

# **DPClus: A density-periphery based graph clustering software mainly focused on detection of protein complexes in interaction networks**

Md. Altaf-Ul-Amin, Hisashi Tsuji, Ken Kurokawa, Hiroko Asahi,  
Yoko Shinbo and Shigehiko Kanaya\*

## **Introduction**

DPClus is a graph clustering software, which we have developed based on a clustering algorithm we published previously in the following paper.

[1] Altaf-Ul-Amin M., Shibo, Y., Mihara, K., Kurokawa, K., Kanaya S., “Development and implementation of an algorithm for detection of protein complexes in large interaction networks,” *BMC Bioinformatics*, **7**, 207 (2006)

The algorithm tends to isolate densely connected regions of a graph as clusters. Users have freedom to choose two input parameters within reasonable range and thus to affect the outcome of the clustering up to certain extent. Though this software can be used for graph clustering in general but it mainly focuses on detection of protein complexes in interaction networks. The proposed software makes it possible to detect and visualize clusters of proteins in interaction networks which mostly represent molecular biological functional units. We believe that the present software can be applied not only to other biological networks but also to networks in many other applications where finding cohesive group is an agenda. The present version of DPClus can be applied to a network consisting of up to 5000 proteins/nodes.

## **Instruction Manual**

### **1. Starting DPClus**

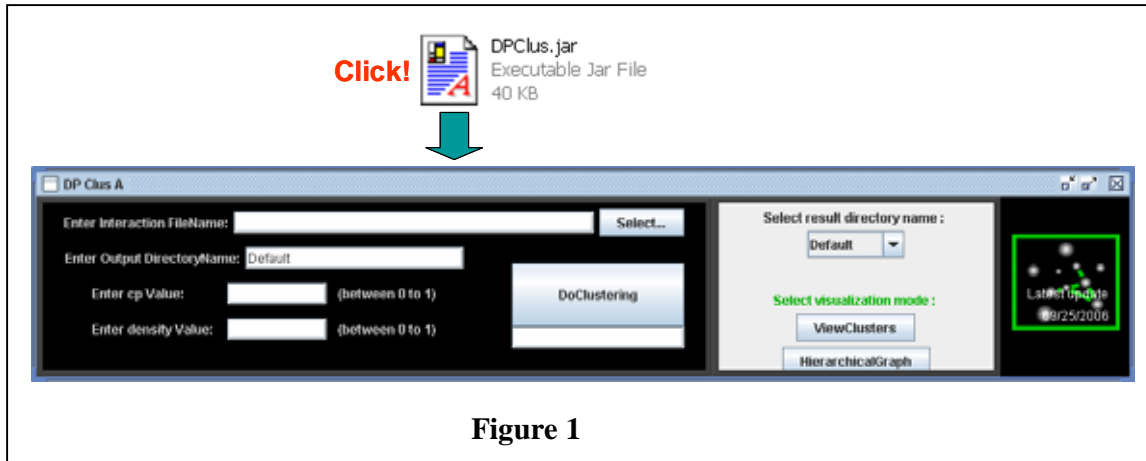
To use DPClus, the user has to install Java j2sdk-1.4.2 in his/her computer. Then the compressed file, DPClus.zip should be downloaded from <http://kanaya.naist.jp/DPClus/>. Uncompressing DPClus.zip produces four files as follows:

**DPClus.jar,**  
**logo.gif,**  
**DPClus\_manual.doc,** and  
**ecoli\_intr.txt.**

The file “**DPClus\_manual.doc**” contains the instruction manual and the file “**ecoli\_intr.txt**” contains the interactions between protein pairs representing the protein-protein interaction network of *E. coli* obtained from <http://dip.mbi.ucla.edu> (DIP). The

---

data in “*ecoli\_intr.txt*” can be used as trial input data while learning the usage of DPPlus. To start the software, user should click on the icon of the file “DPPlus.jar” which brings in front the main window of DPPlus as shown in Fig 1.



## 2. Format of input data

The algorithm of DPPlus receives three inputs: the network, a value of minimum density we allow for the generated clusters ( $d_{in}$ ) and a minimum value for cluster property that determines the nature of periphery tracking ( $cp_{in}$ ). The values for density and cluster property should be within the following range:

$$0 < d_{in} \leq 1, \text{ and}$$

$$0 < cp_{in} \leq 1.$$

The user should prepare the file that describes the network according to the format of Fig. 2. Each line of Fig. 2 is an edge of the network represented by the pair of nodes it connects. Fig. 2 actually shows several interactions between *E. coli* proteins obtained from DIP. The current version of DPPlus allows clustering of a network of maximum 5000 nodes.

```

GroEL PrnP
CarB CarA
MalG MalE
XerD XerC
PntB PntA
SbcC Gam
Gam RecB
Abc2 RecB
RecR RecO
Ssb RecO
FliM FliG
FliF FliG
Hns FliG
-----
-----

```

**Figure 2**

### 3. Execution of Clustering

[1] By clicking the “**Select**” button in the main window (Fig. 3), we can open a file dialog window which can be used to browse to the file that contains the network, then,

[2] Users can get the file name together with its path in the text field beside the label “**Enter Network FileName**”. Alternatively user can also type in the filename together with its path.

[3] Users can **store the results of an execution in a directory of his/her choice** by entering the name in the text field beside the label “Enter Output Directory Name”.

[4] **The desired values of  $d_{in}$  and  $cp_{in}$  should be typed** in the text fields beside the labels “Enter density value” and “Enter  $cp$  value” respectively.

[5] At this point clicking on the “**DoClustering**” button starts the clustering process and when it is over the message “Clustering over” appears in the text field below the “**DoClustering**” button.

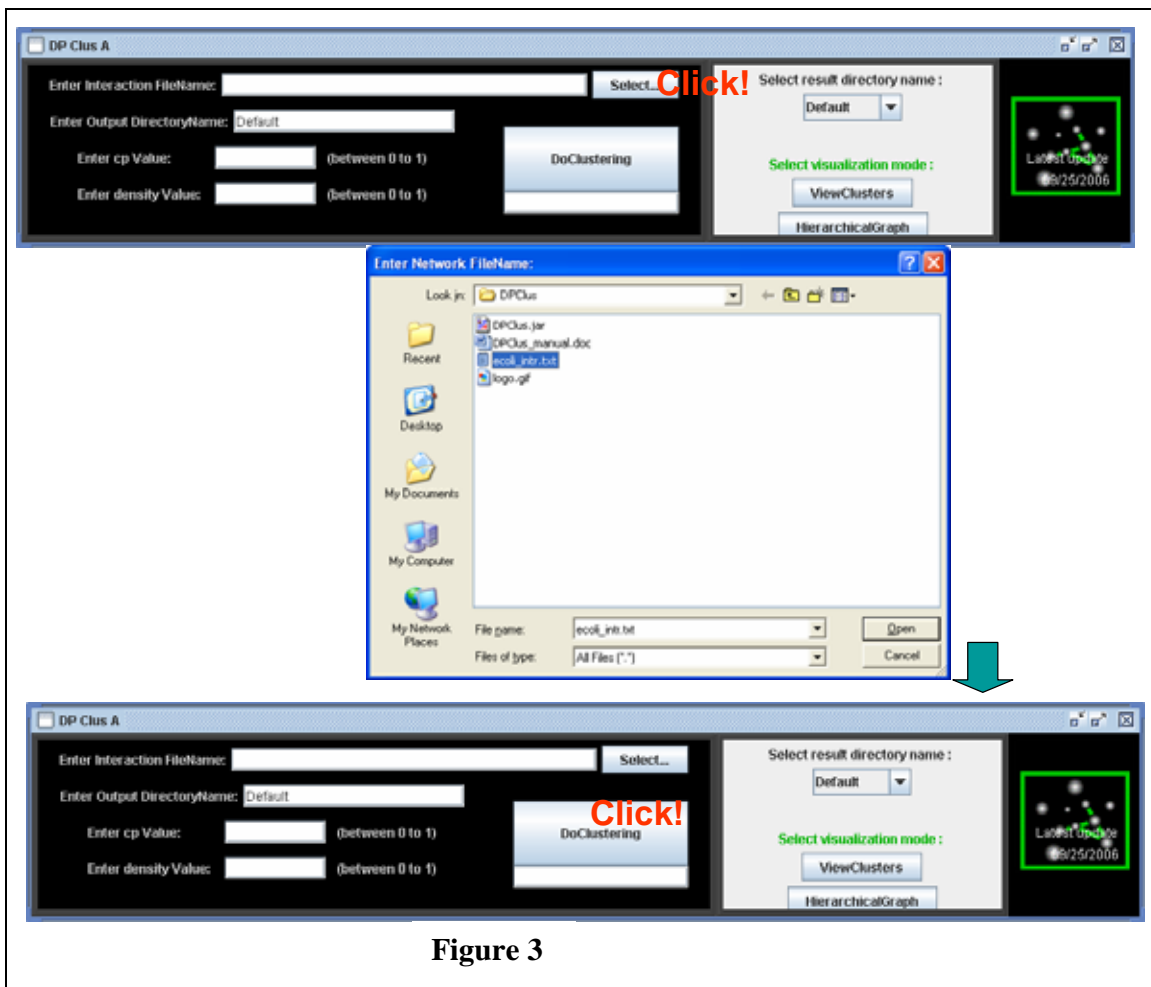


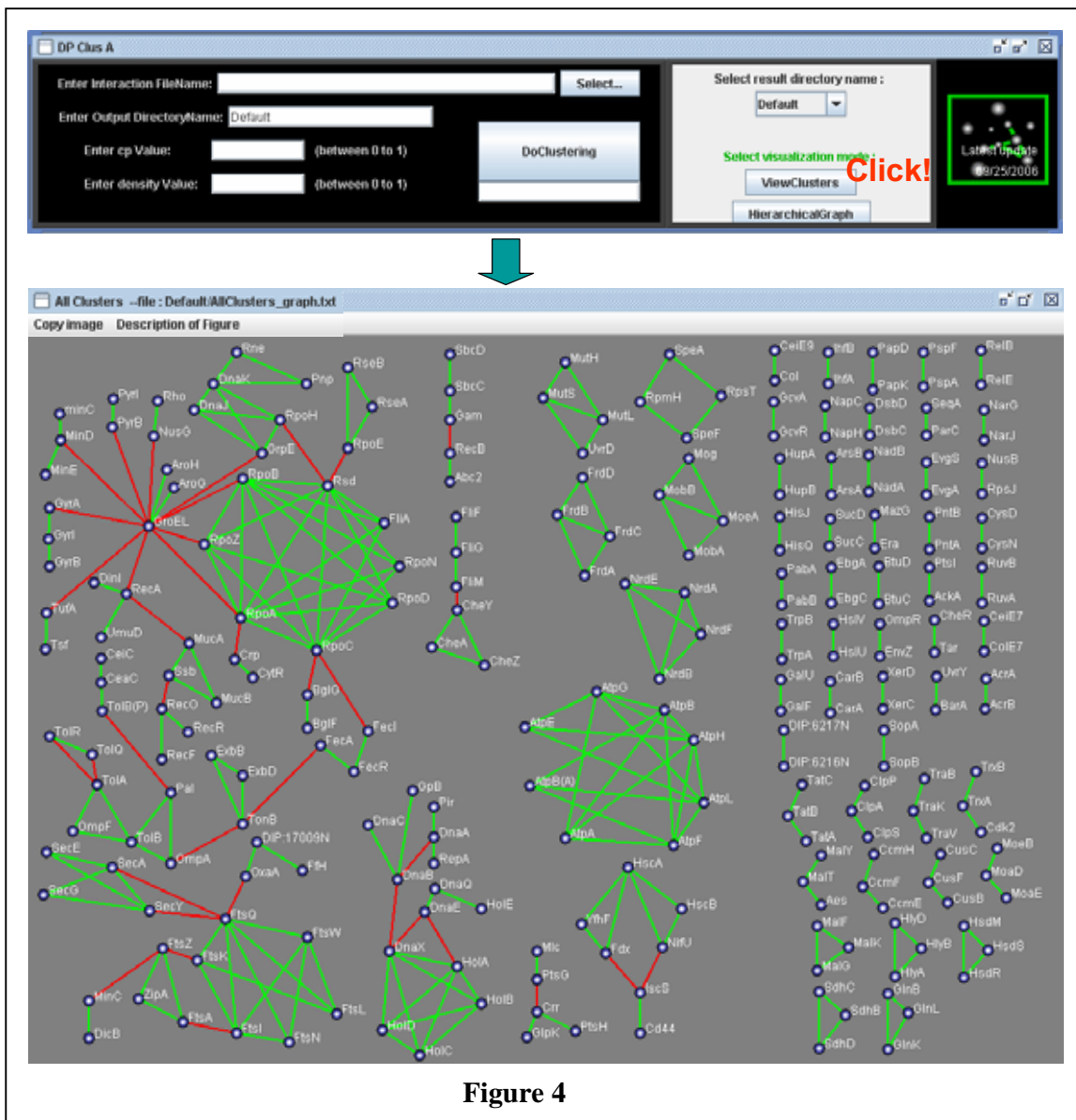
Figure 3

After execution, we have five output files (AllClusters\_graph.txt, cluster.txt, HierarchicalGraph.txt, ClusterRelation.txt and SeperateClusters\_graph.txt). Details of their formats are described in Section 6 (Output files).

## 4. Visualization of clusters

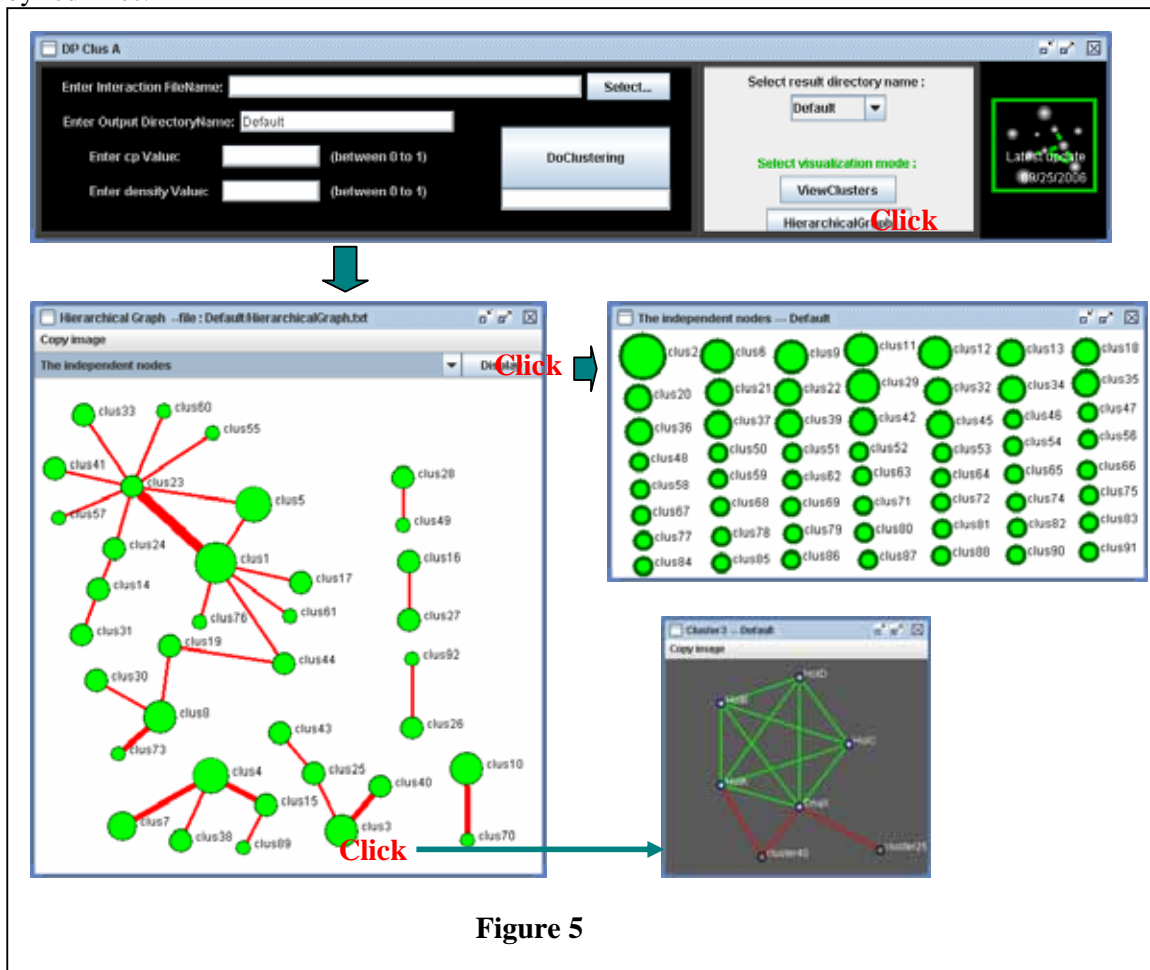
To visualize the clusters, we should click the “**ViewClusters**” button and this make a separate window appear which contains a network showing all the generated clusters in such a way that intra cluster edges are green and inter cluster edges are red.

The user can drag and arrange the nodes in this window according to his/her will and then can save the image using a pull down menu. For example, Fig. 4 (bottom window) shows the clusters (after some arrangements by dragging the nodes) in the *E. coli* PPI network of DIP generated by the present software using  $d_{in}=0.6$  and  $c_{p_{in}}=0.5$ .



## 5. Visualization of hierarchical graph

Also, we can visualize the hierarchical graph where a node represents a cluster and the edges represent the relations between the clusters and to do so we have to click the “**HierarchicalGraph**” button. For example, the bottom left window of Fig. 5 shows part of the hierarchical graph corresponding to the network of Fig. 4 where a cluster is connected to at least another cluster. The independent nodes of the hierarchical graph can be displayed by clicking the “Display” button of the “**HierarchicalGraph**” window. The independent nodes of Fig. 4 are shown in the “**The independent nodes**” window of Fig. 5. The radius of a node in the hierarchical graph is proportional to the logarithm of the number of nodes in the cluster it represents. The width of an edge in the hierarchical graph between a pair of clusters is proportional to the number of edges between those clusters in the original graph. Also in the “**HierarchicalGraph**” window, a user can drag and arrange the nodes and then can save the image using a pull down menu. Users can visualize the network of a single cluster in a separate window by clicking on the corresponding node in the hierarchical graph. For example the bottom right window of Fig. 5 shows the network of cluster3 of the hierarchical graph. Here the internal nodes of the cluster are shown connected by green lines and its neighboring clusters are shown connected by red lines.



## 6. Output files

DPCLUS generates four output files: cluster.txt, AllClusters\_graph.txt, HierarchicalGraph.txt and SeparateClusters\_graph.txt. Descriptions of these files are as follows:

### (i) cluster.txt

The result of clustering is written in text in the file “cluster.txt” according to the format shown in Fig.6. Three types of information concerning each cluster is written consecutively and they are the length of the cluster i.e. the number of proteins in the cluster, the density of the cluster, i.e. the density of the subgraph induced by the cluster and the name of the proteins.

```
ClusterLength of cluster 1 is: 8
Density of cluster 0.8214285714285714
RpoA 7
RpoB 7
RpoC 7
Rsd 7
RpoZ 5
RpoD 5
RpoN 4
FliA 4
ClusterLength of cluster 2 is: 8
Density of cluster 0.6428571428571429
AtpH 6
AtpG 5
-----
-----
```

**Figure 6**

**(ii) AllClusters\_graph.txt**

The format of this file is shown in Fig. 7 and this file is used by DPCLUS for visualization of all the generated clusters together and their inter relations (Fig. 4). In the first line "NODES" is written. The next line shows the total number of nodes included in all the generated clusters and the name of these nodes are written in the following lines. In the next line "EDGES" is written and the following lines contain information on edges. The first two entries of such a line correspond to the serial number of the vertices that are end points of the edge and the third entry is the color code for the color of the edge indicating whether it is an intra cluster edge or an inter cluster edge.

```
NODES
257
RpoA
RpoB
RpoC
Rsd
-----
-----
Mlc
PtsG
EDGES
0    1    FF00
0    2    FF00
0    3    FF00
-----
-----
102  256  FF0000
109  170  FF0000
```

**Figure 7**

### (iii) HierarchicalGraph.txt

The format of this file is shown in Fig. 8 and this file is used by DPCLUS for visualization of the hierarchical graph of the generated clusters (Fig. 5). In the first line "NODES" is written. The next line shows the total number of clusters generated by the algorithm because each cluster is a node in the hierarchical graph. Each of the following 2-column lines contains cluster ID as name of a cluster and a measure for the diameter of the circle that represents the cluster in the hierarchical graph calculated as  $40 * \log(N) / \log(N_{max})$ , where  $N$  is the number of proteins in the cluster and  $N_{max}$  is the number of proteins in biggest generated cluster. In the next line "EDGES" is written and the following lines contain information on edges. The first two entries of such a line correspond to the serial number of the clusters that are end points of the edge and the third entry is the color code for the color of the edge. The fourth entry is a measure of the thickness of the edge calculated as  $10 * E / E_{max}$ , where  $E$  is the number of interactions between the proteins of the clusters connected by the edge and  $E_{max}$  is the maximum value of  $E$  among all the edges in the hierarchical graph.

```
NODES
92
clus1 40.0
clus2 40.0
clus3 30.95904126516483
-----
clus91 13.333333333333334
clus92 13.333333333333334
EDGES
0 4 FF0000 3
0 16 FF0000 3
0 22 FF0000 10
-----
25 91 FF0000 3
27 48 FF0000 3
```

**Figure 8**

**(iv) SeparateClusters\_graph.txt**

This file contains the information on the networks of each of the clusters separately and consecutively. As shown in Fig 9, for each cluster the cluster ID, the number of proteins in it, the name of these proteins and the interactions among these proteins are written. The format is similar to that of “AllClusters\_graph.txt”. This file is used by DPPlus to visualize an individual cluster (bottom right window of Fig. 5).

```
cluster1
8
RpoA
RpoB
-----
-----
RpoN
FliA
0    1    FF00
0    2    FF00
-----
-----
3    7    FF00
4    5    FF00
cluster2
8
AtpH
AtpG
-----
-----
```

**Figure 9**

**(v) ClusterRelation.txt**

This file contains information on how a protein is connected to neighboring clusters other than the cluster it belongs to. The format of this file is shown in Fig. 10. The first column contains the name of the proteins and the subsequent columns contain the IDs of the clusters to which the proteins are connected. This file is used by DPCLus to visualize an individual cluster (bottom right window of Fig. 5).

```
FtsI  cluster15
TonB  cluster8    cluster44
FecA  cluster19
MucA  cluster24
-----
-----
```

**Figure 10**