Clustering Models with Calibur

Background

At times it is advantageous to analyze large sets of Rosetta models by first clustering structurally similar models together. This approach is particularly useful when analyzing output from de novo structural predictions, as the low-scoring model with the greatest number of neighbors is often identified as the representative model. Recently, the Rosetta community has recognized the clustering tool Calibur (Li and Ng, 2010) as a preferred method for performing this analysis, and efforts are underway to implement this tool within Rosetta.

This protocol describes how to use Calibur to perform clustering analysis, and provides a sample data set and command lines for practice.

For further information on Calibur, please reference "Calibur: a tool for clustering large numbers of protein decoys" by Li and Ng, BMC Bioinformatics 2010, 11:25 (http://www.biomedcentral.com/1471-2105/11/25).

Prerequisites

This protocol requires the following:

- Calibur (http://sourceforge.net/projects/calibur/)
- A directory containing structure files in PDB format for clustering
- A text file listing the PDB files for clustering, one file per line

Protocol

Run Calibur

The program is run as follows, with the following flags recommended:

```
/path/to/Calibur/calibur -n -c A filelist.txt 4.0
```

The flags above are defined as follows:

- -n (optional): disables the filtering of outlier decoys
- -c (optional): specifies the chains used, defaults to "AC", i.e. chains A, C and unspecified
- filelist.txt: a text file that specifies the models, with each line representing the path (from the current directory) to a PDB file
- x (optional): specifies the threshold (in this case, 4.0)

Review Results

Calibur will print results of clustering analysis to the window. The largest three clusters will be listed, with the representative model of each cluster listed before the number of models in that cluster, followed by a list of the models, as follows:

```
S_0179_1_0001.pdb 6: S_0150_1_0001.pdb  S_0437_1_0001.pdb   S_0179_1_0001.pdb   S_0205_0001.pdb S_0257_0001
S_0436_1_0001.pdb 3: S_0134_0001.pdb    S_0436_1_0001.pdb   S_0464_0001.pdb
```

In the above example, the largest cluster is represented by "S_0179_1_0001.pdb" and has six members. The second largest cluster is represented by "S_0436_1_0001.pdb" and has three members.