



HAL
open science

jClust: a clustering and visualization toolbox

Georgios A Pavlopoulos, Charalampos N Moschopoulos, Sean D Hooper,
Reinhard Schneider, Sophia Kossida

► To cite this version:

Georgios A Pavlopoulos, Charalampos N Moschopoulos, Sean D Hooper, Reinhard Schneider, Sophia Kossida. jClust: a clustering and visualization toolbox. *Bioinformatics*, 2009, 25 (15), pp.1994-1996. 10.1093/bioinformatics/btp330 . hal-04112451

HAL Id: hal-04112451

<https://hal.science/hal-04112451>

Submitted on 31 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Systems biology

jClust: a clustering and visualization toolbox

Georgios A. Pavlopoulos^{1,*,\dagger}, Charalampos N. Moschopoulos^{2,\dagger}, Sean D. Hooper³, Reinhard Schneider^{1,*,\dagger} and Sophia Kossida^{2,*,\dagger}

¹Structural and Computational Biology Unit, EMBL Meyerhofstrasse 1, Heidelberg, Germany, ²Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens, Soranou Efessiou 4, GR-11527, Athens, Greece and ³Department of Energy Joint Genome Institute (DOE-JGI), Genome Biology Program, 2800 Mitchell Drive, Walnut Creek, CA 94598, US

Received on March 16, 2009; revised on April 27, 2009; accepted on May 15, 2009

Advance Access publication May 19, 2009

Associate Editor: Jonathan Wren

ABSTRACT

jClust is a user-friendly application which provides access to a set of widely used clustering and clique finding algorithms. The toolbox allows a range of filtering procedures to be applied and is combined with an advanced implementation of the Medusa interactive visualization module. These implemented algorithms are *k*-Means, Affinity propagation, Bron–Kerbosch, MULIC, Restricted neighborhood search cluster algorithm, Markov clustering and Spectral clustering, while the supported filtering procedures are haircut, outside–inside, best neighbors and density control operations. The combination of a simple input file format, a set of clustering and filtering algorithms linked together with the visualization tool provides a powerful tool for data analysis and information extraction.

Availability: <http://jclust.embl.de/>

Contact: pavlopou@embl.de; rschneid@embl.de; skossida@bioacademy.gr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

There exists a big variety of clustering algorithms, which are applicable to a wide range of problems. Most of them are available either as source code, as part of a software package like in R or Matlab packages or are available online. Beside the commercially available ones, there are a few web-based or standalone tools like NeAT (Brohee *et al.*, 2008), Cluster 3.0 software (de Hoon *et al.*, 2004) or Cluto (Zhao and Karypis, 2005) which provide access to some of the clustering algorithms. Nevertheless, it requires typically some effort to either implement the source code into own projects, get familiar with a specific software package or prepare the data for a specifically needed input format. A major weakness of most of the currently available tools is that they lack the interactivity and an easy visualization module to explore and navigate through the data. Here, we present the toolbox jClust, which aims to bridge the gap between analysis and visualization by integrating clustering analysis

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors and the last two authors should be regarded as joint Last Authors.

algorithms with tools able to provide these results visually. The tool provides access to a widely used set of clustering algorithms and simultaneously allows the interactive visualization of the data. It reads from a very simple input file format and produces a human readable output file. jClust comes with a user-friendly GUI that makes the functionality and the parameterization of the algorithms easy and we believe that jClust gives the users, the opportunity to analyze and visualize biological data in a fast, easy and efficient way.

2 CLUSTERING

jClust supports a variety of supervised and unsupervised clustering analysis methods. These are *k*-Means (MacQueen, 1967), Spectral clustering (Paccanaro *et al.*, 2006), Affinity propagation (Frey and Dueck, 2007), Restricted neighborhood search cluster algorithms—RNSC (King *et al.*, 2004), Markov clustering—MCL (Enright *et al.*, 2002), MULIC (Andreopoulos *et al.*, 2007a, b) and Bron–Kerbosch (Coen and Joep, 1973). Concerning *k*-Means and the Spectral clustering, the number of clusters needs to be defined by the user. The *k*-Means (MacQueen, 1967) algorithm requires a full, all-against-all distance matrix to run whereas this is not a requirement for the other implemented algorithms. All of the algorithms besides *k*-Means are suitable for sparse graphs and all of the methods are able to analyze large-scale data as long as the local computer memory permits it. The Bron–Kerbosch (Coen and Joep, 1973) algorithms is a very well-known algorithm for finding cliques in a graph, meaning that it isolates strongly connected sub-areas where every node is connected to every other node—all-against-all connections—that belongs to the same clique. All of the aforementioned clustering algorithms assign nodes to only one unique cluster whereas the Bron–Kerbosch (Coen and Joep, 1973) algorithm allows a node to belong to more than one cluster.

3 FILTERING

jCluster gives to the user the opportunity to filter noise from the predicted clusters that have been calculated by one of the previous methods. This way, in a second step, clusters can be enriched by nodes that are important or shrink by removing nodes that should not belong to the cluster. Here, we implemented the following procedures: (i) density, (ii) haircut, (iii) best neighbor and (iv) cutting

- Brohee,S. *et al.* (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res.*, **36**, W444–W451.
- de Hoon,M.J. *et al.* (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
- Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
- Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.
- Gavin,A.C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
- Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.
- Coen,B. and Joep,K. (1973) Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*. Vol. 16, ACM Press, New York, USA.
- King,A.D. *et al.* (2004) Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**, 3013–3020.
- Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.
- MacQueen,J.B. (1967) Kmeans some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 281–297.
- Moschopoulos,C.N. *et al.* (2008) An enhanced Markov clustering method for detecting protein complexes. In *8th IEEE International Conference on Bioinformatics and Bioengineering*, Athens, Greece.
- Paccanaro,A. *et al.* (2006) Spectral clustering of protein sequences. *Nucleic Acids Res.*, **34**, 1571–1580.
- Winter,D.C. *et al.* (1999) Genetic dissection of the budding yeast Arp2/3 complex: a comparison of the in vivo and structural roles of individual subunits. *Proc. Natl Acad. Sci. USA*, **96**, 7288–7293.
- Zhao,Y. and Karypis,G. (2005) Data clustering in life sciences. *Mol. Biotechnol.*, **31**, 55–80.