CSE 6242 Project: Improving the Vectorbase Bioinformatic Data

Jon Gerhart, Will Hutwagner, Bridget Neary, Mani Jain, Joey Gonzales

Introduction - Motivation

New molecular analysis technologies, along with the wide availability of increasing quantities of biological data have allowed researchers an unprecedente vector-pathogen-host interactions involved in the transmission of many infectiou diseases[1]. VectorBase is a resource for studying vectors of human pathogens, wi genomes and other content for over 35 organisms.[2] VectorBase has the stated go valuable resource is one of several Bioinformatics Resource Centers supported b underpinned by community contributions and is crucial in the progress of this fie be to collaborate with VectorBase to further their stated goals of offering users op and interfaces enabling meaningful application and exploration of this data[4].

Problem definition

We aimed to streamline the user interface and data curation process and t interactive visualization tools to help with hypothesis formulation in medical ent VectorBase uses Esri Leaflet to visualize ArcGIS data collected during the norma experimental and control efforts. VectorBase uses a visually busy navigation syst for this service, in large part due to the high dimensionality and sparseness of the

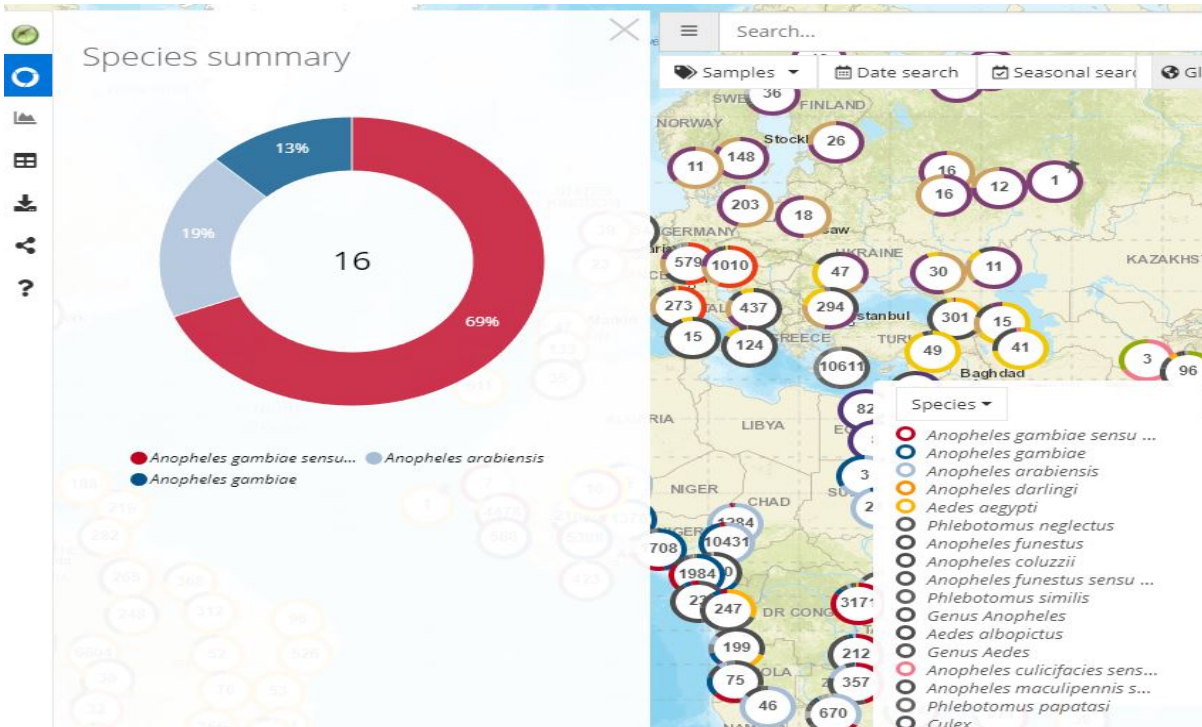

Figure 1: Example of current VectorBase population browser UI

Survey

1. This navigation service would benefit from streamlining and organization int
system. In addition, VectorBase uses violin plots to visualize large sets of phenot
obfuscates some of the key desired dimensions of the data request and merits red

2. Currently, many of the gene annotation links in VectorBase are deprecated wh
usability as a unified reference to other resources. The NCBI refseq database wil
retrieve updated gene annotations with the help of NCBI's Entrez system, by usi
Entrez programming utilities (E-Utilities and Entrez Direct) and bulk transfer via
Further, we plan to fully automate this process and provide it to VectorBase[7].

3. The development of a visualization/clustering tool for quick meta-analysis of e
data, such as large sets of genes. We can use statistical algorithms to arrange thes
similar expressions and view them via a heat map[8]. We plan to take advantage o
machine learning aiming to produce subgroups of related points. These subgroup
organized into dendrograms for analysis (neighbor-joining tree)[9]. We can find n
genes for a specific phenotype by developing a text mining system to look for th
Gene Ontology, PUBMED, and other databases[10].

We will need general software, algorithmic, and visualization-related component
three major aims of the project. For this, we can draw inspiration from BIRDS, a
international project to advance the fields of bioinformatics and information retri
developing new data structures and algorithms. The scope of our project is much
the state-of-the-art techniques developed by the BIRDS project may inspire us to
VectorBase using similar techniques[11]. We can also explore a set visualization te
to the OnSet system by Sadana et al. to analyse the degree of similarity between
clusters[12]. Libbrecht and Noble also provide a useful overview of machine learn
as used in genetics and genomics that we think will support our aims well[13].

Proposed method

Intuition (*Why should it be better than the state of the art?*)

We anticipate that this to be a critical improvement for the field of vector biology
VectorBase is a powerful tool for biologists and the most widely used resource o
user interface requires improvements to make the most of the existing infrastruct

expanding the capabilities of the infrastructure will allow the resource to continu
up-to-date. Harnessing these two improvements we will be able to implement a n
users to find biologically relevant insights to their data based on text annotations
bioinformatics resources and clustering methods that improve on the current data
tool that has static clustering per species. By automating data curation and stream
presentation we improve the impact of the service. Rather than all VectorBase se
organizing data based on user requested subsets of submitted annotations, the ma

and clustering algorithms we plan to implement can expose unexpected correlate
make excellent targets for further research.

Aim 1: Improve existing visualization in vector base

Our updated UI for VectorBase required modifying the existing JavaScrip
implementation, which uses the Esri Leaflet API for the interactive map and D3
data visualization. Currently, when a filtered data selection is made, VectorBase
as a scatterplot, which switches to a violin plot if the selection is larger than a ce
VectorBase reports that their users are confused by the violin plots, and would lik
implement a single visualization style that is easy to read for datasets of any size

Aim 2: Automate gene data curation in vectorbase

VectorBase users reported trouble accessing links to genes recently updat
because VectorBase does not have an automated way to updated HTML links wh
changed by their counterparts. We used lxml and a regex search system to retriev
nuccore hyperlinks and scraped for any hyperlinks in the body that redirected to
and output a master list containing any newer links. This implementation is funct
completed.

Aim 3: Visualizing the relationship between expressed genes

We extracted expressed genes from Vector Base database and used k-mea
cluster genes based on their common features that we extracted from NCBI, EMI
(curated bioinformatics databases). We choose Gene ontologies and molecular pa
number of clusters was the hyperparameter that we tuned and experimented with
dataset of 2500 genes. The optimal number of clusters that gave us good results i
clusters.

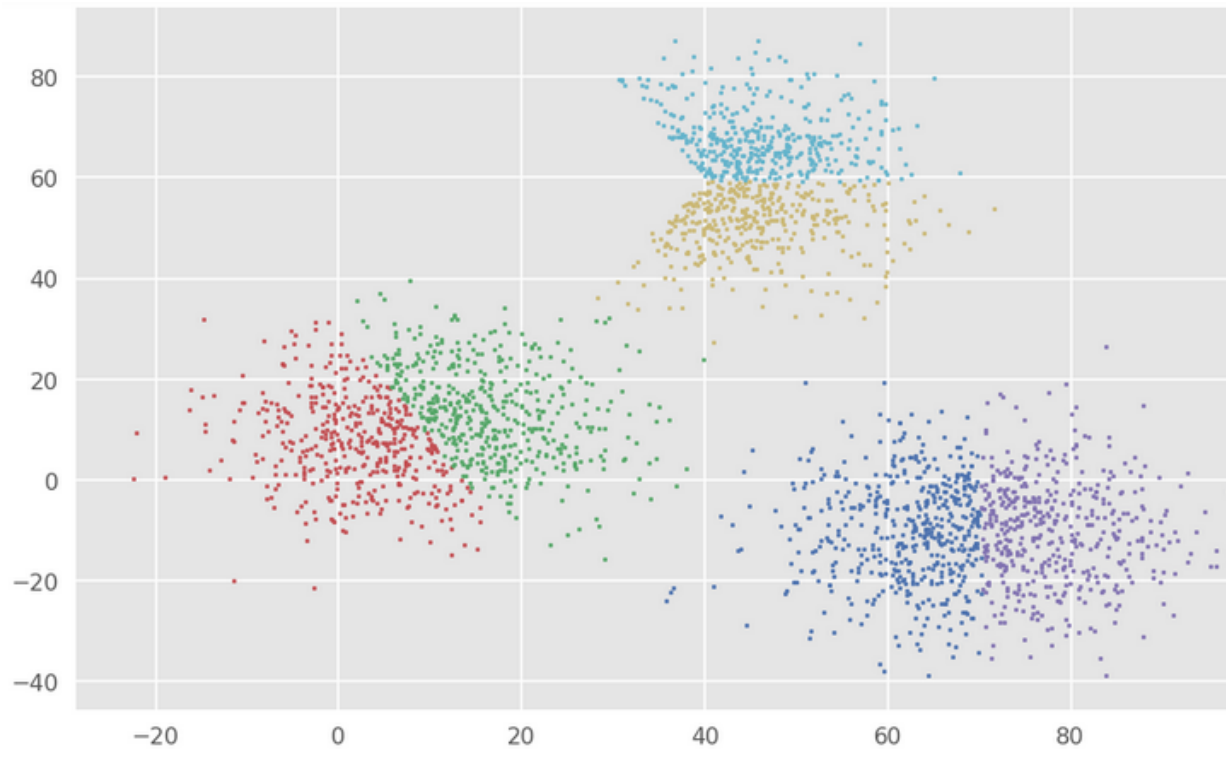

Figure 2: Clustering of expressed genes chosen in the pilot dataset with 6 number of clusters

With the improved population biology map, User selects from a list of co species and enters a list of comma separated Ensembl gene names. We use the Bi obtain gene ontologies for each gene. We use these ontologies to query the same hierarchy of ontologies that range from specific to broad. This hierarchy is then t and relatedness between the list of genes is determined as the number of edges th leaves. This relatedness is visualized as a force directed graph where each node i each edge represents the existence a common ontology term, weighted for the dis lowest common ancestry. A python based user interface web-server was made fo component using Flask. The interface contains a scroll-bar option for species dat text bar for list of genes.

After retrieving data relevant to the gene from the three databases and their class plan to evaluate each of these classes using machine learning algorithms based o distance such as FuzzyWuzzy[17] and using similarity computed from n-gram app we will prepare our own separate small reference set with words having biologic selective terminology. We intended to implement a fuzzy text matching algorithm appearance of each word based on frequencies and matches with the reference se enable us to deliver biologically significant and meaningful results to the user.

enable us to deliver biologically significant and meaningful results to the user.

We then implement a scraper to ensure that the texts are indeed related to the gen
word count, include beneficial language with biochemical and cellular terminolo
language such as “unannotated” and “hypothetically.” One last quality analysis a
performed on the list of texts ensuring that the texts are acceptable and possibly i
machine learning model to find keys or patterns to improve upon the mining and

Finally, we present our gene data through an interactive visualization implemente
JavaScript/D3.

Approaches

We divided the undertaking into three aims, which will be evaluated and
separately:

Aim 1: Improve existing visualization in vector base

VectorBase described issues with visualizing large numbers of samples from thei
browser, when there are multiple species being visualized at the same time. In th
the vectorbase code, all the data points are collapsed in a single violin plot, whic
simple, but loses information relevant to the user's query. We intend to refactor t
code to include a collection of stackable, color coded overlapping fields. Based o
we believe this will be more interpretable.

The map supports different views of the data, from which the user may select ov
open secondary visualizations. VectorBase reported that their component for vie
resistance statistics for selected samples did not clearly communicate data on gro
phenotypes and metrics at the same time, so we developed a mockup component
better supports simultaneous visualization. The example of that D3 code is visua

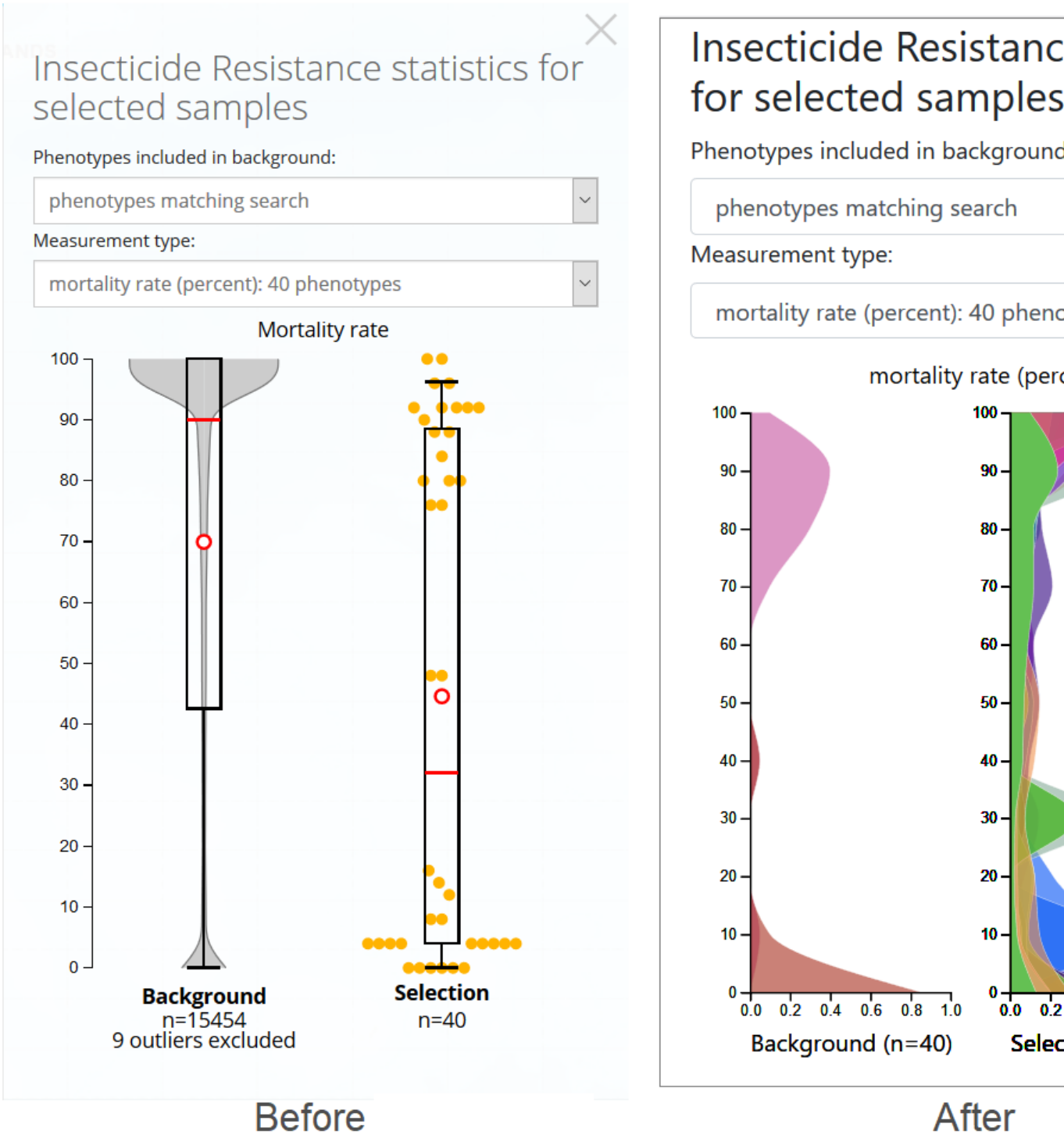

Figure 3: Visualization before and after implementation in D3.

Aim 2: Automate gene data curation in VectorBase

VectorBase users reported trouble accessing links to genes recently updated by N
VectorBase does not have an automated way to updated HTML links whe

changed by their counterparts. We coded a simple script that checks and u
hyperlinks in VectorBase pages when run.

Aim 3: Combine lessons learned from aims 1 and 2 into a working applet (NetG
provide summaries of gene relations from NGS data.

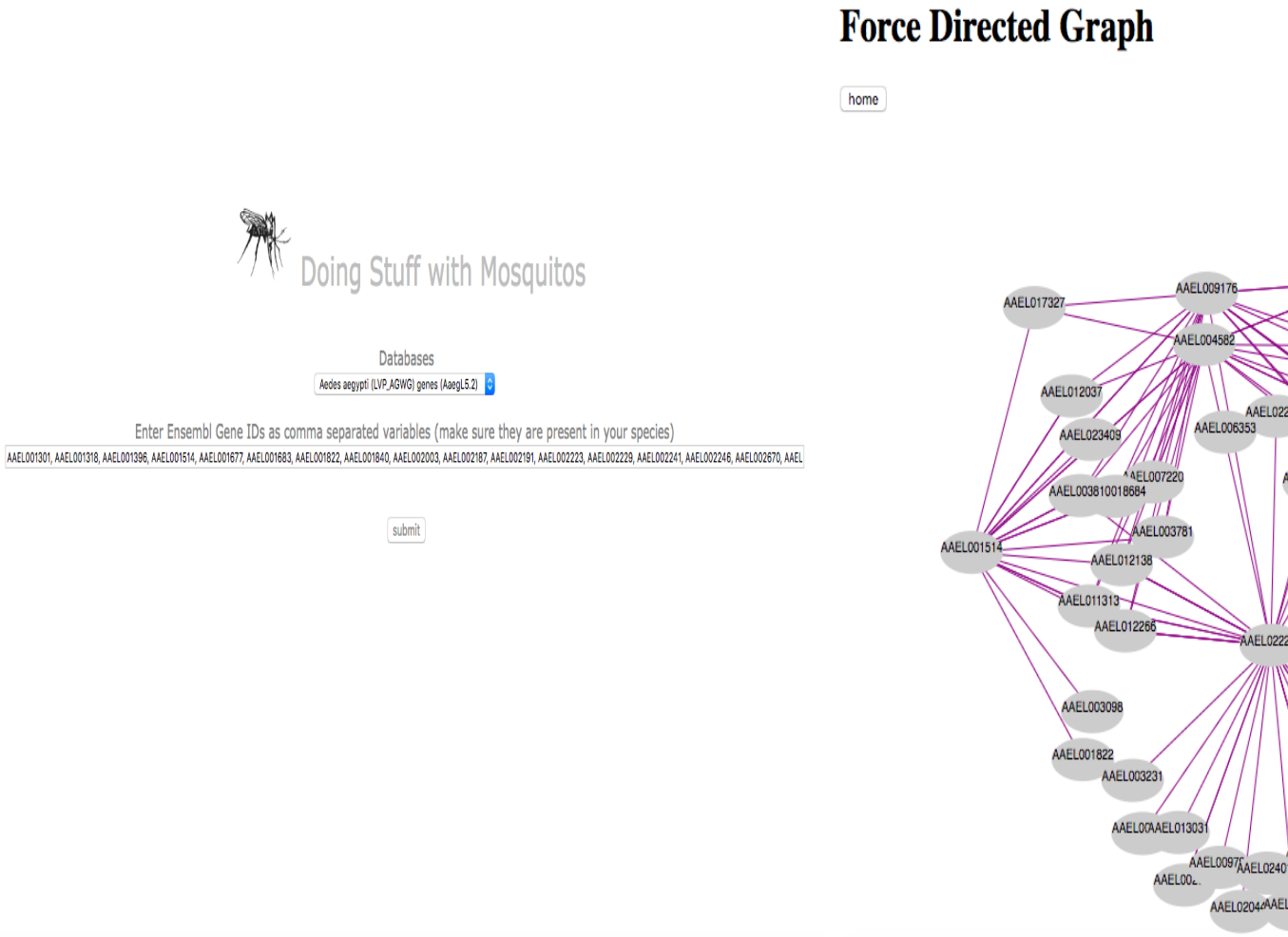

Figure 4: User interface going from species database and gene list to a force dire
visualizes the relatedness between genes.

Experiments/ Evaluation

Description of testbed:

Aim 1 will be evaluated by user feedback through interviews.

Aim 2 outcomes are self-evident.

Aim 3 will be evaluated via surveys of user experiences.

Details of the experiments:

Aim 1 was not completed until April 25th due to difficulties with
implementation of Aim 1 proved to be more challenging than anticipated, so only
a static data selection was completed. Simulations of Aim 1 received favorable re
entomologists interviewed. Aim 2 has been trialed successfully, but Vectorbase w
in implementation due to a recent database update. Aim 3 is not yet feature comp
experimentation, due to delays in Aim 1 information and talent transfer, but mult

experimentation in visualization techniques, edge tracing and culling, and interac
led to a solid concept that could be further refined into a competitive data parsing

Experiments in aim 3 were in 3 phases. The first phase was identifying th
measure distance between genes based on GO terms. Originally, we wanted to w
of the ontology hierarchy according to relationships between the ontology terms,
was unavailable, so each edge was given a weight of 1. We also experimented on
find distances between genes in this hierarchy. Initially we used the Floyd-Warsh
finding shortest paths between all genes. This was too computationally expensive
implement in real time, so we devised a method of comparing lists of parent term
identifying the lowest one for each gene pair, if present. This cut the computation

The second phase of experimentation for aim 3 was defining an edge wei
informative for the user to define the appropriate cutoff for their gene set in term
as an edge. This phase is still in development but the gene distances will likely b
as the distribution tends to be log-normal. Ongoing experimentation also involve
clustering algorithms. We validated our graph clustering using networkx clusterin

Conclusions and discussion:

Due to personnel issues, we did not complete aim 1 within a reasonable S
the full functionality. In addition, the aim 1 delays cascaded into the developmen
front-end and interactivity portions. Further, aim 2 testing was delayed to a point
went through a blanket update of all hyperlinks instead of automating the process
script work purely academic. Aim 3 represents an interesting approach to organiz
presentation of gene IDs by ontology, but is not currently feature complete, missi
interactivity, interface and clustering components. The NetGO team regrets this c

Distribution of team member effort

Jonathan Gerhart coordinated with the VectorBase team, developed proje
reports, vectored and designed poster, designed, tested and wrote aim 2, managed
and contributed to modules 5 and 6 design phase.

Bridget Neary coordinated team efforts, wrote key elements of aim 3, con
phases of development of aim 3, assisted in experimental design, and proofread c
reports.

reports.

Will Hutwagner drafted, developed and debugged the web server and AP
components of aim 3, coordinated aim 3 development and provided logistical su
for the poster presentation.

Mani Jain drafted, tested and reviewed visualization techniques for aim 3 all phases of debugging, contributed significant insight to attempts to implement proofread reports and communications with VectorBase.

Joey Gonzales-Dones implemented the VectorBase PopBio insecticide re for aim 1 and tried to provide general assistance with other tasks where possible. only team member that is not a bioinformatics student, and thus found it more ch overwhelming to contribute to the project, so he wasn't able to contribute as muc as the other team members.