Analyzing and Modeling Time-series Gene Expression Data with STEM and DREM

Jason Ernst
Assistant Professor
University of California, Los Angeles
Time-series Gene Expression Data

- Biological processes occur over time (e.g. immune response, development)
- Frequently studied with time-series of gene expression experiments
- Many methods have focused on modeling and analyzing gene expression data without taking into account time even in cases when applied to time series data

Segal et al., Nature Genetics 2003
Bar-Joseph et al., Nature Biotech 2003
Workman et al., Science 2006
WGCNA (Image source http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/)
Overview

- **Short Time-series Expression Miner (STEM)**

- **Dynamic Regulatory Events Miner (DREM)**
Overview

- Short Time-series Expression Miner (STEM)
- Dynamic Regulatory Events Miner (DREM)
Importance of Clustering Short Time-series Data

- Most time series gene expression data sets are short (3-8 time points)

- Genes that have similar expression patterns are often involved in the same biological process or are co-regulated

Ernst et al. *Bioinformatics*, 2005

Ernst and Bar-Joseph *BMC Bioinformatics*, 2006
Limitations of Standard Clustering Methods for Time Series Data

- Having few time points can pose a challenge for traditional time series models (e.g. autoregressive equations)
- Commonly used methods such as k-means and hierarchical clustering do not use the temporal ordering of experiments
- Thousands of genes and few time points many patterns by random chance
  - Standard clustering methods do not differentiate between real and random patterns

Clusters from K-means on simulated noise (all values drawn independently from the identical distribution)

Ernst et al. ISMB/Bioinformatics, 2005
Method Overview

Approach: Determine temporal patterns with significantly more genes than expected compared to a random ordering of time

STEP 1: Define temporal profiles independent of data
STEP 2: Assign genes to most closely matching profile
STEP 3: Compute expected number of genes per profile
STEP 4: Determine statistically significant profiles
STEP 1: Define a set of distinct and representative model temporal profiles independent of the data.

Ernst et al. ISMB/Bioinformatics, 2005
Start with all temporal profile shapes with at most $c$ unit change between time points (here $c=2$)
Greedily select a set of $k$ profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1 - \text{correlation coefficient})}$.
Greedily select a set of $k$ profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{1-\text{correlation coefficient}}$. 
Greedily select a set of $k$ profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{1 - \text{correlation coefficient}}$. 
Greedily select a set of $k$ profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1 - \text{correlation coefficient})}$.
Greedily select a set of $k$ profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1 - \text{correlation coefficient})}$.
Greedily select a set of $k$ profiles maximizing the minimum distance between any two selected profiles. Here we use the distance between profiles as $\sqrt{(1 - \text{correlation coefficient})}$.
STEP 2: Filter flat genes, and then assign the remaining genes to the most closely matching model profile based on the correlation coefficient.

STEP 1. Define temporal profiles independent of data
STEP 2. Assign genes to most closely matching profile
**STEP 2:** Filter flat genes, and then assign the remaining genes to the most closely matching model profile based on the correlation coefficient.

Data for immune response to a pathogen infection (Guillmen et al PNAS, 2002)

Ernst et al. ISMB/Bioinformatics, 2005

150 genes assigned to profile 43
STEP 3: Compute the expected number of genes assigned to a profile based on a permutation test on the time points.

28 genes expected to be assigned to profile 43

Ernst et al. ISMB/Bioinformatics, 2005
STEP 4: Using the binomial distribution and the counts from steps 2 and 3 associate statistical significance with the number of genes assigned to each profile.

Profiles ordered based on significance; Colored profiles are significant at a 0.05 bonferroni corrected level.

Ernst et al. ISMB/Bioinformatics, 2005
• Demonstration of the Short Time-series Expression Miner
  (http://www.sb.cs.cmu.edu/stem)
Overview

• Short Time-series Expression Miner (STEM)

• Dynamic Regulatory Events Miner (DREM)
Towards a Global View of Gene Regulation Dynamics

- Explicitly model as bifurcation events when sets of genes share an expression pattern and then diverge.

Integrate complementary static transcription factor (TF)-gene interaction information to gain mechanistic insights into TFs driving expression changes.

Static transcription factor-gene interaction data (experimental or computational predictions)

<table>
<thead>
<tr>
<th>Gene</th>
<th>TF A</th>
<th>TF B</th>
<th>TF C</th>
<th>TF D</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GeneB</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GeneC</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GeneD</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Time series gene expression data

<table>
<thead>
<tr>
<th>Gene</th>
<th>0h</th>
<th>1h</th>
<th>2h</th>
<th>4h</th>
<th>6h</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneA</td>
<td>0</td>
<td>1.9</td>
<td>3.2</td>
<td>1.7</td>
<td>1.5</td>
</tr>
<tr>
<td>GeneB</td>
<td>0</td>
<td>-2.7</td>
<td>-1.6</td>
<td>-0.4</td>
<td>-0.2</td>
</tr>
<tr>
<td>GeneC</td>
<td>0</td>
<td>0.3</td>
<td>-0.2</td>
<td>0.4</td>
<td>1.2</td>
</tr>
<tr>
<td>GeneD</td>
<td>0</td>
<td>-0.5</td>
<td>-0.3</td>
<td>-1.1</td>
<td>-2.1</td>
</tr>
</tbody>
</table>

Same Genes
Gene Expression for Heat Shock Response in Yeast

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Dynamic Regulatory Map of Heat Shock Response in Yeast

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Dynamic Regulatory Map of Heat Shock Response in Yeast

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Dynamic Regulatory Map of Heat Shock Response in Yeast
Dynamic Regulatory Map of Heat Shock Response in Yeast

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Dynamic Regulatory Map of Heat Shock Response in Yeast

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Bifurcation Event in the Time Series Gene Expression Data

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Genes Regulated by MBP1 based on Static Input More Likely to be on Higher Path

Transcription factor predictions based on enrichment determined by the hypergeometric distribution

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Bud Counting Experiment Confirms Predictions about Split

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
- Expression data generated from Gaussian distributions
- Transition probabilities between states that can be modeled to depend on the set of regulating TFs for a gene

Expression data generated from Gaussian distributions

Transition probabilities are gene specific and depend on the transcription factors regulating the genes

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
- Expression data generated from Gaussian distributions
- Transition probabilities are gene specific and depend on the transcription factors regulating the genes

Logistic regression classifiers map the set of transcription factors interacting with a gene to transition probabilities

$$x = (x_1, ..., x_n)$$

$$\frac{1}{1 + e^{-\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n}}$$

Specific instance of an Input-Output Hidden Markov Model (Bengio and Frasconi, NIPS 1995)

Ernst et al, Nature-EMBO Molecular Systems Biology 2007
Computational Issues in Reconstructing a Map

- For a given tree structure find the “best” setting of the parameters consistent with the data

\[
\sum_{g \in G} \sum_{q \in Q} \prod_{t=1}^{n-1} f_{q(t)}(o_{g}(t)) \prod_{t=1}^{n-1} P(H_t = q(t)|H_{t-1} = q(t-1), I_g)
\]

- Sum over all genes
- Sum over all paths Q
- Product over all Gaussian emission density values on path
- Product over all transition probabilities on path for a gene determined by logistic regression classifiers

- Exponentially many possible tree structures – hill climbing search over tree structures
- Score models based on consistency with data and its complexity
Amino Acid Starvation Response in Yeast

INO4 not previously known to be involved in the response and only profiled in normal growth conditions

Ernst et al, Nature-EMBO Molecular Systems Biology 2007

Experimental validation by Itamar Simon’s lab
Development in Fly

Dynamic Regulatory Events Miner (DREM) is being used to model a variety of dynamic regulatory responses

Software available at www.sb.cs.cmu.edu/drem

Various Stress Responses in Yeast

Development in Fly

Aerobic-anaerobic shift *E. coli*

With Itamar Simon's lab

Ernst et al, *PloS Computational Biology* 2008
With Zoltan Oltvai's lab

The modENCODE consortium, Roy*, Ernst*, et al, Science 2010

Also many applications of DREM by others
• Demonstration of the Dynamic Regulatory Events Miner
  (http://www.sb.cs.cmu.edu/drem)
Acknowledgements

• Ziv Bar-Joseph
• Oded Vainas
• Itamar Simon
• Zoltan Oltvai
• Qasim Beg
• Continued development of DREM: Marcel Schultz, William E Devanny, Anthony Gitter, Shan Zhong, Aaron Wise...

Funding: NIH, NSF, Siebel Scholarship
Questions