NAME

cluster5D – a C program for clustering 5-dimensional PCA data

SYNOPSIS

cluster5D [-v] [-fract <integer>] [PCA filename]

DESCRIPTION

cluster5D is a program that performs 5-dimensional clustering based on PCA data. It is written in the C programming language and is available for Linux, MACOSX and Windows. The program needs only a PCA file with 5 or more columns of numbers with the format described below in order to run and gives the user the opportunity for more specific clustering options.

USING THE PROGRAM

cluster5D runs either from the command line or automatically via grcarma (https://github.com/pkoukos/grcarma).

Its general command-line usage is:

```
cluster5D [-v] [-fract <integer>] [PCA filename]
```

Each row of the PCA file (frame) must contain at least 6 columns of numbers: the serial number of frame and the PC1, PC2, PC3, PC4 and PC5 values.

If the [PCA filename] argument isn't passed, the program will search automatically the current directory for an appropriate PCA file produced by the MD trajectories analysis program *carma* (http://utopia.duth.gr/~glykos/carma.html) with the name “carma.PCA.fluctuations.dat” (for cartesian PCA clustering analysis) or “carma.dPCA.fluctuations.dat” (for dihedral PCA clustering analysis).

The *[-v]* argument enables the verbose option that makes the program more talkative.

Finally, the *[-fract <integer>]* argument, when it is enabled, gives the user the opportunity to select a fraction by typing an integer

between 0 and 100 that represents the percentage of frames which the program will aim to assign to clusters. Alternatively, the program calculates the fraction automatically.

cluster5D produces an output file with the name “carma.5-D.clusters.dat” that contains the frames with their PC values that took part in clustering and, additionally, one more column (second column) with the number of cluster in which each frame belongs to.

6308803	5	-1.538978	1.053373	-1.384658	-0.660515	-0.261111
6308854	5	-1.615853	0.992672	-1.505216	-0.728250	-0.279509
6308870	5	-1.597823	1.059237	-1.403670	-0.655842	-0.371640
6308898	5	-1.552765	1.163055	-1.414908	-0.719814	-0.400145
6309025	5	-1.531740	1.019562	-1.425150	-0.759172	-0.298672
6309040	5	-1.576121	1.122798	-1.449414	-0.682392	-0.361660
6309161	5	-1.584523	1.194395	-1.348070	-0.699646	-0.397887
6309248	5	-1.584085	1.258206	-1.270004	-0.762941	-0.476194
6309307	5	-1.540676	0.976120	-1.511790	-0.634898	-0.313048
6309352	5	-1.614160	1.030662	-1.408789	-0.654571	-0.321719

Illustration 1: Part of the output file

Alternatively, cluster5D can be used via grcarma, a GUI of the molecular dynamics analysis program carma. In this case, by selecting to do a dihedral or cartesian PCA in a molecular dynamics trajectory, you can also perform a five dimensional clustering analysis with cluster5D by checking the corresponding option, as shown below.

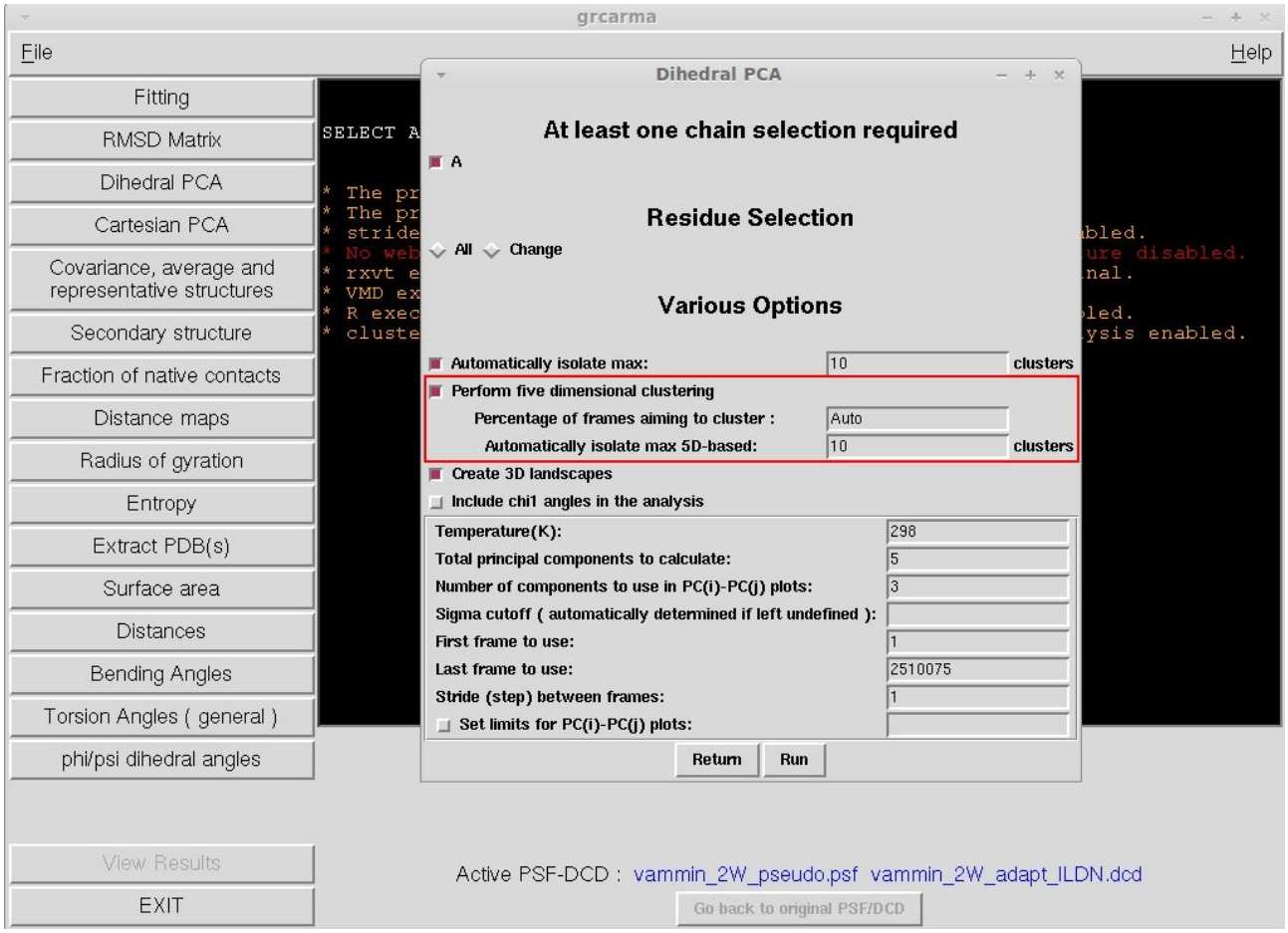

Illustration 2: cluster5D usage example via grcarma

Usage examples:

```
First pass to determine limits ...
Now processing frame  6310325
6310325 frames will enter the calculation.
Second pass to populate the 5D matrix ...
Now processing frame  6310325
Now smoothing ...
Testing density threshold:     443.53
Density threshold set to 443.53.
Clustering now ...
Cluster     1 located, contains    3463020 frames.
Cluster     2 located, contains     302795 frames.
59.68% of frames have been clustered.
All done in 5.1 minutes.
```

Illustration 3: cluster5D output with enabled verbose option and automatically calculated fracrion

```
First pass to determine limits ...
Now processing frame  6310325
6310325 frames will enter the calculation.
Second pass to populate the 5D matrix ...
Now processing frame  6310325
Now smoothing ...
Testing density threshold:    2438.72
Density threshold set to 2438.72.
Clustering now ...
Cluster     1 located, contains    2515919 frames.
All done in 4.9 minutes.
```

Illustration 4: cluster5D output with enabled verbose and fraction options

WHAT IT DOES

The essence of the algorithm is the following : the PC1-PC2-PC3-PC4-PC5 values are used to calculate a (five-dimensional) density distribution function and to map this distribution on a five dimensional matrix. The higher the value of the matrix at a point, the larger the number of frames with corresponding PC values close to that point. The algorithm, then, performs a peak-picking on this 5D density distribution and identifies clusters as sets of frames with similar PC values. The crucial parameter for the peak-picking step is the density threshold above which peaks are picked. Normally, this is calculated automatically from the following equation :

density threshold = mean + pointer * standard deviation

The pointer expresses how many standard deviations above mean the density threshold should be. The algorithm tests different values of pointer starting from 0.0 with a step set to 0.1 and for each corresponding value of the density threshold it calculates the value of a function which has as parameters the percentage of frames, the percentage of explained variance and the number of pixels (points of matrix). The appropriate value of pointer is this which gives the maximum value of the function.

Alternatively, the density threshold is not calculated automatically when the *[-f]* argument is enabled. In this case, the algorithm follows the same testing strategy as described above, but instead of calculating the value of a function, it calculates only the percentage of frames for the corresponding density threshold. This process ends when the percentage of frames calculated become approximately equal to the percentage of frames given by the argument.

The final part of the algorithm is clustering which is performed with the usage of the peak-picking strategy. In more details, the program picks the first peak and, then, searches the values of its nearby pixels. For those pixels with greater or equal values to the density threshold, it searches the values of their nearby pixels and so on. Finally, cluster5D writes in the output file the frames which pixels are marked and, thus, belong to this cluster. The above process is repeated again and again until it is found a peak with a value below the density threshold.

VERSION

Version 0.1, September 2014

AUTHOR

cluster5D has been developed by Athanasios Baltzis, under the supervision of Prof. Nicholas M. Glykos at the Department of Molecular Biology and Genetics of Democritus University of Thrace, Greece.