



HAL
open science

Clust&See3.0: clustering, module exploration and annotation

Fabrice Lopez, Lionel Spinelli, Christine Brun

► **To cite this version:**

Fabrice Lopez, Lionel Spinelli, Christine Brun. Clust&See3.0: clustering, module exploration and annotation. F1000Research, 2024, 13, pp.994. 10.12688/f1000research.152711.1 . hal-04728417

HAL Id: hal-04728417

<https://hal.science/hal-04728417v1>

Submitted on 9 Oct 2024

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.



SOFTWARE TOOL ARTICLE

Clust&See3.0 : clustering, module exploration and annotation

[version 1; peer review: 1 approved]

Fabrice Lopez ¹, Lionel Spinelli^{1,2}, Christine Brun ^{1,3}

¹TAGC (UMR1090), Aix-Marseille Université, INSERM, Turing Centre for Living Systems, Marseille, 13009, France

²INSERM-CNRS, CIML, Turing Centre for Living Systems,, Aix-Marseille Univ, Marseille, France

³CNRS, Marseille, 13009, France

V1 First published: 02 Sep 2024, 13:994
<https://doi.org/10.12688/f1000research.152711.1>
Latest published: 02 Sep 2024, 13:994
<https://doi.org/10.12688/f1000research.152711.1>

Abstract

Background

Cytoscape is an open-source software to visualize and analyze networks. However, large networks, such as protein interaction networks, are still difficult to analyze as a whole.

Methods

Here, we propose Clust&See3.0, a novel version of a Cytoscape app that has been developed to identify, visualize and manipulate network clusters and modules. It is now enriched with functionalities allowing custom annotations of nodes and computation of their statistical enrichments.

Results

As the wealth of multi-omics data is growing, such functionalities are highly valuable for a better understanding of biological module composition, as illustrated by the presented use case.

Conclusions

In summary, the originality of Clust&See3.0 lies in providing users with a complete tool for network clusters analyses: from cluster identification, visualization, node and cluster annotations to annotation statistical analyses.

Open Peer Review


Approval Status 

1

version 1 

02 Sep 2024

[view](#)

1. **Anooja Ali** , REVA University, Bengaluru, India

Any reports and responses or comments on the article can be found at the end of the article.

Keywords

interaction networks, graph partitioning, clustering, visualization, cluster annotations, functional modules, statistical enrichment.



This article is included in the **Cytoscape** gateway.

Corresponding author: Christine Brun (christine-g.brun@inserm.fr)

Author roles: Lopez F: Software; Spinelli L: Formal Analysis, Methodology; Brun C: Conceptualization, Formal Analysis, Investigation, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Funding for article processing charges provided by Human Frontier Science Program grant RGP004/2023 to CB.

Copyright: © 2024 Lopez F *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Lopez F, Spinelli L and Brun C. **Clust&See3.0 : clustering, module exploration and annotation [version 1; peer review: 1 approved]** F1000Research 2024, 13:994 <https://doi.org/10.12688/f1000research.152711.1>

First published: 02 Sep 2024, 13:994 <https://doi.org/10.12688/f1000research.152711.1>

Introduction

Several years ago, we proposed Clust&See, a Cytoscape¹ plug-in that aims to facilitate network clustering and analysis for biologists by providing several original functionalities within a single framework.² Mainly, the tool allows decomposing a network into disjoint or overlapping clusters using several in-house algorithms,³⁻⁵ visualizing those clusters as metanodes linked by several types of edges/relationships, and manipulating the clusters for further detailed visualization, exploration, analyses and comparisons. We have now developed Clust&See3.0 to fit Cytoscape3 new API and added functionalities allowing the user to annotate the nodes and the clusters with orthogonal data as well as to compute statistical enrichments for those annotations. We here provide examples of Clust&See3.0 usage on a protein-protein interaction network, where annotation enrichments are used *(i)* to functionally annotate the clusters with the Gene Ontology terms describing node functions in order to globally investigate the network and proceed with protein function prediction, and *(ii)* to discover features associated to clusters by the integration of data on nodes.

Methods

Implementation

Briefly, the classic use of Clust&See² breaks down into the following phases:

Decompose a network

As in its first version, Clust&See proposes to partition an imported network using three algorithms: FT (Fusion-Transfer), an ascending hierarchical method fusing two clusters iteratively if the fusion results in a modularity gain³; TFIT (iterated Transfer-Fusion),⁴ a multi-level algorithm in which a vertex transfer procedure is performed to the best adjacent cluster while modularity increases, to finally compute a quotient graph. Both algorithms generate strict network partitions where clusters have no node in common. The third one, OCG (Overlapping Cluster Generator)⁵ is an ascending hierarchical method fusing two clusters at each step while modularity increases, starting from an overlapping class system. This leads to overlapping clusters where some nodes belong to several clusters.

Cluster visualization and exploration

For each cluster in a partition, Clust&See provides detailed information about the cluster: a visualization of the cluster's sub-network, and its main topological features (number of nodes, edges, density, etc.)

Each cluster can be viewed and explored independently, either in a compact mode as a metanode, or in an extended mode as the sub-network constituting the cluster (Figure 1). The decomposition of the complete network can be viewed as a set of metanodes linked by edges, the thickness of which is proportional to the number of links between pairs of nodes of the linked metanodes. When clusters are overlapping, another type of link is added between the metanodes to represent the nodes shared between clusters/metanodes. The thickness of this link is then proportional to the number of shared nodes. From this map, the metanodes can be individually switched to the extended mode to visualize the corresponding sub-network, and back. Finally, the user can build a custom map by iteratively adding sub-networks and/or metanodes to obtain a global view of the partition (Figure 1, central panel).

The novel version of Clust&See3.0 now allows the user to annotate and analyze the nodes and the clusters as follows:

Importing node annotations

In order to annotate the clusters, Clust&See3.0 allows importing one or more annotation lists, composed of associations between a network node and one or more terms of any nature (see the Use case).

The annotation file must be a text file and must contain at least two columns: one contains the node identifier (one identifier per line), the second one, the list of terms associated with this identifier, separated by a comma, a semicolon or a tab. The columns can also be separated by a comma, semicolon or tab, as long as the column separator is not the same as the term separator. The annotation import dialog box lets the user select the 2 columns of interest and eliminate any header lines.

After import, Clust&See3.0 provides the number and percentage of annotated nodes in the network. Then the annotation process can be performed.

Annotation rules, statistical enrichment

Two types of enrichment can be computed for each annotation term, using the whole graph as background:

- A one-sided hypergeometric hypothesis test, the null hypothesis of which corresponds to the proportional distribution of the annotation terms between the nodes inside and the nodes outside the cluster. When the

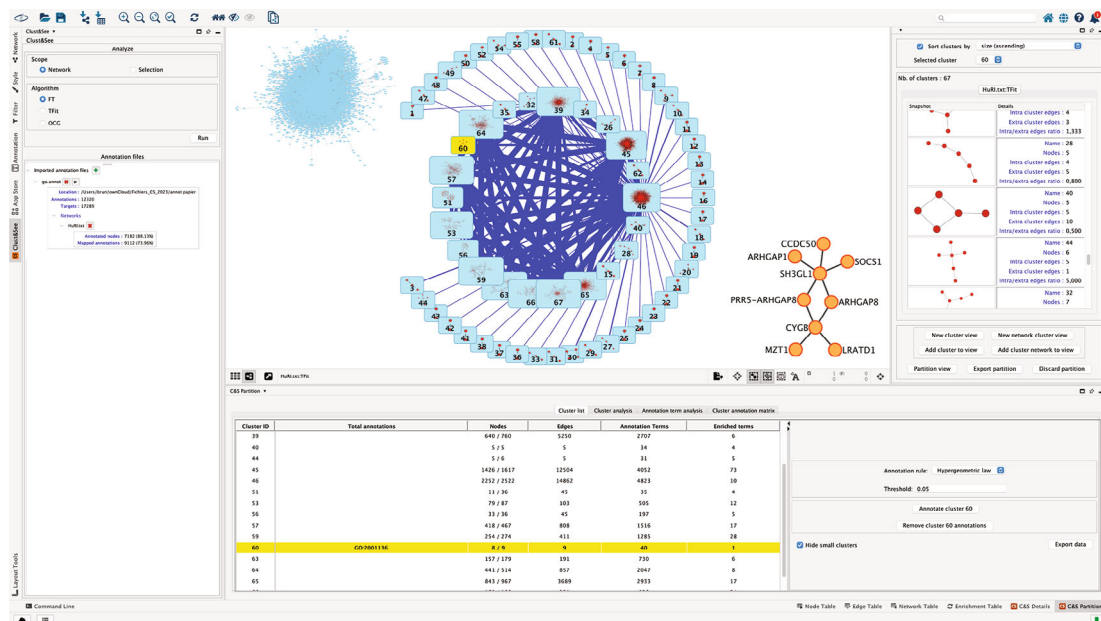


Figure 1. Clust&See3.0 interface. The central panel shows (1) the starting network in the top left corner, (2) the quotient graph after TFit partitioning where modules are represented by blue nodes with the icon of the corresponding sub-network inside. The size of the node reflects the number of nodes contained by the module. The blue links correspond to the edges linking the modules, their width reflecting the number of involved interactions. (3) In the right low corner, module 60 is shown as an expanded network. As in the former version of Clust&See, the right panel shows the details of the subnetworks forming the modules with their characteristics. The lower panel (former Data panel) contains the results of the different Cluster and Annotation analyses.

p-value of the hypergeometric test is sufficiently low to reject the null hypothesis, this indicates that the nodes of the cluster carry the term more frequently than expected by chance, thus pointing toward a potential enrichment. It should be noted that since this test is applied to all clusters, the Benjamini-Hochberg procedure⁶ is used to correct for the multiple testing effect on the p-values. The default value is set to p-values $< 5.10^{-2}$.

- A majority rule, corresponding to a minimum percentage (to be chosen by the user) of nodes annotated to the term among the annotated nodes of the cluster. The default value is set to 50%.

Cluster and annotation analyses

For each annotation list, the user can perform a statistical analysis of the clusters and the annotation terms, with different goals. At first, a global analysis of the annotations of the partition can be performed. When choosing the “Cluster list” tab, Clust&See3.0 provides for each cluster, (i) the number of nodes in the cluster that have received at least one annotation term, (ii) the number of annotation terms appearing at least once in the cluster, (iii) the number of terms that are statistically enriched in the cluster with a hypergeometric test, or that annotate a majority of cluster nodes (*i.e.* with a “majority rule”). For both tests, thresholds are set by the user (Figure 1). Finally, (iv) the terms that are enriched in clusters are shown on demand (Figure 2A). This first type of analysis allows getting a global view of the annotation distribution and to quickly identify clusters that are enriched for annotation terms of interest.

The second approach concerns the details of statistical analyses by cluster by choosing the “Cluster analysis” tab. Here, the user can select a particular cluster for a detailed study of the annotations of its nodes. Clust&See3.0 then lists the terms annotating the cluster proteins and provides (i) the number of nodes annotated to the term, (ii) the percentage of nodes annotated to the selected term among the total number of cluster nodes, (iii) the p-value of the term according to the hypergeometric test, (iv) the percentage of nodes annotated to the selected term among the total number of annotated cluster nodes (Figure 2B).

The third approach concerns the details of the statistical analyses by annotation term. When choosing the “Annotation term analysis” tab, the user can select a particular annotation term and Clust&See3.0 reports for each cluster that contains proteins annotated to this term, the same features than previously (Figure 2C). This type of analysis allows having a detailed view on the distribution of a particular annotation term among all the clusters composing the network.

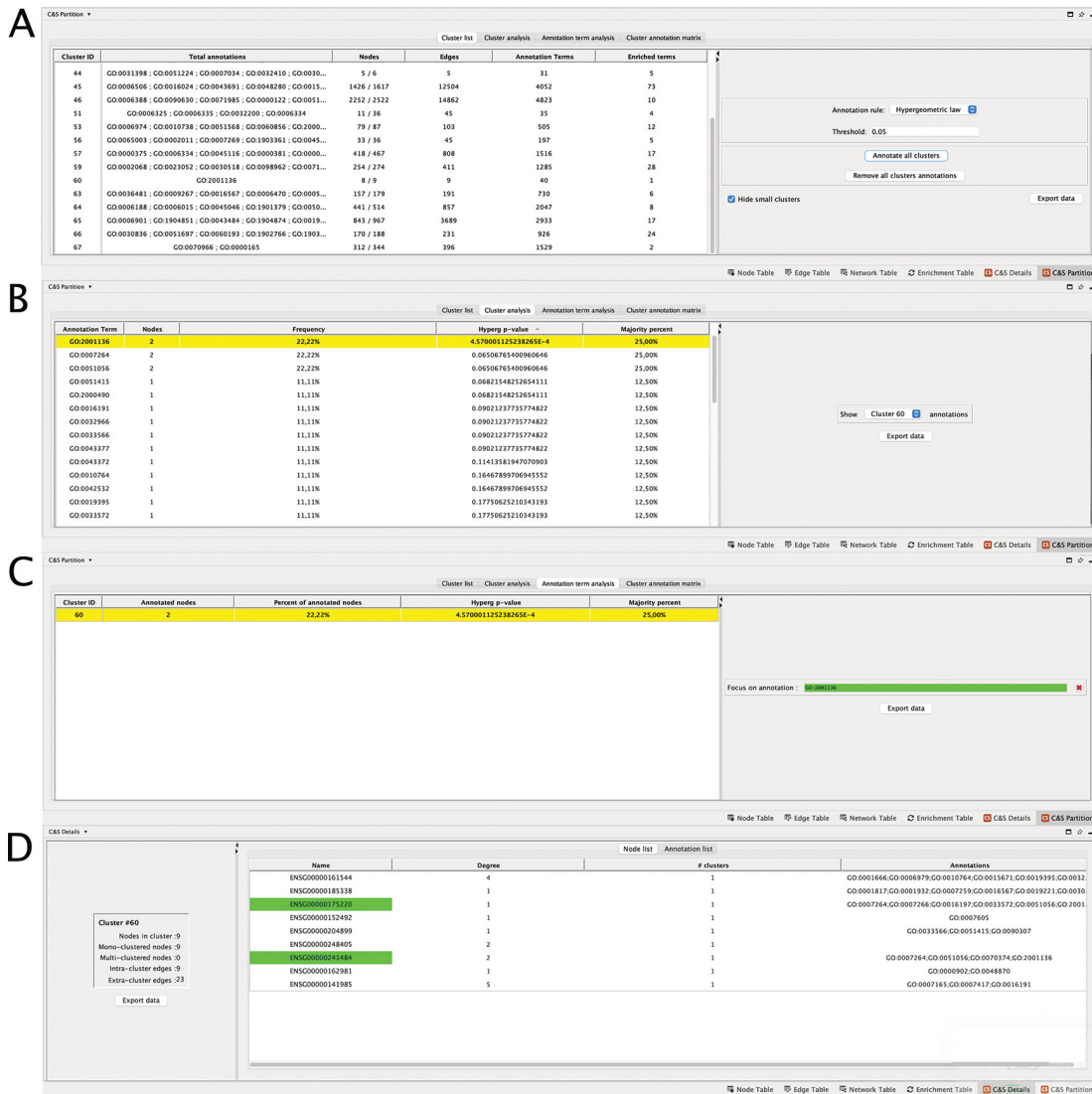


Figure 2. Details of the different tabs of the Data panel. (A) “Cluster List” shows the annotation of all clusters; (B) “Cluster Analysis”, all the annotations of cluster 60; (C) “Annotation term analysis”, all the clusters annotated to GO:2001136; (D) “Node list”, the detail of the annotations of each nodes of the cluster. The node names highlighted in green are those annotated to the term of interest, here GO:2001136.

The tables displaying the results for the clusters are dynamic and can be used to identify the clusters selected in the map and, in the panel detailing the characteristics of the clusters. The reverse is also true: a selected cluster in the map or in the detailed panel is selected in the annotation statistical results tables.

Operation

The minimum system requirements for use of the Clust&See3.0 Cytoscape app include:

Hardware:

Memory: 8 GB

Monitor: 1600×900 (HD+) resolution

Software:

Java 11 and above

Cytoscape 3.8.0 and above

Use Case**Annotation mode for network and cluster exploration, and for function prediction**

We will illustrate how to use Clust&See3.0 by partitioning and annotating the human reference interactome network (HuRI)⁷ with Gene Ontology⁸ (network and annotation files are available at <https://doi.org/10.5281/zenodo.12570870>). We first partitioned the largest connected component of the HuRI network that contains 8149 nodes and 52016 edges, with the TFit algorithm.⁴ Sixty-seven modules were obtained among which 20 contain more than 4 nodes (Figure 1). After loading the annotation file that contains the list of IDs of the Biological Process Gene Ontology (BP GO) associated to all human genes/proteins, the number of annotated nodes in the network is indicated: 88% of the proteins of HuRI have functional annotations in the BP GO.

- “Cluster list” tab

In the “C&S Partition” table, under the “Cluster list” tab, all clusters are shown (Figure 2A). As we empirically consider that the smallest clusters are not suitable for computing statistics on annotations, Clust&See3.0 proposes to hide them for clarity’ sake with a check box “Hide small clusters”.

The number of enriched terms per cluster according to the chosen statistic (“Hypergeometric law” or “Majority rule”) is indicated, and their GO IDs are available in the “Total annotations” column (Figure 2A) (choose “Annotate cluster X/Remove cluster X annotations” or “Annotate all clusters/Remove all cluster annotations”). The list of annotation terms also appears by right clicking on a cluster of interest on the Partition view, under “Clust&See>Annotate cluster”. At this stage, custom annotations can be added manually by the user to tag a cluster of interest. This annotation will also appear in the list of annotation terms in the “Total annotations” column of the “Cluster list” table. This may help having a global quantitative view of the cluster annotations and further analyses.

The individual investigation of a particular cluster starts also under this tab. For instance, Cluster 60 contains 8/9 nodes annotated (number given in the “Nodes” column), and solely 1 term is statistically enriched among the 40 terms (number given in the “Annotation terms” column) that annotate the proteins of the cluster, when the hypergeometric law with a corrected p-value threshold set at 5.10^{-2} is chosen.

- “Cluster analysis” tab

The term GO:2001136 is enriched among the annotations of the proteins of cluster 60 with a corrected p-value of $4.57.10^{-4}$, available in “Hyperg p-value” column under the “Cluster analysis” tab (Figure 2B). The most relevant term (*i.e.* with the lowest p-value) is easily found by using the ranking column containing the corrected p-values for the terms annotating all the proteins of cluster 60. The term GO:2001136 that corresponds to “negative regulation of endocytic recycling”, is annotating 25% of the annotated proteins of the cluster (number in the “Majority percent” column). Two other terms are also annotating 25% of the cluster’s proteins, but with highest p-values *i.e.* less significant ones. These are GO:0007264 “small GTPase-mediated signal transduction” and GO:0051056 “regulation of small GTPase mediated signal transduction”. Then, switching to the “C&S Details” Table (Figure 2D), under the “Node” tab, that shows the detailed annotations of the nodes of the cluster, we see that the three terms are associated to the same two genes encoding the proteins involved in these functions: ENSG00000175220 (ARHGAP1) and ENSG00000241484 (ARHGAP8). By expanding the cluster 60 on the network panel (Figure 1, central panel), it can be seen that these two proteins do not interact directly but with the product of ENSG00000141985 (SH3GL1), a protein regulating endocytosis by recruiting proteins to the membrane.

- “Annotation term analysis” tab

Then, wondering whether other clusters are also annotated to “negative regulation of endocytic recycling”, we switched to the “Annotation term analysis” tab to focus on a particular annotation and find all clusters enriched for this term. By entering the ID GO:2001136 in the dedicated frame, we found that none of the clusters but cluster 60 is annotated to this term, either using the hypergeometric law or the majority rule computations (Figure 2C).

The user can therefore choose to annotate this cluster with this enriched term and, if appropriate, to transfer the annotation/function to any not yet annotated node of the cluster. Notably, the nodes contributing to cluster's annotations are detailed in the "C&S Details" table (Figure 2D).

Export of the annotations and computations as text files are available at each step, as well as a matrix of the whole results under the "Cluster annotation matrix" tab, for further analysis.

Conclusion/Discussion

Clust&See3.0 is a Cytoscape app that allows (i) clustering the nodes of any network, (ii) annotating the clusters with any annotation terms, and (iii) computing their enrichment significance. This versatility of Clust&See3.0 is constituting its advantage compared to other existing Cytoscape enrichment plug-ins. Not only Clust&See3.0 allows as in its previous version loading any partition of the network, even not generated by the app, as long as the graph is completely covered by clusters, but it also allows using any type of data as node annotations, such as user's experimental or curated data. In contrast, most of the existing apps are mainly centered on Gene Ontology terms or other types of classical annotations (i.e.^{8,9}). Second, the results of Clust&See3.0 permit to get a global view of the distribution of the annotations between the whole set of clusters, as well as their statistical value. For all these reasons, we think Clust&See3.0 will be a valuable tool for the community.

Ethics and consent

Ethics and consent not required.

Data availability

Zenodo: Clust&See3.0: clustering, module exploration and annotation. <https://doi.org/10.5281/zenodo.12570870>. The project contains the following underlying data:

- go_annot.txt: The protein annotation file extracted from Gene Ontology Biological Process database.¹⁰
- HuRI_CC.txt: The network file containing the largest connect component of the human reference interactome network.⁷

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0).

Software availability

- Software available from: <https://apps.cytoscape.org/apps/clustnsee3>
- Source code available from: <https://github.com/fafa13/ClustnSee-3>
- Archived source code at time of publication: <https://doi.org/10.5281/zenodo.13220735>.¹¹
- License: GNU General Public Licence, V3

References

1. Shannon P, Markiel A, Ozier O, et al.: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res*. 2003; **13**: 2498–2504. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Spinelli L, Gambette P, Chapple CE, et al.: **Clust&See: a Cytoscape plugin for the identification, visualization and manipulation of network clusters**. *Biosystems*. 2013; **113**: 91–95. [Publisher Full Text](#)
3. Guénoche A: **Consensus of partitions: a constructive approach**. *ADAC*. 2011; **5**: 215–229. [Publisher Full Text](#)
4. Gambette P, Guénoche A: **Bootstrap clustering for graph partitioning**. *RAIRO - Operations Research*. 2012; **45**: 339–352. [Publisher Full Text](#)
5. Becker E, Robisson B, Chapple CE, et al.: **Multifunctional proteins revealed by overlapping clustering in protein interaction network**. *Bioinformatics*. 2012; **28**: 84–90. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995; **57**: 289–300. [Publisher Full Text](#)
7. Luck K, Kim D-K, Lambourne L, et al.: **A reference map of the human binary protein interactome**. *Nature*. 2020; **580**: 402–408. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Ashburner M, Ball CA, Blake JA, et al.: **Gene ontology: tool for the unification of biology**. *The Gene Ontology Consortium*.

- Nat Genet.* 2000; **25**: 25–29.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics.* 2005; **21**: 3448–3449.
[PubMed Abstract](#) | [Publisher Full Text](#)
 10. Merico D, Isserlin R, Stueker O, *et al.*: **Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.** *PLoS One.* 2010; **5**: e13984.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 11. Theories and Approaches of Genomic Complexity: **Cytoscape app ClustnSee 3.** *Zenodo.* 2024.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: 

Version 1

Reviewer Report 10 September 2024

<https://doi.org/10.5256/f1000research.167505.r320335>

© 2024 Ali A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Anooja Ali 

REVA University, Bengaluru, Karnataka, India

The authors propose Clust&See3.0, an improved version of Cytoscape, for visualizing and manipulating the bioinformatic network clusters.

It has additional functionalities, when compared to Cytoscape like custom annotations of nodes. It provided more statistical analysis for cluster evaluation.

How does the user interface of Clust&See3.0 compare to previous versions in terms of usability and accessibility for biologists unfamiliar with network analysis?

What are the key characteristics of clusters identified in large protein-protein interaction networks using Clust&See3.0? How do these characteristics correlate with biological functions?

How does the annotation and enrichment analysis provided by Clust&See3.0 contribute to the accuracy of protein function predictions in complex biological networks?

I hope the user feedback on Clust&See3.0's functionalities inform future updates and enhancements to the software.

The citations refer to the aligners for creating a huge protein network based on functional similarities among proteins. The next citation uses cytoscape tool with KEGG pathway analysis and GO annotation tool for generating biclusters. It deals with the generation of gene ontology clusters for each category of MF, BP and CC from biclusters.

References

1. Ali A, Ajil A, Meenakshi Sundaram A, Joseph N: Detection of Gene Ontology Clusters Using Biclustering Algorithms. *SN Computer Science*. 2023; 4 (3). [Publisher Full Text](#)
2. H V, Ramachandra Anooja, Ali P S, Ambili S, et al.: An Optimization on Bicluster Algorithm for Gene Expression Data. *IEEE explore*. 2023. [Publisher Full Text](#)
3. A, Ali H. V. Ramachandra, Sundaram A. Ajil, Ramakrishnan: Pareto Optimization Technique for Protein Motif Detection in Genomic Data Set. *Springer link*. 2023. [Publisher Full Text](#)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Deep learning , Computer Vision, Data mining, Computational Biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research