

Title: Enabling a Systems Biology Knowledgebase with Gaggle and Firegoose

Project ID: 0016648

Report/Product #: DOE-ISB-4877-1

Prog Mgr: Susan K. Gregurick Phone: 301-903-7672 Division: SC-23.2

PI: Nitin S. Baliga

Final Report (December 12, 2014)

Overview

The overall goal of this project was to extend the existing Gaggle and Firegoose systems to develop an open-source technology that runs over the web and links desktop applications with many databases and software applications. This technology would enable researchers to incorporate workflows for data analysis that can be executed from this interface to other online applications. The four specific aims were to (1) provide one-click mapping of genes, proteins, and complexes across databases and species; (2) enable multiple simultaneous workflows; (3) expand sophisticated data analysis for online resources; and enhance open-source development of the Gaggle-Firegoose infrastructure. Gaggle is an open-source Java software system that integrates existing bioinformatics programs and data sources into a user-friendly, extensible environment to allow interactive exploration, visualization, and analysis of systems biology data. Firegoose is an extension to the Mozilla Firefox web browser that enables data transfer between websites and desktop tools including Gaggle.

In the last phase of this funding period, we have made substantial progress on development and application of the Gaggle integration framework. We implemented the workspace to the Network Portal. Users can capture data from Firegoose and save them to the workspace. Users can create workflows to start multiple software components programmatically and pass data between them. Results of analysis can be saved to the cloud so that they can be easily restored on any machine. We also developed the Gaggle Chrome Goose, a plugin for the Google Chrome browser in tandem with an opencpu server in the Amazon EC2 cloud. This allows users to interactively perform data analysis on a single web page using the R packages deployed on the opencpu server. The cloud-based framework facilitates collaboration between researchers from multiple organizations.

We have made a number of enhancements to the cmonkey₂ application to enable and improve the integration within different environments, and we have created a new tools pipeline for generating EGRIN2 models in a largely automated way.

Research

The goals of this project include development of algorithms to discover gene regulatory networks and software tools to integrate heterogeneous biological data. These computational tools have been extended with new features and applied to new research questions in a wider array of organisms.

Gaggle Architecture. We redesigned the Gaggle architecture as follows. First, we added an

embedded web server to the Gaggle Boss to support websocket communication. This is more robust and flexible than the Java Remote Method Invocation adopted before. And with websocket Gaggle can communicate with any web page directly. Before we had to embed Java Applets on web pages. This is very awkward and insecure. Second, we implemented an application invocation logic. It can automatically discover the paths of the executable of an application and starts it on demand. This allows Gaggle to start virtually any application.

Gaggle Workspace. Researchers gain insights by integrating data from diverse experimental and bioinformatic approaches by working simultaneously with datasets spanning multiple data types (e.g. gene expression, gene annotations, and regulatory network models) and formats (e.g. name lists, matrices, and networks). However, the data analysis process is still overly fragmented and effort-intensive. Users must start applications one-by-one, load datasets manually from many sources (e.g. a user's experimental data and public databases), manually execute analysis steps even in frequently used workflows, and resume interrupted sessions by manually restarting applications and websites and reloading data.

To address these problems, we implemented the Gaggle Workspace to

- capture and organize data from websites and local files;
- manipulate data from the Workspace and via automatically opened applications;
- create workflows automatically or manually which include multiple applications, websites, and datasets;
- save the state of an analysis session across all open websites and desktop applications and reload the state later from any machine;
- automatically generate summary reports.

The Gaggle Workspace enables centralized data organization and flexible automation of data analysis. The cloud-based environment, workflow features, and saving of session states all improve productivity and free users from roaming costs. Figure 1 illustrates some screenshots of using the Gaggle Workspace to capture data and perform data analysis.

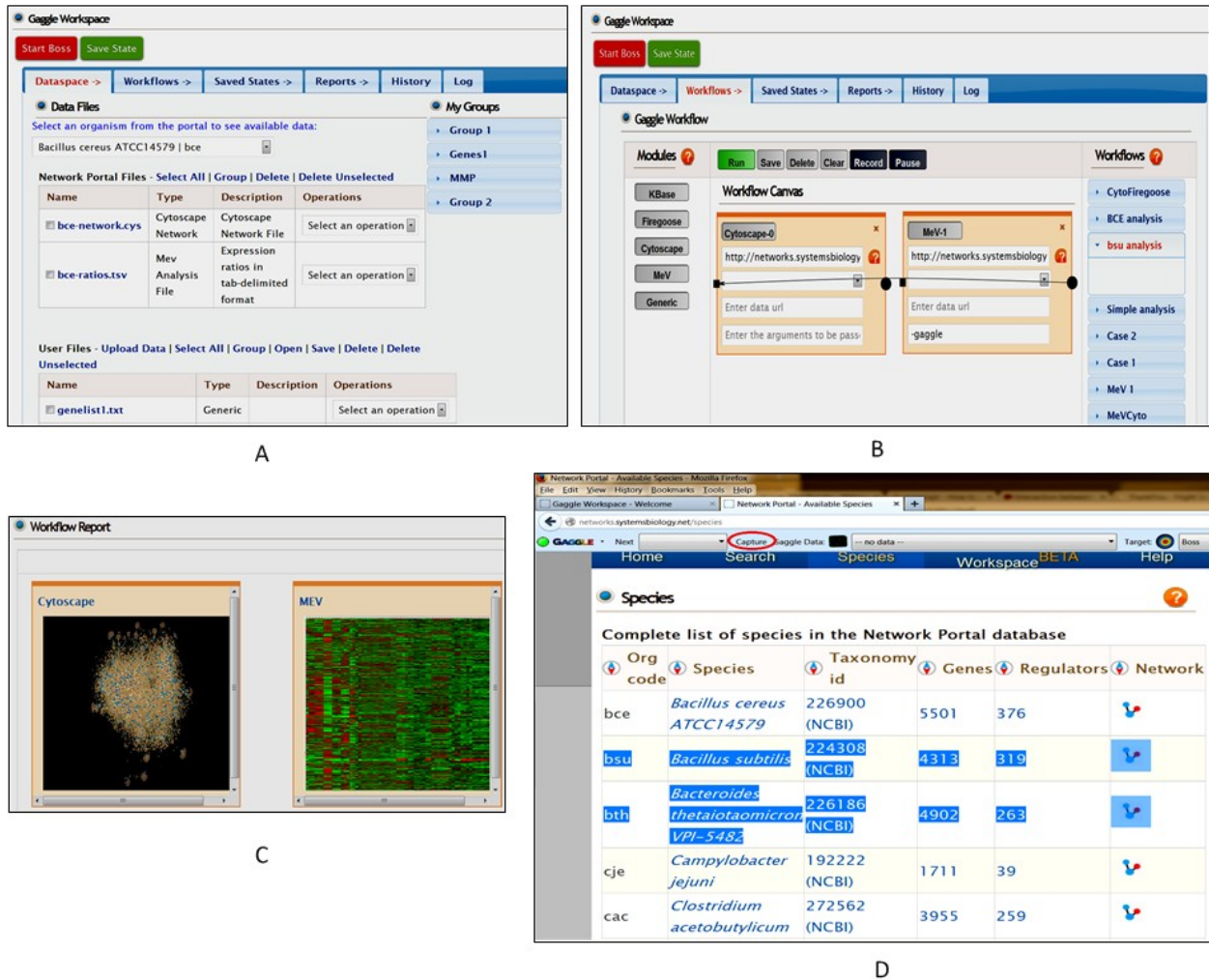


Figure 1. Gaggle Workspace. A: workspace UI. B: User created workflow. C: Reports. D: User captured data from Firegoose.

Advances in Network Inference Algorithm (cMonkey2). There are a number of enhancements and advances for the cmonkey2 software. The efforts that went into providing a generalized platform for building gene co-regulatory modules have been described in a paper that has been submitted to Nucleic Acids Research.

We have improved integration with DOE KBase and provided enhancements to enable RegPredict/RegPrecise integration, described in a publication that is currently in submission. Network portal users can submit cmonkey2/inferelator runs and view their personalized results after finalization.

We also have enhanced the cmonkey2 application to support the creation of EGRIN2 ensembles and are providing an integrated tool chain for the automated generation of EGRIN2 models on a Sun Grid Engine cluster.

To integrate with the Gaggle framework and the central tools Firegoose and the recently developed Chromegoose, cmonkey2 supports micro formats in the cmviewer web application that is provided to monitor and visualize a cmonkey run's intermediate and final results. If desired the user can select a cluster for all available iterations of the run and retrieve the genes and conditions that are associated with the selected cluster. If available, the motif PSSMs are provided for broadcasting as well.

Network Portal. The wider scientific community currently has limited access to tools for network inference, visualization, and analysis because these tasks often require advanced computational knowledge and computing resources. To meet this increasing demand for accessible platforms for the inference, storage, visualization, and analysis of biological networks, we have developed the Network Portal, a world wide web-based comprehensive network resource (<http://networks.systemsbiology.net>) (Turkarslan et al., 2014 Nucleic Acids Res). The portal is integrated into an advanced version of the Gaggle platform to enable interoperability across various desktop and web resources. Given that such integrated data and network analysis has become an essential tool for life science research, this resource should be relevant and exciting to a diverse audience.

Integration of Chrome Goose with Network Portal. We added network portal to the broadcast list of Chrome Goose so users can search data on Network Portal through Chrome Goose. We also added the “explore organisms” feature in Chrome Goose. Users can select an organism and Chrome Goose will open the portal page for the selected organism. We deployed R analysis packages to opencpu to allow users to analyze data using Chrome Goose as a frontend. The R analysis packages interfaces with the Network Portal to perform the analysis.

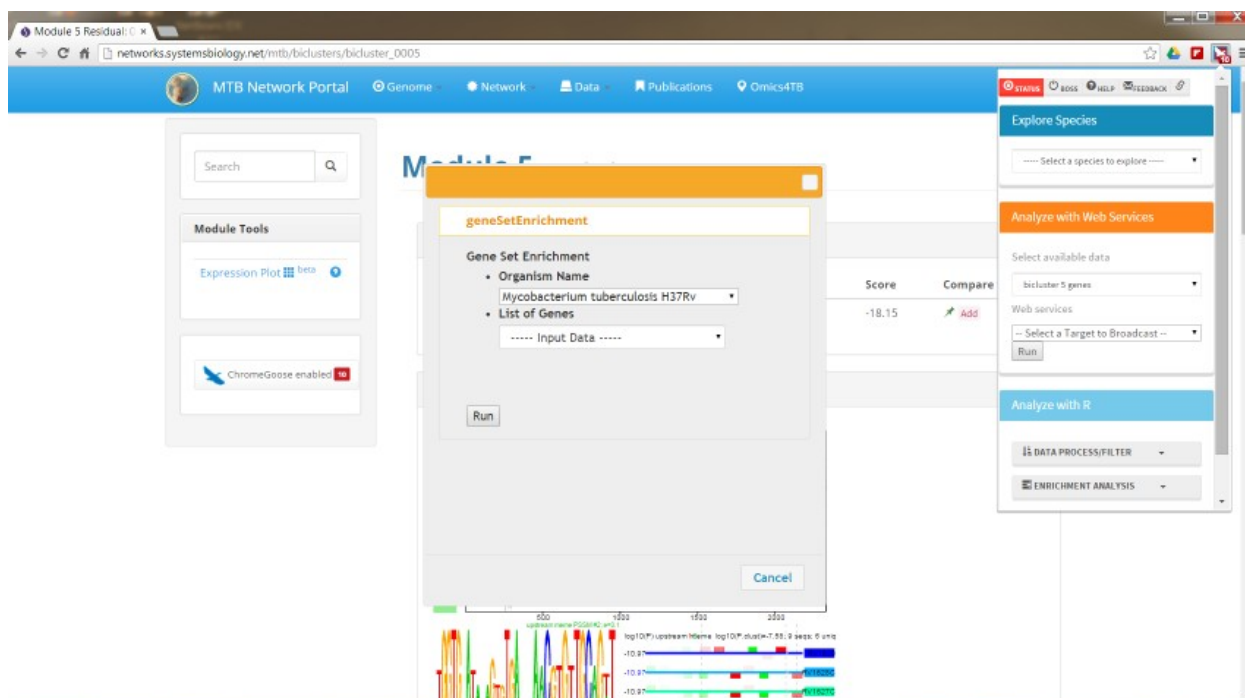
Organism-Specific Portals. In order to handle organism specific data and integrate with targeted resources, we created organism specific portals by developing a Biological Content Management framework and integrated with Gaggle. This framework was developed by using open source Drupal content management system that is widely used by high volume sites such as The Economist and MIT. It provides flexible and easy to use system for content management, advanced security features and extensive tools for collaboration and user management. We also integrate powerful search platform Apache Solr as a backhand for database queries and support faceted searches. In order to maximize the utility of the extensive resources generated by organism specific communities we have opted for a strategy that expands upon existing resources and tools. Well known organism specific resources as well as systems biology resources such as KEGG, String, RefSeq, Ensembl, BioCyc are seamlessly integrated with these portals. We also took advantage of the powerful collaborative tools Drupal provides to facilitate interaction with content and other researchers. In addition, other web and desktop tools such as cytoscape, MEV, R, can further analyze data in the portal by using Gaggle framework. Currently we have created organism specific portals for human pathogen; *Mycobacterium tuberculosis* (<http://networks.systemsbiology.net/mtb>) (Peterson et al., 2014 Nucleic Acids Res), green algae; *Chlamydomonas reinhardtii* (<http://networks.systemsbiology.net/chlamy-portal>), Diatom; *Thalassiosira pseudonana* (<http://networks.systemsbiology.net/diatom-portal>), sulfate-reducing bacteria; *Desulfovibrio vulgaris* Hildenborough (<http://networks.systemsbiology.net/dvh>) and Mouse T-cells

(<http://networks.systemsbiology.net/tcell>).

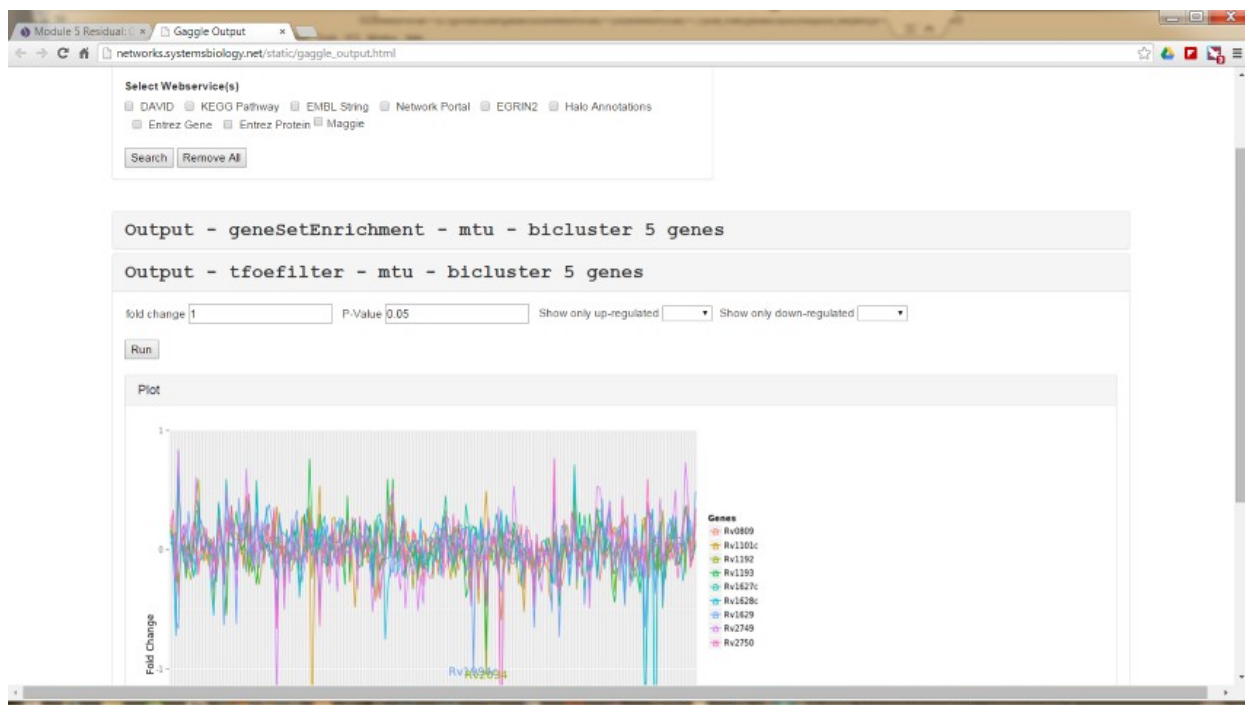
Chrome Goose. We developed the Chrome Goose to make gaggle features available for the Google Chrome browser. We architected the Chrome Goose to employ latest HTML5 technologies such as web sockets, local storage and AngularJS to make the system more robust and flexible. In addition to the features provided by Firegoose, we also implemented cloud-based data analysis. The R analysis tools are hosted on an Opencpu server running on the Amazon EC2 cloud. Users can submit data to the cloud from the Chrome Goose. Data will be processed and results are displayed on a central output web page for user to visualize and compare. The results are also “gaggled” (i.e., tagged on the output web page) so user can further analyze them using the tools provided by Gaggle.

Chrome Goose also interfaces nicely with Network Portal. It allows users to select an organism and open the landing page from network portal corresponding to the selected organism. Network portal pages are gaggled so that Chrome Goose can capture the data for user to perform further investigation. We also included the Network Portal as a destination for user to broadcast lists of genes to explore. Finally, the R packages deployed on the opencpu cloud server call web services offered by the Network Portal to access the data.

Figure 2. Chrome Goose Screenshots



A. Chrome Goose Analysis UI



B. Chrome Goose Analysis Output Page

Other Gaggle Tools. We provided hosting, performed maintenance and bug fixes on Firegoose and all the gaggle applications such as the Genome Browser, MeV, Cygoose, DMV, Translator, etc. given that almost every Java and Firefox update came with significant disruptive changes to the Java platform.

GGBWeb. Biological systems can now be studied across a wide range of molecular types and scales. The high-throughput technologies that allow such studies generate large amounts of complex data. Effective representation, integration, and interaction with this data are critical to drive biological insight. We developed a tool that allows end users to interact with integrated visualizations of dynamic biological processes in the context of the genome. GGBWeb is an interactive web application module that integrates biological data using multiple genomic tracks (Salvanha *et al.*, in preparation). While other genome browsers have been developed, GGBweb includes the unique ability to animate paired data from experiments with a dynamic context (e.g., changes in transcript expression levels and transcription factor binding across a time-course). GGBweb is also readily embedded in pre-existing web resources, encouraging developers to extend their own web approaches using our tool.

The user interface is implemented with HTML5 and a Javascript library that provides a straightforward web environment for end users. Data can be uploaded to the tool using several strategies (e.g. flat-file, database retrieval). An essential missing feature was data interoperability, but was readily solved with the availability of a HTML5 based version of the Gaggle framework, the ChromeGoose.

Summary of Outcomes

Application	Websites	Access Statistics (last 2 months)	Data Sets
Network Portal	http://networks.systemsbiology.net	TS: 399, CS: 4, NU: 262	Gene expression, Regulatory Network, Functions to Module, Functions to Gene, Gene to Module associations for 13 species.
Gaggle Workspace	http://networks.systemsbiology.net/workspace	TS: 5, CS: 0, NU: 2	N/A
cMonkey ₂	http://github.com/baliga-lab/cmonkey2	Forks: 7, 100 views, 9 unique visitors	N/A
Chrome Goose	Chrome extension	TS: 1818, CS: 7, NU: 192	N/A
Firegoose	http://gaggle.systemsbiology.net/geese/firegoose/	TS: 29, CS: 0, NU: 14	N/A
MTB Network Portal	http://networks.systemsbiology.net/mtb	TS: 716, CS: 3, NU: 292	261 ChIP-seq data, 698 Gene expression data, Regulatory network

TS: total sessions, CS: Countries with more than 20 sessions, NU: new users

Publications

- [1] Reiss DJ, Wu WJ, Plaisier CL, Baliga NS. cMonkey₂: Automated, systematic detection of co-regulated modules for any microbe, Nucleic Acids Research Methods Online, in revision.
- [2] Brooks AN, Reiss DJ, Allard A, Wu WJ, Salvanha DM, Plaisier CL, Chandrasekaran S, Pan M, Kaur A, Baliga NS. A system-level model for the microbial regulatory genome. Mol Syst Biol. 2014 Jul 15;10(7):740. PMID: 25028489.
- [3] Wurtmann EJ, Ratushny AV, Pan M, Beer KD, Aitchison JD, Baliga NS. An evolutionarily conserved RNase-based mechanism for repression of transcriptional positive autoregulation. [Mol Microbiol](#). 2014 Apr;92(2):369-82
- [4] Turkarslan S1, Wurtmann EJ, Wu WJ, Jiang N, Bare JC, Foley K, Reiss DJ, Novichkov P, Baliga NS. Network portal: a database for storage, analysis and visualization of biological networks. Nucleic Acids Res. 2014 Jan;42(Database issue):D184-90. Epub 2013 Nov 23. PMID:24271392.
- [5] [Ashworth J](#), [Wurtmann EJ](#), [Baliga NS](#). Reverse engineering systems models of regulation: discovery, prediction and mechanisms. Curr Opin Biotechnol. 2012 Aug;23(4):598-603

References

ChromeGoose at the Google Chrome AppStore and Github:
<https://chrome.google.com/webstore/detail/gaggle-chrome-goose/loehpehjfhoaekdfacfkjbiohohafgb?hl=en-US>
<https://github.com/baliga-lab/ChromeGoose>

cmonkey2 source code at Github
<https://github.com/baliga-lab/cmonkey2>

Network Portal official website and source code:
https://github.com/baliga-lab/network_portal
<http://networks.systemsbiology.net/>

EGRIN 2.0 website
<http://egrin2.systemsbiology.net/>
Gaggle Genome Browser website
<http://igbweb.systemsbiology.net/>