Clustering for multi-crystal data collection

A tool for Synchrotron Serial Crystallography data processing

Gianluca Santoni

ESRF (Grenoble, France)

Users Meeting, February 8th 2016

ESRF

Acknowledgements

- ESRF
  - Alexander Popov
  - Olof Svenson
  - Uli Zander
  - Igor Melnikov

ESRF

Hierarchical cluster analysis and protein crystallography

ESRF

HCA: introduction

A method to cluster large ensambles of data

- Distance matrix
  - Define distance between all pairs of datasets
  - Use unit cell variation
  - Correlation coefficient between datasets

HCA: introduction

A method to cluster large ensambles of data

- Distance matrix
  - Define distance between all pairs of datasets
  - Use unit cell variation
  - Correlation coefficient between datasets
- Linkage method
  - Complete linkage
  - Distance between clusters A and B: $max(d_{a,b}); a \in A, b \in B$
  - Gives upper limit of cluster quality

Definitions of distance

- According to unit cell variation
  - $d(a,b) = max(\frac{|A_a - A_b|}{min(A_a, A_b)}, \frac{|B_a - B_b|}{min(B_a, B_b)}, \frac{|C_a - C_b|}{min(C_a, C_b)})$
  - A,B,C unit cell parameters of each dataset

Definitions of distance

- According to unit cell variation
  - $d(a,b) = max(\frac{|A_a - A_b|}{min(A_a, A_b)}, \frac{|B_a - B_b|}{min(B_a, B_b)}, \frac{|C_a - C_b|}{min(C_a, C_b)})$
  - A,B,C unit cell parameters of each dataset
- According to correlation
  - $d(a,b) = \sqrt{1 - cc_{a,b}^2}$
  - $cc_{a,b}$ correlation coefficient between datasets

Implementation of ccCluster

Unit cell variation

- Requirements
  - Used cctbx.miller module
  - Can run on various file formats
  - Still issues with mtz files

Unit cell variation

- Requirements
  - Used cctbx.miller module
  - Can run on various file formats
  - Still issues with mtz files
- Implementation in python
  - Load all HKL files as cctbx arrays
  - Extract unit cell parameters
  - Calculate distance all possible pairs of datasets
  - Write output file

Unit cell variation

- Requirements
  - Used cctbx.miller module
  - Can run on various file formats
  - Still issues with mtz files
- Implementation in python
  - Load all HKL files as cctbx arrays
  - Extract unit cell parameters
  - Calculate distance all possible pairs of datasets
  - Write output file

```
Correlation coefficients
17   67   0.979925098558
17   50   0.966456682722
17   21   0.558234058272
17   86   0.978797299202
17   62   0.959380884366
17   32   0.976804797012
17   40   0.983532727402
17   51   0.97603906066
17   85   0.923668052721
17   19   0.985738729727
17   65   0.982510403238
17   68   0.97893905106
17   23   0.960741701806
17   74   0.937200562712
17   48   0.986066424517
17   66   0.207813693861
17   69   0.969588892201
```

ESRF

Correlation coefficients calculation

- Requirements
  - Used cctbx.miller module
  - Can run on various file formats
  - Still issues with mtz files

Correlation coefficients calculation

- Requirements
  - Used cctbx.miller module
  - Can run on various file formats
  - Still issues with mtz files
- Implementation in python
  - Load all HKL files as cctbx arrays
  - Extend the reflections
  - Loop over all possible pairs of datasets
  - Calculate correlation coefficients
  - Write output file

Correlation coefficients calculation

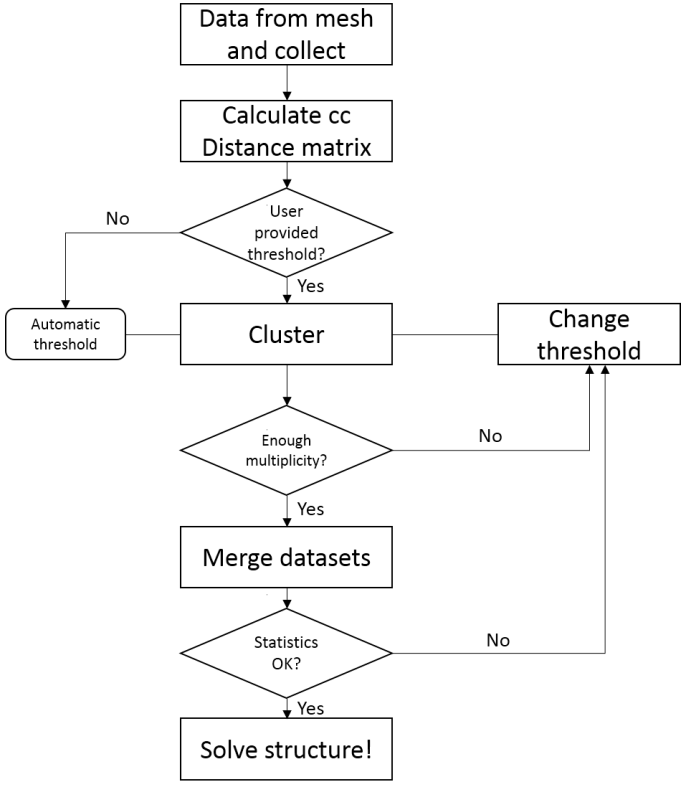
- Requirements
  - Used cctbx.miller module
  - Can run on various file formats
  - Still issues with mtz files
- Implementation in python
  - Load all HKL files as cctbx arrays
  - Extend the reflections
  - Loop over all possible pairs of datasets
  - Calculate correlation coefficients
  - Write output file

```
Correlation coefficients
17   67   0.979925098558
17   50   0.966456682722
17   21   0.558234058272
17   86   0.978797299202
17   62   0.959380884366
17   32   0.976804797012
17   40   0.983532727402
17   51   0.97603906066
17   85   0.923668052721
17   19   0.985738729727
17   65   0.982510403238
17   68   0.97893905106
17   23   0.960741701806
17   74   0.937200562712
17   48   0.986066424517
17   66   0.207813693861
17   69   0.969588892201
```

ESRF

Cluster and merge: implementation

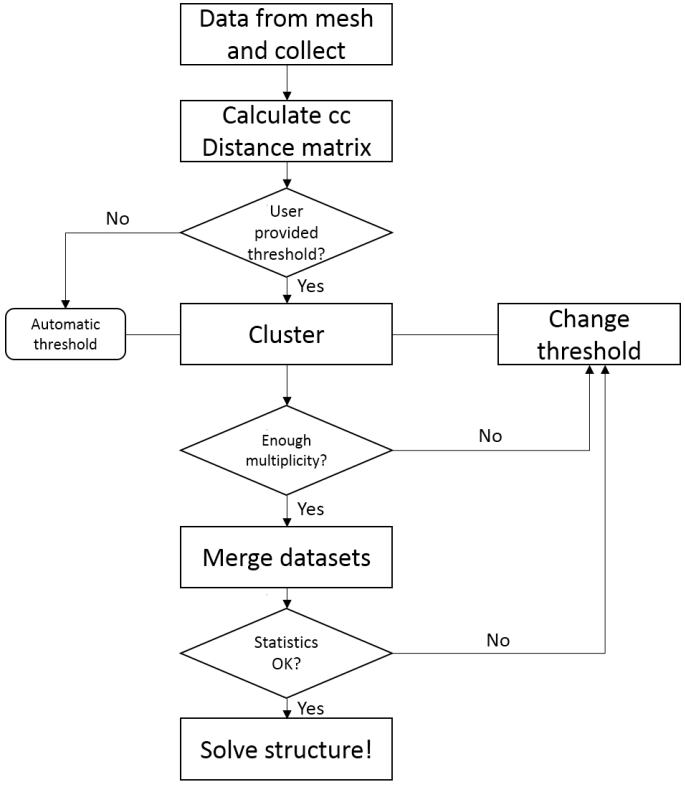
- Clustering
  - Read cc file and generate distance matrix
  - Create a clustering object
  - Methods: plot dendrogram, merge, write log, estimate threshold



ESRF

Cluster and merge: implementation

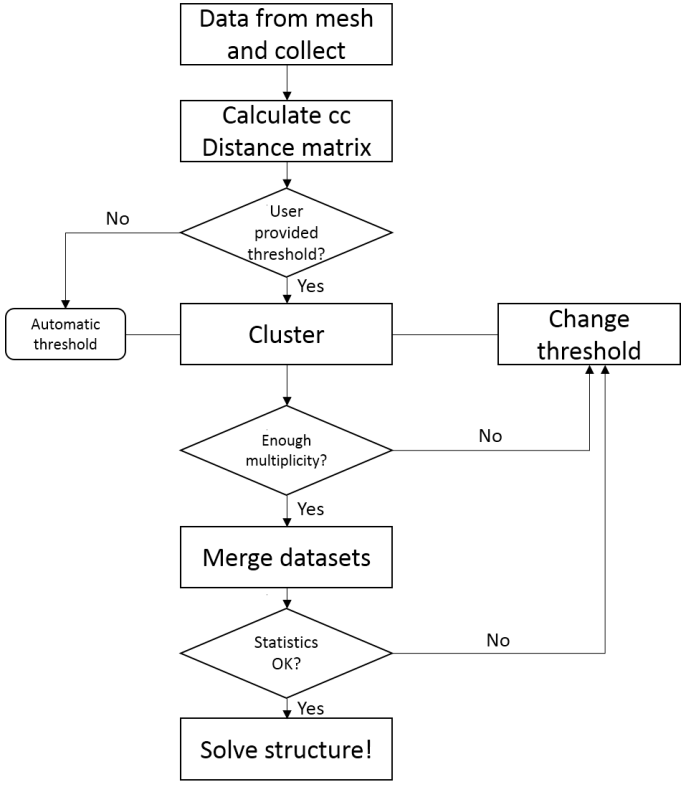
- Clustering
  - Read cc file and generate distance matrix
  - Create a clustering object
  - Methods: plot dendrogram, merge, write log, estimate threshold
- Gui
  - Visualize Dendrogram and run merging
  - Observe the results of merging
  - Visualise summary of statistics



ESRF

Cluster and merge: implementation

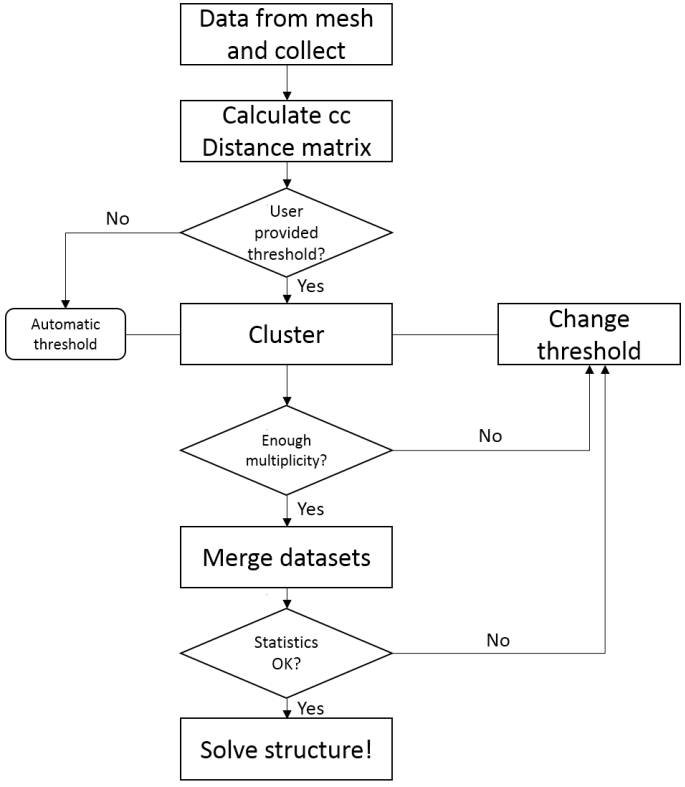
- Clustering
  - Read cc file and generate distance matrix
  - Create a clustering object
  - Methods: plot dendrogram, merge, write log, estimate threshold
- Gui
  - Visualize Dendrogram and run merging
  - Observe the results of merging
  - Visualise summary of statistics
- ResultsTab
  - Show dendrogram
  - Plot statistics vs resolution



ESRF

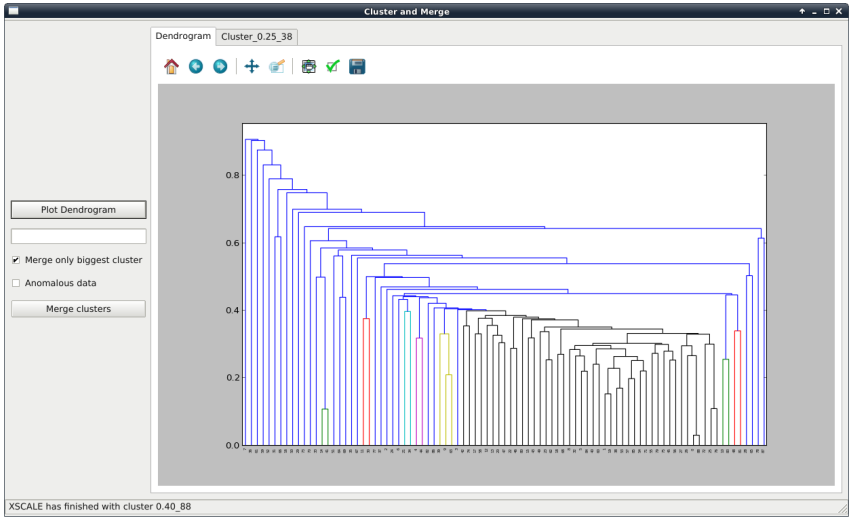
Cluster and merge: implementation

- Clustering
  - Read cc file and generate distance matrix
  - Create a clustering object
  - Methods: plot dendrogram, merge, write log, estimate threshold
- Gui
  - Visualize Dendrogram and run merging
  - Observe the results of merging
  - Visualise summary of statistics
- ResultsTab
  - Show dendrogram
  - Plot statistics vs resolution
- Summary
  - Synthesis of the results
  - Suggests which cluster to chose, according to target



Cluster and merge: gui usage

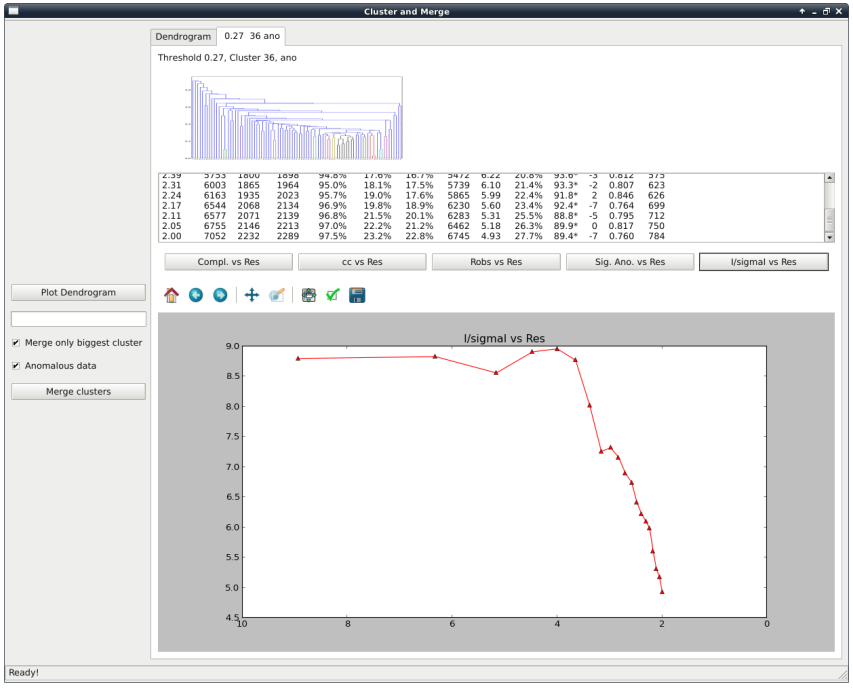
- Main panel
  - Dendrogram. Set threshold by clicking
  - Data merging options and run
  - See results summary



ESRF

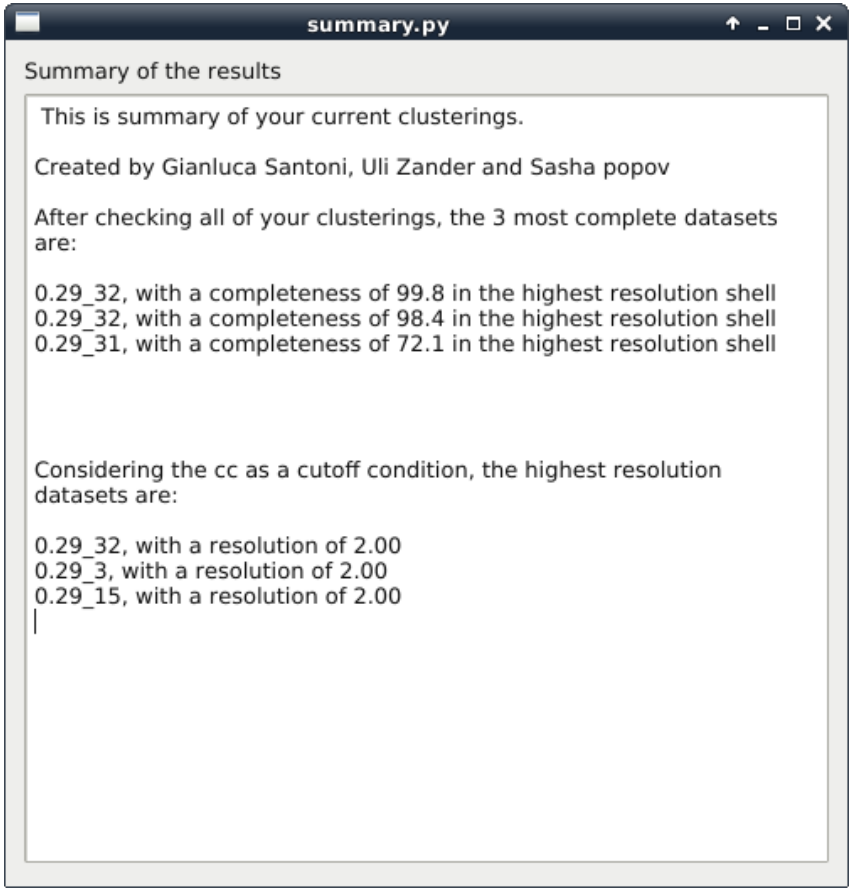
Cluster and merge: gui usage

- Main panel
  - Dendrogram. Set threshold by clicking
  - Data merging options and run
  - See results summary
- Results tab
  - Dendrogram used to cluster
  - Statistics from XSCALE.LP
  - Plot stats as function of resolution



Cluster and merge: gui usage

- Main panel
  - Dendrogram. Set threshold by clicking
  - Data merging options and run
  - See results summary
- Results tab
  - Dendrogram used to cluster
  - Statistics from XSCALE.LP
  - Plot stats as function of resolution
- Summary window



ESRF

Conclusion

What I talked about

- Promising first tests of clustering (as Uli showed)
- We can use cc and unit cell to compare datasets
- Clustering with cc has been tested widely for mesh and collect
- GUI of ccCluster is ready to use (please, ask if interested)

What I did not talk about

- Comparison between the two definitions of distances
  - Few common reflections or error in unit cell determination
  - Lowest angle wedges possible to be determined
- Other applications
  - SAD experiments
  - Weak diffracting / radiation sensitive crystals

Thank you

ESRF