_n2_100m**
2. **THREE-FORWARD**: **3d_ext_den_a2_300_n2_100m**-> **ref_ext_den_a2_300_n3_100m**
3. **COPY-IM:** **ref_no_den_a2_300_n2_100m** -> **new_ref_set**
4. **APP-IM:** **ref_ext_den_a2_300_n3_100m** -> **new_ref_set**
5. **NORM-VAR:** **new_ref_set** -> **new_ref_set_nv**
6. **M_R_A** **E1_data_a2** -> **E1_data_3** *using both translational and rotational alignment with **new_ref_set_nv** as references. After the refinement an aligned data set* **E1_data_3** *will been obtained from which we can extract images corresponding to the different types of references* (Competitive alignment, see Elad et al., 2008, JSB)
7. **ALISELECT** **E1_data_3 -> E1_no_ext** corresponding to references 1-80
8. **ALISELECT** **E1_data_3 -> E1_ext_den** corresponding to references 81-160

However since this job runs only on 5 processors it is time consuming (**~ 3 hours**), so please use the files with aligned and extracted images **E1_no_ext.* and E1_ext_den.***

Check with the command **HEAD** how many images you have got in each subset:

IMAGIC-COMMAND : **head**

```
Options available:
  READ/LOOK  WRITE/SET  HOW_MANY   PLT_OUT    TAKE_OVER
  SHRINK     SORT       COMPARE    HISTOGRAM  MEANINGS
Please specify option [LOOK]                        : how
Input (header) file, loc#s [my_img]                 : E1_ext_den
```

Number of 2D images in file : **XXXX** – *here you will see a number corresponding to the number of images in the file.*

Please check the number of images in the second file **E1_no_ext**
Each data set has to be subjected to statistical analysis to check how well the separation has been done. You have to run

./job_8_msa_wbl_nbl.b > & job_8_msa_wbl_nbl.log &

It takes ~8 min

The job does **MSA, MSA-CLASS, and MSA-SUM** of the subsets and creates two sets of class sums: **clsum_a3_no_ext_400** and **clsum_a3 ext_den_400,** where 400 indicated that we have obtained 400 classes for each subset of data.

The next step will be determination of orientation of these classes:
./ job_9_head_3d_a3.b > & job_9_head_3d_a3.log &

and calculation of reconstructions. This will result in two reconstructions:

3d_a3_no_ext_400_300 and **3d_a3_ext_den_400_300**

The reconstructions were obtained from the best 300 classes out of 400.

Auto masking (step10, *manual*)

IMAGIC-COMMAND : **thr-auto-m**

```
Auto-masking options:
  DO_IT_ALL  DO_IT_ALL  REFINE_THRESHOLD
Please specify option [DO_IT_ALL]                    : DO_IT_ALL
Input 3D volume file [3d_no_den_a2_300_n2_100]       : 3d_a3_no_ext_400_300
Output file containing masked input 3D
[3d_no_den_a2_300_n2_100m]                           : 3d_a3_no_ext_400_300m
Output modulation/variance volume [my_mod_varian]    :
Output file containing 3D mask
[3d_mask_eno_den_m2]                                 : 3d_a3_no_ext_mask
Masking based on local modulation [YES]              :
Band-pass parameters (pixels/fraction) [0.06,0.2]    :
Low-pass parameter (pixels/fraction) [0.04]          :

Threshold options:
  MANUAL  AUTOMATIC
Please specify option [AUTOMATIC]                    : AUTOMATIC
Auto-threshold percentage [5]                        : 6

Image name: 3-D RECONSTRUCTION
Size: 200, 200  Loc: 1  Type: REAL  Cre.Date: 23-Aug-2015  Time: 22:29:57
TRUE_3D RECONSTRUCTION/EXACT-FILTERED BACK-PROJ.

 Number of 3D volumes in input file: 1

 The results are stored in the following files:

 3D modulation/variance volume      : my_mod_varian
 Output plot file                   : my_mod_varian_histogram.plt
 3D mask (based on local modulation): 3d_a3_no_ext_mask
 Masked 3D volume                   : 3d_a3_no_ext_400_300m
```

Do the same for **3d_a3_ext_den_400_300** map with the same parameters..

You will get two masked maps **3d_a3_no_ext_400_300m** and **3d_a3_ext_den_400_300** and two three-dimensional masks **3d_a3_no_ext_mask and 3d_a3_ext_den_mask**

These masks are **BINARY:** they have densities **= 1** in the area of the complex and zero in the outside areas. However, for the next step of assessment of resolution you need the masks with

soft edges. To soften edges you have to use the command **THREED-BLOCK-CONVOLUTION-FILTER**

```
IMAGIC-COMMAND : THR-BLOCK-CON-FIL
Input 3D file, no loc#s                          : 3d_a3_ext_den_mask
Output 3D file                                   : 3d_a3_ext_den_mask_sm

Mode of (real space) block convolution:
  LOWPASS  HIGHPASS
Please specify option [LOWPASS]                  :
Linear size of convolution area (3,5,7...) [3]   : 5
```

Do the same for the other mask

```
IMAGIC-COMMAND : THREED-BLOCK-CONVOLUTION-FILTER

Input 3D file, no loc#s [3d_a3_ext_den_mask]     : 3d_a3_no_ext_mask
Output 3D file [3d_a3_ext_den_mask_sm]           : 3d_a3_no_ext_mask_sm

Mode of (real space) block convolution:
  LOWPASS  HIGHPASS
Please specify option [LOWPASS]                  :
Linear size of convolution area (3,5,7...) [5]   : 5
```

Now we can do the next step

Assessment of the structure resolution

To find out the map resolution means to know which details in the structures are stable. For that the dataset should be split into two (odd and even) in order to have two reconstructions that can be compared. First we sort the **images** according to the 3d-error so that we will then extract two sets of equivalent quality and then to calculate two structures

The script **job_11_3d_no_ext_150_den.b** does the following steps: separation 300 classes into two subsets and calculate two reconstructions without blob:

```
echo "! "
echo "! IMAGIC program: excopy ------------------------------------------
--"
echo "! "
/s/emib/s/imagic/150710/incore/excopy.e <<EOF
2D_IMAGES/SECTIONS
EXTRACT
clsum_a3_no_ext_400_300 -> best 300 images sorted according to three-errors
clsum_a3_no_ext_150a    -> extracted odd images
INT
1-300
INC
2
EOF
echo "! "
echo "! IMAGIC program: excopy ------------------------------------------
--"
echo "! "
/s/emib/s/imagic/150710/incore/excopy.e <<EOF
```

```
2D_IMAGES/SECTIONS
EXTRACT
clsum_a3_no_ext_400_300 -> best 300 images sorted according to three-errors
clsum_a3_no_ext_150b    -> extracted even images
INT
1-300
EVEN
EOF
echo "! "
echo "! IMAGIC program: true_3d ------------------------------------------
--"
echo "! "
/s/emib/s/imagic/150710/openmpi/bin/mpirun    -np    5    -x    IMAGIC_BATCH
/s/emib/s/imagic/150710/threed/true_3d.e_mpi <<EOF
BOTH
5
ALL
C1
YES
clsum_a3_no_ext_150a
ANGREC_HEADER_VALUES
YES
3d_a3_no_ext_150a       -> reconstruction A from 150 images
repr
err
YES
0.75
NO
0.999
0.6
EOF
echo "! "
echo "! IMAGIC program: true_3d ------------------------------------------
--"
echo "! "
/s/emib/s/imagic/150710/openmpi/bin/mpirun    -np    5    -x    IMAGIC_BATCH
/s/emib/s/imagic/150710/threed/true_3d.e_mpi <<EOF
BOTH
5
ALL
C1
YES
clsum_a3_no_ext_150b
ANGREC_HEADER_VALUES
YES
3d_a3_no_ext_150b  -> reconstruction B from 150 images
repr
err
YES
0.75
NO
0.999
0.6
EOF
```

The next job will separate the 300 classes with extra blob into two files: even and odd subsets and calculate from subsets two reconstructions

./job_12_3d_a3_ ext_den_150.b > & job_12_3d_a3_ ext_den_150.log &

The output of two jobs **11** and **12** will be four reconstructions**:**

3d_a3_ext_den_150a
3d_a3_ext_den_150b
3d_a3_no_ext_150a
3d_a3_no_ext_150b

Now we have to mask them with the soft masks using pre-prepared 3D-mask:
3d_a3_ext_den_mask_sm
3d_a3_no_ext_mask_sm

Now you have to run the next job that calculates three dimensional Fourier shell correlation between two masked (with soft edges) reconstructions with the blob.

./job_13_fsc_ext_den_150.b >& job_13_fsc_ext_den_150.log &

The next job will calculate Fourier shell correlation between two masked (with soft edges) reconstructions without the blob.

./job_13_fsc_no_ext.b >& job_13_fsc_no_ext.log &

The resolution of the model can be determined by displaying the Fourier-Shell-Correlation curve with the imagic command **PLOT** to display files **no_ext_fsc.plt** and **ext_den_fsc.plt**. You can also read the values inside the text file using NEDIT

```
 Correlation   Frequency
…….
 0.715941      0.531250E-01         1
 0.626254      0.562500E-01         1
 0.543738      0.593750E-01         1 the threshold of 0.5 for the CCC
 0.530795      0.625000E-01         1 is between these two steps, so the
 0.486770      0.656250E-01         1 corresponding frequency is in between
 0.348930      0.687500E-01         1 them. That is why we take 0.063
 0.236110      0.718750E-01         1
 0.143440      0.750000E-01         1
 0.152714      0.781250E-01         1
…….
```

The resolution can be assessed as 1/Frequency = 1/ 0.063 = ?

Please do the same for **ext_den_fsc.plt**

```
Correlation   Frequency
…….

0.795986      0.531250E-01           1
0.751779      0.562500E-01           1
0.713564      0.593750E-01           1
```

```
0.614754      0.625000E-01          1 The same as above
0.463122      0.656250E-01          1
0.305121      0.687500E-01          1
0.275490      0.718750E-01          1
0.267325      0.750000E-01          1
0.218504      0.781250E-01          1
…….
```

The resolution can be assessed as 1/Frequency = 1/ 0.0538 = ?

Make a surface representation of your model. The threshold 3D density value can be estimated by using the sigma value found by the **SURVEY** command (option **Global**) same as you did for the 3d_1. Then run three-surf to find out the molecular mass that corresponds to that particular sigma value. If the mass corresponding to this volume is too big, raise the threshold and check the surface calculation again. The molecular mass of the portal protein is ~ **450** kDa. To look at the surface use the command **DISPLAY**

To check what we have obtained (without rounds of refinement) we can do the following;

```
IMAGIC-COMMAND : THREE-SUBTRACT

Mode of operation                                    : TWO_SUB
1st input file, 3D loc#s [3d_a3_ext_den_150b]        : 3d_a3_ext_den_400_300m
2nd input file, NO loc#s [3d_a3_ext_den_mask_sm]     : 3d_a3_no_ext_400_300m
Output file, NO loc#s [3d_a3_ext_den_150bm]          : 3d_the_blob

Image name: 3-D RECONSTRUCTION
Size: 200, 200  Loc: 1  Type: REAL  Cre.Date: 24-Aug-2015  Time: 18:16:35
TRUE_3D RECONSTRUCTION/EXACT-FILTERED BACK-PROJ.
```

The next step to asses if there are any differences quickly is to calculate orthogonal projections of the difference:

```
IMAGIC-COMMAND : thr-forw

Options available                                    : FORWARD
Input 3D image file [3d_ext_den_a2_300_n3_100m]      : 3d_the_blob
Output file for forward projections [jnk]            : ort_blob
Threshold 3D density value [-99999]                  :
Use default interpolation (WIDENING) [YES]           :

Choose Euler angles option:
  PLT_FILE              ANOTHER_HEADER_FILE
  INTERACTIVE           ASYM_TRIANGLE
  RANDOM                UNIFORM
  STEREO                TOMOGRAPHY
  CONICAL               ORTHOGONAL
  SPIRAL                TETRAHEDRON
  ICOSAHEDRON
Please specify option [ASYM]                         : ort

Full output of all parameters [YES]                  :

Use MPI parallelisation [YES]                        : no

Now display the projections using DISPLAY, what will you see?
```

Then use the command **MOVIE:**

```
IMAGIC-COMMAND : movie

Input file, image loc#s [surf]                    : surf
Mode of greyvalue scaling:
  INTERACTIVE  HEADER  LOCAL_SURVEY  GLOBAL_SURVEY
Please specify option [GLOBAL_SURVEY]             : GLOBAL
Time between two images in seconds [.5]           : 0.5
Do you need help how to control MOVIE [NO]        : yes
```

You could make a movie using option **TOM**ogaraphy, number of views **54, beta = -180,180, gamma = 12.6,** the output file name **movie_Blob**. To see the movie use the command **MOVIE** with the input file **movie_Blob**

Another possibility is to convert the reconstruction files **3d*** into the **SPIDER** or **CCP4** format by the command **EM2EM** and to display it via **WEB, Chimera or Pymol.**

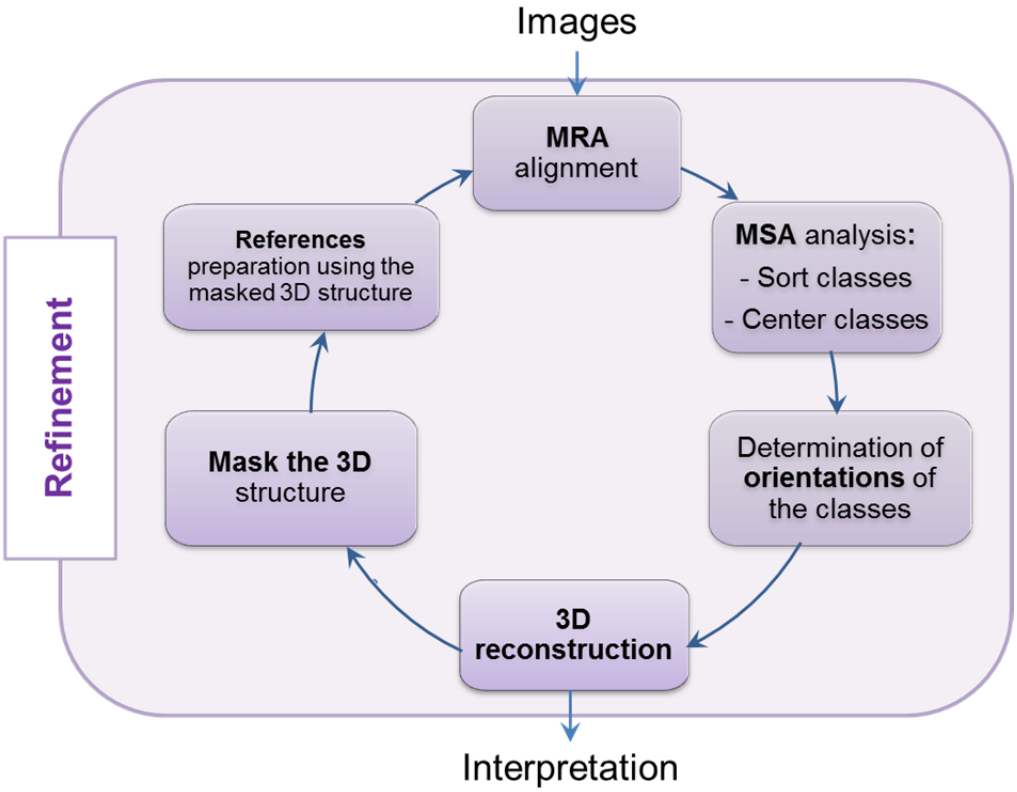
To visualize the reconstruction in software such as Pymol and Chimera the format of the file can be converted from IMAGIC to CCP4. This can be done with the command **EM2EM**

```
IMAGIC-COMMAND : em2em

Data format of the input to be converted:
  BROOKHAVEN_STEM     CCP4                DATA_ONLY
  DIGITAL_MICROGRAPH  EM                  FABOSA
  FORMATTED           GIF                 IMAGIC
  KONTRON             MDPP                MRC
  OFFSET              PIC                 PIF
  PGM                 RAW_IMAGE           RAWIV
  SEMPER              SHF                 SPIDER
  SUPRIM              TIFF                TVIPS
  VOLUMETRIC
Please specify option                     IMAGIC
Export to which data format:
  BROOKHAVEN_STEM     CCP4                DATA_ONLY
  DIGITAL_MICROGRAPH  EM                  FORMATTED
  GIF                 IMAGIC              KONTRON
  MDPP                MRC                 OFFSET
  PIC                 PIF                 PGM
  POSTSCRIPT          RAW_IMAGE           RAWIV
  SEMPER              SHF                 SPIDER
  SUPRIM              TIFF                TVIPS
  X-PLOR
Please specify option                     CCP4
Input 2D image(s) or a 3D volume:
  2D  3D
Please specify option                     3D
Input image file                          EM structure_in IMAGIC
Output image file                         EM structure.ccp4
Pixel spacing in Angstroem (X,Y,Z)        1.6
Consider different coordinate systems     YES
```

Now you can display your results in CHIMERA

Refinement of structures



The structure can be refined if the whole round of processing is repeated again with the improved model obtained from the best images (based on 3d_error), but they have to be well distributed on the Euler sphere (evenly distributed around equator, since it is important to have side views for the reconstruction). From this step, one can use a projection matching procedure to determine orientations of original images in **IMAGIC, EMAN, SPIDER,** and **Relion**

Flow chart of the second practical , IMAGIC