Superclusteroid: a Web tool dedicated to data processing of protein-protein interaction networks



Athina Ropodi[1,2,#], Nikolaos Sakkos[2,#], Charalampos Moschopoulos[2,3]


George Magklaras[4], Sophia Kossida[2,*]

[1]Department of Informatics, University of Athens, GR-15784, Athens, Greece.
[2]Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens, Soranou Efesiou 4, GR-11527, Athens, Greece.
[3]Department of Computer Engineering & Informatics, University of Patras, GR-26500, Rio, Greece.
[4]The Biotechnology Centre of Oslo, University of Oslo, P.O. Box 1125 Blindern, 0317 Oslo, Norway

*To whom correspondence should be addressed.
#Equal contribution to this work.

Abstract

The study of proteins and the interactions between them, known as Protein-Protein Interactions (PPI), is extremely important in interpreting all biological cellular functions. In this article, a new web tool called Superclusteroid is presented which can analyse PPI data, in order to detect protein complexes or characterise the functionality of unknown proteins. The tool is essentially an intuitive PPI data processing pipeline. It supports various input file formats and provides services such as clustering, PPI network visualisation and protein cluster function prediction. Each Superclusteroid service can be used in a sequential manner or on an individual basis. In order to assess the reliability of our tool to infer PPIs, the results of the tool were compared to already known MIPS database complexes and a case scenario is presented where a known protein complex is predicted and the functionality of some of its proteins is revealed.

Availability: Superclusteroid is freely available online at http://superclusteroid.uio.no/.

Background

In the recent era, high-throughput detection methods (Ito *et al.*, 2001; Gavin *et al.*, 2002; Stoll *et al.*, 2005; Willats, 2002) have produced a vast amount of biological data to be analysed using computational methods. Proteomics is the discipline with the objective to analyse and understand all data concerning proteins. A proteome-wide approach of understanding protein function is very important, as it is widely known that proteins rarely act alone at a biochemical level and they interact with other proteins (Bu *et al.*, 2003). This type of protein-protein interactions can easily be described as a protein-protein interaction network (PPI network), where the nodes represent proteins and the edges the interactions among them.

As protein-protein interactions are a crucial part of cellular processes, it is understandable that the processing of large-scale experiment data is extremely useful. In fact, the identification of smaller groups of proteins (clusters) which share more interactions among themselves and fewer with the remaining proteins of the network can lead to the discovery of protein complexes or functional modules (Spirin and Mirny, 2003). It is reasonable to assume that proteins appearing to be more closely connected must share a common function.

Until now, various computational approaches have been proposed in the academic world in the form of web-based or stand-alone software tools. Examples include applications such as NEAT (Brohee *et al.*, 2008) and jClust (Pavlopoulos *et al.*, 2009). However, most of these tools lack vital software application properties. In particular, we believe that the user should be able to execute various algorithms interactively. In addition, the ability to explore and navigate through PPI data visually is an important one. Interactive algorithm execution and visualisation of resulting PPI data make the interpretation of results easier for the scientist.

By using the Superclusteroid tool, the user can apply different clustering, visualisation and prediction methods in a continuous manner, which embraces user interaction. From the moment the user uploads the input data, all resulting files can be further manipulated in discrete stages, as the tool was specifically designed to bridge the compatibility gap amongst the various methods

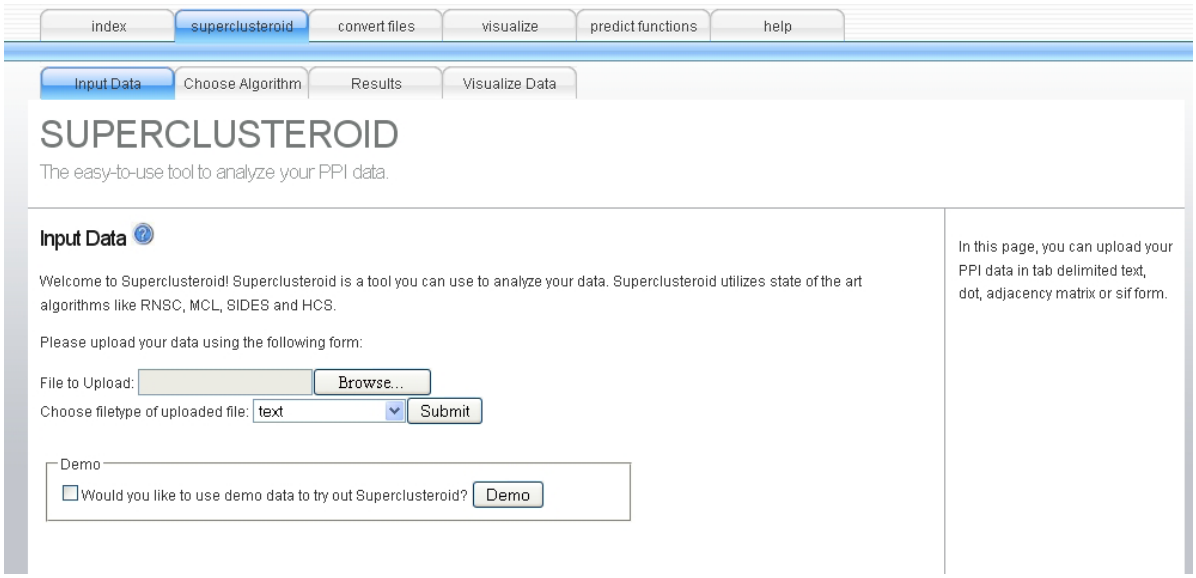

Figure 1. Algorithm selection module of the Superclusteroid tool.

of each of the PPI data manipulation stages. Moreover, a variety of available PPI visualization modules are employed, in order to facilitate an intuitive result interpretation, where applicable.

Implementation

Design

Superclusteroid is a web-based application written in Perl and can be accessed using any internet browser able to execute a Java applet (note: although Java compatibility is not required for all operations, some of the visualisation tools do require the execution of a Java applet in the browser). It utilises already available clustering algorithms. As different algorithms provide different results (Brohee and van Helden, 2006; Li *et al.*, 2009), the user can choose among a set of widely used clustering algorithms to process the input data (Figure 1). These algorithms are: (i) MCL (Markov Cluster), an algorithm that computes the graph of random walks of an input graph, yielding a stochastic matrix (Enright *et al.*, 2002); (ii) Restricted Neighbourhood Search Clustering Algorithm (RNSC), a cost-based local search algorithm based loosely on the tabu search metaheuristic (King *et al.*, 2004); (iii) Highly Connected Subgraphs Algorithm (HCS), based on the detection of highly connected subgraphs (Hartuv and Shamir, 2000); (iv) SideS, a variation of HCS which uses a statistical model to express the statistical significance of a cluster (Koyuturk *et al.*, 2007). It has to be noted that the additional algorithms (SideS and HCS) are not available on any other online tool, despite their efficiency on protein complex detection. The resulting files are tab-delimited data with two columns, one for the name of the cluster and one for the protein belonging to that cluster.

The above results can be automatically visualised or can be downloaded for later use. Additionally, the original network or other DOT files can be viewed by choosing the "visualize" tab on the home page, as it is shown in Figure 2. In either case, a java applet named "ZGRViewer" (Pietriga, 2005) is used to support the "fdp" and "twopi" GraphViz/DOT tools (http://www.graphviz.org/Documentation.php) for spring model and radial layouts respectively. ZGRViewer is designed to handle large graphs, and offers a zoomable user interface (ZUI), which enables smooth zooming and easy navigation in the visualised structure. Furthermore, the user is able to visualise on a new tab of his/her browser a specific cluster. This dynamic visualisation module of Superclusteroid makes it easier for users to explore and analyse the clustering results, contrary to the static module of other web tools such as NEAT.

By choosing a specific protein, the user may continue with the analysis by implementing the Majority Vote Prediction Algorithm (MVPA) (Bu *et al.*, 2003) or the Hypergeometric Distribution

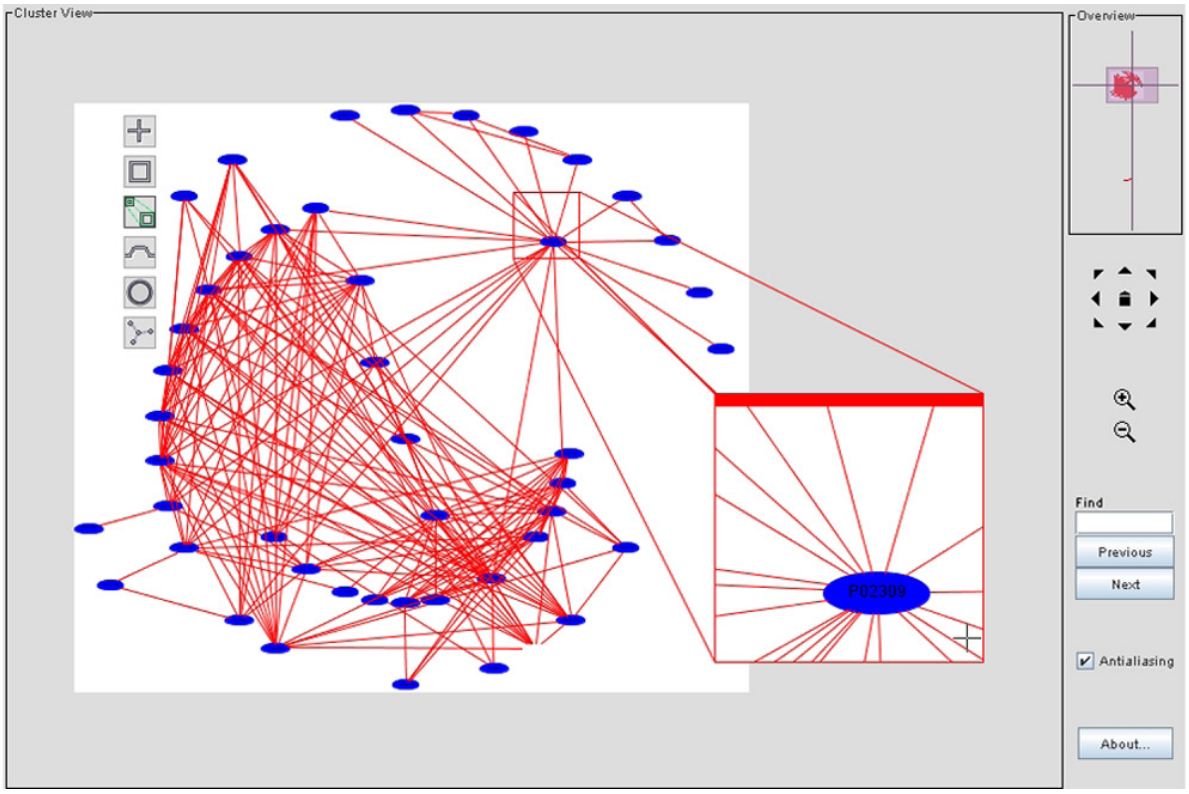

Figure 2. Visualisation module of Superclusteroid tool.

Prediction Algorithm (HDPA) (Enright *et al.*, 2002). Both methods apply only for PPI data with Uniprot IDs and for the S. cerevisiae organism. The functional categories used are those provided in the FunCat database (Ruepp et al., 2004).

Input

The web tool can manipulate different input formats. Specifically, Superclusteroid supports tab-delimited text files, adjacency matrices in text files, DOT files-using the DOT network description languages and SIF files, a popular tab-delimited text file mostly used in Cytoscape (Shannon *et al.*, 2003). The input file can be uploaded in an easy and quick manner in a user-friendly web page. Multiple identical PPIs are removed from further analysis. More information about the input data format is available at Superclusteroid help pages.

For the purpose of presenting Superclusteroid, the Gavin 2006 dataset (Gavin *et al.*, 2006) is used for all four algorithms available and produces the required clustering results using the default parameters. In order to prove the tool's ability to predict protein complexes, the four different clustering results are compared with the recorded protein complexes stored in the MIPS database concerning the *S. cerevisiae* organism (Mewes *et al.*, 2002). The recorded complexes of the MIPS database are used as a golden standard in order to compare the results of the each time applied algorithm (Brohee and van Helden, 2006; Li *et al.*, 2009).

Table 1. The MVPA function category scores of the protein P38334.

#	Category	Score
1	Cellular transport, transport facilities and transport routes	9
2	Metabolism	2
3	Biogenesis of Cellular Components	1
4	Cell Type Differentiation	1
5	Cell Cycle and DNA processing	1
6	Energy	1

Table 2. The HDPA function category scores of the protein P38334

#	Category	Score
1	Cellular transport, transport facilities and transport routes	1.10 E-05
2	Energy	0.504003
3	Cell Type Differentiation	0.647883
4	Metabolism	0.761164
5	Biogenesis of Cellular Components	0.902691
6	Cell Cycle and DNA processing	0.95261

Results

In order to prove the efficiency of Superclusteroid compared to other similar clustering tools, we performed experiments using the Gavin 2006 dataset (Gavin *et al.*, 2006). This dataset consists of 1,430 proteins and 6,531 interactions which derived from Tandem Affinity Purification method (Puig et al., 2001) and Mass Spectrometry (Ho *et al.*, 2002).

We chose to use the MCL algorithm which according to (Brohee and van Helden, 2006) and (Li *et al.*, 2009), is one of the best clustering algorithms. The initial dataset was divided into 188 clusters, where each of these can be visualised and manipulated independently. Then we chose randomly a protein, the one called P38334, and we tried to determine its functionality by using the corresponding Superclusteroid module. Tables 1 and 2 show the results of the MVPA and the HDPA algorithms.

In both cases, the function category "Cellular transport, transport facilities and transport routes", according to the FUNCAT database, is the most likely for the protein P38334.

By using the UniProt database (Magrane and Consortium , 2011), it can be seen that P38334 is part of the TRAPP complex (Sacher *et al.*, 1998), which according to Gene Ontology data (Barrell et al., 2009), is a large complex on the cis-Golgi that mediates vesicle docking and fusion. It is divided into two parts: TRAPP I, which is a multisubunit complex that consists of seven subunits, and TRAPP II, which has three additional subunits and that functions as a tether at latter stages of the transport pathway. Therefore, the Superclusteroid successfully predicted the functionality of the P38334 protein, which is a service that is not provided by other similar clustering tools.

Conclusion

Our results prove that Superclusteroid is capable of predicting protein complexes in an easy-to-use way. Additionally, data formats can be easily manipulated and clustering results can be cross-referenced as the tool provides four different clustering algorithms. Superclusteroid also detects complexes that do not match any confirmed complex in MIPS database. As we cluster the complete interactome, of which the confirmed complexes provide only partial coverage, we speculate that complexes detected by our method could match yet unknown or unconfirmed protein complexes. However, it must be emphasised that protein complexes are not the only ones that can be detected. As explained earlier, the clustering algorithms provide protein groups that are more "connected" among themselves. This statistical significance does not apply specifically to protein complexes, but it is also applicable to functional modules. This term is used for proteins that participate in a common cellular process while binding each other at a different time and place (Spirin and Mirny, 2003).

To sum up, Superclusteroid: (i) uploads and manipulates input of PPI data; (ii) performs clustering on PPI data using four different algorithms; (iii) visualises PPI networks and clustering results; (iv) predicts protein function. It can be used for the prediction of protein complexes in a user-friendly way. Superclusteroid also provides a help page that contains explicit instructions describing its services and a comprehensive list of the web services available, along with their description and the access URL for each of them. Additionally, the web tool provides demo data to help the user to understand its functionality.

The tool is implemented in the GNU/Linux environment and is written in Perl[1]. In addition to the website, web services utilising the SOAP protocol[2] are also available in order to design workflows and integrate them with other available resources.

Acknowledgements

We would like to thank the University Biotechnology Center of Oslo for hosting the Superclusteroid web tool. We also would like to thank Erik Bongcam-Rudloff for organising the joint EMBnet-EMBRACE workshop on creating web services for Bioinformatics (Uppsala, Sweden, 2008).

References

1. Barrell D, Dimmer E, Huntley R P, Binns D, O'Donovan C,Apweiler R (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res 37*, (Database issue) D396-403.
2. Brohee S, Faust K, Lima-Mendez G, Sand O, Janky R, Vanderstocken G, Deville Y,van Helden J (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res* **36**, W444-51.
3. Brohee S,van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488.
4. Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, Li G,Chen R (2003) Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res* 31, 2443-50.
5. Enright A J, Van Dongen S,Ouzounis C A (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30, 1575-84.
6. Gavin A C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen L J, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier M A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick J M, Kuster B, Bork P, Russell R B,Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-6.
7. Gavin A C, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J M, Michon A M, Cruciat C M, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier M A, Copley R R, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G,Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7.
8. Hartuv E,Shamir R (2000) A clustering algorithm based on graph connectivity. *Information Processing Letters* **76**, 175-181.
9. Ho Y, Gruhler A, Heilbut A, Bader G D, Moore L, Adams S L, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems A R, Sassi H, Nielsen P A, Rasmussen K J, Andersen J R, Johansen L E, Hansen L H, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen B D, Matthiesen J, Hendrickson R C, Gleeson F, Pawson T, Moran M F, Durocher D, Mann M, Hogue C W, Figeys D,Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3.
10. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M,Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Science* **98**, 4569-4574.
11. King A D, Przulj N,Jurisica I (2004) Protein complex prediction via cost-based clustering. *Bioinformatics* **20**, 3013-20.
12. Koyuturk M, Szpankowski W,Grama A (2007) Assessing significance of connectivity and conservation in protein interaction networks. *J Comput Biol* **14**, 747-64.
13. Li X, Wu M, Kwoh C K,Ng S K (2009) Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics* **11** (Suppl 1), S3.
14. Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database* (Oxford) bar009.

1 http://www.perl.com/
2 http://www.w3.org/TR/soap/

15. Mewes H W, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S,Weil B (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-4.
16. Pavlopoulos G A, Moschopoulos C N, Hooper S D, Schneider R,Kossida S (2009) jClust: a clustering and visualization toolbox. *Bioinformatics* **25**, 1994-6..
17. Pietriga E (2005) A Toolkit for Addressing HCI Issues in Visual Language Environments. *EEE Symposium on Visual Languages and Human-Centric Computing* (VL/HCC'05), 145-152.
18. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M,Seraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218-29.
19. Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M,Mewes H W (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* **32**, 5539-45.
20. Sacher M, Jiang Y, Barrowman J, Scarpa A, Burston J, Zhang L, Schieltz D, Yates J R, 3rd, Abeliovich H,Ferro-Novick S (1998) TRAPP, a highly conserved novel complex on the cis-Golgi that mediates vesicle docking and fusion. *EMBO J* **17**, 2494-503.
21. Shannon P, Markiel A, Ozier O, Baliga N S, Wang J T, Ramage D, Amin N, Schwikowski B,Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504.
22. Spirin V,Mirny L A (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* **100**, 12123-8.
23. Stoll D, Templin M F, Bachmann J,Joos T O (2005) Protein microarrays: applications and future challenges. *Curr Opin Drug Discov Devel* **8**, 239-52.
24. Willats W G (2002) Phage display: practicalities and prospects. *Plant Mol Biol* **50**, 837-54.