Computational Advances in High-Throughput Biological Data Analysis

Mike Langston
Professor
Department of Electrical Engineering and Computer Science
University of Tennessee
USA
7 March 2011
Outline of Talk

Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning
Outline of Talk

Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning
Clustering
A Classic Application

Toolchain

- cDNA or mRNA Microarrays
- normalization
- Correlation Computation
- Real-Valued Matrix
- Graph Transforms
- Unweighted Incomplete Graph
- Clique-Centric Methods
- Unsupervised Methods
- k-Connected Components
- HCS Subgraphs
- Clique-Centric Methods
- k-Cores
- High-Pass Filtering
- Thresholding

Raw Data
Gene Expression Profiles

Increasing Edge Density (and Increasing Problem Complexity)

- Principal Component Analysis
- k-Means Clustering

FPT VC Codes
HPC & Novel Methods

NP-complete Problems
## Clustering

### Algorithms Ranked by Quartile Comparisons

<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>Average Quartile</th>
<th>Small (3-10 genes)</th>
<th>Medium (11-100 genes)</th>
<th>Large (101-1000 genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Quartile</td>
<td>BAT5 Jaccard</td>
<td>Quartile</td>
</tr>
<tr>
<td>K-Clique Communities</td>
<td>1.00</td>
<td>1</td>
<td>0.7531</td>
<td>1</td>
</tr>
<tr>
<td>Maximal Clique</td>
<td>1.00</td>
<td>1</td>
<td>0.8433</td>
<td>1</td>
</tr>
<tr>
<td>Paraclique</td>
<td>1.00</td>
<td>1</td>
<td>0.7576</td>
<td>1</td>
</tr>
<tr>
<td>Ward (H)</td>
<td>1.33</td>
<td>2</td>
<td>0.5782</td>
<td>1</td>
</tr>
<tr>
<td>CAST</td>
<td>1.67</td>
<td>1</td>
<td>0.7455</td>
<td>3</td>
</tr>
<tr>
<td>QT Clust</td>
<td>2.00</td>
<td>2</td>
<td>0.5473</td>
<td>2</td>
</tr>
<tr>
<td>Complete (H)</td>
<td>2.33</td>
<td>3</td>
<td>0.3933</td>
<td>2</td>
</tr>
<tr>
<td>NNN</td>
<td>2.67</td>
<td>2</td>
<td>0.5521</td>
<td>2</td>
</tr>
<tr>
<td>K-Means</td>
<td>3.00</td>
<td>4</td>
<td>0.2573</td>
<td>3</td>
</tr>
<tr>
<td>SOM</td>
<td>3.00</td>
<td>4</td>
<td>0.3260</td>
<td>2</td>
</tr>
<tr>
<td>WGCNA</td>
<td>3.00</td>
<td>3</td>
<td>0.4391</td>
<td>3</td>
</tr>
<tr>
<td>Average (H)</td>
<td>3.33</td>
<td>3</td>
<td>0.4087</td>
<td>4</td>
</tr>
<tr>
<td>McQuitty (H)</td>
<td>3.33</td>
<td>3</td>
<td>0.4594</td>
<td>3</td>
</tr>
<tr>
<td>SAMBA</td>
<td>3.50</td>
<td>0.0000</td>
<td>4</td>
<td>0.1860</td>
</tr>
<tr>
<td>CLICK</td>
<td>4.00</td>
<td>4</td>
<td>0.0339</td>
<td>4</td>
</tr>
</tbody>
</table>
Coexpression Analysis

There's a high probability that somewhere in here is a polymorphism controlling this trait.

Transcript abundance can be the phenotype!
Coexpression Analysis

Two Paracliques

Concentrated Parental Alleles
Thresholding
## Thresholding

<table>
<thead>
<tr>
<th>Method</th>
<th>Anoxia</th>
<th>Reoxygenation</th>
<th>Alpha</th>
<th>Absolute deviations from GO threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Functional Similarity</td>
<td>0.97</td>
<td>0.92</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Spectral Clustering</td>
<td>0.93</td>
<td>0.97</td>
<td>0.89</td>
<td>0.04+0.05+0.04=0.13</td>
</tr>
<tr>
<td>Maximal Clique-2</td>
<td>0.90</td>
<td>0.91</td>
<td>0.74</td>
<td>0.07+0.01+0.11=0.19</td>
</tr>
<tr>
<td>Power</td>
<td>0.88</td>
<td>0.94</td>
<td>0.96</td>
<td>0.09+0.02+0.11=0.22</td>
</tr>
<tr>
<td>Bonferroni adjustment</td>
<td>0.85</td>
<td>0.93</td>
<td>0.95</td>
<td>0.12+0.01+0.10=0.23</td>
</tr>
<tr>
<td>Control-Spot</td>
<td>0.93</td>
<td>0.83</td>
<td>0.70</td>
<td>0.04+0.09+0.15=0.28</td>
</tr>
<tr>
<td>Maximal Clique-3</td>
<td>0.87</td>
<td>0.89</td>
<td>0.60</td>
<td>0.10+0.03+0.25=0.38</td>
</tr>
<tr>
<td>Top 1 Percent</td>
<td>0.81</td>
<td>0.81</td>
<td>0.72</td>
<td>0.16+0.11+0.13=0.40</td>
</tr>
</tbody>
</table>

Estimated threshold for each dataset, sorted by performance of the methods. GO functional similarity thresholds are the standard against which the methods are compared, summing absolute deviations across datasets (thresholds above GO are in bold).
Pioneering approach going back twenty-five years
- Well-Quasi-Order theory
- nonuniform measure of complexity

Exploit knowledge of the solution space
- Consider an algorithm with a time bound such as $O(2^{kn})$.
- And now one with a time bound more like $O(2^kn)$.
- Both are exponential in parameter value(s).
- But what happens when $k$ is fixed?
- Fixed-Parameter Tractable (FPT) iff $O(f(k)n^c)$
- Confines superpolynomial behavior to the parameter

Duality
- We solve **vertex cover**, clique’s complementary dual
- $O(1.2738^k k^{1.5} + kn)$ time

Key features
- Kernelization, branching and interleaving
Outline of Talk

Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning
A Clique Compute Engine

Input Graph → Parametric Tuning, Decomposition and Refinement → Highly Parallel Computation → Cliques for Post-Processing

Preprocessing and Kernelization

Distilled Genesets, Models and Testable Hypotheses

Prioritized by GO, CREs, pathways, literature, etc

Transcriptomic Context

Branching and Interleaving

Works well with synthetic data. But with real data, dynamic workload balancing is required. And that can be very tricky!

GrAPPA, NERSC and the TeraGrid
Now also using new ORNL-UT Cray XT5 system, Kraken
  • currently the world’s largest academic (non defense) computer
  • $10^5$ processor cores (and expanding)
  • nearly $10^{12}$ calculations per second (a petaflop)
  • quite a beast to harness, at least for combinatorial work
Workload Balancing and Speedup

- optimum (linear) speedup
- dynamic load balancing (estra-30)
- dynamic load balancing (folic-30)
- dynamic load balancing (avg)
Differential Analysis

Gene (vertex) comparisons:

- differential expression
- does not require multiple conditions
- compare the two lists of gene expression levels
Correlate (edge) comparisons

- differential correlation
- requires multiple conditions in control versus stimulus
- compare two lists of gene-gene correlations
Putative network (clique) comparisons

- differential topology
- compare dense subgraphs, sort by ontology, CREs, etc
- consider granularity, for example, with the clique intersection graph
Outline of Talk

Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning
Data Description

- Mikael Benson, Göteborg, Sweden, 56 patients and 39 controls
- Affymetrix HU133 arrays
- roughly 33,000 genes
- nasal secretions, lymphocytes, skin
- hay fever, eczema

Preprocessing

- MAS5.0
- log transformed
- replicates averaged
- centered around zero with z scores
- probesets with consistently low expression levels removed

Threshold Selection

- chosen to balance graph densities
- AFFX spots retained for quality control
Clique profiles using the five most highly represented genes:

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Clique membership</th>
<th>Gene Symbol</th>
<th>Clique membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBE1C</td>
<td>29%</td>
<td>FGFR2</td>
<td>66%</td>
</tr>
<tr>
<td>RANBP6</td>
<td>27%</td>
<td>NFIB</td>
<td>65%</td>
</tr>
<tr>
<td>DKFZP564O123</td>
<td>26%</td>
<td>PPL</td>
<td>64%</td>
</tr>
<tr>
<td>SLC25A13</td>
<td>24%</td>
<td>FGFR3</td>
<td>64%</td>
</tr>
<tr>
<td>GTPBP4</td>
<td>21%</td>
<td>CDH3</td>
<td>56%</td>
</tr>
</tbody>
</table>

Applied differential screens, then ChIP-chip technologies, etc.

**Sample Result:** Discovered a novel and key role for **ITK** (IL2-inducible T-cell kinase)
Data Inhomogeniety

- huge problem without model organisms
- no recombinant inbred human populations
- tumors and other diseases are often not uniform
- Pablo Moscato, Newcastle, Australia, prostate cancer data

Creative Use of Graph Algorithms

- perform multiple data views
- drive correlations with both persons and genes
- exclude outliers with clique-centric tools
- perform differential analysis to distill biomarkers from genome
Application, Cancer

Genes Drive Person-Person Correlations

Select Thresholds, Extract Cliques

Classify Subtypes and Eliminate Outliers

Sample Result: Putative Prostate Cancer Biomarkers

- KLK3 = PSA
- ETS1
- MAZR
- KROX
- NFKB

Perform Assorted Forms of Differential Analysis to Identify Network Differences

Persons Drive Gene-Gene Correlations

Classify Subtypes and Eliminate Outliers

Select Thresholds, Extract Cliques

Perform Assorted Forms of Differential Analysis to Identify Network Differences
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
Sample Result: 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 adipose tissue 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?
Outline of Talk

Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning
Algorithmic Training

- Raw Data
  - Gene Scoring
    - Eliminate Poorly Discriminating Genes
  - Dominating Set
    - Eliminate Poorly Covering Genes
  - Calculate Sample Similarities
    - Apply Threshold
      - Verify by Classification
        - Set of Discriminatory Genes
        - Gene Scores
Gene Scoring

\[
score(gene_i) = |m_{classA} - m_{classB}| - |\sigma_{classA} + \sigma_{classB}|
\]

Followed by edge weighting.
Edge Weight Spectrum
Allergic Rhinitis: Top 100 Genes

Number of Sample Pairs (Edges)

Weight

-0.029 -0.028 -0.027 -0.026 -0.025 -0.024 -0.023 -0.022 -0.021 -0.02 -0.019 -0.018 -0.017 -0.016 -0.015 -0.014 -0.013 -0.012 -0.011 -0.01 -0.009 -0.008 -0.007 -0.006 -0.005 -0.004 -0.003 -0.002 -0.001 0

-0 10 20 30 40 50 60 70 80 90 100

Patient-Patient
Patient-Control
Control-Control
Allergic Rhinitis: Top 100 μRNAs

Number of Sample Pairs (Edges) vs Weight

- Patient-Patient
- Patient-Control
- Control-Control
Allergic Rhinitis: Top 100 Methylation Sites

- Patient-Control
- Patient-Patient
- Control-Control

Number of Sample Pairs (Edges)

Weight

0.2772 0.2777 0.2782 0.2787 0.2792 0.2797 0.2802 0.2807 0.2812 0.2817 0.2822 0.2827 0.2832 0.2837 0.2842 0.2847 0.2852
Where We Are, Where We're Going

Raw Data

Dense Subgraphs

Expanded Graphs

Directed Graphs

Verified Pathways

- Full and Partial Correlation, Thresholding, Power of Abstraction, Graph Theory, HPC, Spectral Methods, Hermert Analysis
- Graph Expansion, Text Mining, Paraclique, Neighborhoods, Anchored Subgraphs, GO, PPI, String, Ingenuity, Cytoscape
- Bayesian Methods, KEGG, QTLs, Structural Equation Modeling
- Knock Outs, Knock Downs, RNAi, μRNA
The Langston Lab

Computer Science, Mathematics, Molecular Biology, Statistics