Data and text mining

Medusa: a simple tool for interaction graph analysis

Sean D. Hooper[1,*] and Peer Bork[1,2]

[1]European Molecular Biology Laboratory, Heidelberg, Germany and [2]Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany

ABSTRACT

Summary: Medusa is a Java application for visualizing and manipulating graphs of interaction, such as data from the STRING database. It features an intuitive user interface developed with the help of biologists. Medusa is optimized for accessing protein interaction data from STRING, but can be used for any type of graph from any scientific field.

Availability: Medusa, along with sample datasets and instructions, can be downloaded from http://www.bork.embl.de/medusa

Contact: hooper@embl.de

There are many graph analysis applications available, including Pajek (Batagelj and Mrvar, 1998), Cytoscape (Shannon *et al.*, 2003), Osprey (Breitkreutz *et al.*, 2003), ProViz (Iragne *et al.*, 2005), CNPlot (Batada, 2004) and SHARKview (Pinney *et al.*, 2005), each with their strengths but also drawbacks. Here, we present Medusa, a tool which addresses many of these issues. A quick rundown of its features and other programs that lack them (shown in parentheses) are as follows: Medusa displays up to 10 multiple edges concurrently between nodes using Bezier curves (all above applications) and allows users to add and delete nodes and edges by simply clicking the mouse (Pajek, Cytoscape, others). Background images can be inserted to enhance figure quality (Osprey, Cytoscape). Edges can be hidden or shown depending on the type of interaction (Osprey, Cytoscape, Pajek). Node properties can be described, such as color, annotation, position and shape (Osprey, Cytoscape). Medusa requires no additional packages (ProViz, CNPlot) and runs on any machine with Java 1.4.2 installed (ProViz, Pajek). It runs as standalone (SHARKView) and as an applet for use in web interfaces (most other applications). Graphs can be exported to image or postscript files. Overall, Medusa (screenshot shown as Fig 1) is designed to be a simple and an intuitive tool for customization of interaction graphs of any kind.

With the ever-increasing mass of biological data, as exemplified by large-scale cross-species comparisons of proteins and genes, an easy visualization of the often highly complex linkage between, e.g. proteins becomes more and more important. For instance, the interactions between core proteins and alternative functional modules of complexes (work in progress) can be represented as a graph of nodes (proteins) and edges (interactions between core and modules). Other graphs of protein interaction may be more abstract; for instance, the STRING ('search tool for the retrieval of interacting genes/proteins') database predicts protein–protein associations and includes a variety of indirect (non-physical) evidence types (von Mering *et al.*, 2005). The results from the STRING database can be studied using its web interface, or alternatively handled directly by the Medusa application, which is run on the client side.

The handling of large networks is made easier by layout algorithms and the option to hide or show certain edge types. Moreover, nodes and edges can be deleted or added either manually via mouse clicks, or added directly from STRING. Users can create their own data files which can then be easily appended to existing graphs. These data files are simply tab-delimited text fields describing edge relationships.

Used solely as a graph visualization tool, Medusa is designed to be immediately accessible, focusing on construction of figures. For graph analysis of a more mathematical nature, Pajek is highly recommended, although it has a much steeper learning curve. Medusa can export graphs to Pajek format.

Medusa is also available as an applet version for use directly in web pages. This version has a lower degree of functionality, since the main goal is to allow an easy display of graphs from a web-based database. It is already implemented in the STRING website and is planned to be incorporated into other web databases (work in progress). The graph is simply passed to the applet as parameters, along with other preferences, making it highly suitable as an interactive enhancement of an existing web server.

Medusa has been used in a number of EMBL projects as an in-house tool, ranging from yeast cell cycle studies (http://www.cbs.dtu.dk/cellcycle/yeast_complexes/figure1.html; de Lichtenberg *et al.*, 2005) to genotype–phenotype links (Korbel *et al.*, 2005). It is now publicly available and is designed to be a simple but efficient tool for quickly analysing graphs and producing figures. Medusa is easy to install and requires only Java 1.4.2, making it accessible to a wide variety of platforms. Medusa is free for academic use.
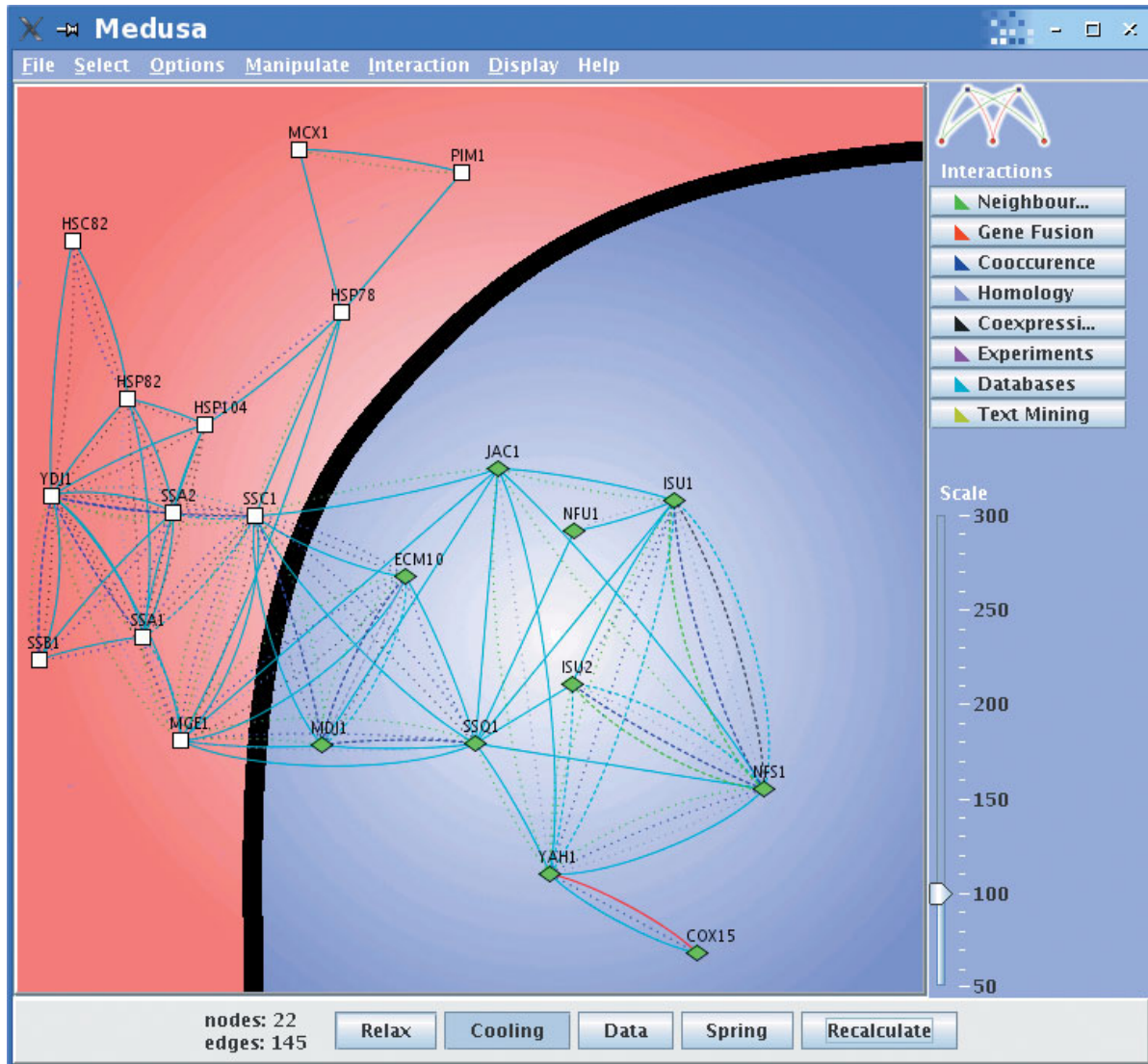
REFERENCES

Batada,N.N. (2004) CNplot: visualizing pre-clustered networks. *Bioinformatics*, **20**, 1455–1456.

Batagelj,V. and Mrvar,A. (1998) Pajek—program for large network analysis. *Connections*, **2**, 47–57.

---

*To whom correspondence should be addressed.

Fig. 1. Investigating the evidential and physical links between proteins transversing the mitochondrial-cytoplasmic boundary in Yeast. Details from work in progress (Lars Steinmetz, personal communication). For instance, ECM10 (a homolog to heat shock protein 70) is linked by gene neighbourhood, co-expression and text mining to the mitochondrial precursor protein MDJ1.

Breitkreutz,B.J. *et al.* (2003) Osprey: a network visualization system. *Genome Biol.*, **4**, R22.

de Lichtenberg,U. *et al.* (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.

Iragne,F. *et al.* (2005) ProViz: protein interaction visualization and exploration. *Bioinformatics*, **21**, 272–274.

Korbel,J.O. *et al.* (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.*, **3**, e134.

Pinney,J.W. *et al.* (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Res.*, **33**, 1399–1409.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

von Mering,C. *et al.* (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33** (Database Issue), D433–D437.