

Biomarker Identification using Graph Theoretic and Particle Swarm Optimization-based Support Vector Machine Analysis of the Phase I ToxCast™ Dataset

Lyle D. Burgoon, Ph.D.

Visiting Assistant Professor

Director, Computational Biology and Bioinformatics
Laboratory, Gene Expression in Development and
Disease Initiative

Quantitative Biology Initiative

Department of Biochemistry & Molecular Biology

National Food Safety & Toxicology Center

Center for Integrative Toxicology

E-mail: burgoonL@msu.edu

<http://www.msu.edu/~burgoonl>

MICHIGAN STATE

UNIVERSITY

Overview



- Objective
 - Particle Swarm Optimization (PSO) Coupled to Support Vector Machines (SVMs)
 - Apriori Algorithm
 - Apriori + Network Traversal
 - Conclusions
-

Objective



- To identify putative agglomerative *in vitro* biomarkers predictive of *in vivo* toxicity
 - Putative agglomerative biomarkers should have high sensitivity and specificity
-

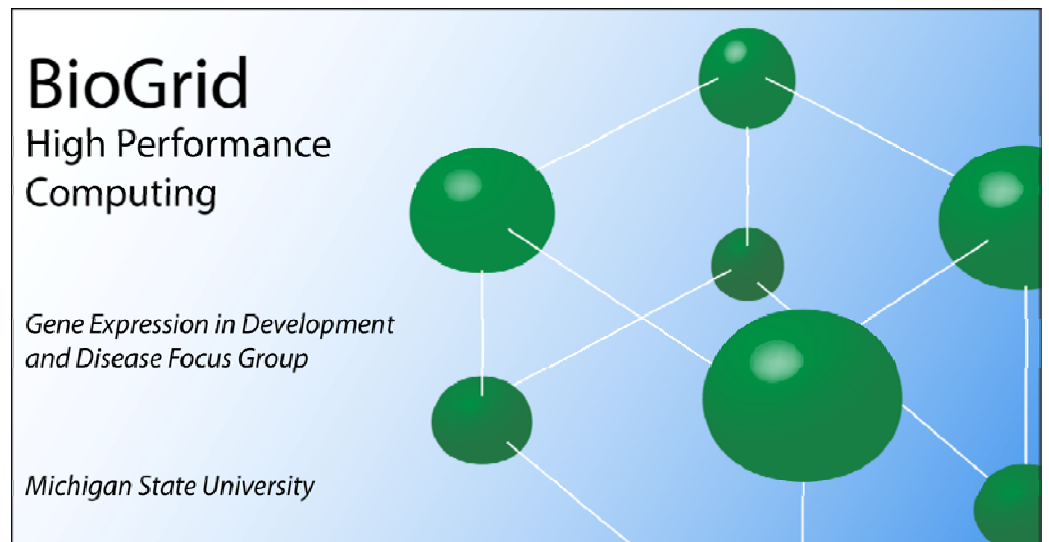


Particle Swarm Optimization Coupled Support Vector Machines
POWERED BY BIOGRID

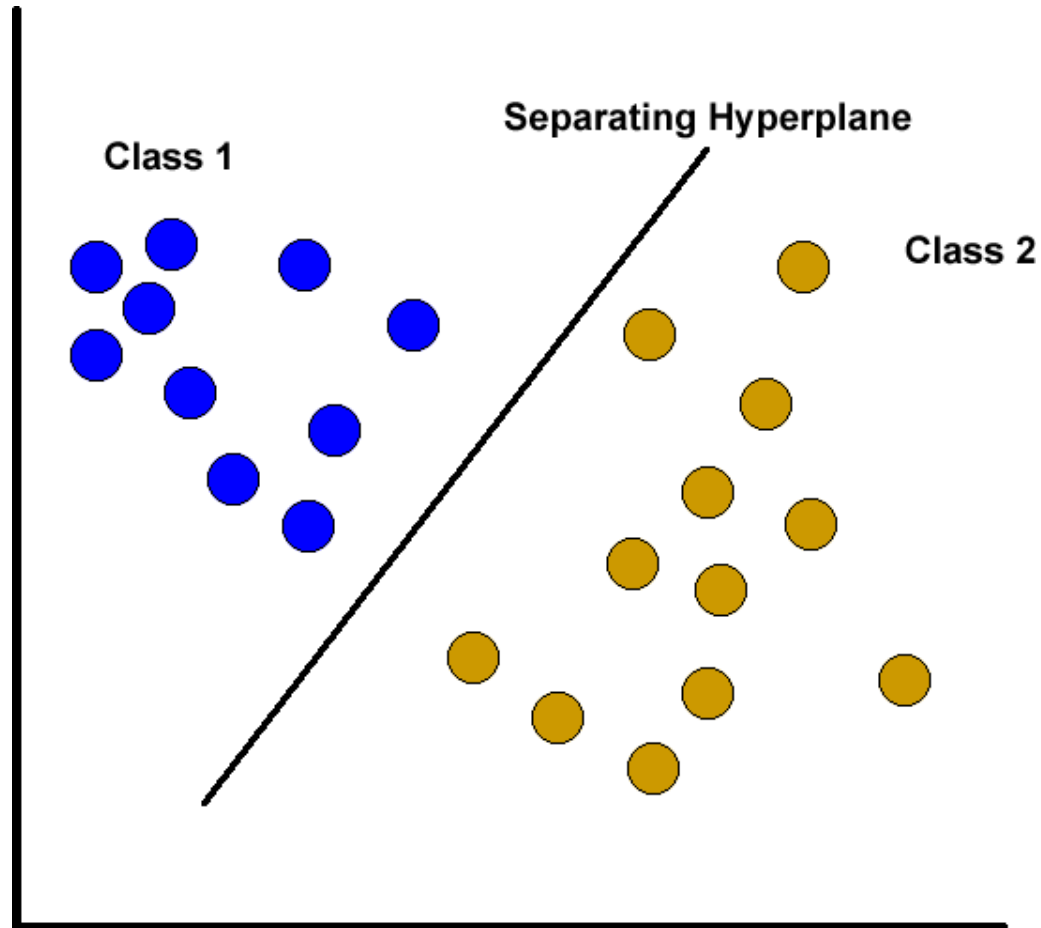


High Performance Computing

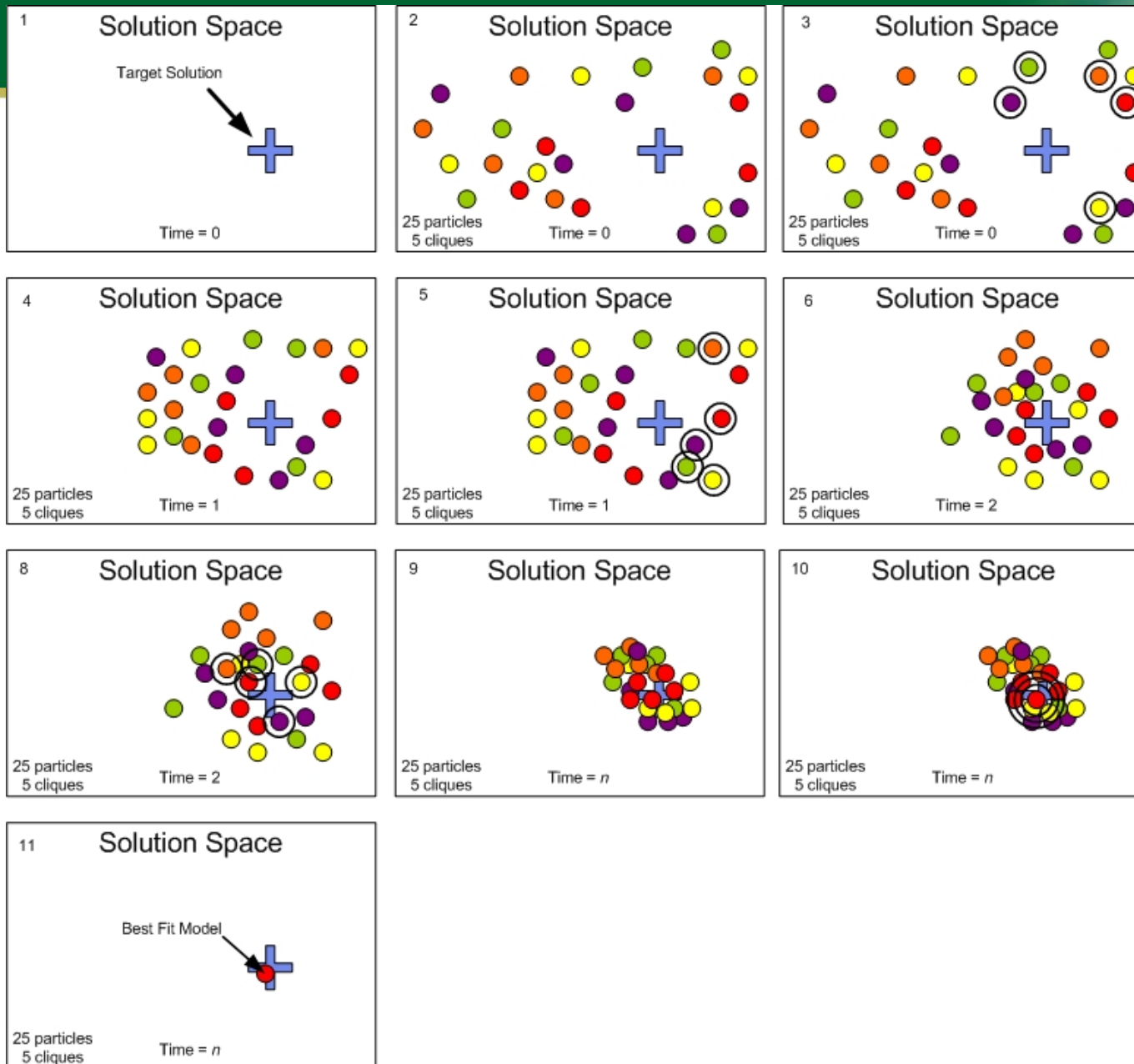
- Virtual Super Computer
- Currently an 11 node (computer) grid
 - 1 computer: 2 dual processors (essentially 4 processors)
 - 1 computer: 2 quad processors (essentially 8 processors)
 - 1 computer: 4 Intel Xeon processors (full-time)
 - 1 computer: 4 processors
 - 1 computer: 2 processors
 - 6 computers: single processors
- Equivalent to the processing power of 28 computers



Support Vector Machines



Particle Swarm Optimization (cont'd)



Method...in practice

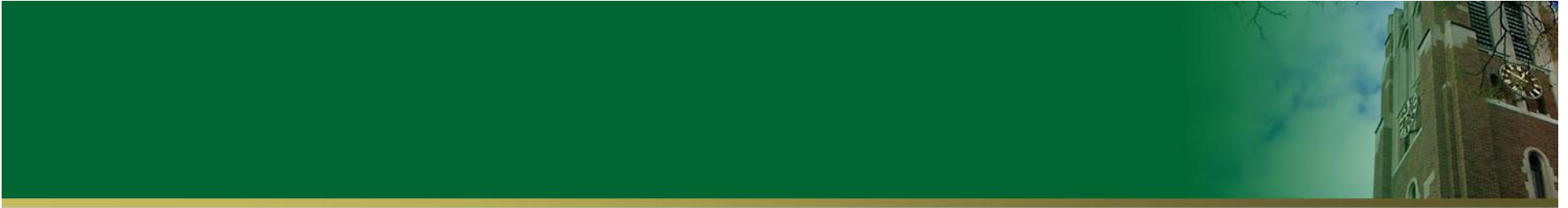


- Performed the PSO-SVM
 - All ToxCast *in vitro* data
 - Performed a PSO-SVM for each ToxRefDB phenotype separately
 - Ran on BioGrid
 - PSO
 - 1,000 SVM models
 - 10,000 iterations aiming to minimize the objective function
 - Ruby
 - SVM
 - Objective function for the PSO
 - 10-fold cross-validation
 - Mean of the sensitivity
-

PSO-SVM Results



- We have identified and constructed models for every ToxRefDB phenotype
 - Models averaged 20 *in vitro* assays
 - Still working on identifying if any of these make biological sense...
 - Liver Hypertrophy, proliferative lesions, and liver necrosis appear to make sense biologically
 - Need to validate these models further
-



Apriori Algorithm + Network Traversal



Apriori Algorithm

- Frequent Itemset Mining
- Learns association rules
 - {Chocolate, laundry detergent, diapers} => {"Save the kitty cats" promotion}
 - {Frozen pizza, Pepsi, Ramen} => {Shaving cream, men's razors}
- Common method for identifying items that people purchase together
 - Used by grocery stores to determine items in store that may "drive" purchase of other items
 - {Milk} => {Cereal}

Apriori Calculations



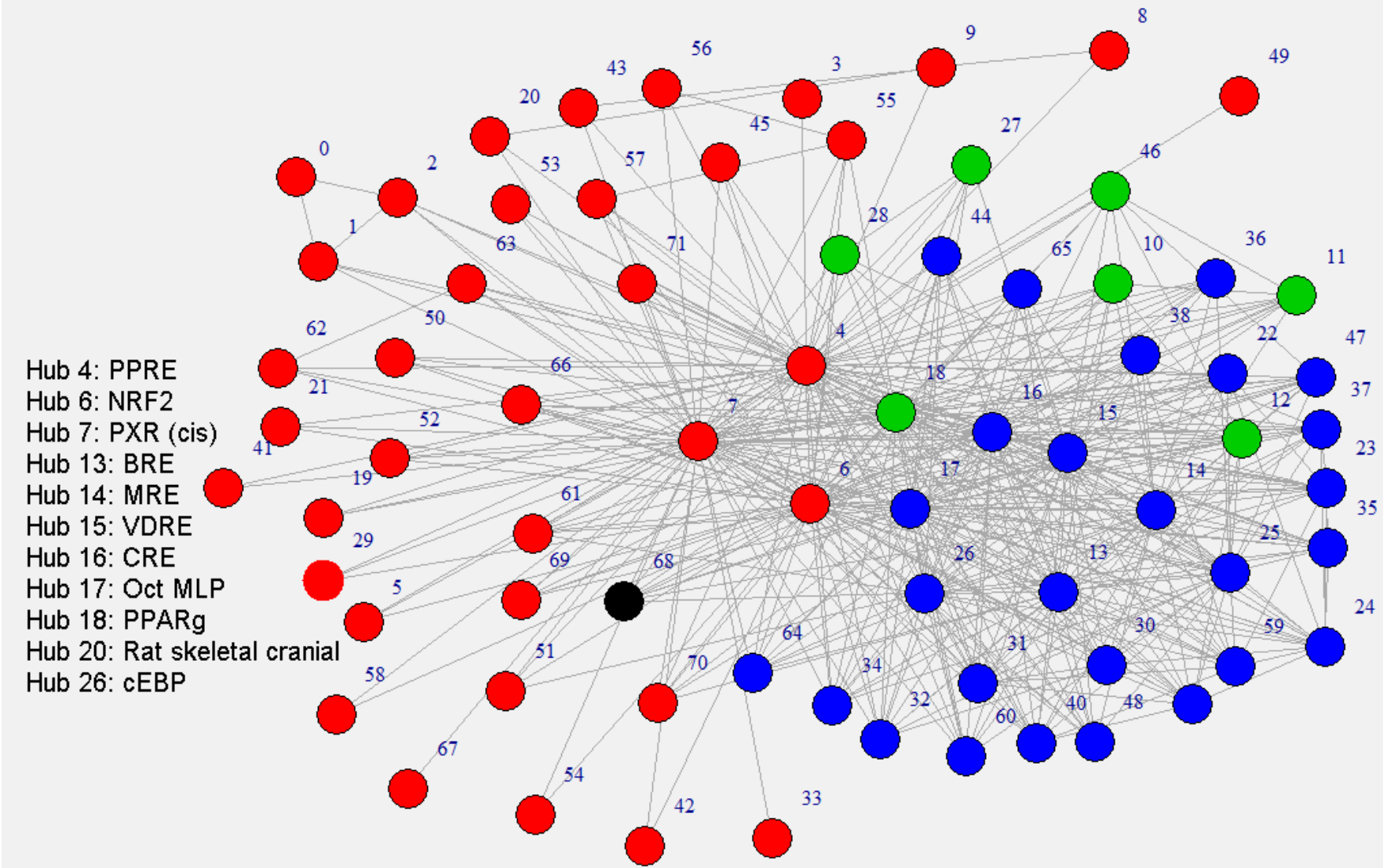
- Support
 - Number of transactions that include all items in both the antecedent and the consequent
 - Confidence
 - $\text{Support} / \text{number of transactions that include all items in antecedent}$
 - Lift
 - $\text{Number of transactions containing the consequent} / \text{total number of transactions}$
-

Apriori Analysis



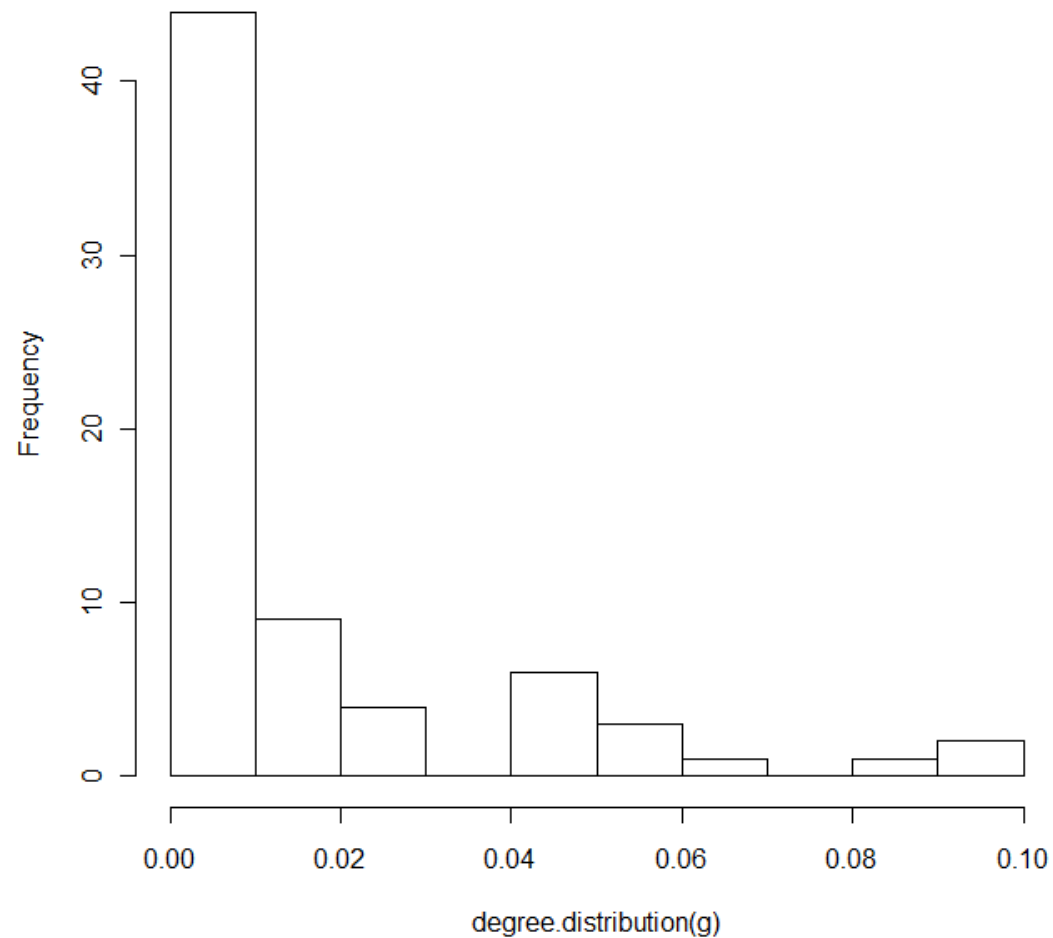
- R using the arules package
 - Minimum support level: 80%
 - Minimum confidence level: 80%
 - Discretized the ToxCast and ToxRefDB data
 - ToxCast: $< 100 \Rightarrow \text{ACTIVE}$; else $\Rightarrow \text{INACTIVE}$
 - ToxRefDB: $< 1000 \Rightarrow \text{ACTIVE}$; else $\Rightarrow \text{INACTIVE}$
 - Only calculate all 2-member rules
 - $\{\text{Item 1}\} \Rightarrow \{\text{Item 2}\}$
-

Network Construction (Attagene)

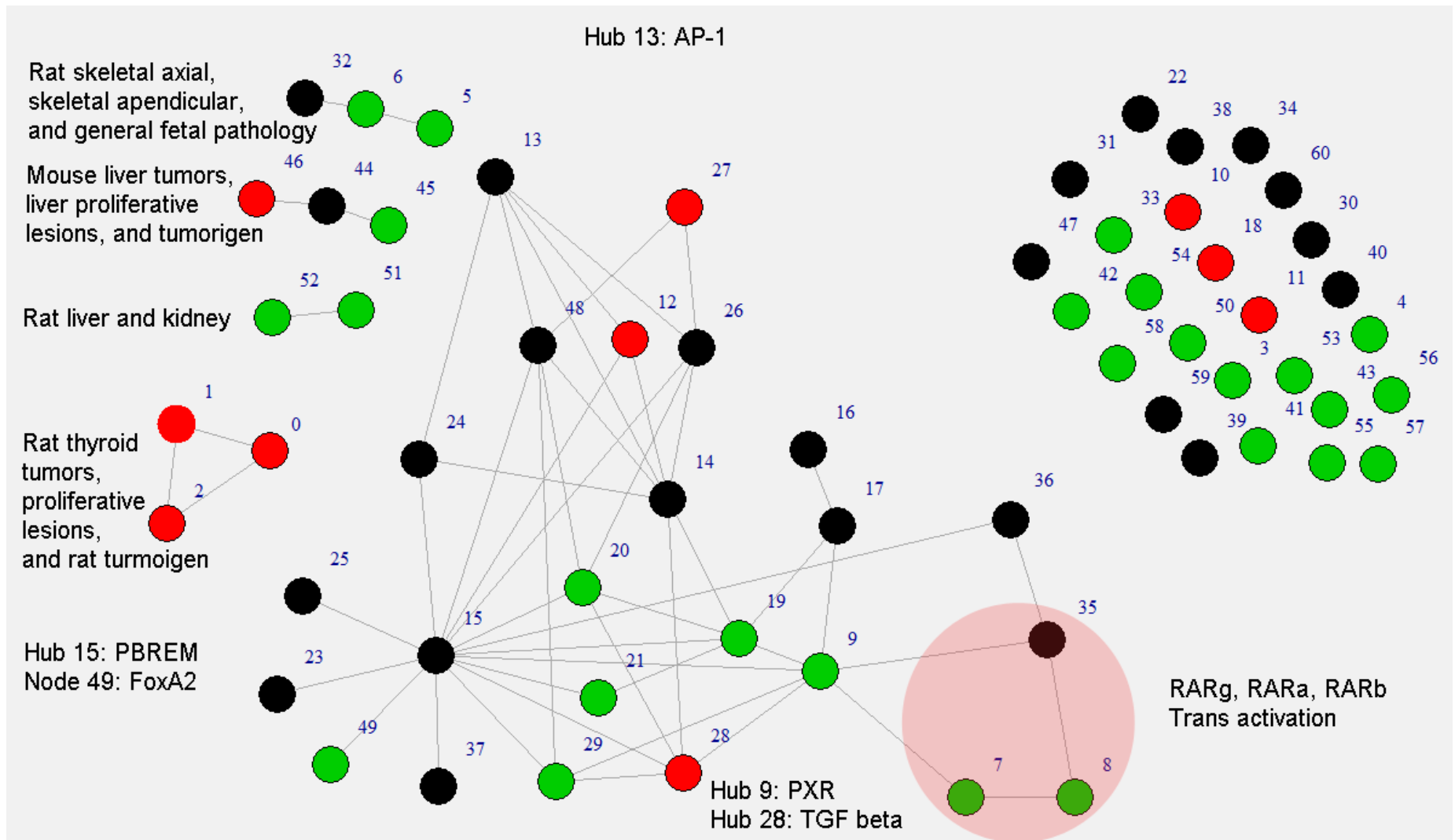


Networks Are Scale Free

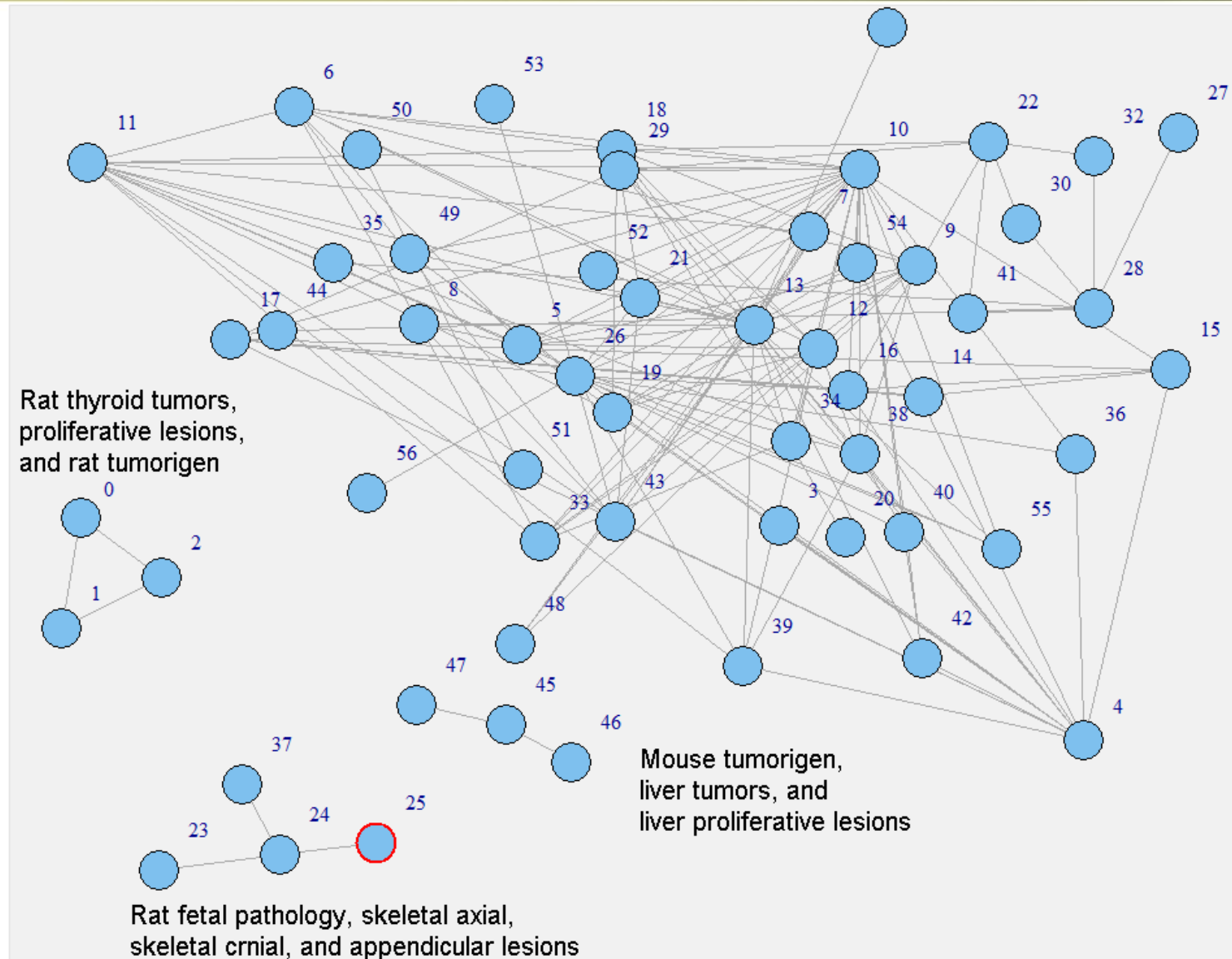
Histogram of degree.distribution(g)



Network Construction (Attagene)



Network Construction (BioSeek)



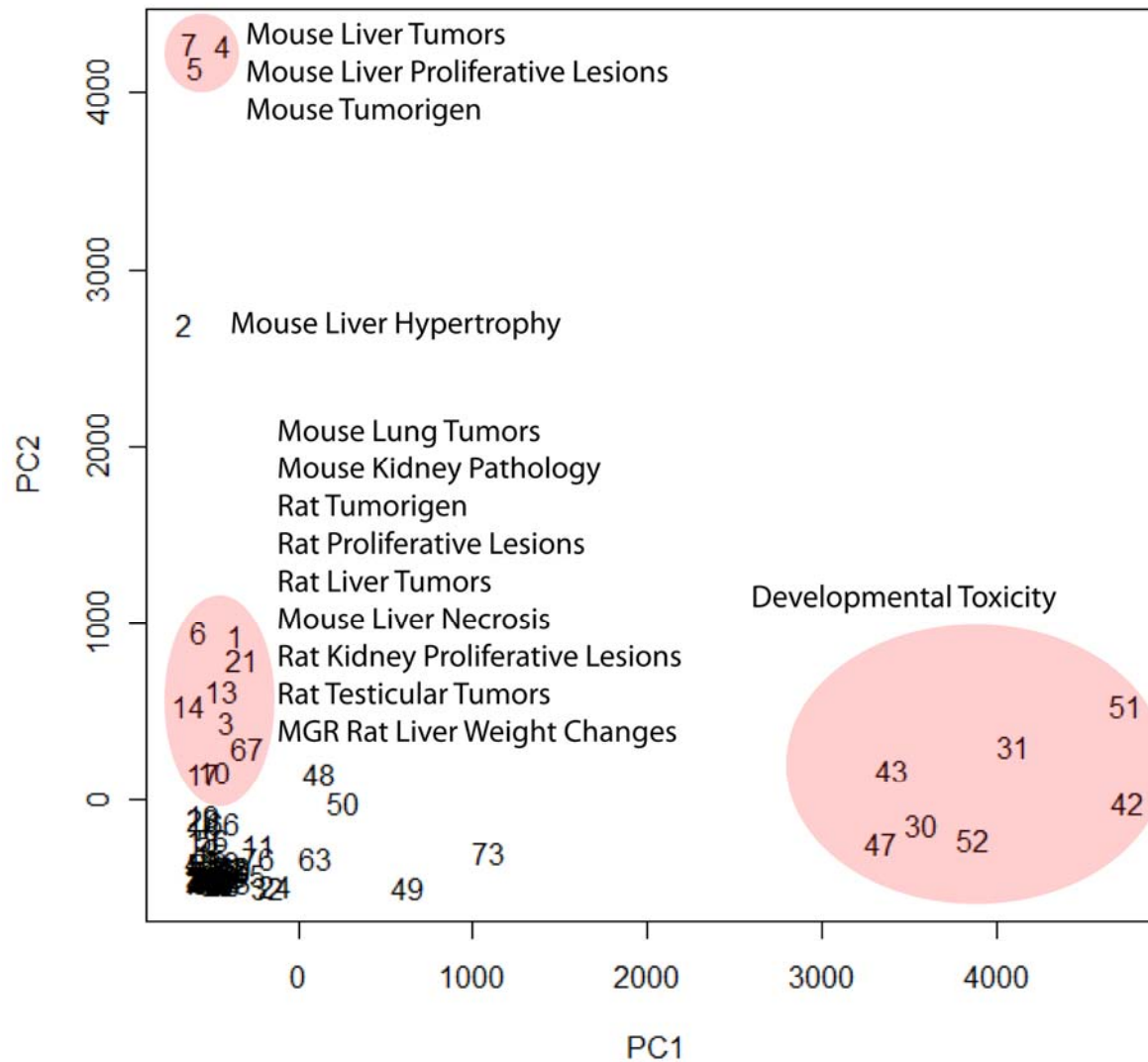


Application of Marketing Principles and Network Analysis to ToxCast

ANALYSIS OF CHEMICAL-CLASS SPECIFIC BIOMARKERS



Looking for Clustering in ToxRefDB...



Fused Mouse Liver Tumorigen Data



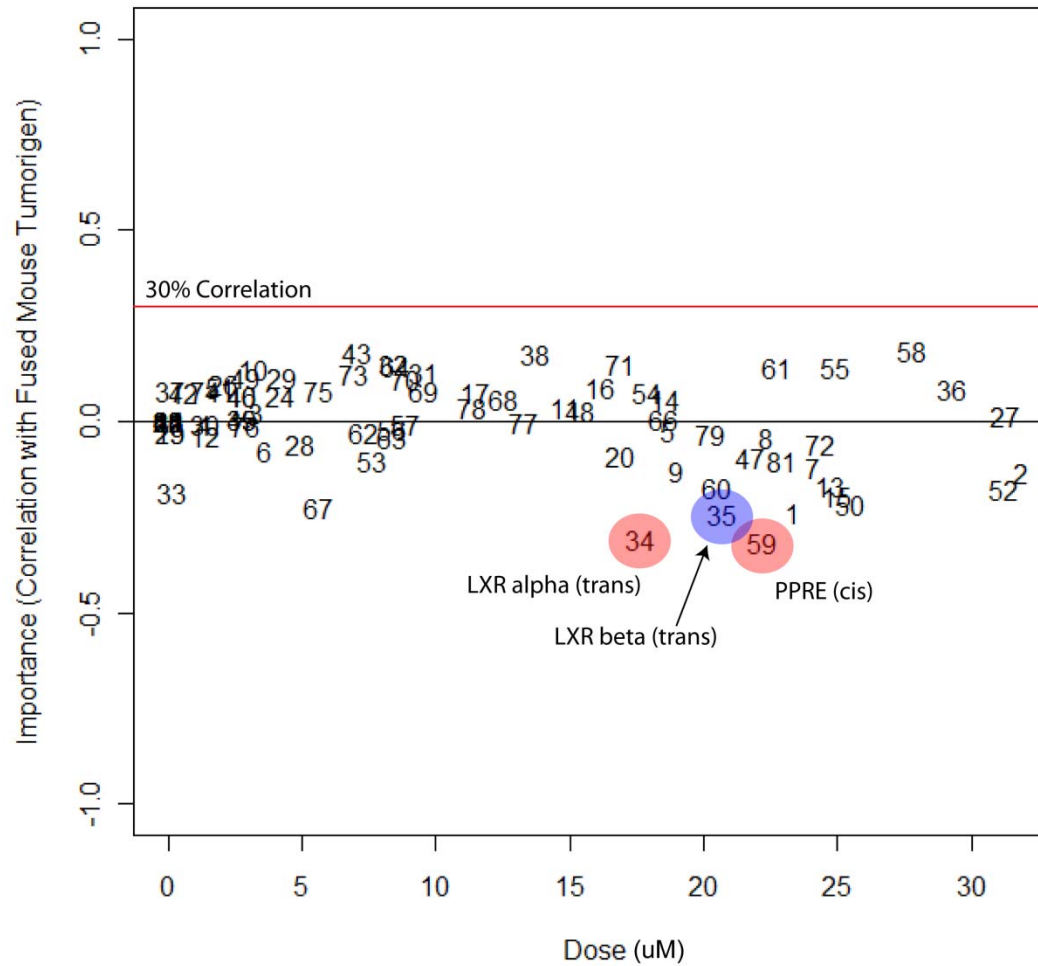
- ToxRefDB
 - If the mouse liver tumor data are “ACTIVE” for...
 - Mouse liver tumors
 - Mouse liver proliferative lesions
 - Mouse tumorigen
 - Then it is considered a mouse liver tumorigen
-

Key-Driver Analysis

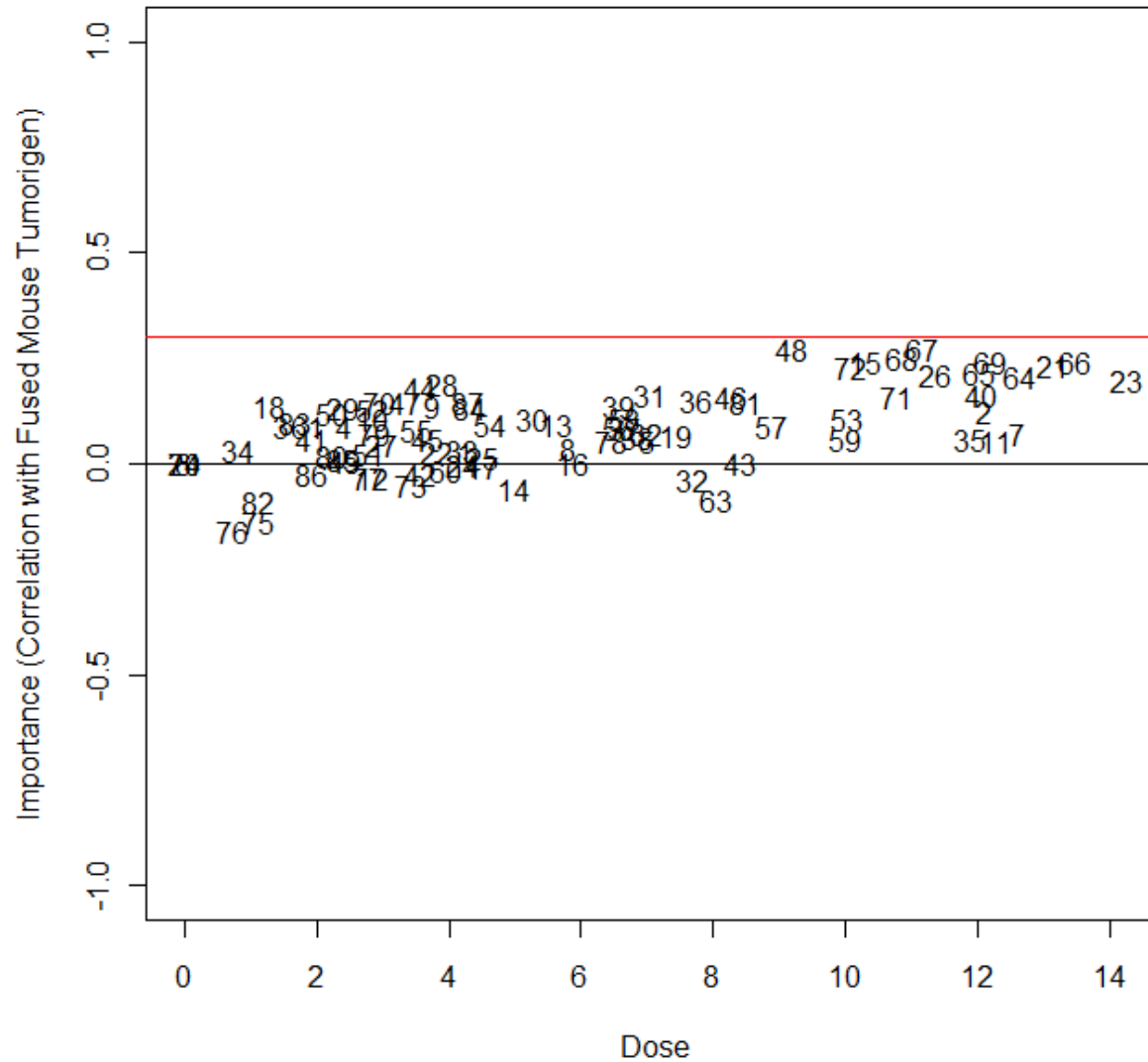


- AKA: Importance-Performance Analysis
 - Generally applied to customer-satisfaction surveys
 - Identify the “Importance” factor
 - Marketing: “Your Overall Satisfaction”
 - Toxicology: “Toxicity Phenotype Data”
 - Correlation of phenotype data to the “Importance” factor
 - Plot correlation (y-axis) and the
 - Marketing: mean or median satisfaction score
 - Toxicology: mean or median ED₅₀, or similar
-

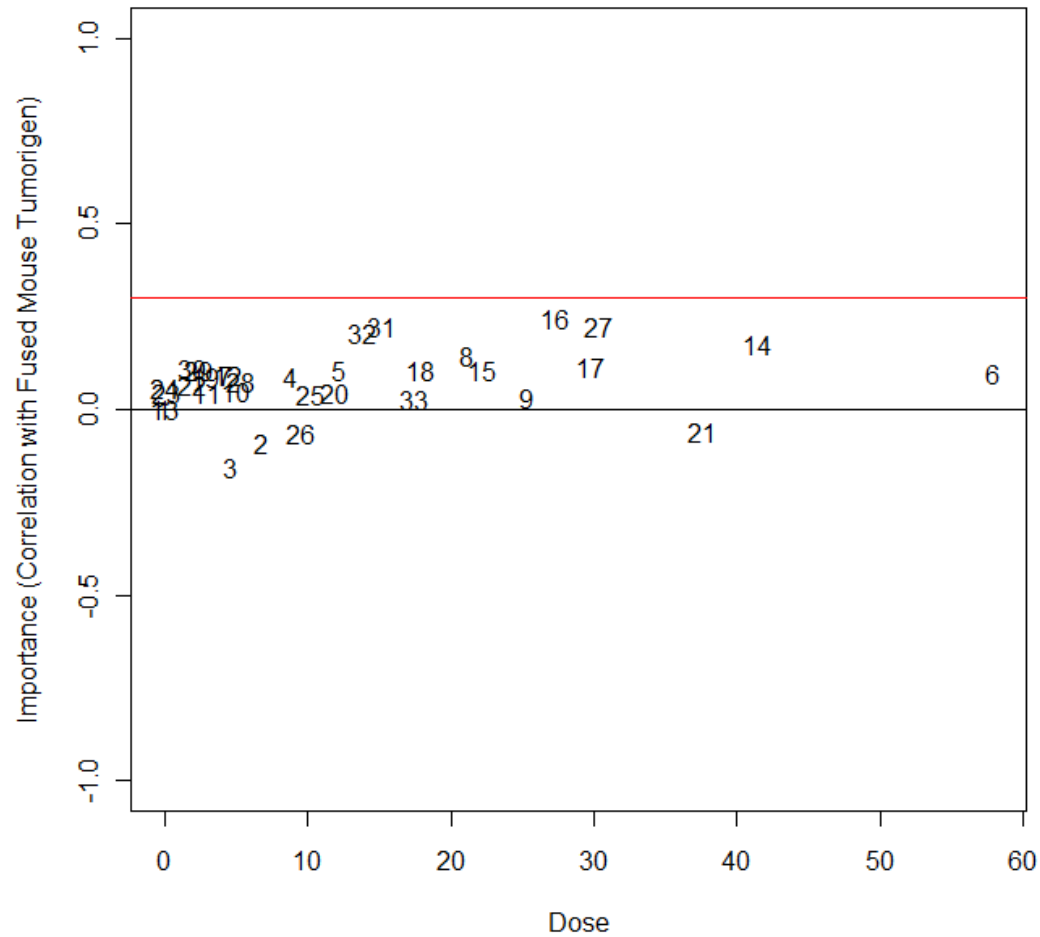
Key-Driver Analysis (Attagene)



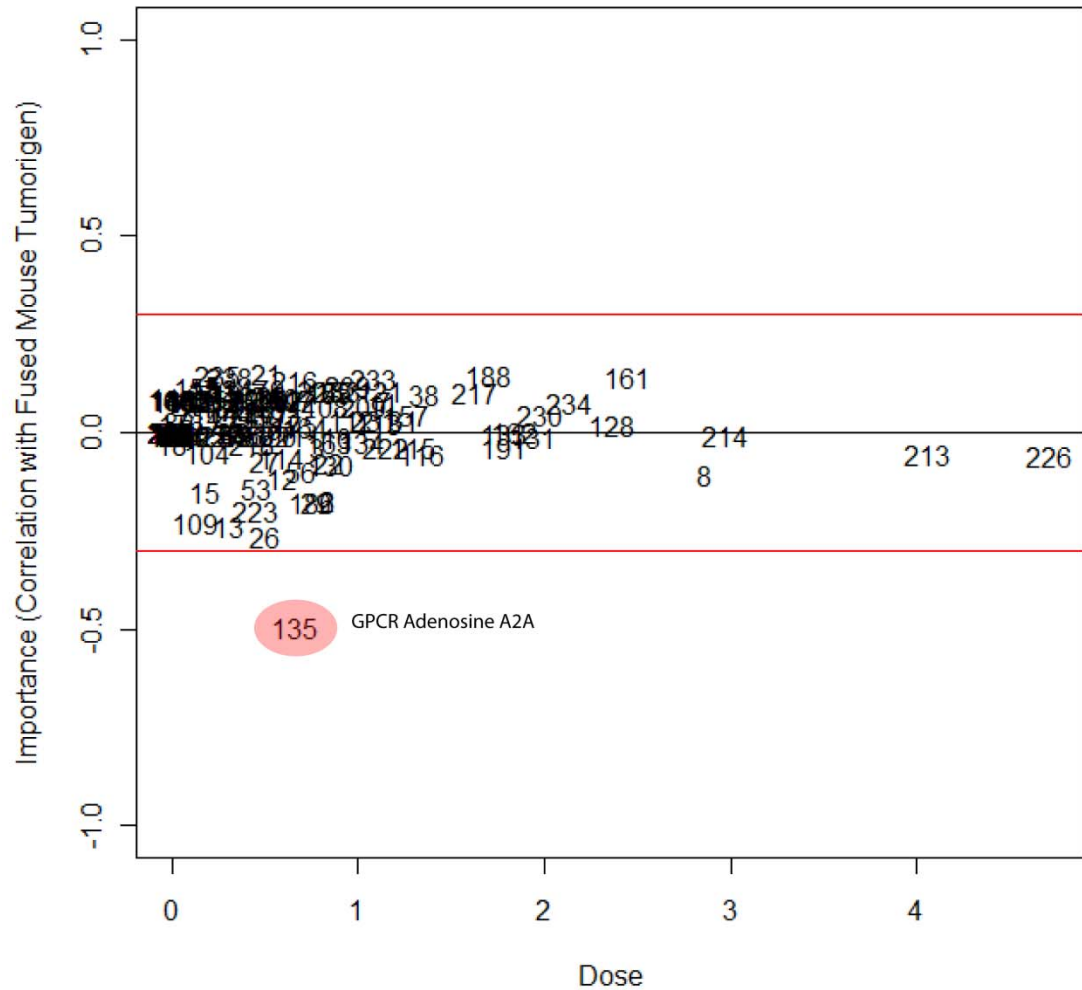
Key-Driver Analysis (BioSeek)



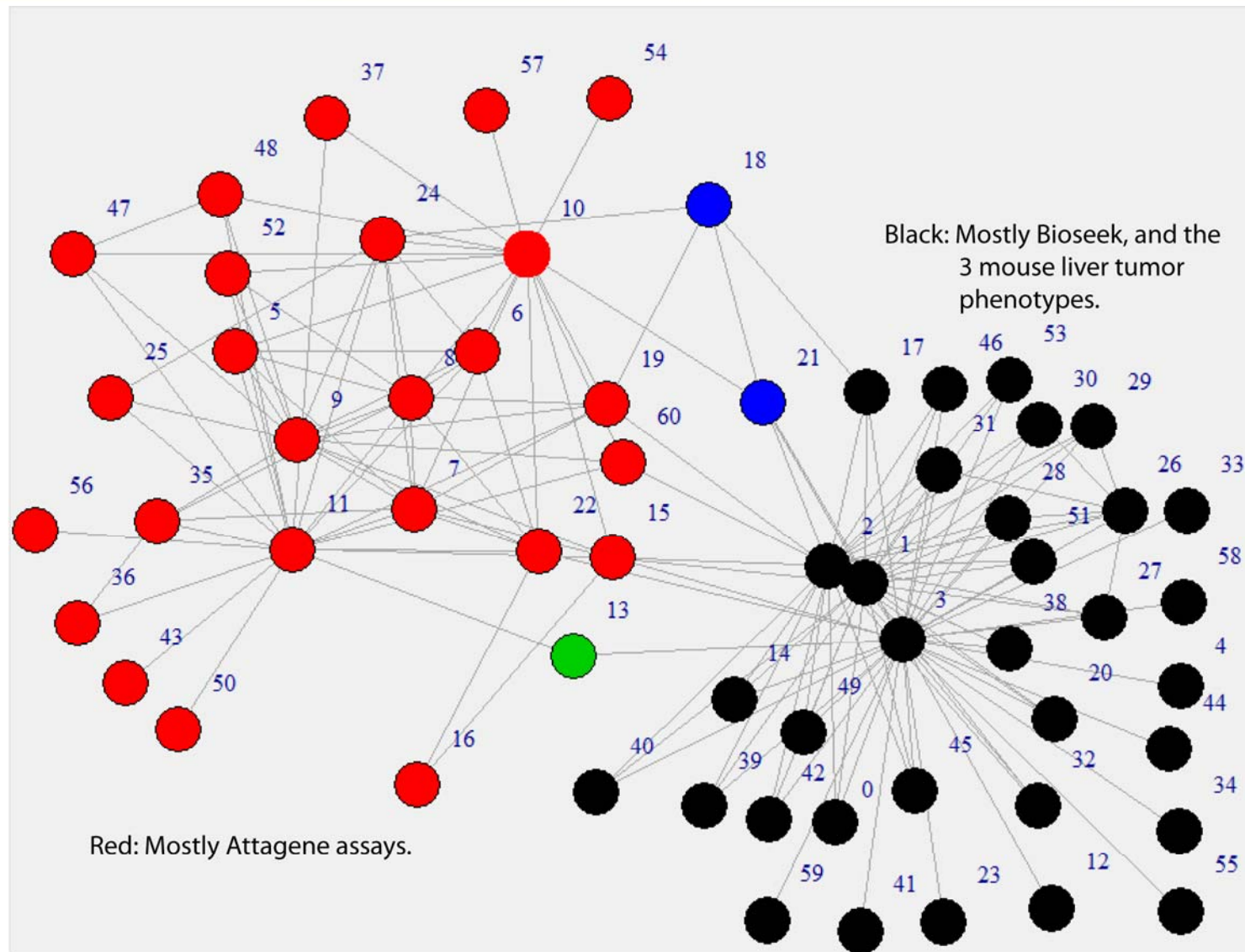
Key-Driver Analysis (Cellumen)



Key-Driver Analysis (NovaScreen)



Back to Network Traversal...



Future Directions



- Data fusion with Structure Activity Relationship (SAR) data
 - Apriori + network traversal may allow for data fusion and identification of relationships between HTS methods and SAR
 - Weighting the nodes with confidence from Apriori may create more informative network paths
 - May lead to more informative predictive biomarkers
-

Conclusions



- “Traditional” data mining approaches do not work “wholesale”
 - My analyses suggest dividing ToxRefDB data into chemical classes increases success
 - Apriori + network traversal may hold promise in identifying putative agglomerative biomarkers
-

Acknowledgements



- EPA NCCT
 - Michigan State University
 - Strategic Partnership Grant
 - Computational Biology and Bioinformatics Laboratory
 - Anna Merkoulouitch
-