BLEND: managing, scaling and merging multiple datasets

James Foadi[(a)] and Pierre Aller[(b)]

(a) Imperial College London (j.foadi@imperial.ac.uk)
Diamond Light Source Ltd (james.foadi@diamond.ac.uk)
(b) Diamond Light Source Ltd (pierre.aller@diamond.ac.uk)

Main aim of this tutorials is to get users acquainted with the program BLEND (Foadi et al, 2013) and to show how it can be used effectively to obtain complete data sets out of several partial or complete ones. The reader interested in exploring pros and cons of using data from multiple crystals can refer, for instance, to Liu et al. (2011), Giordano et al. (2012), Axford et al. (2012), Hanson et al. (2012).

DATA SETS INCLUDED WITH THIS TUTORIALS

There are two groups of data included in the sub-directory called "data". One group from crystals of insulin and the second from crystals of lysozyme. It is assumed that the "blend_tutorials.tgz" file has been unzipped and untarred in your $HOME directory. The file can be unzipped and untarred anywhere, but appropriate modifications will have to be applied to the paths printed in this document. If you type:

```
ls $HOME/blend_tutorial
```

you should get the following:

```
BLEND_tutorial_guide.pdf  data
```

and, by typing,

```
ls $HOME/blend_tutorial/data
```

you should get:

```
insulin  lysozyme
```

The first tutorial will be illustrated with and without use of CCP4 interface; for the other only the interface will be used and the user should have no problem re-running them without interface.

TUTORIAL 1.

14 DATA SETS OF CRYO-COOLED INSULIN CRYSTALS
(data collected by Liz Carpenter)
Easy tutorial to get you started with BLEND

a) Using CCP4I

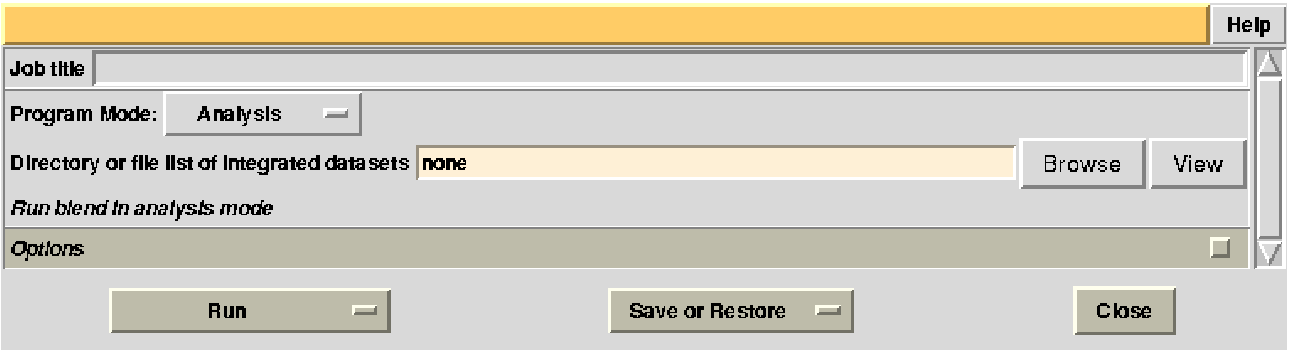
First we need to create a project directory. The CCP4 interface associates a project directory to an already-existing physical directory in your computer file system. We need to create a directory first; let's call this directory "insulin" (we are assuming you are in the blend_tutorials directory):

```
mkdir insulin
```

Next we need to associate a project directory called “BLEND_TUTORIAL_INSULIN” to the insulin directory just created. You can do this using “Directories & ProjectDir” in ccp4i. Now that the project directory has been created let us change to the “BLEND_TUTORIAL_INSULIN” one using “Change Project” in ccp4i. Everything is now ready to start practising with BLEND.

The insulin data have been collected at the Diamond Light Source synchrotron and integrated using *MOSFLM*. Angular rotation range varies from data set to data set, as these have been collected from different crystals. Please, explore some of these integrated data clicking “View Any File” in ccp4i and selecting the “mtz MTZ” field in “File type”.

In the CCP4 GUI, BLEND can be found either in the “Data Reduction and Analysis” or the “Program List” section on the left. When started, BLEND section looks like the following:


Job title

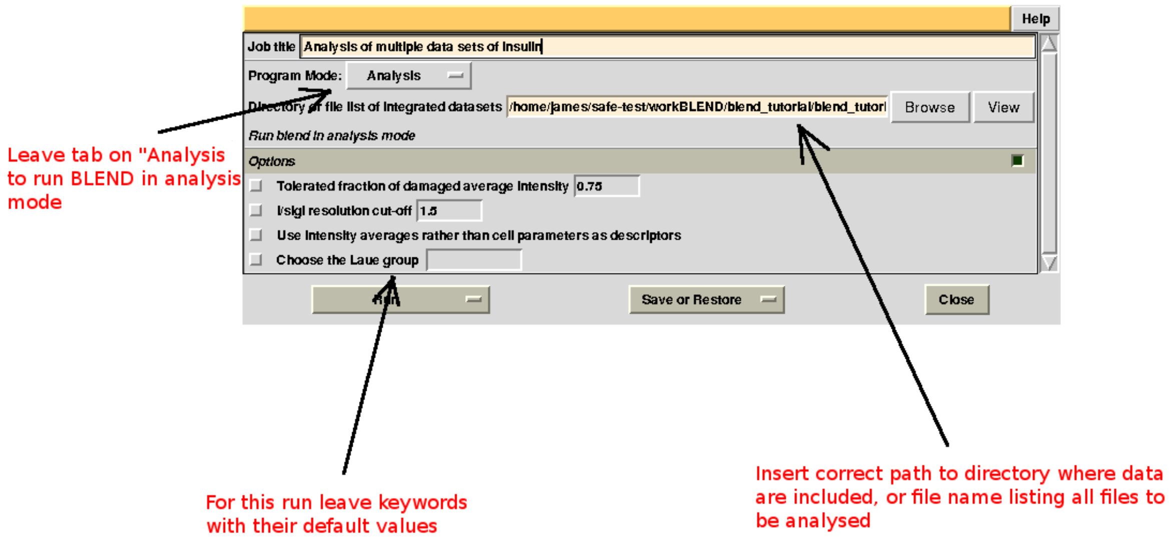
Program Mode: Analysis

Directory or file list of integrated datasets: none

Run blend in analysis mode

Options □

To begin with we run BLEND in analysis mode, with the main goal of creating clusters of data that can, potentially, merge well together. After having filled the blank fields with the appropriate paths and keywords, the above interface will resemble this:


Job title: Analysis of multiple data sets of insulin

Program Mode: Analysis

Directory or file list of integrated datasets: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Run blend in analysis mode

Options ■

□ Tolerated fraction of damaged average intensity 0.75

□ I/sigI resolution cut-off 1.5

□ Use intensity averages rather than cell parameters as descriptors

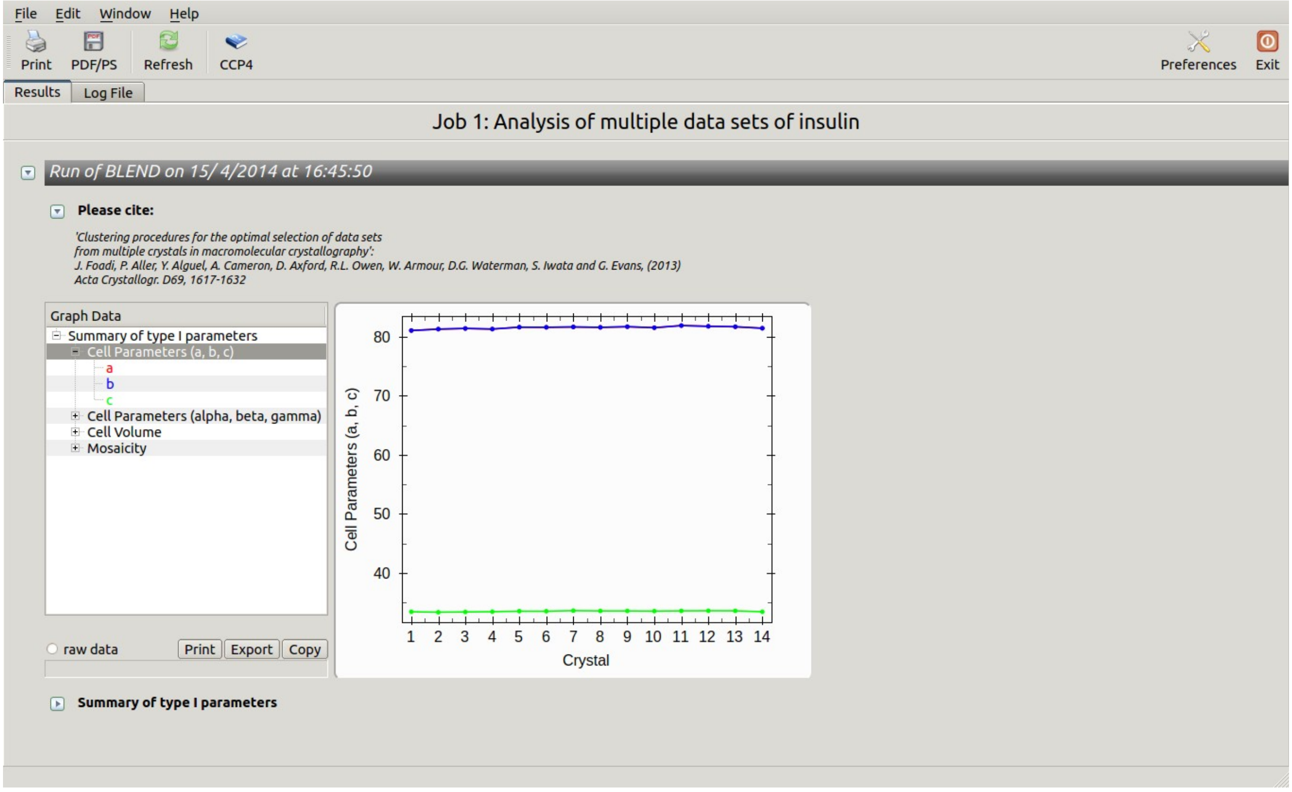
□ Choose the Laue group

Leave tab on "Analysis to run BLEND in analysis mode

For this run leave keywords with their default values

Insert correct path to directory where data are included, or file name listing all files to be analysed

All keywords with their meaning are explained in BLEND documentation. After the run in analysis mode, we can look at the output using ccp4i “View Files from Job → View Job Results (new style)”:



Job 1: Analysis of multiple data sets of insulin

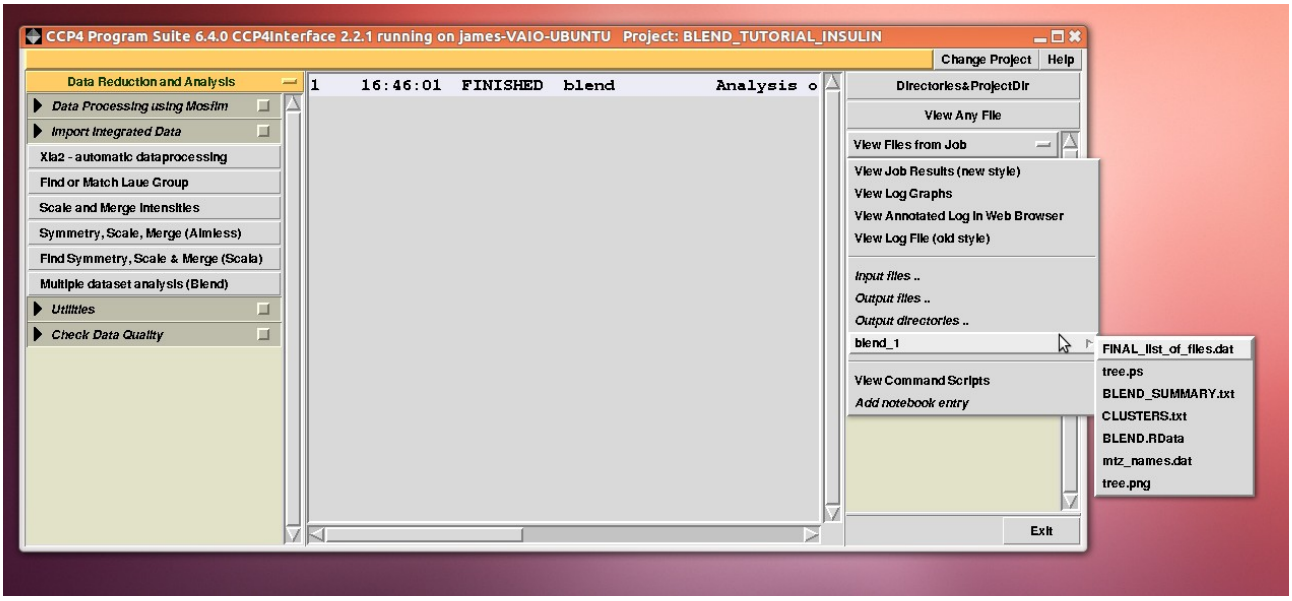
Run of BLEND on 15/ 4/2014 at 16:45:50

Please cite:

'Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography':
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Summary of type I parameters

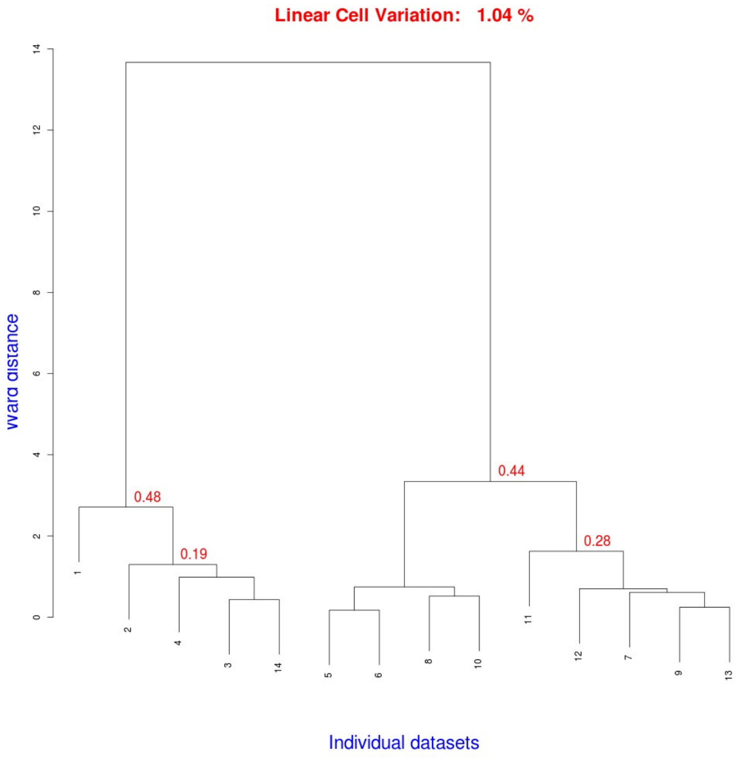
Cell sides, cell angles, cell volume and mosaicity can be viewed for all the 14 data sets included in this insulin tutorial. Numerical values can be explored within the same view. More files have been produced by this run of BLEND in analysis mode. These are listed by selecting “View Files from Job” → “blend_1”:



CCP4 Program Suite 6.4.0 CCP4Interface 2.2.1 running on james-VAIO-UBUNTU Project: BLEND_TUTORIAL_INSULIN

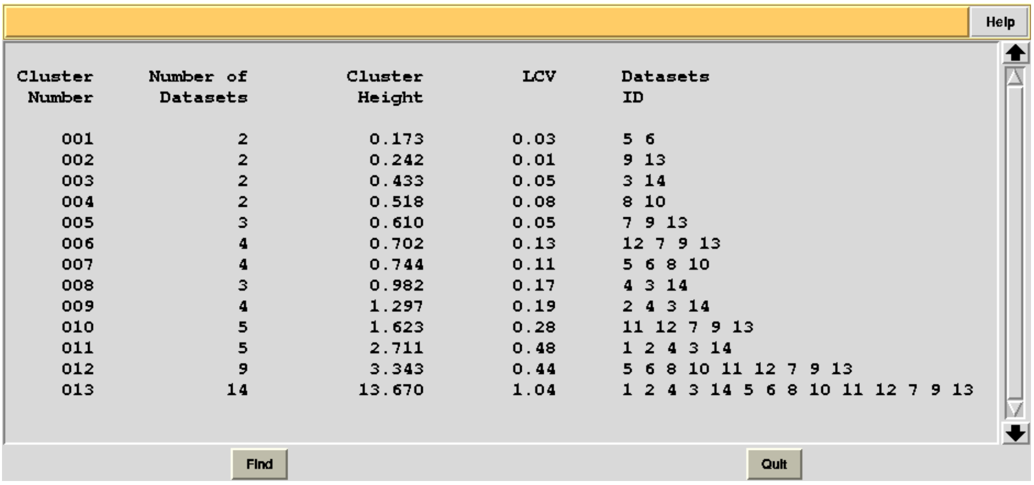
```
1    16:46:01  FINISHED  blend          Analysis o
```

In fact “blend_1” is a directory under the “insulin” directory, containing all files produced by this run of BLEND. The most important among these files is possibly the dendrogram (file “tree.ps” and file “tree.png”). Both the PNG and PS files display the same dendrogram. They can be viewed using any graphical tool, but, at present, only the PS file can be visualised using the CCP4 GUI:



The five numbers in red are the Linear Cell Variation (LCV) values for the top five clusters. They give an indication of cell similarity among all crystals included in the specific cluster; thus, ultimately, they can be associated to isomorphism between different data sets. In the above dendrogram we can see two clusters with similar cell variability (0.44 and 0.48), while such variability is more than doubled when these clusters merge into the overall cluster containing all 14 data sets; this is probably an indication of some minor form of non-isomorphism. Tests done so far with BLEND show that some structural dissimilarities can be noticed for values of LCV higher than 1 – 1.5 %.

LCV values for all clusters are reported in the file “CLUSTERS.txt”:

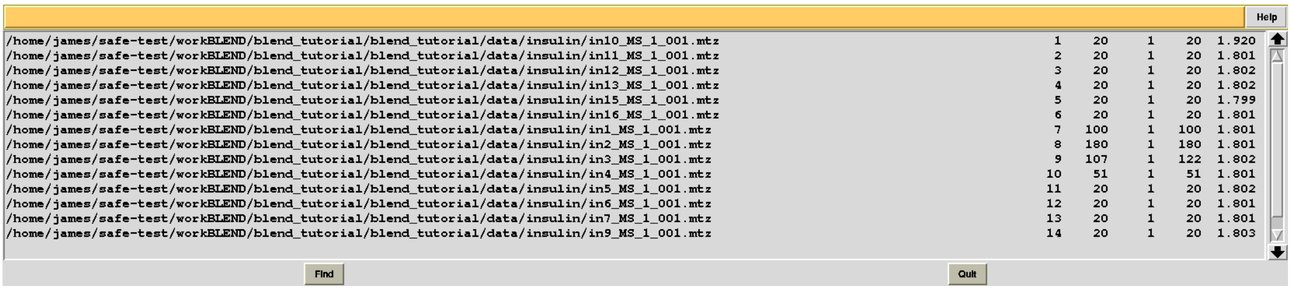


```
Cluster      Number of        Cluster        LCV      Datasets
 Number       Datasets         Height                  ID

    001              2          0.173       0.03      5 6
    002              2          0.242       0.01      9 13
    003              2          0.433       0.05      3 14
    004              2          0.518       0.08      8 10
    005              3          0.610       0.05      7 9 13
    006              4          0.702       0.13      12 7 9 13
    007              4          0.744       0.11      5 6 8 10
    008              3          0.982       0.17      4 3 14
    009              4          1.297       0.19      2 4 3 14
    010              5          1.623       0.28      11 12 7 9 13
    011              5          2.711       0.48      1 2 4 3 14
    012              9          3.343       0.44      5 6 8 10 11 12 7 9 13
    013             14         13.670       1.04      1 2 4 3 14 5 6 8 10 11 12 7 9 13
```

This is, essentially, a numerical version of the dendrogram and, as we will see, is often useful to run BLEND in synthesis mode.

Another useful file in directory “blend_1” is “BLEND_SUMMARY.txt”, which is essentially a table reporting cell parameters and other quantities for each data set.

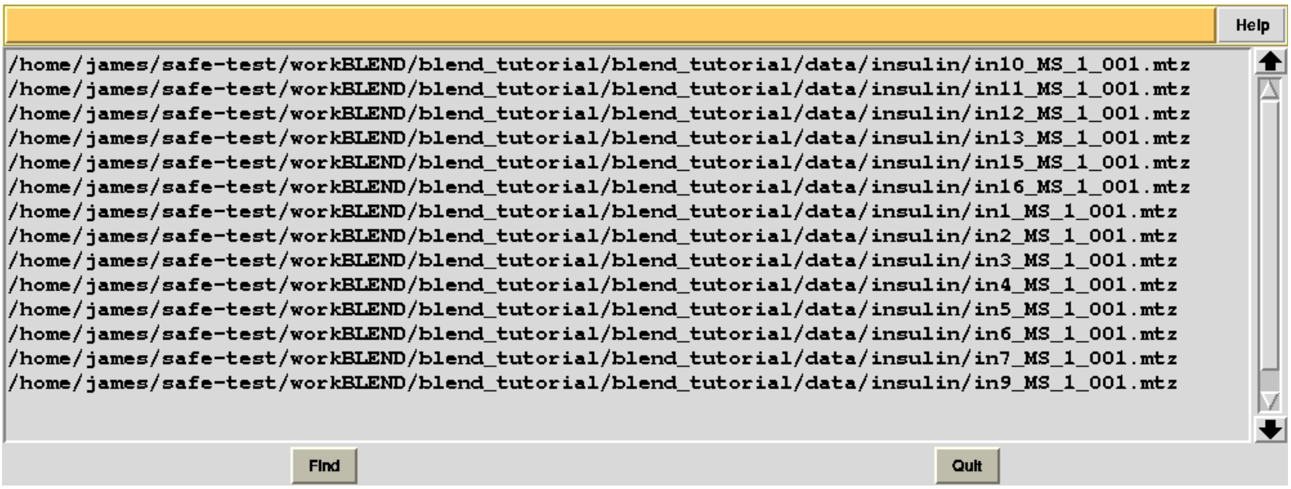
The unique serial number assigned by BLEND to each data set can be found in the “FINAL_list_of_files.dat” file. For us this file looks like the following:



```
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in10_MS_1_001.mtz     1   20    1   20  1.920
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in11_MS_1_001.mtz     2   20    1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in12_MS_1_001.mtz     3   20    1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in13_MS_1_001.mtz     4   20    1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in15_MS_1_001.mtz     5   20    1   20  1.799
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in16_MS_1_001.mtz     6   20    1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in1_MS_1_001.mtz      7  100    1  100  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in2_MS_1_001.mtz      8  180    1  180  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in3_MS_1_001.mtz      9  107    1  122  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in4_MS_1_001.mtz     10   51    1   51  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in5_MS_1_001.mtz     11   20    1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in6_MS_1_001.mtz     12   20    1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in7_MS_1_001.mtz     13   20    1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in9_MS_1_001.mtz     14   20    1   20  1.803
```

The first column contains full path to all input data files; the second contains the serial number assigned to input files; the fourth and fifth column list initial and final image numbers, while the third lists the last image BLEND will include for all subsequent scaling and merging jobs. Numbers in these column have been calculated through the procedure to get rid of radiation damaged images. In the specific case the only data set where images will be discarded is data set 9, for which BLEND suggests to remove from image 108 to image 122. The last column include resolution cuts suggested by BLEND for all subsequent scaling and merging jobs. These are worked out using intensity averages decay with resolution, and are controlled by keyword ISIGI.

All input data files were assigned a serial number; this is not always the case, because sometimes individual data sets are excluded as they contain multiple wedges or other irregular features, and BLEND does not accept them in the group to be further analysed. There might also be reasons for which the user does not want to consider some of the files included in the data directory (for example if they belong to an unwanted experiment). In this case the user can make use of the file “mtz_names.dat”, also included in directory “blend_1”, to re-run BLEND using less input data. File “mtz_names.dat” looks like the following:

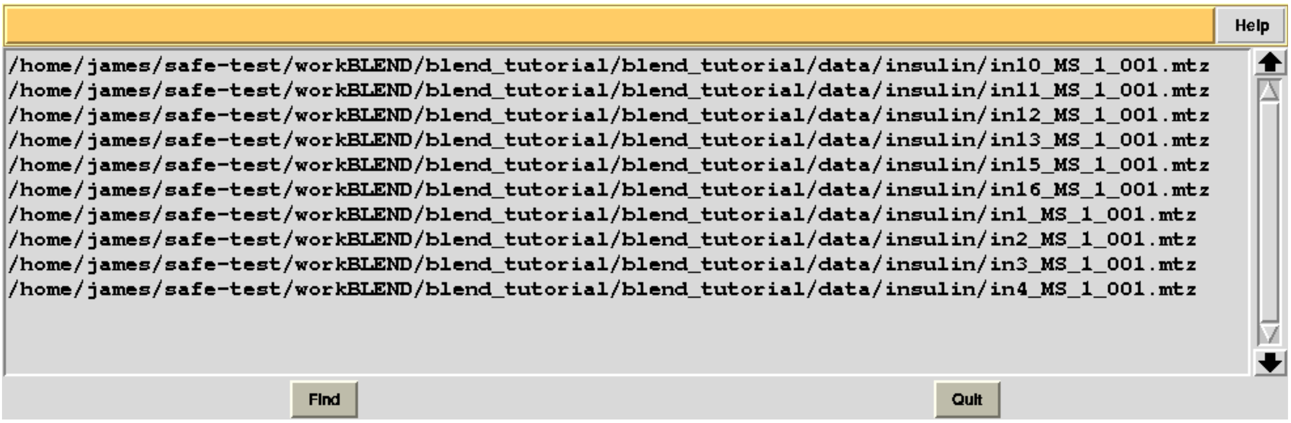


```
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in10_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in11_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in12_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in13_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in15_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in16_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in1_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in2_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in3_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in4_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in5_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in6_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in7_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in9_MS_1_001.mtz
```

In order to use it, suppose we are not any longer interested in the last 4 files. We first copy “mtz_names.dat” into a file called “original.dat” in a newly-created directory called “amended_insulin”:

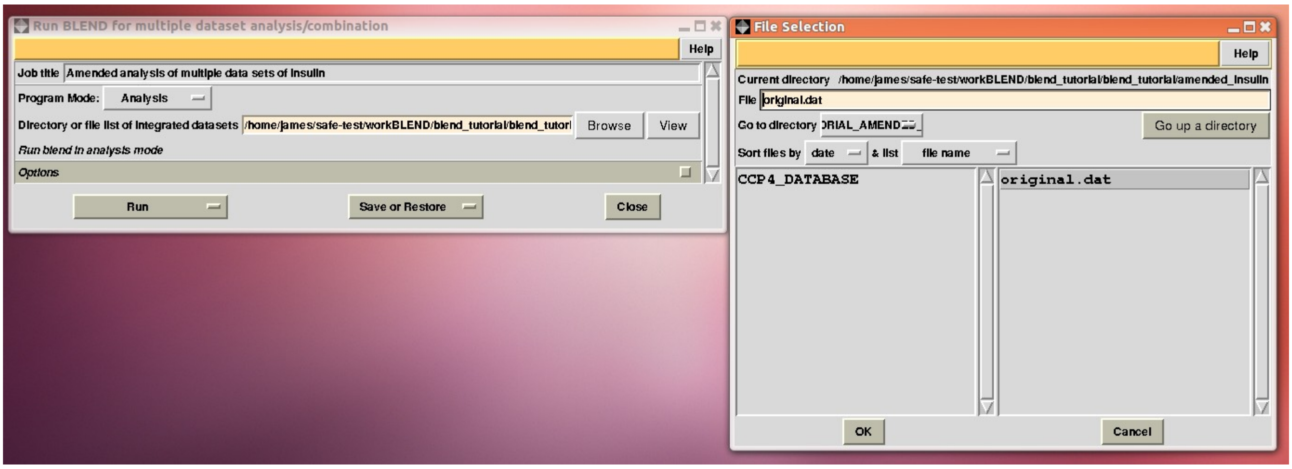
```
mkdir amended_insulin
cp insulin/blend_1/mtz_names.dat amended_insulin/original.dat
```

At this point the last 4 files can be deleted from “original.dat” file which, now, looks like:



```
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in10_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in11_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in12_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in13_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in15_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in16_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in1_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in2_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in3_MS_1_001.mtz
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in4_MS_1_001.mtz
```

Next, we associate this directory to a new project directory in ccp4i, called “BLEND_TURORIAL_AMENDED_INSULIN”, and select this project directory. BLEND can now be executed again in analysis mode, this time selecting the “original.dat” file as starting point, rather than the whole directory containing all 14 insulin data sets:



Run BLEND for multiple dataset analysis/combination

Job title: Amended analysis of multiple data sets of insulin

Program Mode: Analysis

Directory or file list of integrated datasets: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Run blend in analysis mode

Options

File Selection

Current directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/amended_insulin

File: original.dat

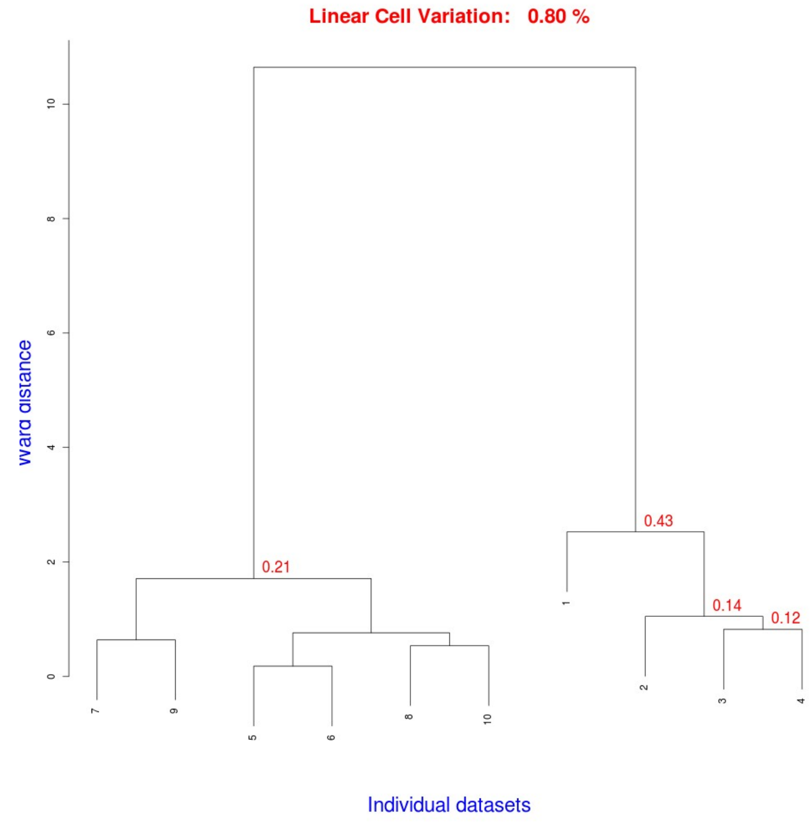
Go to directory: ORIAL_AMEND

Sort files by: date & list: file name

CCP4_DATABASE

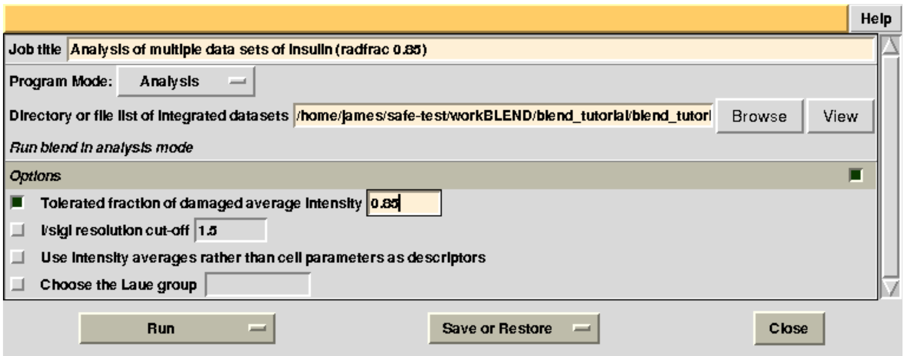
original.dat

The usual files will be produced (still in directory “blend_1”, but inside directory “amended_insulin”), although this time *BLEND* will have dealt with 10, rather than 14 data sets, and the dendrogram will look different from the one previously obtained:



As numbering of the 10 data sets has not been modified, one can see that clusters for these data include the same elements of corresponding clusters in the previous example, with the exclusion of data sets numbered 11 to 14.

Having shown how *BLEND* in analysis mode can be re-executed using a subset of data, let us go back to the previous working directory by selecting “BLEND_TUTORIAL_INSULIN” project directory in ccp4i. Let's try and increase or decrease image cutting due to radiation damage, by using the RADFRAC keyword. An input value of 0.85 means that we would like to keep only images for which intensities have, on average, been dampened by less than 15% of their initial value. Input for RADFRAC can be included via ccp4i as shown here:



Job title Analysis of multiple data sets of insulin (radfrac 0.85)

Program Mode: Analysis

Directory or file list of integrated datasets /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Run blend in analysis mode

Options ■

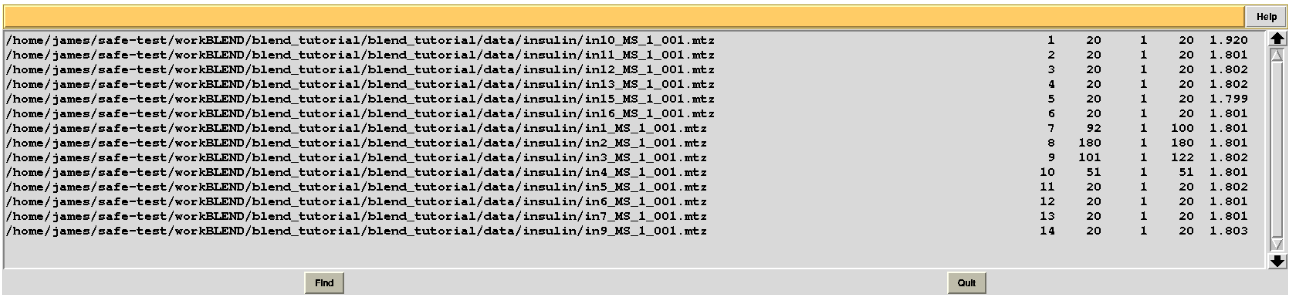
■ Tolerated fraction of damaged average intensity 0.85

□ I/sigI resolution cut-off 1.5

□ Use intensity averages rather than cell parameters as descriptors

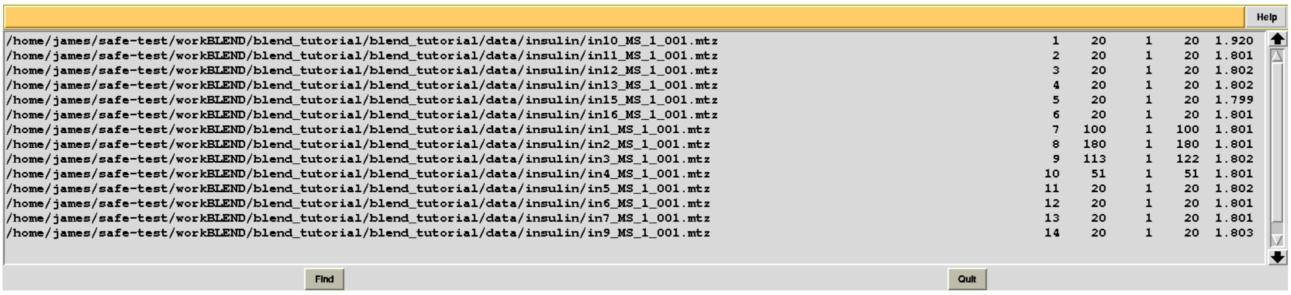
□ Choose the Laue group

This time “FINAL_list_of_files.dat” shows some difference in the number of images discarded. For data set 9 discarded images range 102 to 122, while before it ranged 108 to 122, and images 93 to 100 are now discarded from data set 7, while no images had bee discarded in the previous case:



```
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in10_MS_1_001.mtz    1   20   1   20  1.920
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in11_MS_1_001.mtz    2   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in12_MS_1_001.mtz    3   20   1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in13_MS_1_001.mtz    4   20   1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in15_MS_1_001.mtz    5   20   1   20  1.799
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in16_MS_1_001.mtz    6   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in1_MS_1_001.mtz     7   92   1  100  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in2_MS_1_001.mtz     8  180   1  180  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in3_MS_1_001.mtz     9  101   1  122  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in4_MS_1_001.mtz    10   51   1   51  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in5_MS_1_001.mtz    11   20   1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in6_MS_1_001.mtz    12   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in7_MS_1_001.mtz    13   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in9_MS_1_001.mtz    14   20   1   20  1.803
```

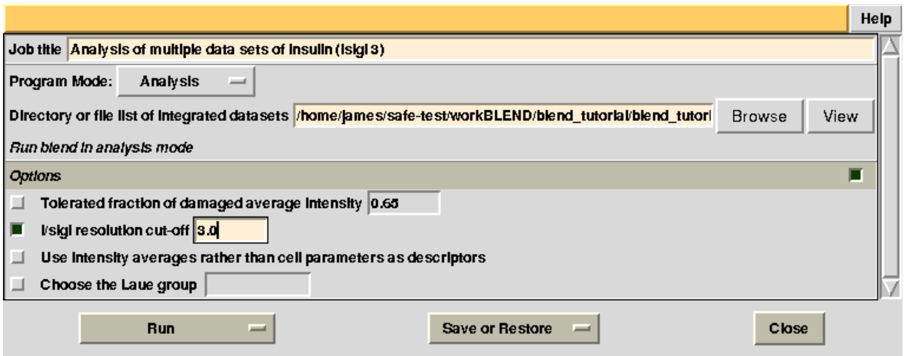
Let's modify RADFRAC once more and use 0.65, rather than 0.85. As a result:



```
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in10_MS_1_001.mtz    1   20   1   20  1.920
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in11_MS_1_001.mtz    2   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in12_MS_1_001.mtz    3   20   1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in13_MS_1_001.mtz    4   20   1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in15_MS_1_001.mtz    5   20   1   20  1.799
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in16_MS_1_001.mtz    6   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in1_MS_1_001.mtz     7  100   1  100  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in2_MS_1_001.mtz     8  180   1  180  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in3_MS_1_001.mtz     9  113   1  122  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in4_MS_1_001.mtz    10   51   1   51  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in5_MS_1_001.mtz    11   20   1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in6_MS_1_001.mtz    12   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in7_MS_1_001.mtz    13   20   1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in9_MS_1_001.mtz    14   20   1   20  1.803
```

Images discarded are now less because we have allowed for intensities to be dampened, on average, by up to 35% of their starting value.

A similar data exclusion can be applied in resolution terms, by changing value to the keyword ISIGI. For instance, let's re-run BLEND in analysis mode (with RADFRAC 0.75 - default value), fixing ISIGI to 3:



Job title: Analysis of multiple data sets of insulin (Isigi 3)

Program Mode: Analysis

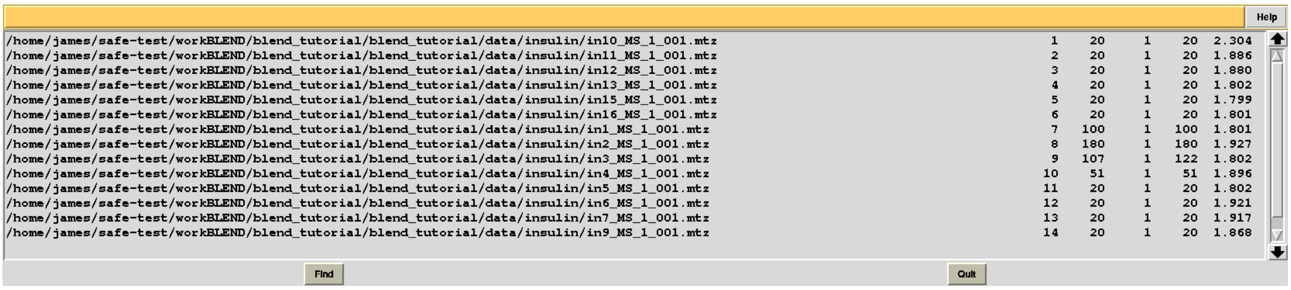
Directory or file list of integrated datasets: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Run blend in analysis mode

Options

- □ Tolerated fraction of damaged average intensity: 0.65
- ■ I/sigI resolution cut-off: 3.0
- □ Use intensity averages rather than cell parameters as descriptors
- □ Choose the Laue group

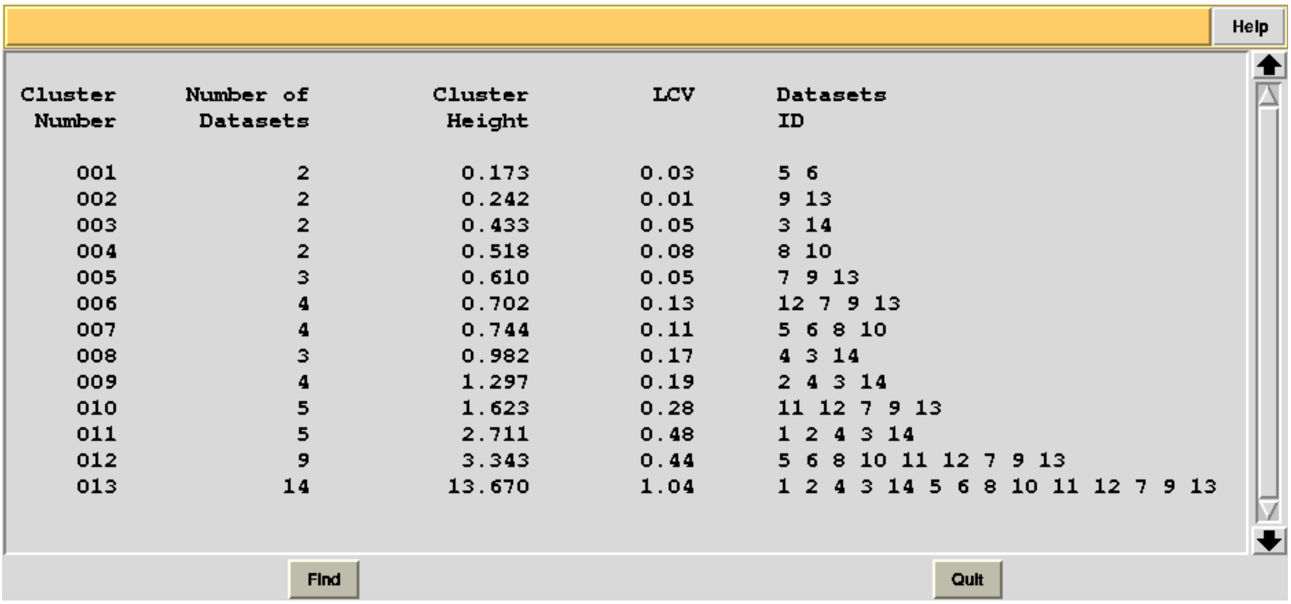
Results should be produced in directory “blend_4”. The “FINAL_list_of_files.dat” looks as follows:



```
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in10_MS_1_001.mtz     1   20    1   20  2.304
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in11_MS_1_001.mtz     2   20    1   20  1.886
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in12_MS_1_001.mtz     3   20    1   20  1.880
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in13_MS_1_001.mtz     4   20    1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in15_MS_1_001.mtz     5   20    1   20  1.799
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in16_MS_1_001.mtz     6   20    1   20  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in1_MS_1_001.mtz      7  100    1  100  1.801
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in2_MS_1_001.mtz      8  180    1  180  1.927
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in3_MS_1_001.mtz      9  107    1  122  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in4_MS_1_001.mtz     10   51    1   51  1.896
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in5_MS_1_001.mtz     11   20    1   20  1.802
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in6_MS_1_001.mtz     12   20    1   20  1.921
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in7_MS_1_001.mtz     13   20    1   20  1.917
/home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/data/insulin/in9_MS_1_001.mtz     14   20    1   20  1.868
```

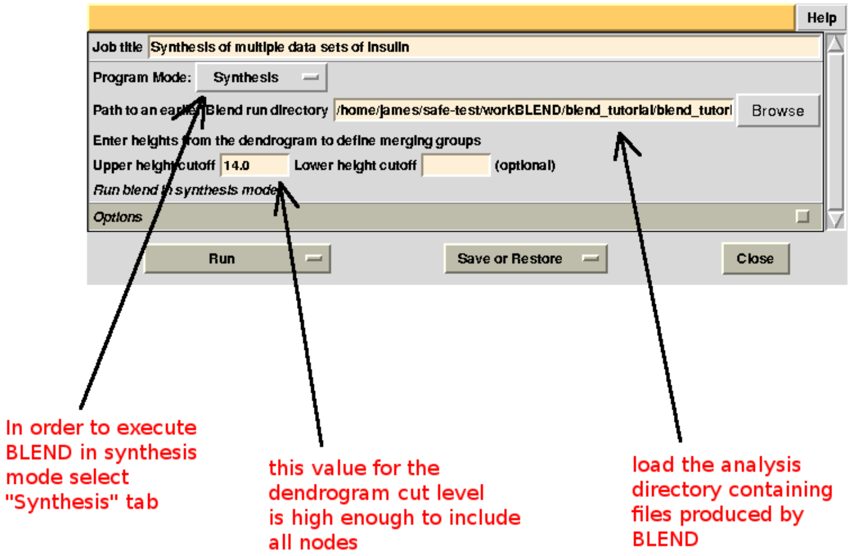
Resolution cutoffs have now changed. For the first data set the highest is 2.304, while previously it was 1.920. We expect an increase if the value of ISIGI is lowered. This is normally the case. But in the specific insulin example data are exceptionally good, and their signal-to-noise ratio is quite high to the very highest resolution. Thus, results will not change here if values lower than 1.5 are used for ISIGI.

Let's pay a closer look at the results of cluster analysis, as described in file “CLUSTERS.txt”:



```
 Cluster     Number of        Cluster         LCV      Datasets
  Number      Datasets         Height                   ID

     001             2          0.173        0.03      5 6
     002             2          0.242        0.01      9 13
     003             2          0.433        0.05      3 14
     004             2          0.518        0.08      8 10
     005             3          0.610        0.05      7 9 13
     006             4          0.702        0.13      12 7 9 13
     007             4          0.744        0.11      5 6 8 10
     008             3          0.982        0.17      4 3 14
     009             4          1.297        0.19      2 4 3 14
     010             5          1.623        0.28      11 12 7 9 13
     011             5          2.711        0.48      1 2 4 3 14
     012             9          3.343        0.44      5 6 8 10 11 12 7 9 13
     013            14         13.670        1.04      1 2 4 3 14 5 6 8 10 11 12 7 9 13
```

Cluster numbers and corresponding data sets are listed in column 1 and column 5 respectively. The second column reports how many data sets form the specific cluster, while in column 3 it is found the exact height at which data sets merge to form the cluster. Column 4 lists all LCV values. The dendrogram encourages people to look for grouping of individual data sets. Larger groups attract one's attention more than smaller groups. In this sense the presence of two clusters, cluster (1 2 4 3 14) and cluster (5 6 8 10 11 12 7 9 13), is immediately perceived. Very soon, though, we realise that the second cluster is made up of two smaller clusters, cluster (5 6 8 10) and cluster (11 12 7 9 13). Finer details and even smaller clusters can be singled out by proceeding to the lower part of the dendrogram. This way of proceeding through the dendrogram interpretation can have some validity, but is highly subjective. In *BLEND* it is found that a more logical, straightforward and objective way of interpreting the dendrogram is to assume that each node can represent a valid data set. Subsequent operations of space group assignment (*POINTLESS*) and scaling (*AIMLESS*) determine the degree of validity. This interpretation is accomplished with *BLEND synthesis* execution mode. From “CLUSTERS.txt” we read that everything merges at height 13.670. If the user wishes to produce results for all nodes under a certain height, it will suffice to specify this in the GUI. For instance, to produce results for all nodes in the insulin example, the BLEND section of the GUI will look like this:



Job title: Synthesis of multiple data sets of insulin

Program Mode: Synthesis

Path to an earlier Blend run directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Enter heights from the dendrogram to define merging groups

Upper height cutoff: 14.0 Lower height cutoff: (optional)

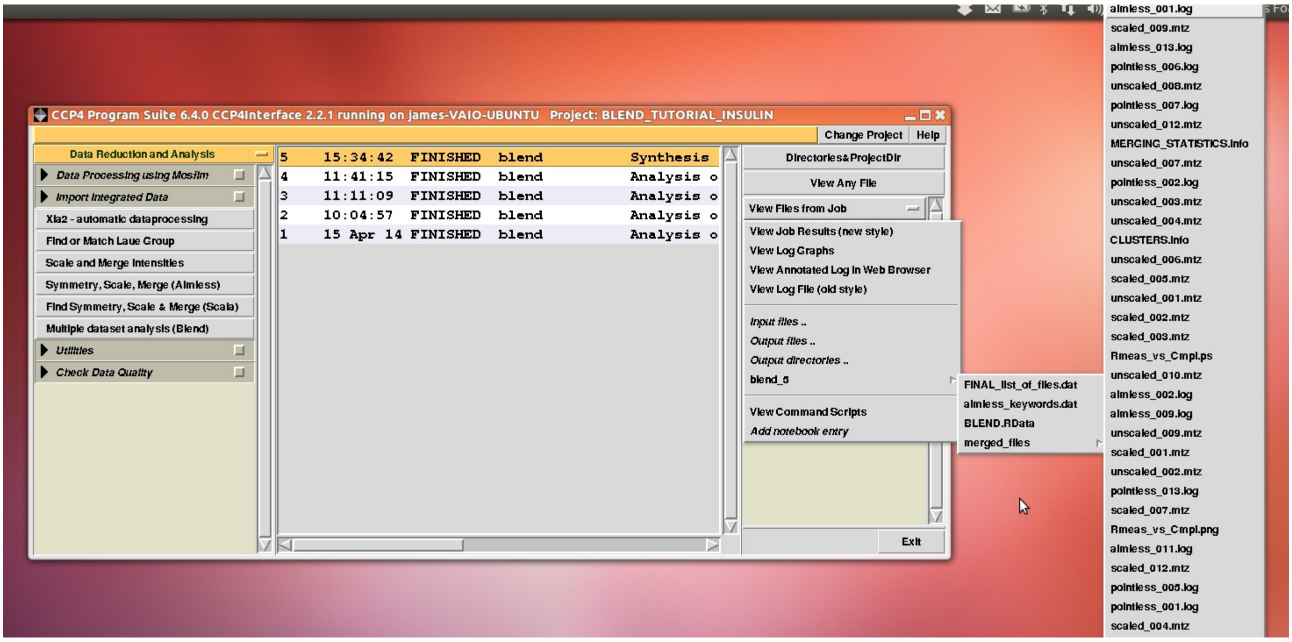
Run blend in synthesis mode

In order to execute BLEND in synthesis mode select "Synthesis" tab

this value for the dendrogram cut level is high enough to include all nodes

load the analysis directory containing files produced by BLEND

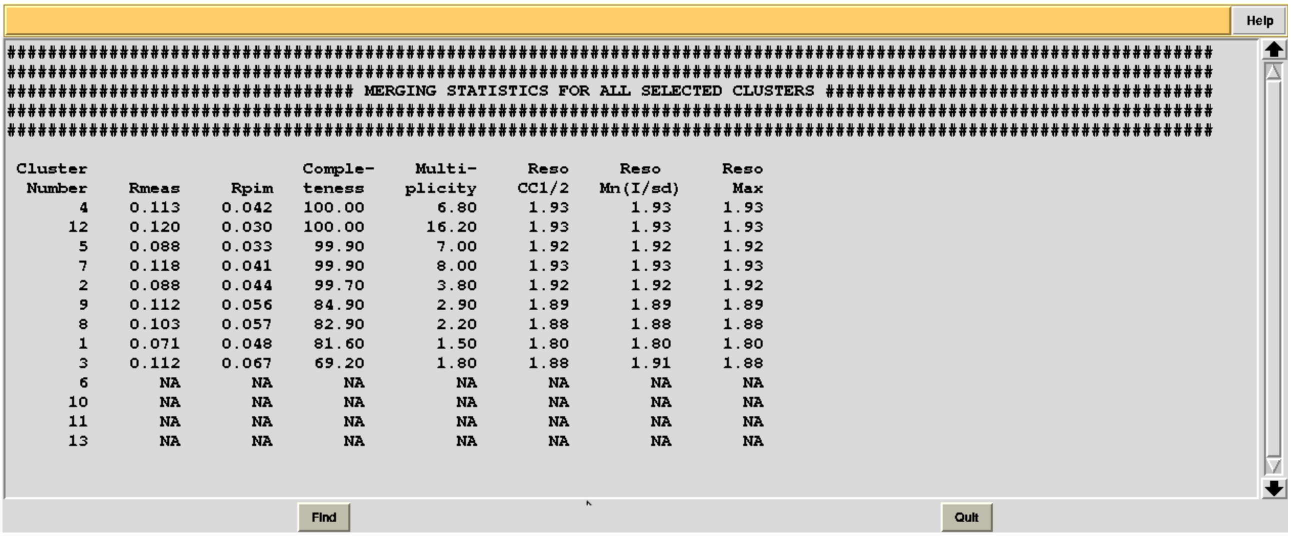
After execution, several files will be produced in directory “blend_5” (this is the fifth job for the GUI under project directory “BLEND_TUTORIAL_INSULIN”), including another directory packed with files, called “merged_files”:



CCP4 Program Suite 6.4.0 CCP4Interface 2.2.1 running on james-VAIO-UBUNTU Project: BLEND_TUTORIAL_INSULIN

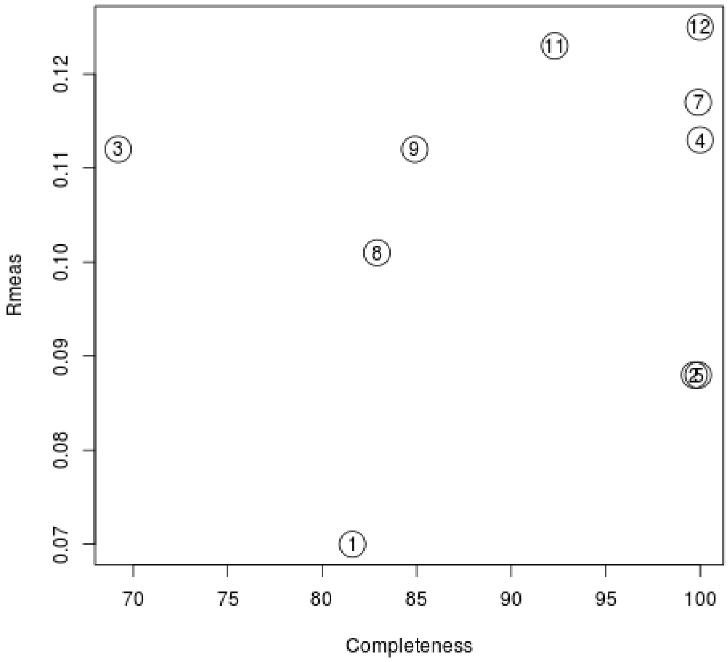
```
5    15:34:42  FINISHED  blend          Synthesis
4    11:41:15  FINISHED  blend          Analysis o
3    11:11:09  FINISHED  blend          Analysis o
2    10:04:57  FINISHED  blend          Analysis o
1    15 Apr 14 FINISHED  blend          Analysis o
```

In this directory one can find unscaled (produced by POINTLESS) and scaled (produced by AIMLESS) files corresponding to all nodes selected in the dendrogram. Final overall merging statistics can be found in the “MERGING_STATISTICS.info” file:

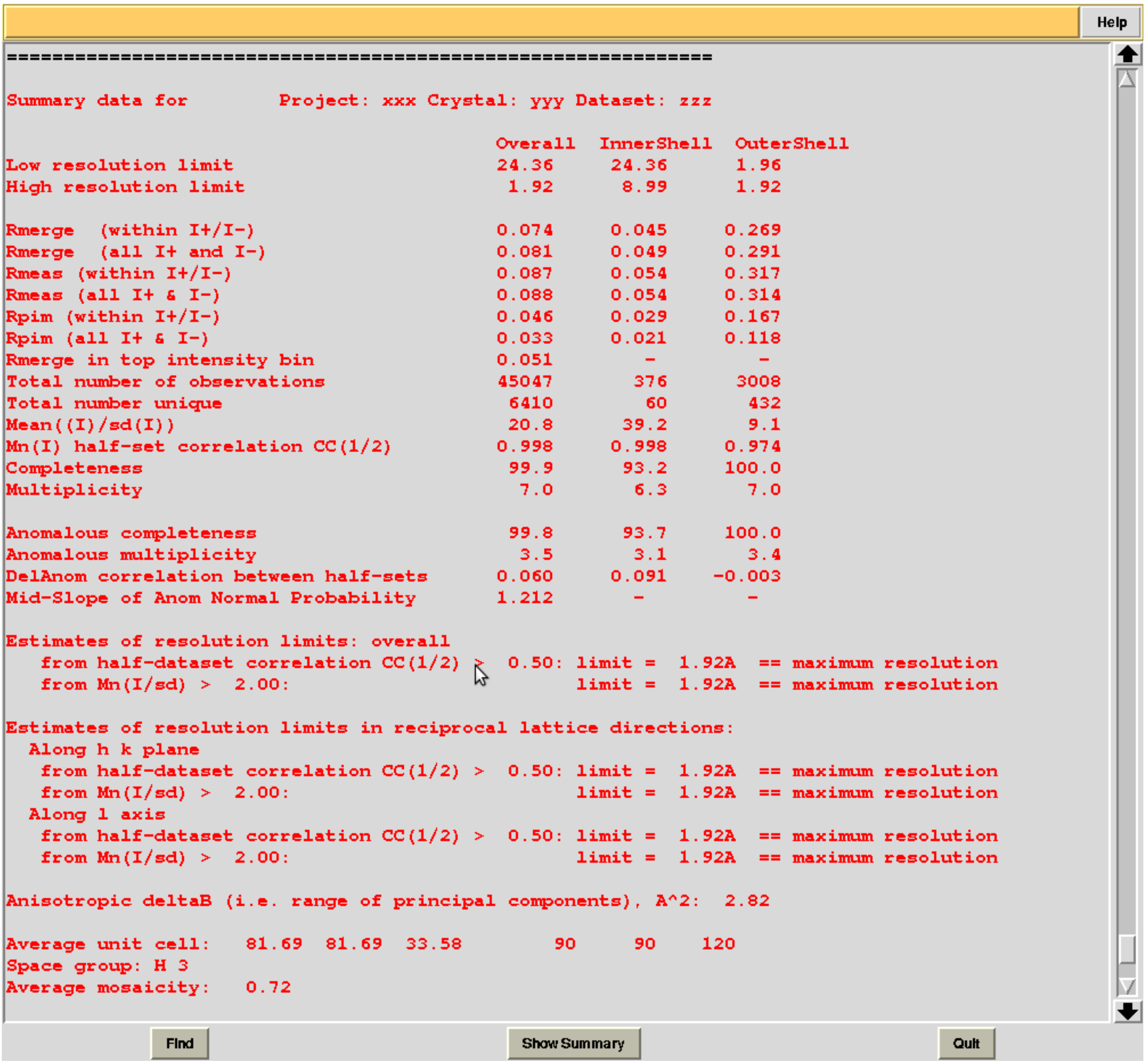


```
###############################################################################################################
###############################################################################################################
################################ MERGING STATISTICS FOR ALL SELECTED CLUSTERS #################################
###############################################################################################################
###############################################################################################################

 Cluster                       Comple-     Multi-      Reso      Reso       Reso
  Number     Rmeas      Rpim   teness     plicity     CC1/2   Mn(I/sd)      Max
       4     0.113     0.042   100.00        6.80      1.93       1.93      1.93
      12     0.120     0.030   100.00       16.20      1.93       1.93      1.93
       5     0.088     0.033    99.90        7.00      1.92       1.92      1.92
       7     0.118     0.041    99.90        8.00      1.93       1.93      1.93
       2     0.088     0.044    99.70        3.80      1.92       1.92      1.92
       9     0.112     0.056    84.90        2.90      1.89       1.89      1.89
       8     0.103     0.057    82.90        2.20      1.88       1.88      1.88
       1     0.071     0.048    81.60        1.50      1.80       1.80      1.80
       3     0.112     0.067    69.20        1.80      1.88       1.91      1.88
       6        NA        NA       NA          NA        NA         NA        NA
      10        NA        NA       NA          NA        NA         NA        NA
      11        NA        NA       NA          NA        NA         NA        NA
      13        NA        NA       NA          NA        NA         NA        NA
```

Values are sorted on increasing completeness, followed by lower Rmeas. Higher completeness is often associated with high redundancy (multiplicity), but also with poor isomorphism, because many more data sets are needed for both completeness and multiplicity to increase. For example, cluster 12 is 100% complete, as it is cluster 4, but this last cluster has lower Rmeas, quite certainly because it is the union of just two data sets, while cluster 12 includes nine data sets. A good result is represented by cluster 2, which has Rmeas = 0.088 and it is 99.7% complete. Cluster 5 is an even better choice; it has same Rmeas than cluster 2, but a much lower Rpim. It is a nearly complete set and has high multiplicity. This is what it should probably be used. A visual version of the above table is provided by the “Rmeas_vs_Cmpl.png” or “Rmeas_vs_Cmpl.ps” plot:

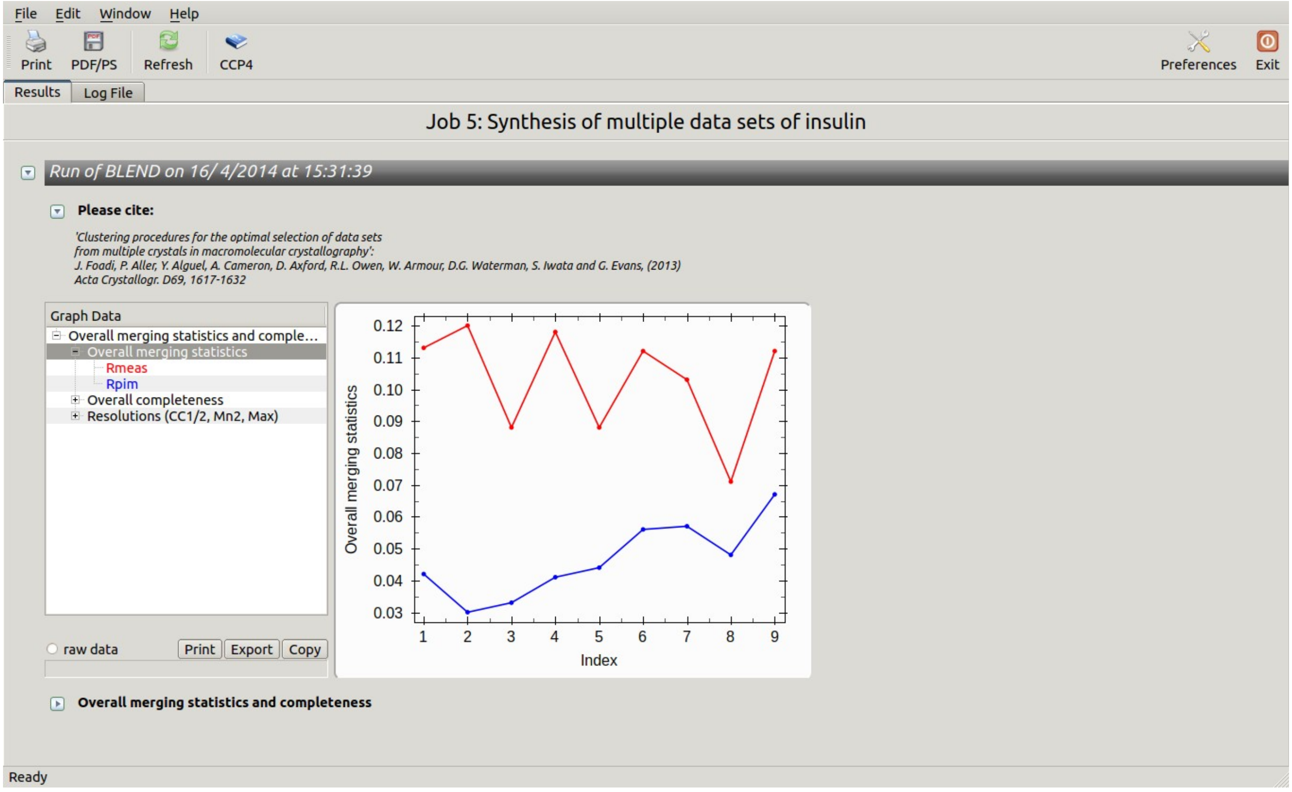


From the overall view provided by this plot we can see that 2 and 5 are the best data sets in terms of statistics and completeness (they are barely visible because they overlap on top of each other). Unscaled and scaled files related to all results, in the “merged_files” directory, can be inspected with the usual CCP4 tools (try, for instance, to view “scaled_005.mtz”). POINTLESS and AIMLESS logs for all unscaled and scaled files are also included in this directory. For instance, the final part of the *AIMLESS* log file for cluster 5 shows:



```
==============================================================

Summary data for        Project: xxx Crystal: yyy Dataset: zzz

                                           Overall  InnerShell  OuterShell
Low resolution limit                        24.36     24.36      1.96
High resolution limit                        1.92      8.99      1.92

Rmerge  (within I+/I-)                      0.074     0.045     0.269
Rmerge  (all I+ and I-)                     0.081     0.049     0.291
Rmeas (within I+/I-)                        0.087     0.054     0.317
Rmeas (all I+ & I-)                         0.088     0.054     0.314
Rpim (within I+/I-)                         0.046     0.029     0.167
Rpim (all I+ & I-)                          0.033     0.021     0.118
Rmerge in top intensity bin                 0.051        -         -
Total number of observations                45047       376      3008
Total number unique                          6410        60       432
Mean((I)/sd(I))                              20.8      39.2       9.1
Mn(I) half-set correlation CC(1/2)          0.998     0.998     0.974
Completeness                                 99.9      93.2     100.0
Multiplicity                                  7.0       6.3       7.0

Anomalous completeness                       99.8      93.7     100.0
Anomalous multiplicity                        3.5       3.1       3.4
DelAnom correlation between half-sets       0.060     0.091    -0.003
Mid-Slope of Anom Normal Probability        1.212       -         -

Estimates of resolution limits: overall
   from half-dataset correlation CC(1/2) >  0.50: limit =  1.92A  == maximum resolution
   from Mn(I/sd) >  2.00:                         limit =  1.92A  == maximum resolution

Estimates of resolution limits in reciprocal lattice directions:
  Along h k plane
   from half-dataset correlation CC(1/2) >  0.50: limit =  1.92A  == maximum resolution
   from Mn(I/sd) >  2.00:                         limit =  1.92A  == maximum resolution
  Along l axis
   from half-dataset correlation CC(1/2) >  0.50: limit =  1.92A  == maximum resolution
   from Mn(I/sd) >  2.00:                         limit =  1.92A  == maximum resolution

Anisotropic deltaB (i.e. range of principal components), A^2:  2.82

Average unit cell:   81.69   81.69   33.58        90      90     120
Space group: H 3
Average mosaicity:   0.72
```

Overall merging statistics can also be explored selecting “View Files from Job” → “View Job Results (new style)”:



Job 5: Synthesis of multiple data sets of insulin

Run of BLEND on 16/ 4/2014 at 15:31:39

Please cite:

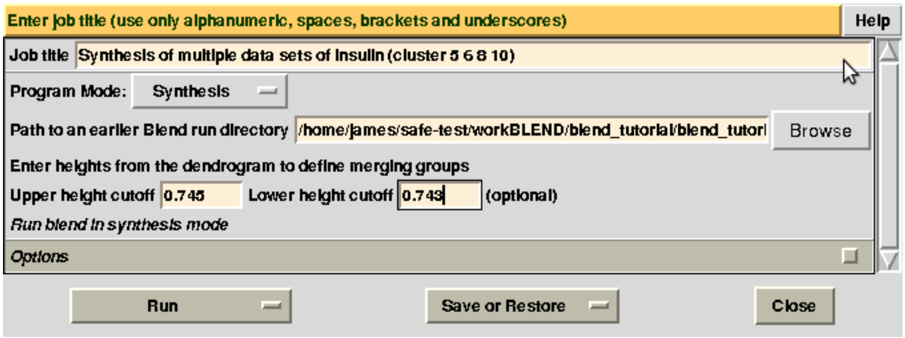
'Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography':
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Graph Data
Overall merging statistics and comple...
Overall merging statistics
Rmeas
Rpim
Overall completeness
Resolutions (CC1/2, Mn2, Max)
raw data
Print
Export
Copy
Overall merging statistics
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
1
2
3
4
5
6
7
8
9
Index

Overall merging statistics and completeness

Ready

There might be reasons for which the user wishes to calculate results for certain nodes only, rather than for all nodes of the dendrogram. This can be done using two numerical values, rather than one, in *BLEND* synthesis. Suppose we only want to calculate data for cluster (5 6 8 10). This cluster merges at a height of 0.744 (check “CLUSTERS.txt”). Two numerical values in between which 0.744 and no other nodes is included are 0.745 and 0.743. Thus, to produce a synthesis just for this cluster, one can run *BLEND* as follows:


Enter job title (use only alphanumeric, spaces, brackets and underscores)

Job title: Synthesis of multiple data sets of insulin (cluster 5 6 8 10)

Program Mode: Synthesis

Path to an earlier Blend run directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

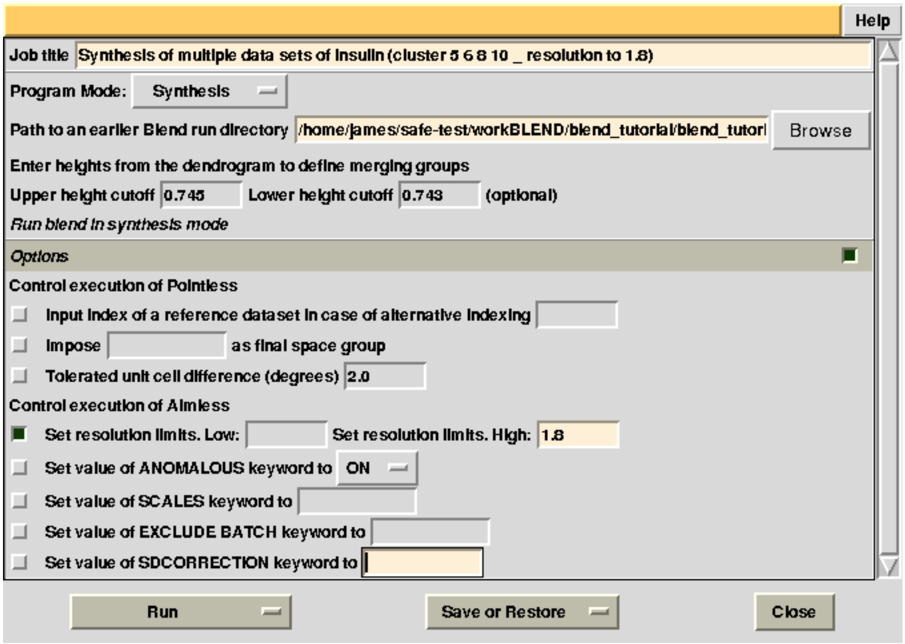
Enter heights from the dendrogram to define merging groups

Upper height cutoff: 0.745 Lower height cutoff: 0.743 (optional)

Run blend in synthesis mode

Statistics and related files for only one data set are contained within the “merged_files” directory.

Suppose we want to extend resolution of the previous result from 1.93 to 1.80. *BLEND* can be again executed in synthesis mode using the appropriate “RESOLUTION” keyword:



Job title Synthesis of multiple data sets of insulin (cluster 5 6 8 10 _ resolution to 1.8)

Program Mode: Synthesis

Path to an earlier Blend run directory /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Enter heights from the dendrogram to define merging groups

Upper height cutoff 0.745 Lower height cutoff 0.743 (optional)

Run blend in synthesis mode

Options

Control execution of Pointless

□ Input index of a reference dataset in case of alternative indexing

□ Impose as final space group

□ Tolerated unit cell difference (degrees) 2.0

Control execution of Aimless

☒ Set resolution limits. Low: Set resolution limits. High: 1.8

□ Set value of ANOMALOUS keyword to ON

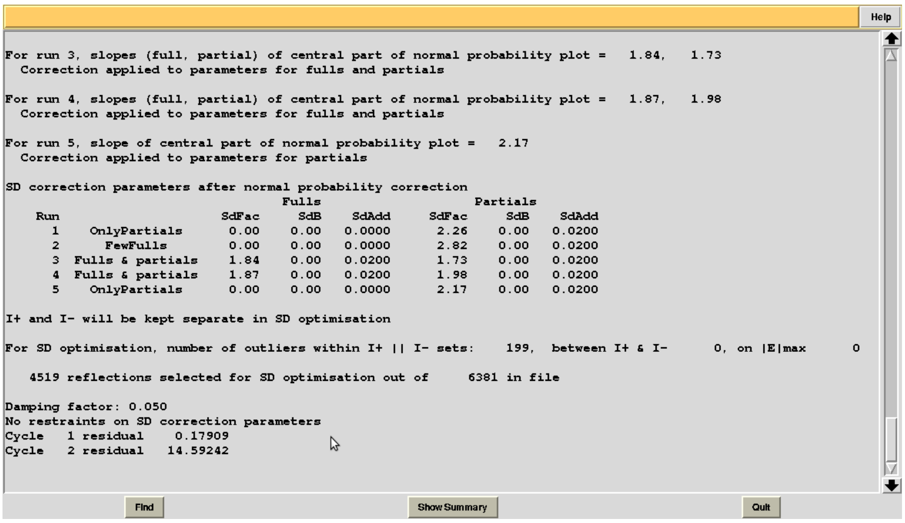
□ Set value of SCALES keyword to

□ Set value of EXCLUDE BATCH keyword to

□ Set value of SDCORRECTION keyword to

Check statistics under the new “merged_files” directory, they should have changed to include the desired resolution cutoff.

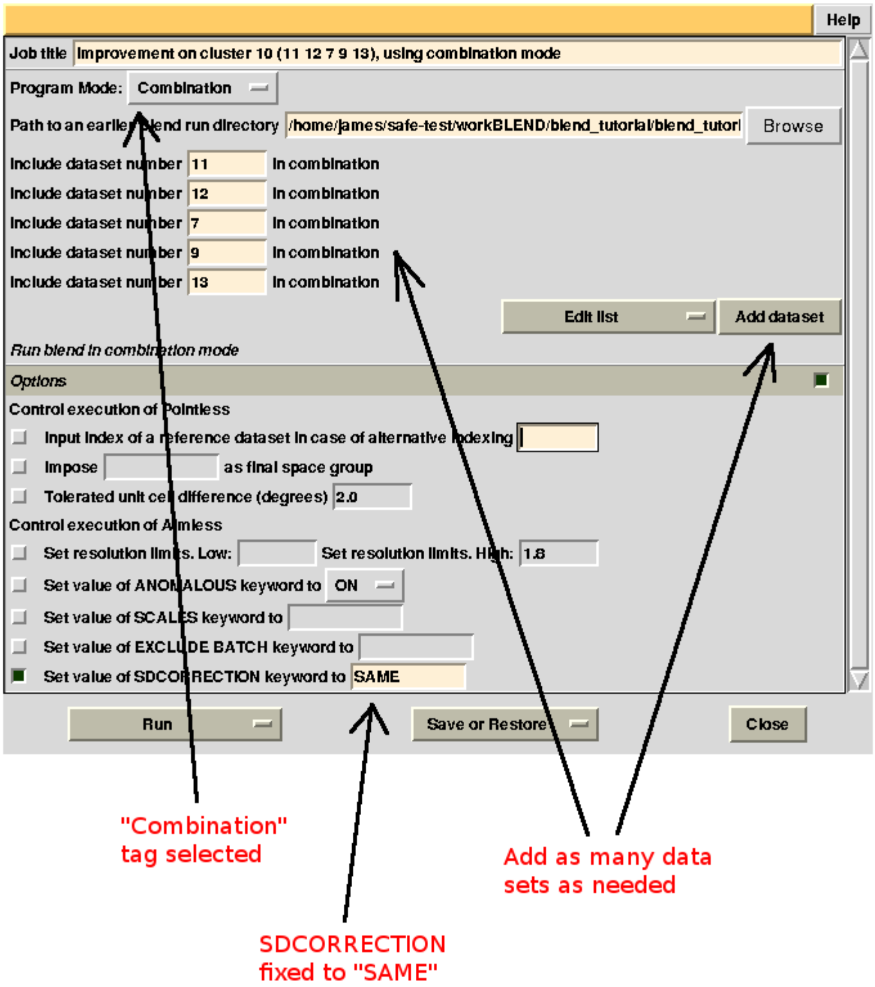
Reading “MERGING_STATISTICS.info” allows people to gain a better understanding of the ability individual data sets have to form larger, more complete and isomorphous groups. For example, in job number 5, it clearly emerges that cluster 5, composed of data sets 7, 9, 13, has good merging statistics and it is very complete. The two additional data sets joining 7, 9, 13 and forming cluster 10, i.e. data sets 11 and 12, make the situation worse, because scaling with *AIMLESS* fails (results given as “NA”). In cases like this it is useful to re-run *BLEND* using specific keywords, only for that specific node, using *BLEND* in combination mode. The “NA” for cluster 10 can be investigated by looking at the *AIMLESS* log inside the “merged_files_all” directory (file “aimless_010.log”):



```
For run 3, slopes (full, partial) of central part of normal probability plot =   1.84,   1.73
  Correction applied to parameters for fulls and partials

For run 4, slopes (full, partial) of central part of normal probability plot =   1.87,   1.98
  Correction applied to parameters for fulls and partials

For run 5, slope of central part of normal probability plot =   2.17
  Correction applied to parameters for partials

SD correction parameters after normal probability correction
                                    Fulls                      Partials
    Run                       SdFac     SdB    SdAdd     SdFac     SdB    SdAdd
      1    OnlyPartials        0.00    0.00   0.0000      2.26    0.00   0.0200
      2      FewFulls          0.00    0.00   0.0000      2.82    0.00   0.0200
      3  Fulls & partials      1.84    0.00   0.0200      1.73    0.00   0.0200
      4  Fulls & partials      1.87    0.00   0.0200      1.98    0.00   0.0200
      5    OnlyPartials        0.00    0.00   0.0000      2.17    0.00   0.0200

I+ and I- will be kept separate in SD optimisation

For SD optimisation, number of outliers within I+ || I- sets:    199,  between I+ & I-      0, on |E|max      0

   4519 reflections selected for SD optimisation out of     6381 in file

Damping factor: 0.050
No restraints on SD correction parameters
Cycle   1 residual    0.17909
Cycle   2 residual   14.59242
```

AIMLESS stopped abruptly while trying to compute standard errors (SD). This is, at present, a potentially frequent occurrence in *AIMLESS*, whenever combinations from multiple data sets are involved. Most of the time it is caused by the failed refinement of standard errors. A possible way out is through the addition of a same set of SD parameters for all data sets (SDCORR SAME).

Alternatively, it is possible to give up with SD refinement (thus having poor errors estimate) altogether (SDCORR NOREFINE). Let's re-run *BLEND* using keyword SDCORR SAME:



Job title: Improvement on cluster 10 (11 12 7 9 13), using combination mode

Program Mode: Combination

Path to an earlier blend run directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Include dataset number 11 in combination

Include dataset number 12 in combination

Include dataset number 7 in combination

Include dataset number 9 in combination

Include dataset number 13 in combination

Run blend in combination mode

Options

Control execution of Pointless

□ Input index of a reference dataset in case of alternative indexing

□ Impose as final space group

□ Tolerated unit cell difference (degrees) 2.0

Control execution of Aimless

□ Set resolution limits. Low: Set resolution limits. High: 1.8

□ Set value of ANOMALOUS keyword to ON

□ Set value of SCALES keyword to

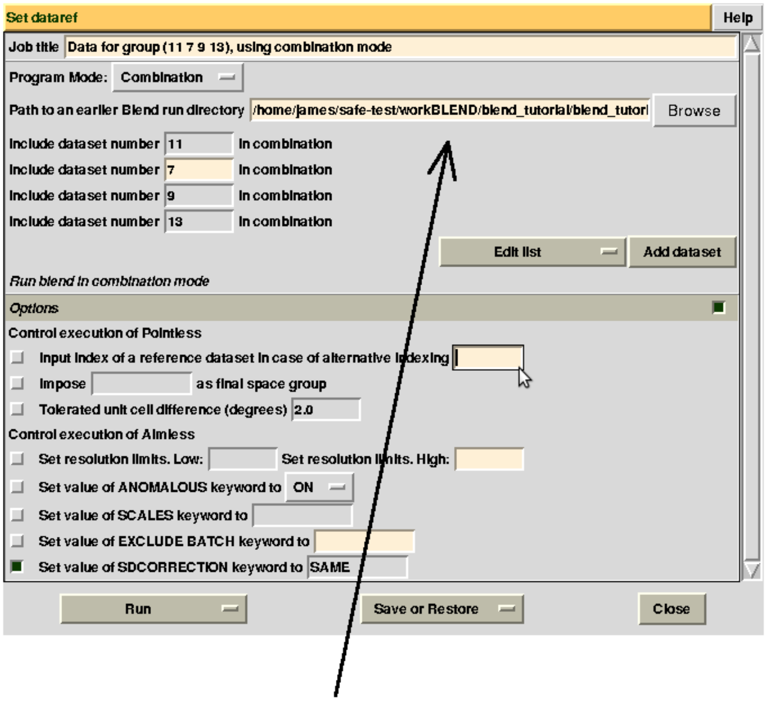
□ Set value of EXCLUDE BATCH keyword to

■ Set value of SDCORRECTION keyword to SAME

Help
Browse
Edit list
Add dataset
Run
Save or Restore
Close
"Combination" tag selected
Add as many data sets as needed
SDCORRECTION fixed to "SAME"

Now time *BLEND* returns a scaled data set with valid merging statistics (Rmeas = 0.119, Rpim = 0.042). The result can be found in the newly-created directory “combined_files”. This directory can be progressively populated with many different combinations. Data sets and merging statistics for other failed *BLEND* synthesis jobs can be attempted in a similar way (try yourself clusters 6 and 13).

Running *BLEND* in combination mode it is also possible to compute merging statistics for groups of data sets not corresponding to any cluster. Thus, for example, we might want to investigate whether data set 11 or data set 12 make statistics for cluster 5 worse. First we try with 11 (remember to use always the keyword line SDCORR SAME). In order to have results from this run included in the same “combined_files” directory from a previous run in combination mode, we will have to select that specific directory in the GUI:



Job title: Data for group (11 7 9 13), using combination mode

Program Mode: Combination

Path to an earlier Blend run directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Include dataset number 11 in combination

Include dataset number 7 in combination

Include dataset number 9 in combination

Include dataset number 13 in combination

Run blend in combination mode

Options

Control execution of Pointless

□ Input index of a reference dataset in case of alternative indexing

□ Impose as final space group

□ Tolerated unit cell difference (degrees) 2.0

Control execution of Aimless

□ Set resolution limits. Low: Set resolution limits. High:

□ Set value of ANOMALOUS keyword to ON

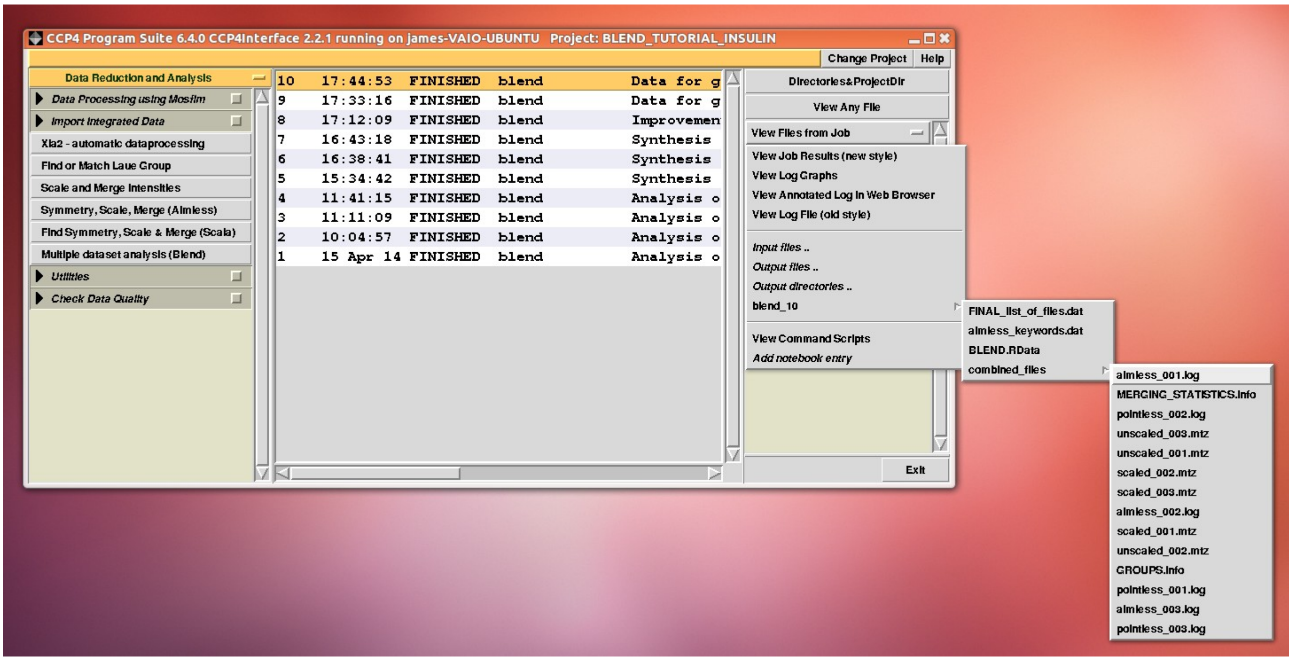
□ Set value of SCALES keyword to

□ Set value of EXCLUDE BATCH keyword to

■ Set value of SDCORRECTION keyword to SAME

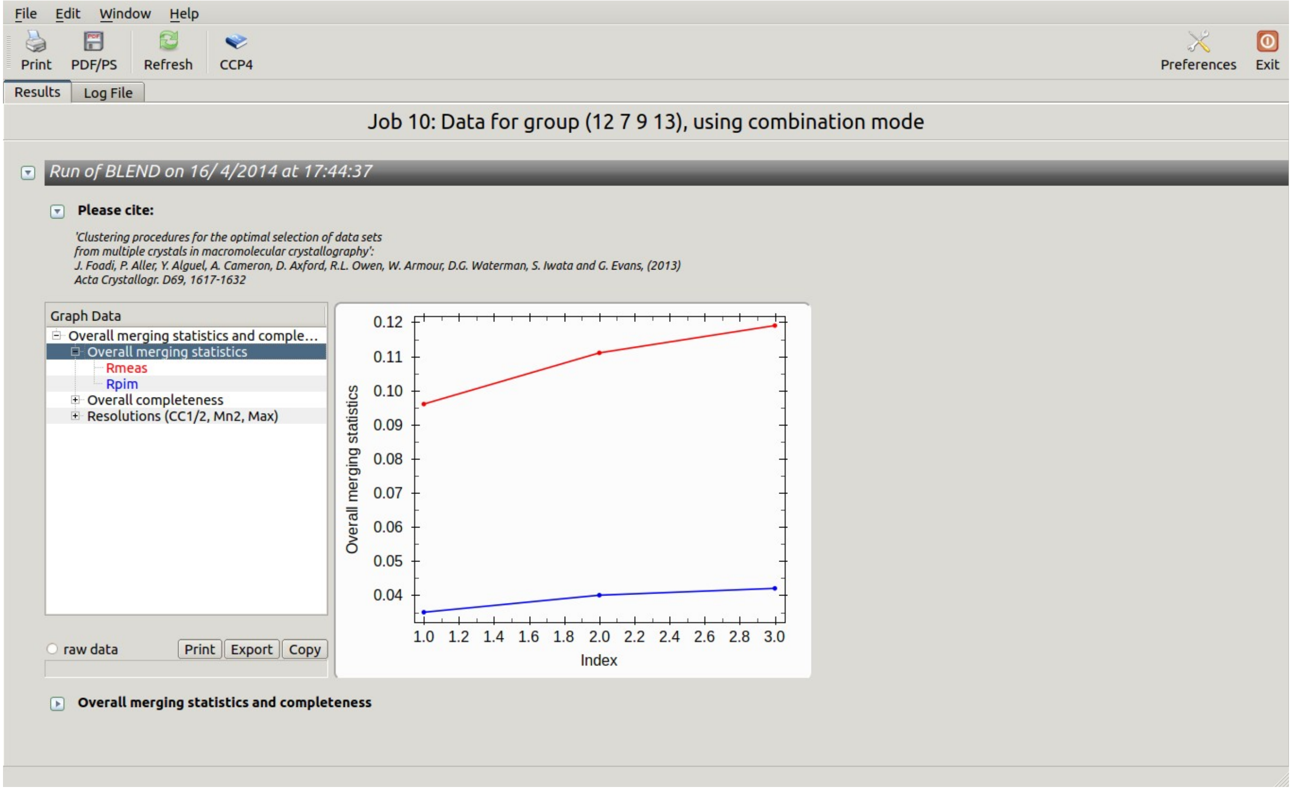
By selecting directory "blend_8"
we are implicitly asking all results
contained in this directory to be copied
into the new "combined_files" directory

Next we try with 12 (don't forget to use “blend_9” as path to earlier BLEND directory). After these last two runs, the population for the last “combined_files” directory include files for three groups:



```
10  17:44:53 FINISHED blend     Data for g
9   17:33:16 FINISHED blend     Data for g
8   17:12:09 FINISHED blend     Improvemen
7   16:43:18 FINISHED blend     Synthesis
6   16:38:41 FINISHED blend     Synthesis
5   15:34:42 FINISHED blend     Synthesis
4   11:41:15 FINISHED blend     Analysis o
3   11:11:09 FINISHED blend     Analysis o
2   10:04:57 FINISHED blend     Analysis o
1   15 Apr 14 FINISHED blend    Analysis o
```

Checking file “MERGING_STATISTICS.info” leads to the conclusion that statistics are better with data set 11. Thus we can decide to “filter out” data set 12 from cluster 10, in order to achieve a better composite data set. Results can also be viewed using the “new style” interface:



Job 10: Data for group (12 7 9 13), using combination mode

Run of BLEND on 16/ 4/2014 at 17:44:37

Please cite:

'Clustering procedures for the optimal selection of data sets
from multiple crystals in macromolecular crystallography':
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Overall merging statistics and completeness

b) Without CCP4I

Like any other CCP4 program, BLEND can be executed without using ccp4i, simply typing commands at the keyboard in an interactive command shell. This, quite often, makes the program more flexible.

Here we will briefly repeat all steps previously carried out with ccp4i. First a new directory, named for example “insulin_no_ccp4i” will be created in order to keep this part of the tutorial separate from the rest:

```
mkdir insulin_no_ccp4
cd insulin_no_ccp4
```

Next we run BLEND in analysis mode using all 14 files in the “data/insulin” directory:

```
blend -a ../data/insulin
```

After start the program immediately halts, waiting for keywords input; in this specific instance we simply push the “Enter” key once, thus forcing BLEND to use default values. At completion the program produces the following files:

BLEND.RData
BLEND_SUMMARY.txt
CLUSTERS.txt
FINAL_list_of_files.dat
mtz_names.dat

```
tree.png
tree.ps
```

whose meaning and content has been previously explained.

To show how “mtz_names.dat” can be re-used for a different run of BLEND in analysis mode, and similarly to what done previously in this tutorial, let's create a directory called “amended_insulin” and copy “mtz_names.dat” in it:

```
mkdir amended_insulin
cd amended_insulin
cp ../mtz_names.dat original.dat
```

Then “original.dat” is edited to remove the last four lines, thus creating a different input, consisting only of 10 out of 14 data sets. Next, the program is executed using “original.dat” as input:

```
blend -a original.dat
```

Results of this run are similar to what described before. To carry on let's move back one level to “insulin_no_ccp4i”:

```
cd ../
```

If a log file is needed, then a keyword file has to be prepared before execution. Let's call this file “blend_keywords.dat”, and play both with RADFRAC and ISIGI keywords, changing from their default values 0.75 and 1.5 to the new values 0.65 and 3.0. We can also add the “SDCORR SAME” keyword so to avoid some scaling jobs to fail (see earlier parts of this tutorial). The file looks like this:

```
RADFRAC     0.65
ISIGI       3.0
SDCORR      SAME
```

Then BLEND is executed again using “mtz_names.dat” as input:

```
blend -a mtz_names.dat < blend_keywords.dat > blend_analysis.log
```

After completion, “blend_analysis.log” can be either read using any editor, or viewed graphically with the CCP4 program “logview”:

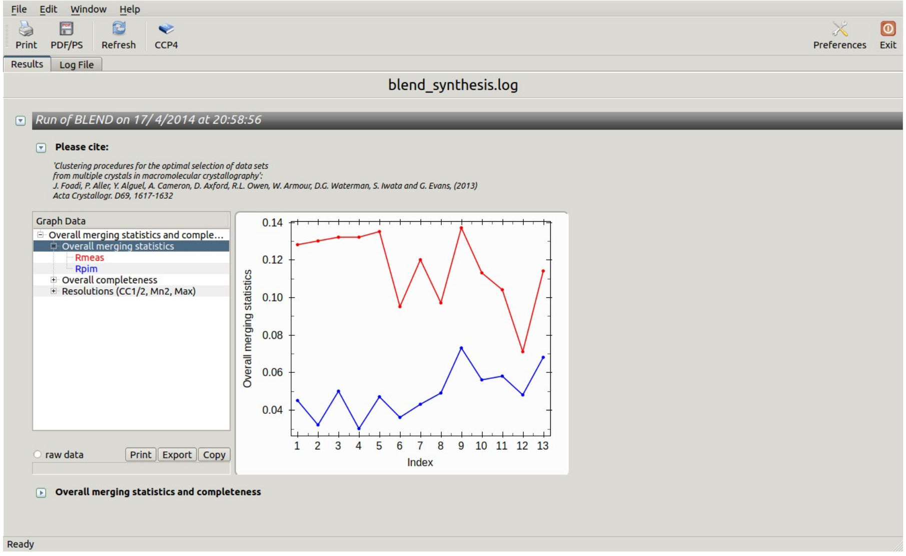
```
logview blend_analysis.log
```

To run BLEND in synthesis mode and produce scaled data for all nodes of the dendrogram, type:

```
blend -s 14 < blend_keywords.dat > blend_synthesis.log
```

At completion the result can, again, be viewed using “logview”:

```
logview blend_synthesis.log
```



blend_synthesis.log

Run of BLEND on 17/ 4/2014 at 20:58:56

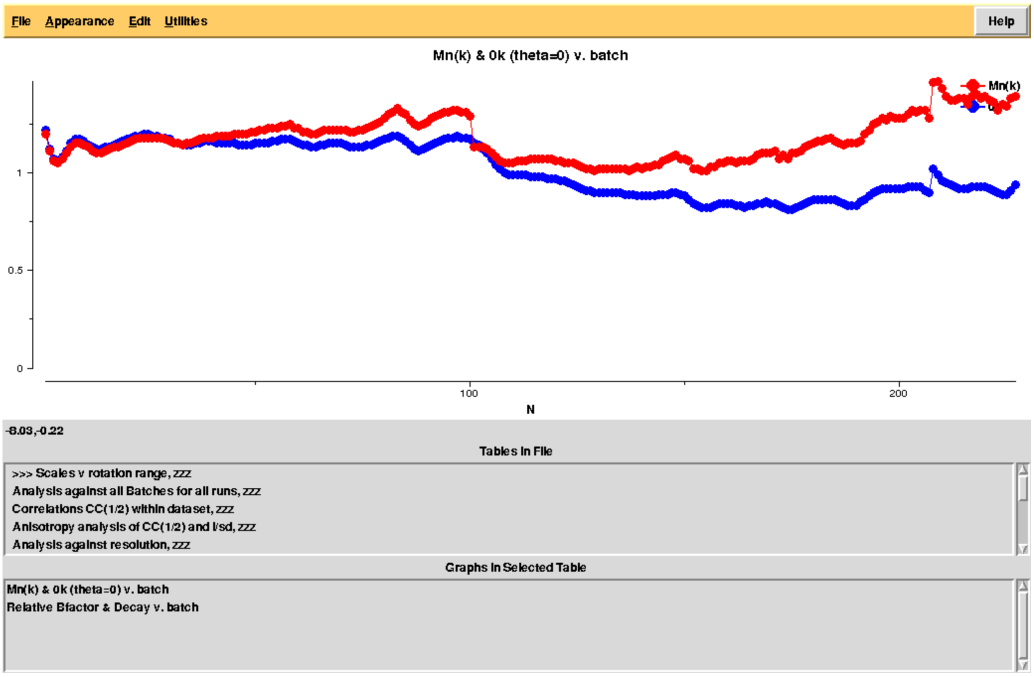
Please cite:

'Clustering procedures for the optimal selection of data sets
from multiple crystals in macromolecular crystallography':
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Overall merging statistics and completeness

All unscaled and scaled mtz files, and all POINTLESS and AIMLESS logs, are contained in the "merged_files" directory. Logs can be viewed using "logview" or "loggraph". For example:

```
loggraph merged_files/aimless_005.log
```



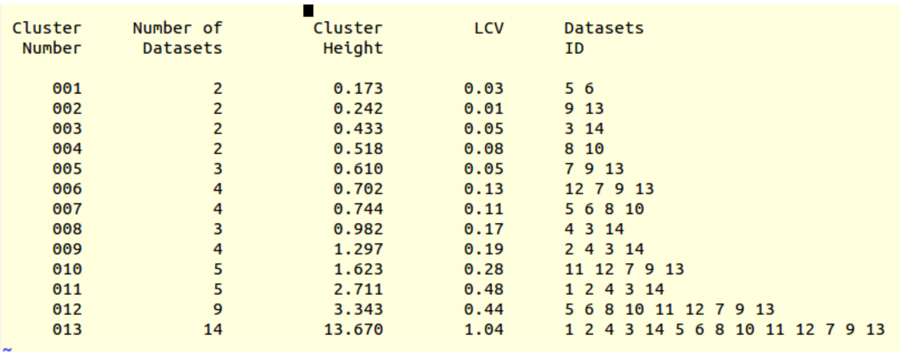
Tables In File

>>> Scales v rotation range, zzz
Analysis against all Batches for all runs, zzz
Correlations CC(1/2) within dataset, zzz
Anisotropy analysis of CC(1/2) and I/sd, zzz
Analysis against resolution, zzz

Graphs In Selected Table

Mn(k) & 0k (theta=0) v. batch
Relative Bfactor & Decay v. batch

In order to execute BLEND in combination mode, so to scale groups of data sets not corresponding to any cluster, it is handy to copy and paste numeric strings from "CLUSTERS.txt":



```
Cluster      Number of          Cluster          LCV        Datasets
 Number       Datasets           Height                      ID

    001              2            0.173         0.03        5 6
    002              2            0.242         0.01        9 13
    003              2            0.433         0.05        3 14
    004              2            0.518         0.08        8 10
    005              3            0.610         0.05        7 9 13
    006              4            0.702         0.13        12 7 9 13
    007              4            0.744         0.11        5 6 8 10
    008              3            0.982         0.17        4 3 14
    009              4            1.297         0.19        2 4 3 14
    010              5            1.623         0.28        11 12 7 9 13
    011              5            2.711         0.48        1 2 4 3 14
    012              9            3.343         0.44        5 6 8 10 11 12 7 9 13
    013             14           13.670         1.04        1 2 4 3 14 5 6 8 10 11 12 7 9 13
~
```

For instance, string “7 9 13” from cluster 6 and string “8 10” from cluster 4 can be both copied, pasted and joined into a new string (one needs to have good reasons for doing it!):

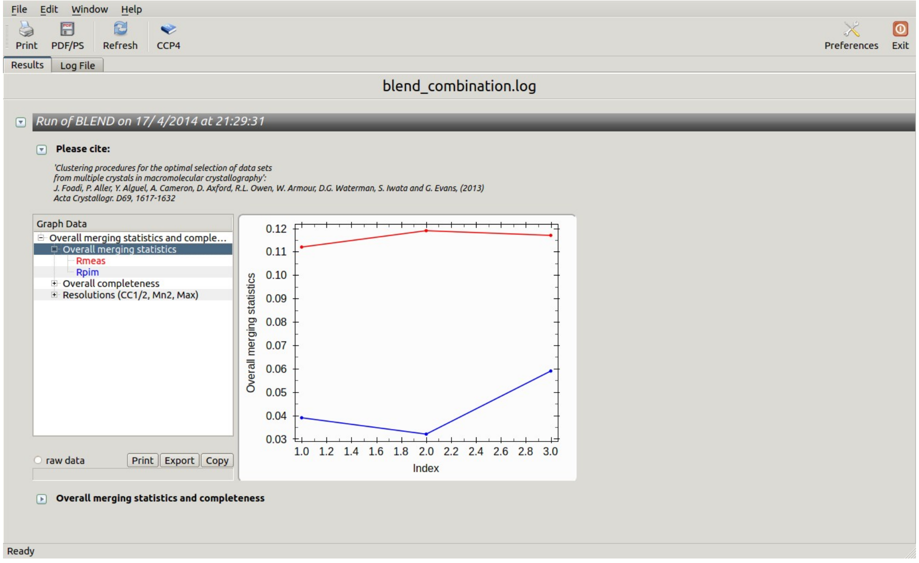
```
blend -c 7 9 13 8 10 < blend_keywords.dat
```

Next, we can try combination “4 3 14” + “5 6”, and combination “3 14” + “7 9 13”:

```
blend -c 4 3 14 5 6 < blend_keywords.dat
blend -c 3 14 7 9 13 < blend_keywords.dat > blend_combination.log
```

All files produced will be contained in directory “combined_files”. The final log file will include whatever is contained in “combined_files”:

```
logview blend_combination.log
```



blend_combination.log

Run of BLEND on 17/ 4/2014 at 21:29:31

Please cite:

'Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography':
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Graph Data
Overall merging statistics and comple...
Overall merging statistics
Rmeas
Rpim
Overall completeness
Resolutions (CC1/2, Mn2, Max)
Overall merging statistics
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
Index
○ raw data
Print
Export
Copy
Overall merging statistics and completeness
Ready

TUTORIAL 2.

28 SWEEPS OF LYSOZYME CRYSTALS
(data collected by Pierre Aller)

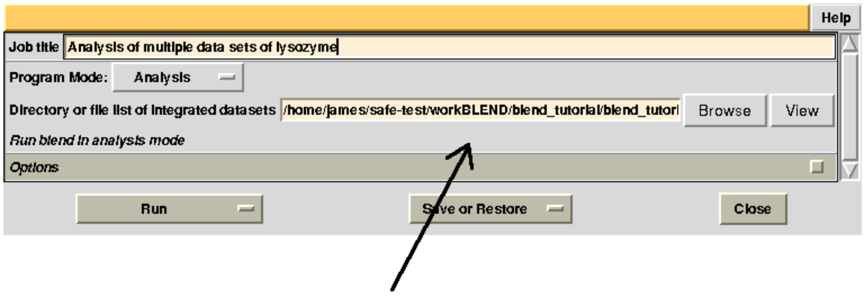
How to obtain complete data sets, while preserving isomorphism with *BLEND*

These data come from hen egg lysozyme with two different procedures. The first group of 12 10-degree sweeps has been collected with pure lysozyme crystals at room temperature from crystallisation plates. The second group of 16 10-degree sweeps has been obtained from crystals soaked in a solution of sodium bromide; data have been collected at the Diamond Light Source synchrotron with a wavelength close to the bromide peak. Data collection went further than 10 degrees for each data set, but we have limited sweeps in order to replicate similar situations met when radiation damage acts fast. For this tutorial in situ data have been stacked with data collected at low temperature and from a slightly different structure (additional bromide content) in order to demonstrate the interplay between completeness and non-isomorphism.

Let's create a new directory, called “lysozyme”, in the “blend_tutorial” directory, and associate the new project directory “BLEND_TUTORIAL_LYSOZYME” to “lysozyme”:

```
mkdir lysozyme_test
```

Next, we run *BLEND* in analysis mode using all mtz files in “data/lysozyme”:



Job title: Analysis of multiple data sets of lysozyme

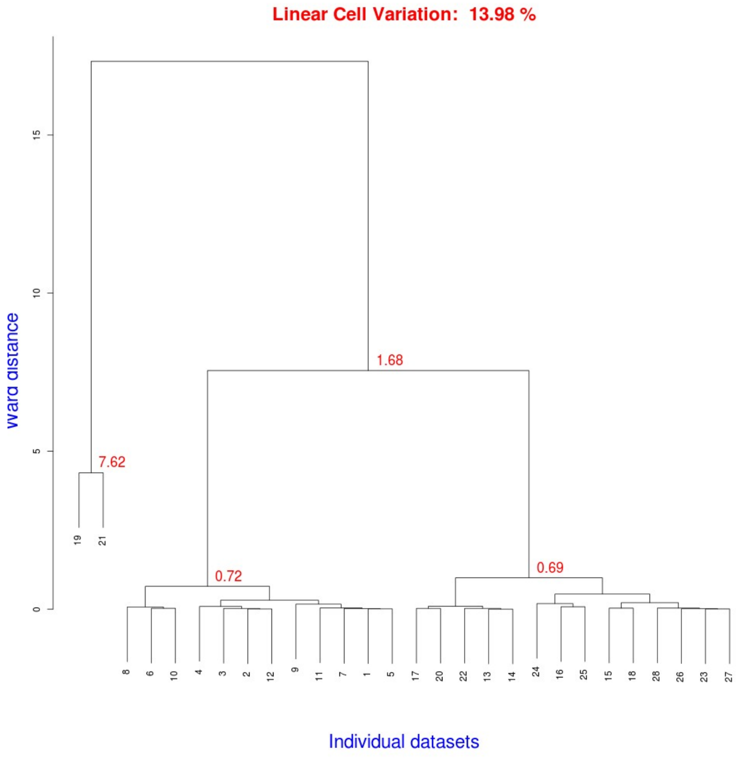
Program Mode: Analysis

Directory or file list of integrated datasets: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

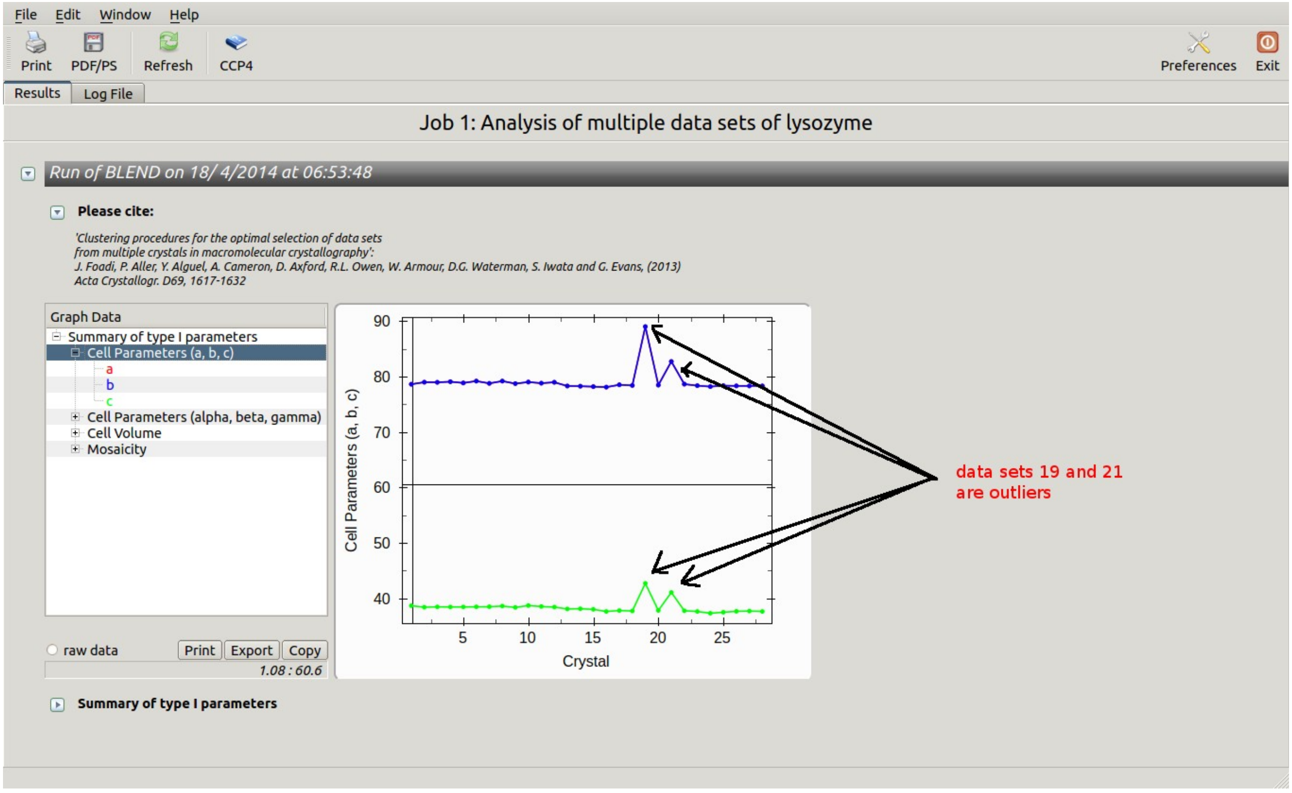
Run blend in analysis mode

select data in directory "data/lysozyme"

In the group of files produced by this run the following dendrogram (file “tree.ps” or “tree.png”) is found:



The overall Linear Cell Variation (LCV) has a high value. This is certainly caused by the couple of crystals in the isolated branch on the left of the dendrogram, as it can be confirmed by looking at cell parameters plots in BLEND log file (new style):



Job 1: Analysis of multiple data sets of lysozyme

Run of BLEND on 18/ 4/2014 at 06:53:48

Please cite:

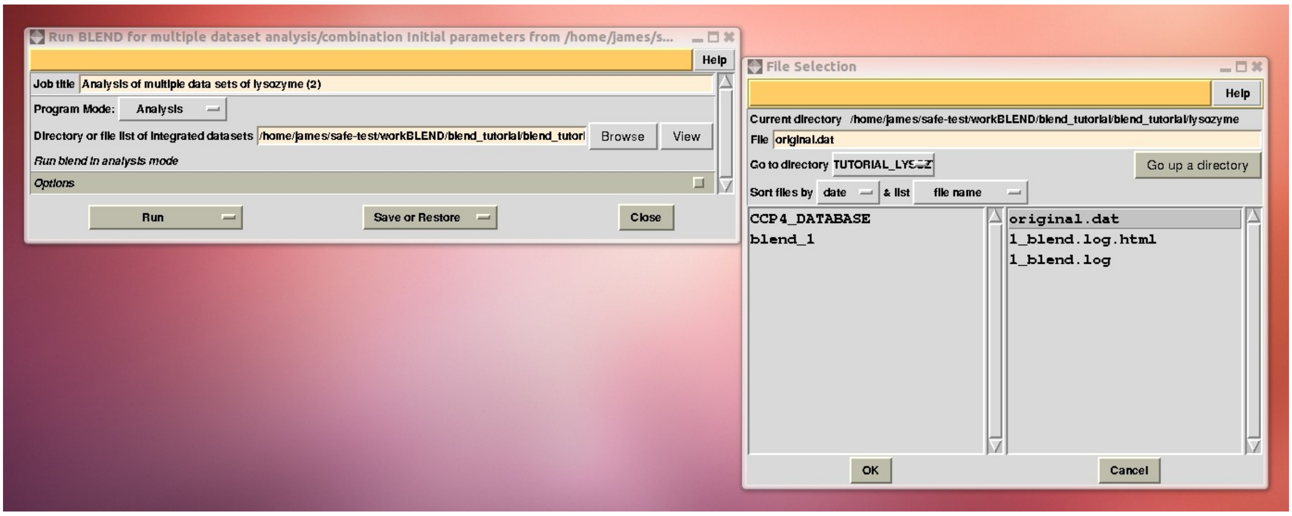
'Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography':
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Summary of type I parameters

Data consist of many short sweeps (not useful to form a complete data set), therefore we can get rid of sweeps 19 and 21 without loosing much data and re-run *BLEND* in analysis mode. One way of executing *BLEND* using the same data without sweeps 19 and 21 is by renaming file “mtz_names.dat” as “original.dat” and deleting the lines pointing to these sweeps:

```
mv lysozyme/blend_1/mtz_names.dat lysozyme/original.dat
[edit “lysozyme/original.dat” and delete lines containing “dataset_019.mtz” and
        “dataset_021.mtz”]
```

BLEND is now executed using “original.dat” as input:



Run BLEND for multiple dataset analysis/combination Initial parameters from /home/james/s...

Job title: Analysis of multiple data sets of lysozyme (2)

Program Mode: Analysis

Directory or file list of integrated datasets: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Run blend in analysis mode

Options □

File Selection

Current directory /home/james/safe-test/workBLEND/blend_tutorial/blend_tutorial/lysozyme

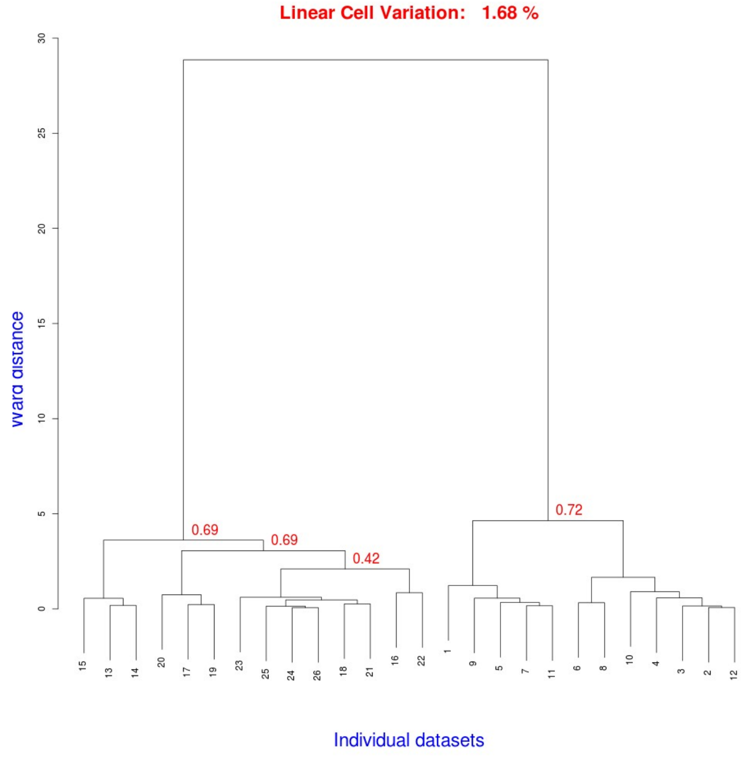
File: original.dat

Go to directory: TUTORIAL_LYSOZ

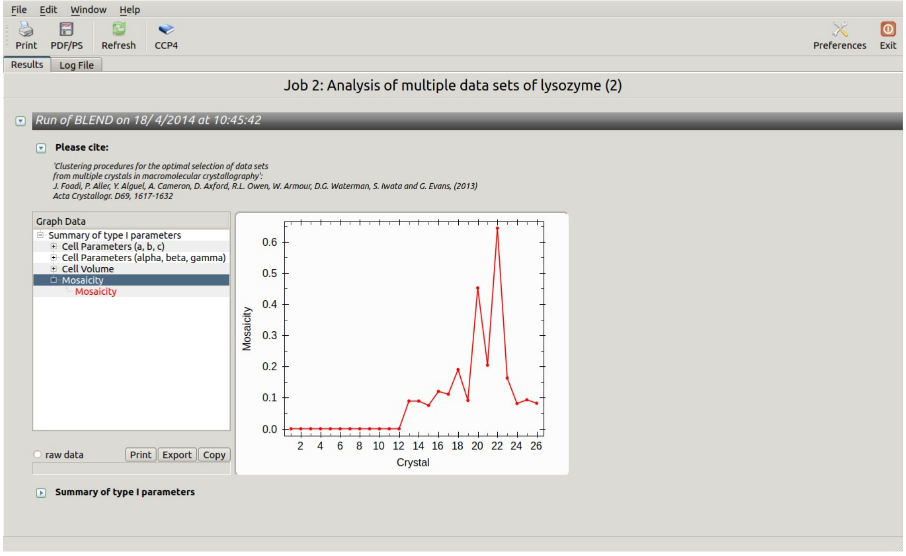
Sort files by: date & list: file name

```
CCP4_DATABASE        original.dat
blend_1              1_blend.log.html
                     1_blend.log
```

The ensuing dendrogram is:



As we can see, the new LCV is much smaller, although the value 1.68% still indicates some non-isomorphism among all crystals. We know this is the case because crystals have been assembled from differently-prepared structures. Indeed, LCV values computed separately for the left branch and for the right branch of the dendrogram return 0.69% and 0.72%, respectively. Incidentally, data sets group 1 – 12 has different features from group 13 – 26 (remember, two sweeps missing from the initial group!) because it corresponds to a different collection from a different group of crystals. This coarse division, observed in the two separate dendrogram's branches, can also be hinted by looking at the different mosaicity trends in BLEND log file:



Job 2: Analysis of multiple data sets of lysozyme (2)

Run of BLEND on 18/ 4/2014 at 10:45:42

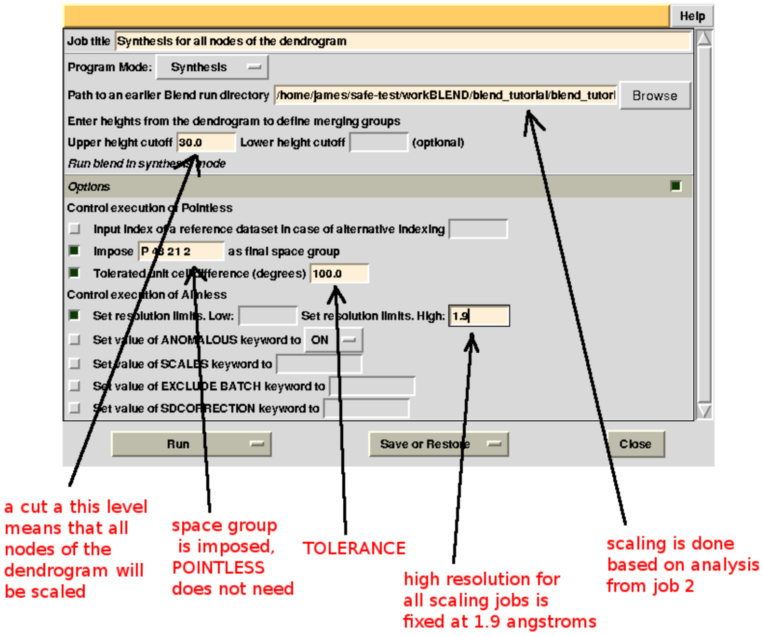
Please cite:

'Clustering procedures for the optimal selection of data sets
from multiple crystals in macromolecular crystallography':
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Scaled data for all nodes of the new dendrogram can be computed running *BLEND* in synthesis mode. We know that all data should belong to space group P 4_3 2_1 2. We also would like to use the same resolution for all data, say 1.9 Å. Accordingly, the program is executed using the following keywords:

```
CHOOSE SPACEGROUP P 43 21 2
TOLERANCE 100
RESO HIGH 1.9
```

Keyword TOLERANCE is set to a high value to bypass the default tendency of *POINTLESS* to stop execution when cell parameters differ too much. Considering now the dendrogram, using only one value for the height means that the program would process all nodes below this value. In the “CLUSTERS.txt” file it can be seen that these have values smaller or equal to 28.861; thus any value larger than this produces the desired scaled data:



Job title: Synthesis for all nodes of the dendrogram

Program Mode: Synthesis

Path to an earlier Blend run directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutorl

Enter heights from the dendrogram to define merging groups

Upper height cutoff: 30.0 Lower height cutoff: (optional)

Run blend in synthesis mode

Control execution of Pointless

□ Input index of a reference dataset in case of alternative indexing

■ Impose P 43 21 2 as final space group

■ Tolerated unit cell difference (degrees) 100.0

Control execution of Aimless

■ Set resolution limits. Low: Set resolution limits. High: 1.9

□ Set value of ANOMALOUS keyword to ON

□ Set value of SCALES keyword to

□ Set value of EXCLUDE BATCH keyword to

□ Set value of SDCORRECTION keyword to

a cut a this level means that all nodes of the dendrogram will be scaled

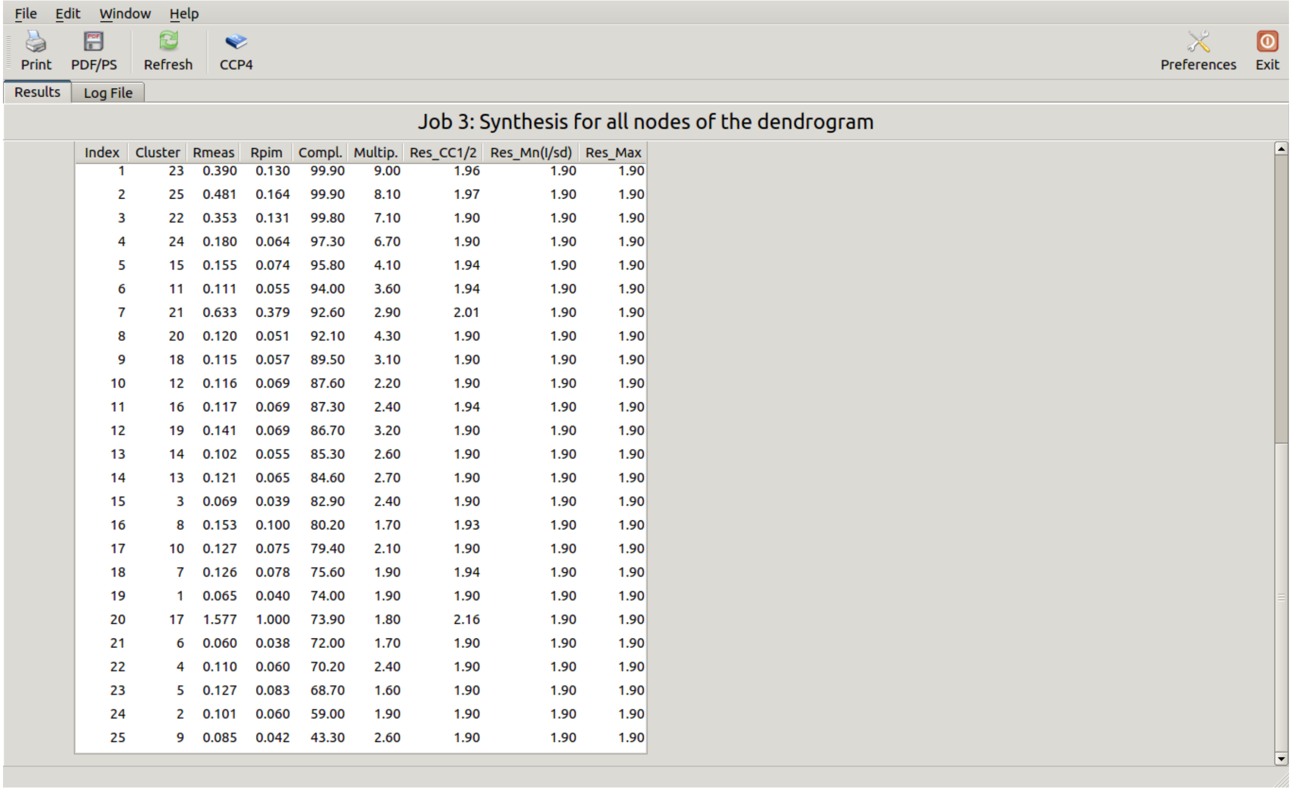
space group is imposed, POINTLESS does not need

TOLERANCE

high resolution for all scaling jobs is fixed at 1.9 angstroms

scaling is done based on analysis from job 2

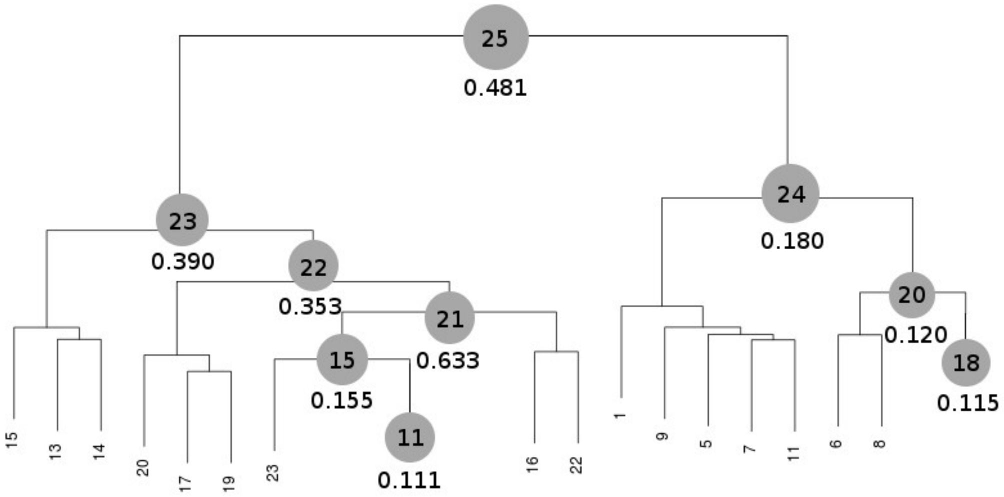
This run will take a while to complete (many clusters!). At the end of execution a new “merged_files” directory is created with scaled files and merging statistics tabulated. Here is a view of the final statistics in BLEND log:



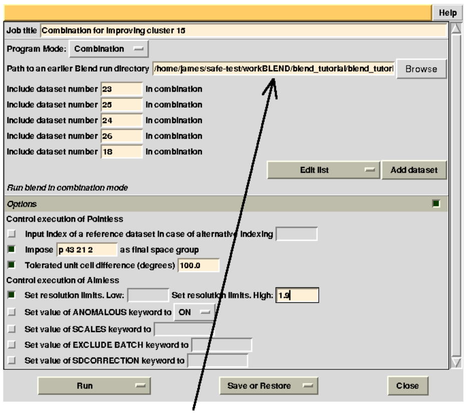
Job 3: Synthesis for all nodes of the dendrogram

Index	Cluster	Rmeas	Rpim	Compl.	Multip.	Res_CC1/2	Res_Mn(I/sd)	Res_Max
1	23	0.390	0.130	99.90	9.00	1.96	1.90	1.90
2	25	0.481	0.164	99.90	8.10	1.97	1.90	1.90
3	22	0.353	0.131	99.80	7.10	1.90	1.90	1.90
4	24	0.180	0.064	97.30	6.70	1.90	1.90	1.90
5	15	0.155	0.074	95.80	4.10	1.94	1.90	1.90
6	11	0.111	0.055	94.00	3.60	1.94	1.90	1.90
7	21	0.633	0.379	92.60	2.90	2.01	1.90	1.90
8	20	0.120	0.051	92.10	4.30	1.90	1.90	1.90
9	18	0.115	0.057	89.50	3.10	1.90	1.90	1.90
10	12	0.116	0.069	87.60	2.20	1.90	1.90	1.90
11	16	0.117	0.069	87.30	2.40	1.94	1.90	1.90
12	19	0.141	0.069	86.70	3.20	1.90	1.90	1.90
13	14	0.102	0.055	85.30	2.60	1.90	1.90	1.90
14	13	0.121	0.065	84.60	2.70	1.90	1.90	1.90
15	3	0.069	0.039	82.90	2.40	1.90	1.90	1.90
16	8	0.153	0.100	80.20	1.70	1.93	1.90	1.90
17	10	0.127	0.075	79.40	2.10	1.90	1.90	1.90
18	7	0.126	0.078	75.60	1.90	1.94	1.90	1.90
19	1	0.065	0.040	74.00	1.90	1.90	1.90	1.90
20	17	1.577	1.000	73.90	1.80	2.16	1.90	1.90
21	6	0.060	0.038	72.00	1.70	1.90	1.90	1.90
22	4	0.110	0.060	70.20	2.40	1.90	1.90	1.90
23	5	0.127	0.083	68.70	1.60	1.90	1.90	1.90
24	2	0.101	0.060	59.00	1.90	1.90	1.90	1.90
25	9	0.085	0.042	43.30	2.60	1.90	1.90	1.90

Scaled files corresponding to these results are included in the “merged_files” directory, and they are ready for further investigations. Some of the statistics can be improved by filtering out certain sweeps from specific clusters. A visualization of the above table is provided in the following picture, where clusters with completeness around 90% or better are represented by grey circles and the corresponding R_{meas} value is typed underneath:



Clusters 21, 22, 23 and 25 have clearly unacceptable R_{meas} values, while values for clusters 15 and 24 can, possibly, be improved. The main way to improve statistics at this point is by running *BLEND* in combination mode. Let's try, for instance, to improve the statistics for cluster 15. As the R_{meas} value jumps from 0.111 for cluster 11 to 0.155 for cluster 15 and as cluster 15 is simply cluster 11 with the addition of sweep 23, it is straightforward to impute the reason for the increase to the bad quality of sweep 23. But there might also be other possibilities. For example sweep 23 does not match well some of the sweeps composing cluster 11; by excluding these sweeps, then, statistics might improve. In conclusion we can try and couple sweep 23 with cluster 11, excluding in turn one of the sweeps forming cluster 11. For all combinations after the first we will have to make sure that the path to an earlier BLEND directory points to the results directory of the previous combination. So for the first combination:



Job title: Combination for improving cluster 15

Program Mode: Combination

Path to an earlier Blend run directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Include dataset number 23 in combination
Include dataset number 25 in combination
Include dataset number 24 in combination
Include dataset number 26 in combination
Include dataset number 18 in combination

Run blend in combination mode

Options ☒

Control execution of Pointless

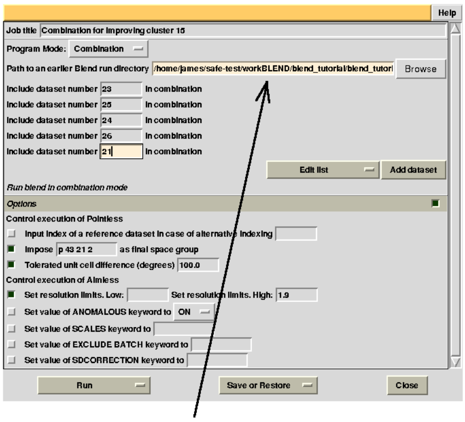
□ Input index of a reference dataset in case of alternative indexing
☒ Impose p 43 21 2 as final space group
☒ Tolerated unit cell difference (degrees) 100.0

Control execution of Aimless

☒ Set resolution limits. Low: Set resolution limits. High: 1.9
□ Set value of ANOMALOUS keyword to ON
□ Set value of SCALES keyword to
□ Set value of EXCLUDE BATCH keyword to
□ Set value of SDCORRECTION keyword to

this points to directory
"lysozyme/blend_2"

For the second combination:



Job title: Combination for improving cluster 15

Program Mode: Combination

Path to an earlier Blend run directory: /home/james/safe-test/workBLEND/blend_tutorial/blend_tutori

Include dataset number 23 in combination
Include dataset number 25 in combination
Include dataset number 24 in combination
Include dataset number 26 in combination
Include dataset number 21 in combination

Run blend in combination mode

Options

Control execution of Pointless

□ Input index of a reference dataset in case of alternative indexing
☑ Impose p 43 21 2 as final space group
☑ Tolerated unit cell difference (degrees) 100.0

Control execution of Aimless

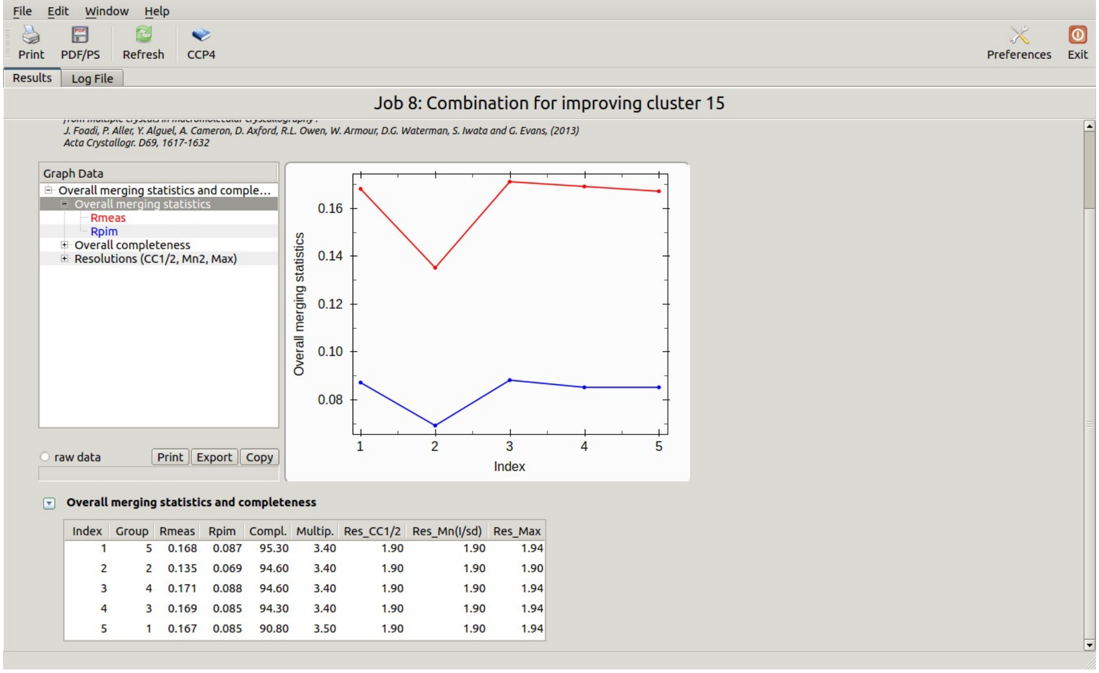
☑ Set resolution limits. Low: Set resolution limits. High: 1.9
□ Set value of ANOMALOUS keyword to ON
□ Set value of SCALES keyword to
□ Set value of EXCLUDE BATCH keyword to
□ Set value of SDCORRECTION keyword to

this points to directory
"lysozyme/blend_4"

To summarize, for all combinations tried to improve statistics for cluster 15, we have:

combination 23 25 24 26 18 points to directory “lysozyme/blend_2”
combination 23 25 24 26 21 points to directory “lysozyme/blend_4”
combination 23 25 24 18 21 points to directory “lysozyme/blend_5”
combination 23 25 26 18 21 points to directory “lysozyme/blend_6”
combination 23 24 26 18 21 points to directory “lysozyme/blend_7”

Final statistics can be viewed in the “MERGED_STATISTICS.info” file under “lysozyme/blend_8/combined_files” or simply using the interface:



Job 8: Combination for improving cluster 15

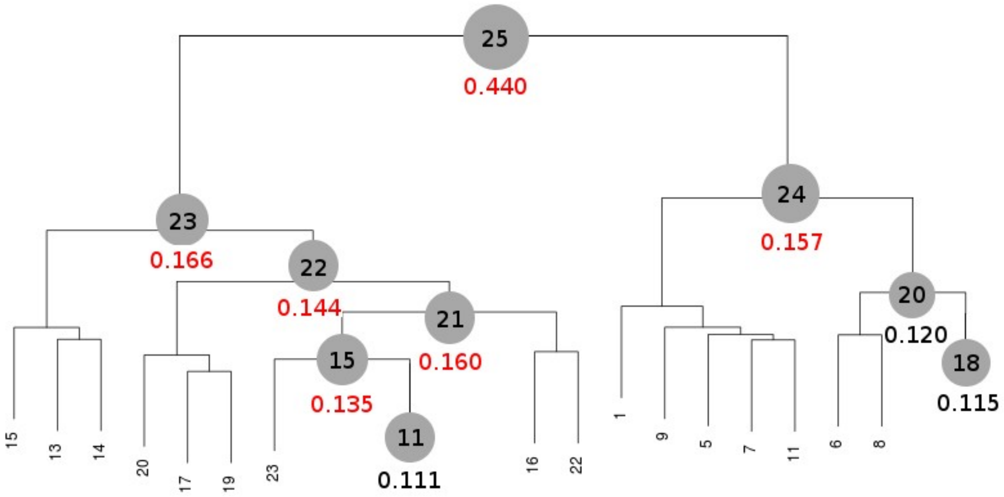
J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, (2013)
Acta Crystallogr. D69, 1617-1632

Overall merging statistics and completeness

Index	Group	Rmeas	Rpim	Compl.	Multip.	Res_CC1/2	Res_Mn(I/sd)	Res_Max
1	5	0.168	0.087	95.30	3.40	1.90	1.90	1.94
2	2	0.135	0.069	94.60	3.40	1.90	1.90	1.90
3	4	0.171	0.088	94.60	3.40	1.90	1.90	1.94
4	3	0.169	0.085	94.30	3.40	1.90	1.90	1.94
5	1	0.167	0.085	90.80	3.50	1.90	1.90	1.94

The lowest Rmeas is the one for group 2. Thus, discarding sweep 18 from cluster 15 causes Rmeas to improve from 0.155 to 0.135, while completeness decrease just of a little from 95.8% to 94.6%.

Similar combinations can be tried on all other clusters. Results are shown in the following picture:



Values shown in red correspond to those cases where filtering out one or more sweeps has made it possible to lower Rmeas.

Try to reproduce the same results by executing *BLEND* in combination mode. As a hint you can find it useful to know that the sweeps discarded are 11, 15, 18 and 22.

To conclude this tutorial, use *BLEND* in combination mode with just one element at a time to reproduce scaled data and statistics for individual sweeps. It is interesting to observe that the 4 data sets excluded from the previous filtering job do not always correspond to those having worst Rmeas values.

A final remark: all scaled mtz files do not contain structure factors, but only scaled intensities. To create structure factors and proceed with phasing, model building and refinement, you need to run the CCP4 program *TRUNCATE* or any other software that does an equivalent job.

REFERENCES

(2011) Q. Liu, Z. Zhang and W.A. Hendrickson, “Multi-crystal anomalous diffraction for low-resolution macromolecular phasing”, Acta Cryst. **D**67, 45-49

(2012) R. Giordano, R.M.F. Leal, G.P. Bourenkov, S. McSweeney and A.N. Popov, “The application of hierarchical cluster analysis to the selection of isomorphous crystals”, Acta Cryst. **D**68, 649-658

(2012) D. Axford, R.L. Owen, J. Aishima, J. Foadi, A.W. Morgan, J.I. Robinson, J.E. Nettleship, R.J. Owens, I. Moraes, E.E. Fry, J.M. Grimes, K. Harlos, A. Kotecha, J. Ren, G. Sutton, T.S. Walter, D.I. Stuart and G. Evans, “In situ macromolecular crystallography using microbeams”, Acta Cryst. **D**68, 592-600

(2012) M.A. Hanson, C.B. Roth, E. Jo, M.T. Griffith, F.L. Scott, G. Reinhart, H. Desale, B. Clemons, S.M. Cahalan, S.C. Schuerer, M.G. Sanna, G.W. Han, P. Kuhn, H. Rosen, R.C. Stevens, “Crystal structure of a lipid G protein-coupled receptor

(2013) J. Foadi, P. Aller, Y. Alguel, A. Cameron, D. Axford, R.L. Owen, W. Armour, D.G. Waterman, S. Iwata and G. Evans, “Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography”, Acta Cryst. **D**69, 1617-1632, Science **335**, 851-855