Implementation of ccCluster 0000

Conclusion 00

Clustering for multi-crystal data collection A tool for Synchrotron Serial Crystallography data processing

Gianluca Santoni

ESRF (Grenoble, France)

Users Meeting, February 8th 2016



Acknowledgements

Implementation of ccCluster 0000

Conclusion 00

ESRF

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



Implementation of ccCluster 0000

Conclusion 00

Hierarchical cluster analysis and protein crystallography



Hierarchical cluster analysis and protein crystallography ${ \bullet } { \bigcirc }$

HCA: introduction

Implementation of ccCluster 0000

Conclusion 00

A method to cluster large ensambles of data

 $\hfill\square$ Distance matrix

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



Hierarchical cluster analysis and protein crystallography ${ \bullet } { \bigcirc }$

Implementation of ccCluster 0000

Conclusion 00

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
- $\hfill\square$ 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

Implementation of ccCluster 0000

Conclusion 00

 $\hfill\square$ 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



Implementation of ccCluster 0000

Conclusion 00

Definitions of distance

 $\hfill\square$ 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

 $\hfill\square$ According to correlation

$$\bullet \ d(a,b) = \sqrt{1 - cc_{a,b}^2}$$

■ *cc_{a,b}* correlation coefficient between datasets



Implementation of ccCluster

Conclusion 00

Implementation of ccCluster



Unit cell variation

Implementation of ccCluster

•000

Conclusion 00

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



Unit cell variation

Implementation of ccCluster

•000

Conclusion 00

- Used cctbx.miller module
- Can run on various file formats
- Still issues with mtz files
- $\hfill\square$ 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

Implementation of ccCluster

Conclusion 00

- Used cctbx.miller module
- Can run on various file formats
- Still issues with mtz files
- $\hfill\square$ 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



Implementation of ccCluster OOOO Conclusion 00

Correlation coefficients calculation

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



Implementation of ccCluster

Conclusion 00

Correlation coefficients calculation

- Used cctbx.miller module
- Can run on various file formats
- Still issues with mtz files
- $\hfill\square$ 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



Implementation of ccCluster $\bigcirc \bigcirc \bigcirc \bigcirc$

Conclusion 00

Correlation coefficients calculation

- Used cctbx.miller module
- Can run on various file formats
- Still issues with mtz files
- $\hfill\square$ 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

on coefficients 0.979925098558 0.966456682722
0.979925098558 0.966456682722
0.966456682722
0 550004050070
0.008234008272
0.978797299202
0.959380884366
0.976804797012
0.983532727402
0.97603906066
0.923668052721
0.985738729727
0.982510403238
0.97893905106
0.960741701806
0.937200562712
0.986066424517
0.207813693861
0.969588892201



Implementation of ccCluster $\bigcirc \bigcirc \bullet \bigcirc$

Conclusion 00

Cluster and merge: implementation

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





Implementation of ccCluster $\bigcirc \bigcirc \bullet \bigcirc$

Conclusion 00

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





Implementation of ccCluster $\bigcirc \bigcirc \bullet \bigcirc$

Conclusion 00

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





Implementation of ccCluster $\bigcirc \bigcirc \bullet \bigcirc$

Conclusion 00

Cluster and merge: implementation

- \Box 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





Implementation of ccCluster

Conclusion 00

Cluster and merge: gui usage

□ Main panel

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





Cluster and merge: gui usage

Implementation of ccCluster

Conclusion 00

$\hfill\square$ 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

summary.py 🔶 🗖 Summary.py			
Summary of the results			
This is summary of your current clusterings.			
Created by Gianluca Santoni, Uli Zander and Sasha popov			
After checking all of your clusterings, the 3 most complete datasets are:			
0.29 , $22,$ with a completeness of 99.8 in the highest resolution shell 0.29 , $22,$ with a completeness of 98.4 in the highest resolution shell 0.29 _31, with a completeness of 72.1 in the highest resolution shell			
Considering the cc as a cutoff condition, the highest resolution datasets are:			
0.29_32, with a resolution of 2.00 0.29_3, with a resolution of 2.00 0.29_15, with a resolution of 2.00			



Implementation of ccCluster 0000 Conclusion

Conclusion



Implementation of ccCluster

Conclusion • O

What I talked about

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



Implementation of ccCluster 0000

Conclusion O

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
- $\hfill\square$ Other applications
 - SAD experiments
 - Weak diffracting / radiation sensitive crystals



Implementation of ccCluster 0000 Conclusion 00

Thank you

