Scalable Computational Methods for the Analysis of High-Throughput Biological Data

Mike Langston

Professor, Department of Electrical Engineering and Computer Science University of Tennessee

21July 2009

Outline of Talk

Biological Foundations

Correlation Coefficients

Graph Algorithms

High Performance Computations

Radiological Data Analysis

Ongoing Work

Foundations

Systems Biology

- **How do biological entities function in unison and at all levels of scale?**
  - **genes**
  - **proteins**
  - **cells**
  - **tissues**
  - **organisms**
- **Linkage, communication and networks (*graphs!*)**

Foundations

Omics:

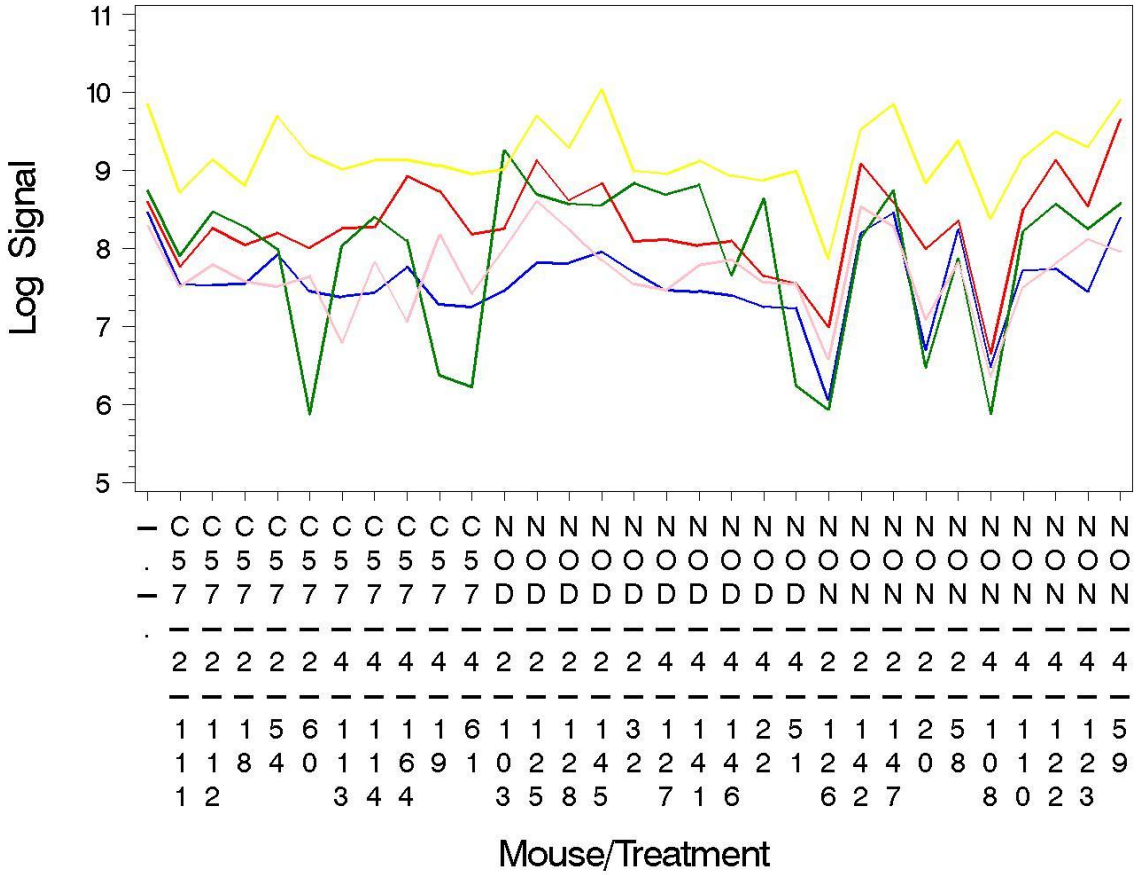
- **Key to deciphering complex systems**
- **Humans: 10^{13}+ cells, 200+ cell types**
- **Genome (blueprint, 20K+ genes, 10M+ polymorphisms)**
- **Proteome (functional units, unknown # of proteins)**
- **Transcriptome**
  - **Translation (tRNA) via transcription (mRNA)**
  - **Function and Signaling (siRNA, miRNA, etc)**
- **Other: metabalome, lipidome, interactome, ...**

Correlation

Correlation Coefficients

Here are five mouse Genes with Pearson correlations of at least 0.65. What of

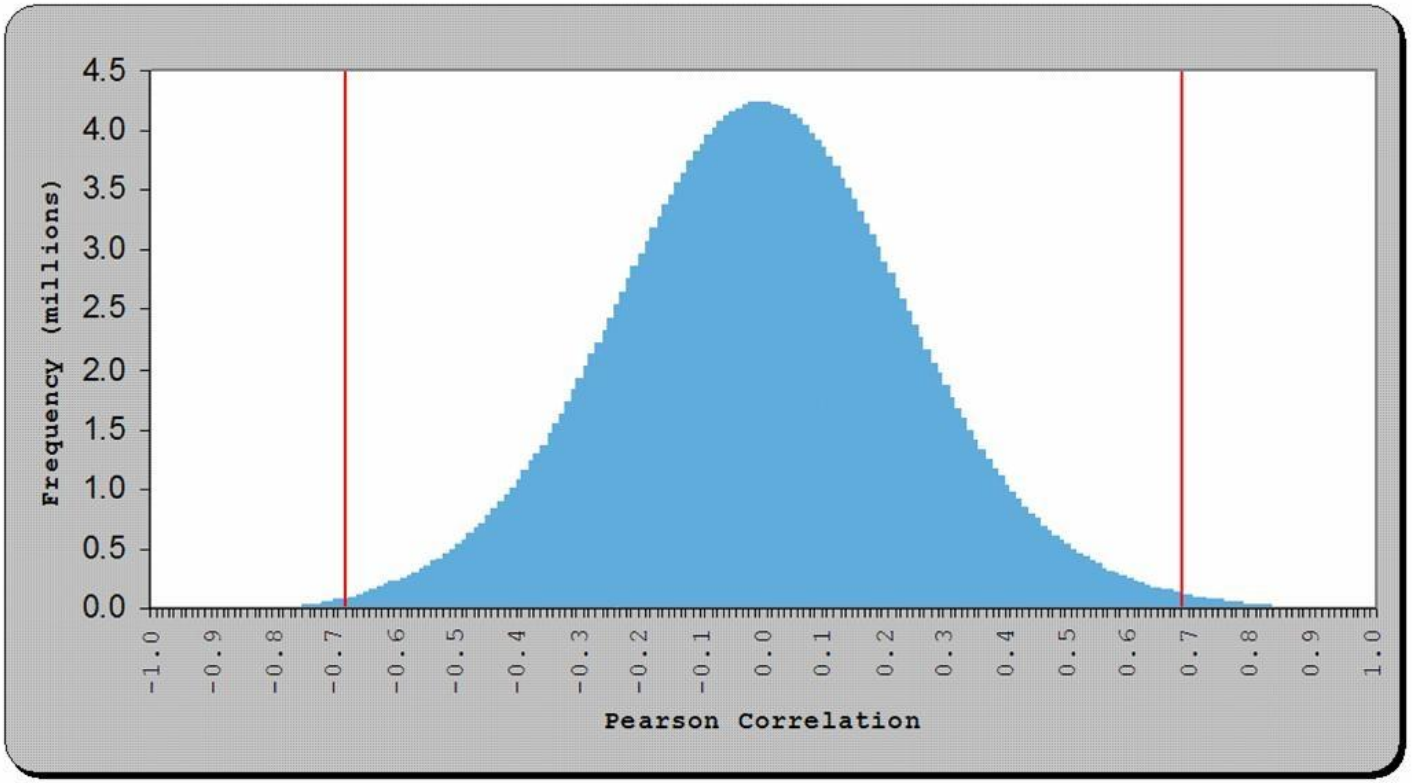
- **noise?**
- **experimental design?**
- **circadian rhythms?**
- **sex, tameness, etc?**
- **other confounds?**
- **other metrics?**



Correlation

Coefficient Profiles, High-Pass Filtering and Graphs

- Pearson, Spearman, Mutual Information, ...
- p-values, q-values, Bonferroni corrections, false discovery rates, ...

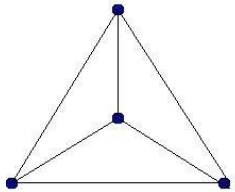


Dense Subgraph Extraction

Computational Tools

Maximum Clique

K_4



- must run often
- time is a limiting factor
- exploit fixed-parameter tractability (FPT)
  - *NP*-complete problem on n and k
  - solve in time $f(k)n^c$ not $n^{f(k)}$
  - Vertex Cover (dual) in time $O(1.2759^k k^{1.5}+kn)$

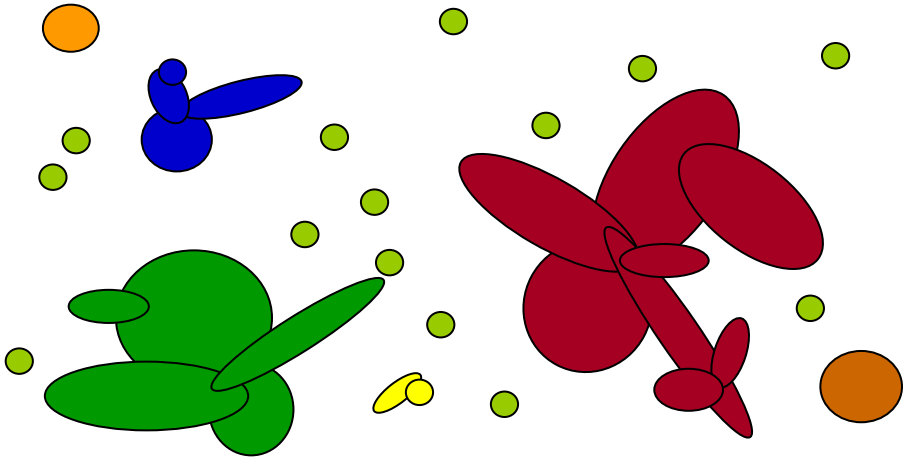
Dense Subgraph Extraction

Computational Tools

Maximum Clique

Maximal Clique

- huge outputs
- various orderings
- memory is often the limiting factor



Dense Subgraph Extraction

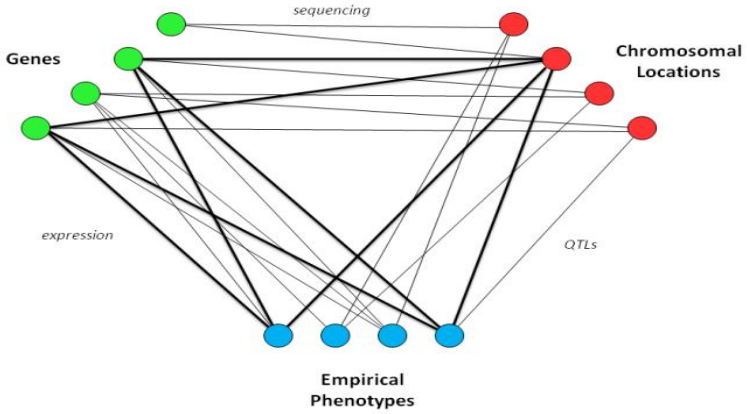
Computational Tools

Maximum Clique

Maximal Clique

Other Clique Variants

- *k*-clique
  - new algorithms
  - *k*-partite graphs
- Paraclique
  - form of soft thresholding for noisy data



A 2-2-1 three-clique

Supercomputer Implementations

Codes ported to platforms such as the SGI Altix (NERSC)

- **256 dual-CPU processors, each clocked at 1.5 GHz**
- **two terabytes of shared memory**

Supercomputer Implementations

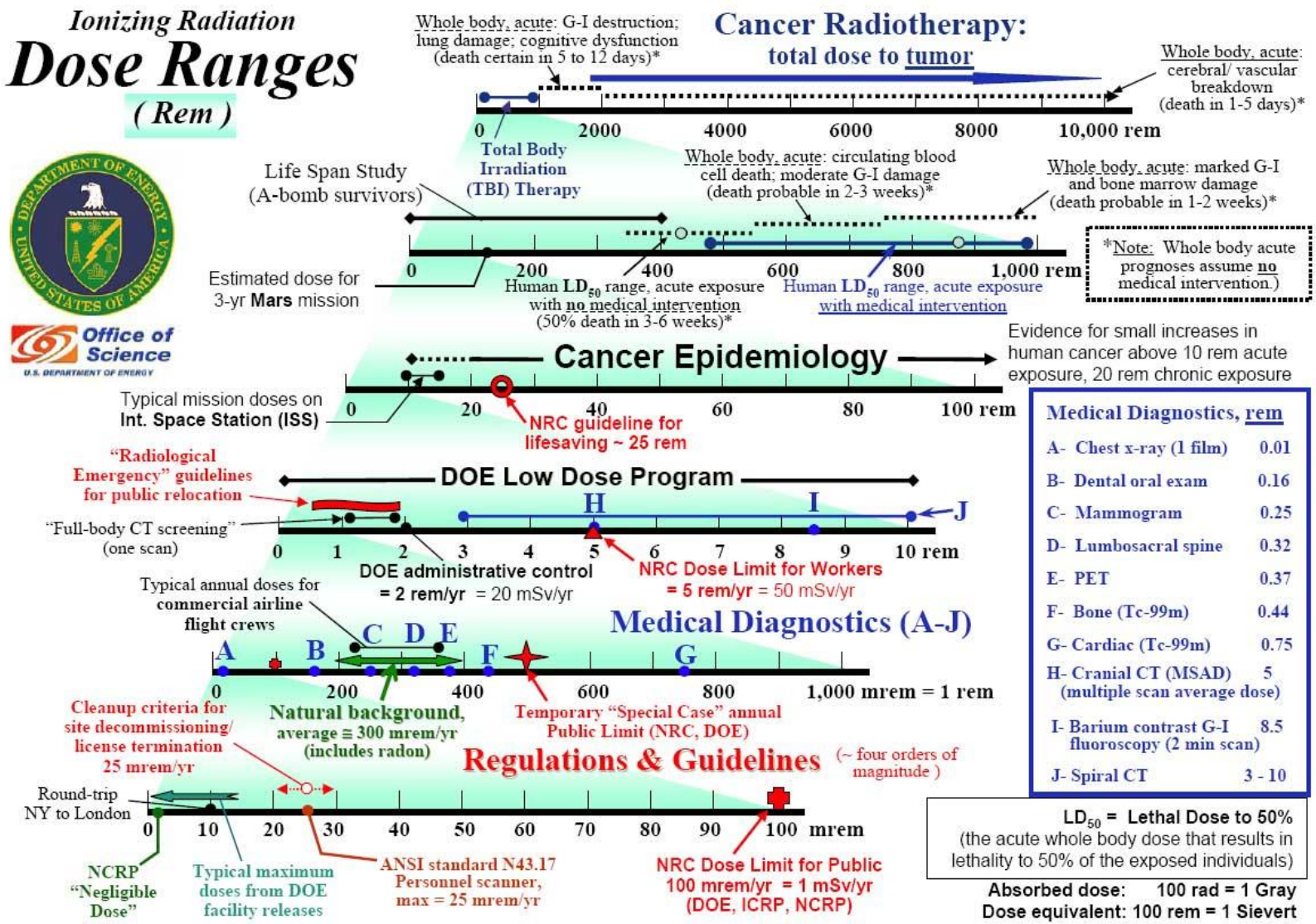
And Kraken, the new Cray XT5 system at ORNL

- currently the world's largest academic (non defense) computer
- 10^5 processor cores, nearly 10^{12} calculations/second (a petaflop)

Sample Application

Low dose ionizing radiation and its impact on human health

- Sources of low dose radiation exposures
  - medical diagnostics
  - hazardous waste abatement
  - handling materials for nuclear weapons and power systems
  - even terrorist acts such as dirty bombs
- In all these the major type of exposures will be low dose IR (primarily X- and gamma-radiation) from fission products
- Are low doses safe, perhaps even therapeutic?
- Identify biological pathways that are activated or repressed by IR
- Understand the risks so that we may protect the workforce



http://lowdose.tricity.wsu.edu

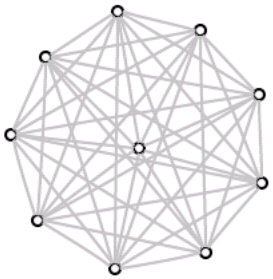
Sample Application

Case-control mouse data obtained at ORNL

- six strains, two durations
- acute dose of ten centigrays (rads)
- harvest skin, spleen, thymus, testis
- measure 21,547 transcripts

Sample Results

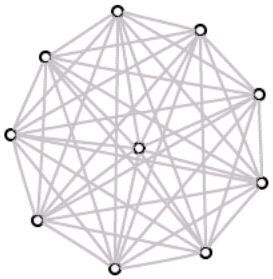
Example 1: a striking clique of size ten in dose



- populated with genes whose function is either associated with oxidative stress response or is unknown
- such an overexpressed ontology has an astronomically low p-value (unlikely to have occurred by mere chance)
- yet the same ten genes in control are only tenuously connected
- share a common transcription factor associated with IR?

Sample Results

Example 1: a striking clique of size ten in dose



- populated with genes whose function is either associated with oxidative stress response or is unknown
- such an overexpressed ontology has an astronomically low p-value (unlikely to have occurred by mere chance)
- yet the same ten genes in control are only tenuously connected
- share a common transcription factor associated with IR?

Example 2: gene for Carboxylesterase 2 (*Ces2*)

- catalyzes hydrolysis of xenobiotics
- differentially expressed with IR
- differentially correlated with many immune function genes
- present in 31% of dose cliques, but only 0.7% of control cliques
- plays some central role in low dose IR response?

Sample Results

Example 3: gene for Tubby-like Protein 4 (*Tulp4*)

- a nucleus of six genes are putatively coregulated in dose
- in fact they appear together in 5765 dose cliques
- yet no more than two occur together in any control clique
- this nucleus includes genes known to be involved in
    - immune function
    - stress mediation
    - and so these are consistent with IR response

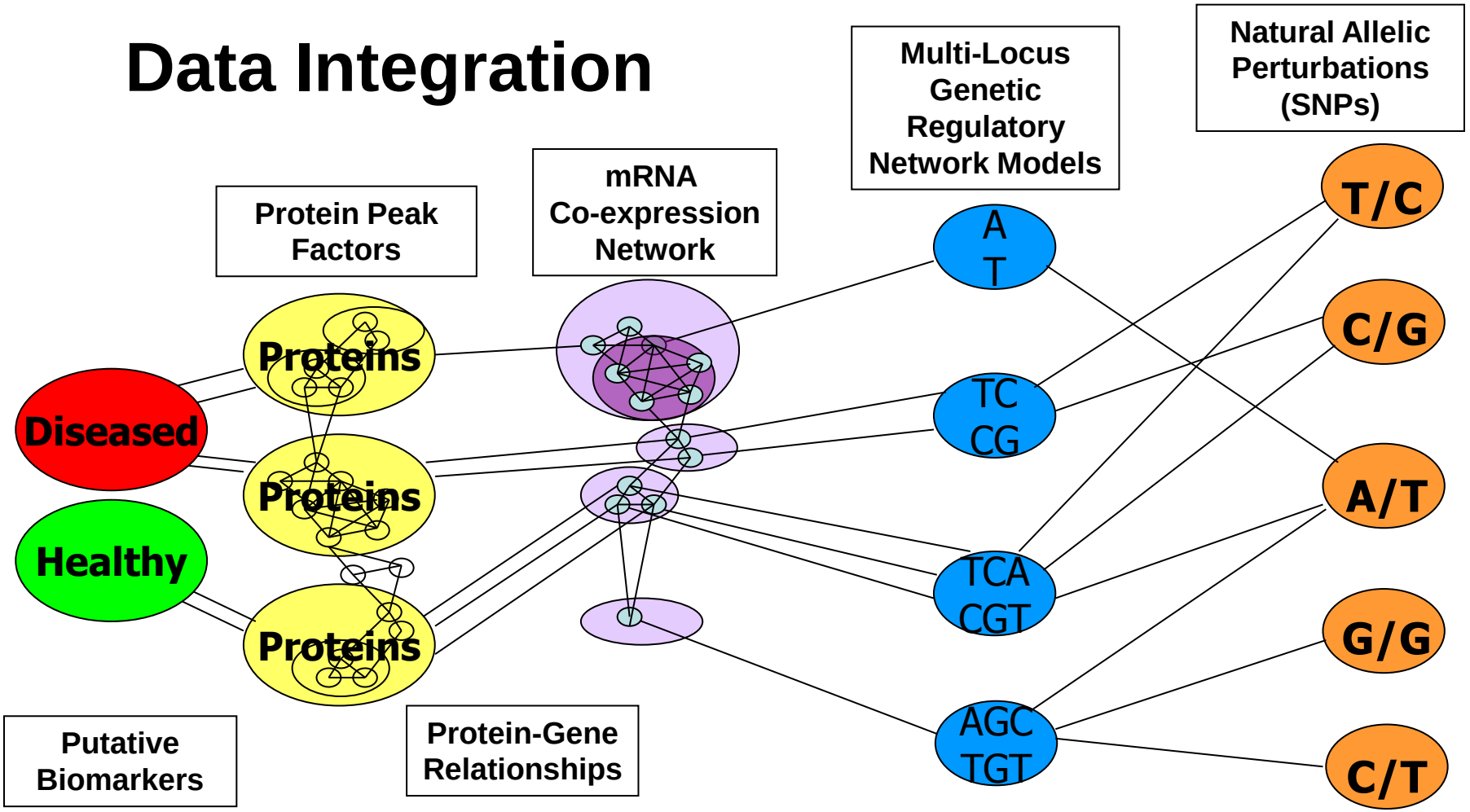
Sample Results

Example 3: gene for Tubby-like Protein 4 (*Tulp4*)

- a nucleus of six genes are putatively coregulated in dose
- in fact they appear together in 5765 dose cliques
- yet no more than two occur together in any control clique
- this nucleus includes genes known to be involved in
    - immune function
    - stress mediation
    - and so these are consistent with IR response
- but one of these is Tulp4...why is a tubby-like protein here?
- original classification
    - based on sequence similarity to Tub, an obesity protein
    - responsive to oxidative stress
- it's in 4.7% of the dose cliques and only 0.01% of control
- novel role for Tulp4 as a transcriptional regulator of immune response to IR?

Sample Ongoing Work

Data Integration



Collaborators

Research Scientists (Incomplete!):

- Elissa Chesler (ORNL)
- Arnold Saxton (UT)
- Brynn Voy (ORNL)

Current Students:

- Patricia Carey
- John Eblen
- Boloye Gomero
- Jeremy Jay
- Sudhir Naswa
- Clinton Nolan
- Charles Phillips
- Gary Rogers
- Dinesh Weerapurage

Recent PhD Graduates:

- Faisal Abu-Khzam
- Xinxia Peng
- Andy Perkins
- Yun Zhang

Recent MS Graduates:

- Nicole Baldwin
- Bhavesh Borate
- Jordan Lefebvre
- Zuopan Li
- Jon Scharff

Come Join Us!

The University of Tennessee and Oak Ridge National Laboratory

- Applications
  - Alcoholism, Allergy, Cancer, Diabetes, Neuroscience, Radiology...
- Combinatorial Analysis of Algorithms
- Complex Cross (2^{10} RI Mouse Lines)
- Database and Website Development
- High Performance Computing
- Metabolomics, Proteomics, Transcriptomics, ...
- RNAi and µRNA Interference
- Work Alongside Both Computer and Domain Scientists!

Collaboratory Science at the Langston Lab

Computer Science, Mathematics, Molecular Biology, Statistics