

Brief Bio and (PR)²: Problems & Pitches – Rants & Raves by Jason Montojo

Self Introduction



Jason Montojo is a bioinformatics software developer for Gary Bader's and Quaid Morris' labs at the Donnelly Centre for Cellular and Biomolecular Research at the University of Toronto. He is a core developer for the Cytoscape project, which provides an open source platform for network analysis and visualization. He is also a member of the GeneMANIA development team, which develops tools for predicting the function of and exploring the relationships between genes.

Jason has mentored numerous student projects for the Google Summer of Code program on behalf of the Eclipse Foundation and the National Resource for Network Biology. He has also served as an instructor and contributor to the Software Carpentry project, which aims to help scientists become more productive by teaching them basic computing skills. Jason holds a Master's degree in Computer Science from the University of Toronto.

Publications

GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, Morris Q, Bader GD. *Bioinformatics* 2010 26: 22.

The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q. *Nucleic Acids Res.* 2010 Jul 1;38 Suppl:W214-20.

Practical Programming: An Introduction to Computer Science Using Python. Campbell J, Gries P, Montojo J, Wilson GV. Pragmatic Bookshelf, January 2009.

Tools

<http://cytoscape.org>

<http://genemania.org>

General Questions

1) What are your main interests in attending the workshop?

Since this will be my first time attending the workshop, I'm interested in learning more about other scientific software development efforts, especially those with large user bases from diverse research communities. In particular, I am interested in learning how scientists are using these tools and identifying opportunities for integration and interoperation.

2) What are the tools or services you would like to share at the workshop?

I'd like to present two tools. The first is Cytoscape, which is an extensible open source platform for network graph analysis and visualization. Although many of the plugins available for Cytoscape target

systems biology researchers—in particular, those in the fields of genomics and proteomics—the software is generic enough to be used in other fields such as social networking research.

The second is GeneMANIA, which is a tool for predicting gene function through the principle of guilt-by-association. Given a gene list, GeneMANIA returns the connections between those genes using a large collection of functional association data, including protein and genetic interactions, pathways, co-expression, co-localization, and protein domain similarity. These connections are then visualized through Cytoscape Web, or the Cytoscape desktop application.

3) Please list three features or functions of your tools or services that are most important for the users.

- Accessibility: Our primary audience is computational biologists who have little or no software development experience. We designed Cytoscape and GeneMANIA so that ordinary researchers can leverage our technology without having to learn to program. We strive to give our users easy access to the latest data sets and analysis tools so they can focus on research instead of computational issues.
- Extensibility: Both Cytoscape and GeneMANIA were built as modular platforms for supporting scientific research with the expectation that other software developers will integrate our technology within their projects, or vice-versa.
- Scalability: Even though Cytoscape and GeneMANIA are capable of producing publication-quality network diagrams, both are designed to handle complex networks. Cytoscape is capable of visualizing and analyzing networks with millions of edges while GeneMANIA performs real-time predictions using a collection of almost 300 million interactions.

4) What are the major concerns for the architect design of the tools and services?

For Cytoscape, maintaining a simple but flexible and sustainable API is one of our largest concerns. In version 2.X, the API was unwieldy, inconsistent and unnecessarily exposed a lot of implementation details. This prevented the team from re-architecting the code into something more efficient without breaking the existing API. The plugin author community became frustrated that each new point release of Cytoscape required them to issue a new release of their plugin to fix those breakages. Version 3.0 is a rewrite of the entire platform to mitigate the risk of API breakage by implementing strict rules for API usage and evolution, which is enforced by our runtime environment, OSGi.

5) Are you aware of especially innovative approaches to plug-and-play feature where algorithms and plugins can be shared between different tools? If yes, please list down the approaches.

Both Cytoscape and GeneMANIA use a modular approach to tool development in which the core algorithms and business logic is cleanly separated from everything else. This allows us to write libraries for conducting domain-specific analysis that aren't tied to a specific platform. We then write platform-specific integrations that leverage those libraries. For example, GeneMANIA offers a library for gene function prediction that has been integrated into a set of command-line tools, an interactive web service, a plugin for Cytoscape 2.X, and an app for Cytoscape 3.0, all using the same code base. Likewise, even though Cytoscape is primarily a desktop application, all its analysis tools are designed to run without a graphical user interface so they can be leveraged by scripts or integrated into other software projects.

6) What are the challenges in developing the tools and services?

- Building a strong community: A software project can only be successful if both the users and the project members are happy. In the open source world, developers often come from different organizations with different agendas. This can lead to conflicting requirements within the project team and can significantly impede development, or even result in attrition. Without a concerted effort from the project members, the quality of the software may decline and lead users to seek other alternatives. To mitigate this risk, we have established a number of social protocols for the various phases in the software release cycle.
- Ease of use: Many of the software developers in the Cytoscape and GeneMANIA teams are not computational biologists so the user interfaces we design may not always be easy for them to use. Fortunately, we have many collaborations with actual researchers who provide us with critical feedback on how to improve our software.
- Cross-platform compatibility: Our users conduct research on many different platforms. We need to ensure our tools behave consistently on each one. Fortunately, our development teams use a variety of different operating systems as their primary development environment. We supplement this by testing our software on virtual machines configured to run the operating systems we claim to support.

7) Are you or your group working on any of these challenges? If yes, please explain.

Yes, this is explained above.

8) Does your development team contain volunteer developers? If yes, please explain how they are involved.

We occasionally have volunteer developers but they usually are not involved in an on-going capacity. In most cases, they develop a specific module or plugin, which we then integrate into our codebase.

9) What would you like to learn and achieve at the workshop?

I'm interested in learning more about CISHell and HUBzero and connecting with their development teams. I can already see some overlap between CISHell and Cytoscape, as well as between HUBzero and GeneMANIA.