Computational Genomics

Reconstructing dynamic regulatory networks in multiple species
Methods for reconstructing networks in cells

Amit et al, Science 2009


Gerstein et al, Science 2010
Key problem: Most high-throughput data is static

Time-series measurements

Static data sources

DNA

motif

CHIP-chip

microarray

PPI
Method: Integrating time series expression and static protein-DNA interaction data
### a Time Series Expression Data

- **Expression Level**
- **Time**

### b Static TF-DNA Binding Data

- **TF A**
- **TF B**
- **TF C**
- **TF D**

### c Model Structure

- **Expression Level**
- **Time**

### d IOHMM Model

- **Expression Levels**
Things are a bit more complicated: Real data
A Hidden Markov Model

Hidden States

Observed outputs (expression levels)

\[
L(H, O; \Theta) = \prod_{i=1}^{n} \left[ \prod_{t=1}^{T} p(O_t(i) \mid H_t(i)) \right] \prod_{t=2}^{T} p(H_t(i) \mid H_{t-1}(i))
\]

Schliep et al Bioinformatics 2003
Input – Output Hidden Markov Model

Bengio and Frasconi, *NIPS* 1995

Input (Static transcription factor-gene interactions)

Hidden States Variables
(We constrain transitions between states to form a tree structure)

Output State Variables
(Gaussian distribution for expression values)

Log Likelihood

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

Product over all transition probabilities on path
Results: Yeast response pathways
Application to AA starvation in yeast with condition specific data

Expression data: Gasch et al Mol. Bio. Cell. 2000,

Chip-chip data: Harbison et al Nature 2004
Application to AA starvation in yeast with condition specific data

- Amino acid transport
  - Ribosome biogenesis and assembly
    - Cellular Carbohydrate Metabolism
    - Protein biosynthesis
      - Nucleotide Biosynthesis
      - Nitrogen Compound Metabolism
    - Ribosome biogenesis and assembly
      - Amino acid transport
Application to AA starvation in yeast with general binding and motif data

new predictions
Validating predicted interactions for Ino4

Ino4 Occupancy in Gene Promoter Region at 0h and 4h

Gene Intensity

SCD

AA Starvation

Intensity

0
2
4
6
8
10
12

YDR497C
YNL169C
YGR196C
YHR123W

Gene

0h
4h

Gene

YDR497C
YNL169C
YGR196C
YHR123W

Intensity

SCD

AA Starvation
Validating predicted interactions for Ino4

- Ino4 regulates phospholipid biosynthesis
- Many genes in the path are known lipid metabolism genes (GO p-value 6*10^{-5}).
- May be connected to the need of membrane used for the autophagocytosis process which regulates equilibrium between proteins and the diminishing set of amino acids
Stress and hormone response

Stem cells differentiation

Immune response

Fly development


IRF7
microRNAs
miRNA-target detection using expression data

- Pearson-correlation (Cheng et al. 2008, Huang et al. 2011)
- Regression with graphical models, GenMiR++ (Huang et al. 2007)

**Disadvantages**
- methods often assume linear correlation
- time is not implicitly modeled
- no prediction of time-specific regulation
- joint modeling of TF regulation seldom possible
Incorporating miRNAs into DREM

- Incorporate miRNA expression ratios with logistic function to generate dynamic input map for miRNAs
- Enforce positivity constraints for miRNAs coefficients in the logistic regression model (convex optimization, still global optimum)
Dynamic regulatory models for lung development in mice

- Down regulated miRNA
- Up regulated miRNA
Validating miRNAs controlling Week 1 – Week 2 transition

<table>
<thead>
<tr>
<th>miRNA</th>
<th>Sign. paths opposite direction</th>
<th>Corrected enrichment p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-125a-5p</td>
<td>A, B, C+E</td>
<td>0.0108, 0.00008, 0.003872</td>
</tr>
<tr>
<td>miR-337-5p</td>
<td>B, C+E</td>
<td>0.0332, 0.00008</td>
</tr>
<tr>
<td>miR-467c</td>
<td>D</td>
<td>&lt; 10⁻⁶</td>
</tr>
<tr>
<td>miR-466a-3p</td>
<td>D</td>
<td>0.05152</td>
</tr>
<tr>
<td>miR-466d-3p</td>
<td>D</td>
<td>0.03904</td>
</tr>
<tr>
<td>miR-30d</td>
<td>H</td>
<td>0.01456</td>
</tr>
<tr>
<td>miR-30a</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

Fig
From development to disease

Analysis of significant miRNAs in patients with Idiopathic Pulmonary Fibrosis (IPF)

<table>
<thead>
<tr>
<th>miR</th>
<th>10 control vs 10 IPF (TissueBank)</th>
<th>28 control vs 33 IPF (TissueBank)</th>
<th>142 control vs 162 IPF (LGRC)</th>
<th>bleomycin at 14d (3 replicates)</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-125a</td>
<td>down</td>
<td>-</td>
<td>-</td>
<td>down</td>
</tr>
<tr>
<td>miR-30a</td>
<td>down</td>
<td>down</td>
<td>down</td>
<td>down</td>
</tr>
<tr>
<td>miR-30d</td>
<td>down</td>
<td>down</td>
<td>down</td>
<td>down</td>
</tr>
<tr>
<td>miR-337</td>
<td>-</td>
<td>up</td>
<td>up</td>
<td>-</td>
</tr>
<tr>
<td>miR-466a-3p</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>down</td>
</tr>
<tr>
<td>miR-466d-3p</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>down</td>
</tr>
</tbody>
</table>

Schulz et al PNAS 2013
Analysis of significant miRNAs in patients with Idiopathic Pulmonary Fibrosis (IPF)

10 control vs 10 IPF (Tissue Bank)

28 control vs 33 IPF (Tissue Bank)

142 control vs 162 IPF (LGRC)

bleomycin at 14d (3 replicates)

Liu et al data J. Ex. Med. 2010

miR-125a down

miR-30a down

miR-30d down

miR-337 up

miR-467c down

miR-466a-3p down

miR-466d-3p down

From development to disease

Development

Disease

Schulz et al PNAS 2013
E. coli
Escherichia coli Regulatory Network

- Comprehensive genome wide binding data is currently not available for most transcription factors
- Key transcription factor activate some genes and repress others
  - Direction of interaction should also be part of the input to DREM
- 25% of genes have at least one known regulator based on curated small scale experiments
  - No confirmed negative data
- Expression and motif data also available
Using unlabeled data helps supervised learning

Consider setting:

- Set $X$ of instances drawn from unknown distribution $P(X)$
- Wish to learn target function $f: X \rightarrow Y$

Given:

- iid labeled examples $L = \{\langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle \}$
- iid unlabeled examples $U = \{x_{m+1}, \ldots, x_{m+n}\}$

Determine:

$$\hat{f} \leftarrow \arg\min_{h \in H} \Pr_{x \in P(X)} [h(x) \neq f(x)]$$
### SEmi-supervised REgulatory Network Discoverer (SEREND)

<table>
<thead>
<tr>
<th>Motif</th>
<th>Exp 1</th>
<th>Exp 2</th>
<th>...</th>
<th>Exp p</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>8.0</td>
<td>1.2</td>
<td>-0.5</td>
<td>...</td>
<td>0.4</td>
</tr>
<tr>
<td>Gene 2</td>
<td>6.2</td>
<td>-0.4</td>
<td>1.0</td>
<td>...</td>
<td>2.0</td>
</tr>
<tr>
<td>Gene 3</td>
<td>7.0</td>
<td>-0.8</td>
<td>1.2</td>
<td>...</td>
<td>3.2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene N</td>
<td>2.2</td>
<td>0.4</td>
<td>1.4</td>
<td>...</td>
<td>-1.4</td>
</tr>
</tbody>
</table>

**Self-training rule to change ‘unknown’ labels to either activated or repressed**

\[
P(\text{regulated}) > 2 \times (\#\text{activated} + \#\text{repressed})/N
\]

- **Binary Logistic Regression Classifier**
  - \(P(\text{regulated})\)
  - \(P(\text{activated}) + P(\text{repressed})\)

- **3-Way Logistic Regression Classifier**
  - \(P(\text{regulated})\)

- **“Meta” Classifier**


Expression Data from (Faith et al, PloS Biology 2007); Curated Interactions from EcoCyc; Motifs from Regulon DB
The Self-Training Step

Legend
+ regulated
? unknown
0 unregulated

No Self-Training

Input Labels
+ 
? 

Labels for Final Classification
+ 
0 

+ 
+++ ???
+++ +++

With Self-Training

Input Labels
+ 
? 

After Self-Training

+ 
++ 

Labels for Final Classification
+ 
0 

+ 
+++ +++
+++ +++

+ regulated
? unknown
0 unregulated
Aerobic-anaerobic shift response in *E. coli*

Expression Level (log base 10)

1 activator; -1 repressor

DREM is useful, but several questions remain …

Who controls the master regulators?