• L12 - Introduction to Protein Structure; Structure Comparison & Classification
• L13 - Predicting protein structure
• L14 - Predicting protein interactions
• L15 - Gene Regulatory Networks
• L16 - Protein Interaction Networks
• L17 - Computable Network Models
Outline

• Bayesian Networks for PPI prediction
• Gene expression
  – Distance metrics
  – Clustering
  – Signatures
  – Modules
    • Bayesian networks
    • Regression
    • Mutual Information
    • Evaluation on real and simulated data
Predictions

Last time: protein structure

Now: protein interactions

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/


© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Bayesian Networks

Predict unknown variables from observations

A “natural” way to think about biological networks.

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/. Source: Jansen, Ronald, Haiyuan Yu, et al. "A Bayesian Networks Approach for Predicting Protein-protein Interactions from Genomic Data." Science 302, no. 5644 (2003): 449-53.
Bayesian Networks

• Bayesian Networks are a tool for reasoning with probabilities
• Consist of a graph (network) and a set of probabilities
• These can be “learned” from the data
Graphical Structure Expresses our Beliefs

Cause (often “hidden”)

Effects (observed)
How do we obtain a BN?

• Two problems:
  – learning graph structure
    • NP-complete
    • approximation algorithms
  – probability distributions
Goal

• What other data could help?

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Properties of real interactions: correlated expression
Expression Profile Reliability (EPR)

INT = high confidence interactions from small scale experiments

$d = "distance"$ that measures the difference between two mRNA expression profiles

Note: proteins involved in “true” protein-protein interactions have more similar mRNA expression profiles than random pairs. Use this to assess how good an experimental set of interactions is.

Co-evolution

Which pattern below is more likely to represent a pair of interacting proteins?

More likely to interact

Courtesy of Cokus et al. License: CC-BY.

Rosetta Stone

• Look for genes that are fused in some organisms
  – Almost 7,000 pairs found in *E. coli*.
  – >6% of known interactions can be found with this method
  – Not very common in eukaryotes
Integrating diverse data

A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data

Ronald Jansen,¹* Haiyuan Yu,¹ Dov Greenbaum,¹ Yuval Kluger,¹ Nevan J. Krogan,⁴ Sambath Chung,¹,² Andrew Emili,⁴ Michael Snyder,² Jack F. Greenblatt,⁴ Mark Gerstein¹,³†

SCIENCE VOL 302 17 OCTOBER 2003

http://www.sciencemag.org/content/302/5644/449.abstract
Requirement of Bayesian Classification

• Gold standard training data
  – Independent from evidence
  – Large
  – No systematic bias
Positive training data: MIPS
• Hand-curated from literature
Negative training data:
• Proteins in different subcellular compartments
# Integrating diverse data

<table>
<thead>
<tr>
<th>Data type</th>
<th>Dataset</th>
<th># protein pairs</th>
<th>Used for ...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental interaction data</strong></td>
<td>In-vivo pull-down</td>
<td>31,304</td>
<td>Integration of experimental interaction data (PIE)</td>
</tr>
<tr>
<td></td>
<td>Gavin et al.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ho et al.</td>
<td>25,333</td>
<td></td>
</tr>
<tr>
<td>Yeast two-hybrid</td>
<td>Uetz et al.</td>
<td>981</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ito et al.</td>
<td>4,393</td>
<td></td>
</tr>
<tr>
<td><strong>Other genomic features</strong></td>
<td>mRNA Expression</td>
<td></td>
<td>De novo prediction (PIP)</td>
</tr>
<tr>
<td></td>
<td>Rosetta compendium</td>
<td>19,334,806</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cell cycle</td>
<td>17,467,005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biological function</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO biological process</td>
<td>3,146,286</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MIPS function</td>
<td>6,161,805</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Essentiality</td>
<td>8,130,528</td>
<td></td>
</tr>
<tr>
<td><strong>Gold standards</strong></td>
<td>Positives</td>
<td>8,250</td>
<td>Training &amp; testing</td>
</tr>
<tr>
<td></td>
<td>Proteins in the same MIPS complex</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negatives</td>
<td>2,708,746</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proteins separated by localization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

likelihood ratio =

\[
\frac{P(\text{true PPI} | \text{Data})}{P(\text{false PPI} | \text{Data})} = \frac{P(\text{Data} | \text{true PPI})P(\text{true PPI})}{P(\text{Data} | \text{false PPI})P(\text{false PPI})}
\]

log likelihood ratio =

\[
\log \left[ \frac{P(\text{true PPI} | \text{Data})}{P(\text{false PPI} | \text{Data})} \right] = \log \left[ \frac{P(\text{true PPI})}{P(\text{false PPI})} \right] + \log \left[ \frac{P(\text{Data} | \text{true PPI})}{P(\text{Data} | \text{false PPI})} \right]
\]

Prior probability is the same for all interactions
--does not affect ranking

Ranking function =

\[
\log \left[ \frac{P(\text{Data} | \text{true PPI})}{P(\text{Data} | \text{false PPI})} \right] = \prod_{i=1}^{M} \frac{P(\text{Observation}_i | \text{true PPI})}{P(\text{Observation}_i | \text{false PPI})}
\]
Protein pairs in the essentiality data can take on three discrete values (EE, both essential; NN, both non-essential; and NE, one essential and one not)

\[ \text{Likelihood} = L = \frac{P(f \mid pos)}{P(f \mid neg)} \]

| Essentiality | # protein pairs | Gold-standard overlap | P(Ess|pos) | P(Ess|neg) | L |
|--------------|----------------|-----------------------|-----------|-----------|---|
| EE           | 384,126        | 1,114                 | 81,924    | 1,114     | 5.18E-01 | 1.43E-01 | 3.6 |
| NE           | 2,767,812      | 924                   | 285,487   | 1,738     | 367,411   | 0.005    | 2.90E-01 | 4.98E-01 | 0.6 |
| NN           | 4,976,590      | 12                    | 206,313   | 2,150     | 573,724   | 0.004    | 1.92E-01 | 3.60E-01 | 0.5 |
| Sum          | 8,130,528      | 2,150                 | 573,724   | -         | -         | 1.00E+00 | 1.00E+00 | 1.0 |

\[ \frac{1,114}{2150} \]

\[ \frac{81,924}{573,734} \]
| Essentiality | # protein pairs | Gold-standard overlap |            |            |              | P(Ess|pos) | P(Ess|neg) | L  |
|--------------|-----------------|-----------------------|------------|------------|----------------|----------|-----------|-----|
|              |                 | pos                   | neg        | sum(pos)   | sum(neg)       | sum(pos)/sum(neg) |           |           |     |
| EE           | 384,126         | 1,114                 | 81,924     | 1,114      | 81,924         | 0.01        | 5.18E-01 | 1.43E-01 | 3.6 |
| NE           | 2,767,812       | 624                   | 285,487    | 1,738      | 387,411         | 0.005       | 2.90E-01 | 4.98E-01 | 0.6 |
| NN           | 4,978,590       | 412                   | 206,313    | 2,150      | 573,724         | 0.004       | 1.92E-01 | 3.60E-01 | 0.5 |
| Sum          | 8,130,528       | 2,150                 | 573,724    |            |                |            | 1.00E+00 | 1.00E+00 | 1.0 |

| Expression correlation | # protein pairs | Gold standard overlap |            |            |              | P(exp|pos) | P(exp|neg) | L  |
|------------------------|-----------------|-----------------------|------------|------------|----------------|----------|-----------|-----|
|                        |                 | pos                   | neg        | sum(pos)   | sum(neg)       | sum(pos)/sum(neg) |           |           |     |
| 0                      | 18,773,128      | 7,614                 | 2,635,273  |            |                |            | 1.00E+00 | 1.00E+00 | 1.0 |

| MIPS function similarity | # protein pairs | Gold standard overlap |            |            |              | P(MIPS|pos) | P(MIPS|neg) | L  |
|--------------------------|-----------------|-----------------------|------------|------------|----------------|----------|-----------|-----|
|                          |                 | pos                   | neg        | sum(pos)   | sum(neg)       | sum(pos)/sum(neg) |           |           |     |
| 1 -- 9                   | 6,584           | 171                   | 1,094      | 171        | 1,094          | 0.16      | 2.12E-02 | 8.33E-04 | 25.6 |
| 10 -- 99                 | 25,823          | 584                   | 4,229      | 755        | 5,323          | 0.14      | 7.25E-02 | 3.22E-03 | 22.5 |
| 100 -- 1000              | 88,548          | 688                   | 13,011     | 1,443      | 18,334         | 0.08      | 8.55E-02 | 9.91E-03 | 8.6  |
| 1000 -- 10000            | 255,096         | 6,146                 | 47,126     | 7,539      | 65,460         | 0.12      | 7.83E-02 | 3.59E-02 | 21.3 |
| 10000 -- Inf             | 5,785,754       | 462                   | 1,248,119  | 8,051      | 1,313,579      | 0.01      | 5.74E-02 | 9.50E-01 | 0.1  |
| Sum                      | 6,161,805       | 8,051                 | 1,313,579  |            |                |            | 1.00E+00 | 1.00E+00 | 1.0  |

| GO biological process similarity | # protein pairs | Gold standard overlap |            |            |              | P(GO|pos) | P(GO|neg) | L  |
|----------------------------------|-----------------|-----------------------|------------|------------|----------------|----------|-----------|-----|
|                                  |                 | pos                   | neg        | sum(pos)   | sum(neg)       | sum(pos)/sum(neg) |           |           |     |
| 1 -- 9                           | 3,146,286       | 7,520                 | 645,025    |            |                |            | 1.00E+00 | 1.00E+00 | 1.0  |

The values in red indicate the lowest L values for each section.
A Bayesian Networks Approach for Predicting Protein-protein Interactions from Genomic Data.

What do we mean by fully connected?

\[ P(X_1 \ldots X_n | \text{PPI}) = \prod_i [P(X_i | \text{PPI})] \]

\[ P(X_1 \ldots X_n | \text{PPI}) \neq \prod_i [P(X_i | \text{PPI})] \]
Fully connected $\rightarrow$
Compute probabilities for all 16 possible combinations

$$P(X_1 \ldots X_n|\text{PPI}) \neq \prod_i [P(X_i|\text{PPI})]$$
Fully connected $\rightarrow$
Compute probabilities for all 16 possible combinations

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Interpret with caution, as numbers are small.

How many gold-standard events do we score correctly at different likelihood cutoffs?

\[
\log \left[ \frac{P(\text{Data} \mid \text{true}_\text{PPI})}{P(\text{Data} \mid \text{false}_\text{PPI})} \right]
\]

Prediction based on single data type all have TP/FP<1

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Outline

• Bayesian Networks for PPI prediction
• Gene expression
  – Distance metrics
  – Clustering
  – Signatures
  – Modules
    • Bayesian networks
    • Regression
    • Mutual Information
    • Evaluation on real and simulated data
Gene Expression Data

- Identify co-expressed genes
- Classify new datasets
- Discover regulatory networks

DATA DUMP
The number of gene-expression data sets in publicly available databases has climbed to nearly one million over the past decade.

Clustering

• Text Section 16.2
• Multiple mechanisms could lead to up-regulation in any one condition
• Goal: Find genes that have “similar” expression over many condition.
• How do you define “similar”? 
Distance Metrics

![Graph showing distance metrics over time](image-url)
Expression data as multidimensional vectors

Gene A

expression in experiment 1

$X_A = (1, 0.5, -1, 0.25, ...)$

Gene B

expression in experiment 2

$X_B = (0.2, 0.4, -1.2, 0.05, ...)$

$X_{A,1}$ $X_{B,1}$ $X_{A,2}$ $X_{B,2}$ ...

What is a natural way to compare these vectors?
• $X_{i,j} = \text{Expression of gene } i \text{ in condition } j$

\[
d(X_A, X_B) = \sqrt{\sum_{k=1}^{N} (X_{A,k} - X_{B,k})^2}
\]
Distance

• Metrics have a formal definition:
  – \( d(x, y) \geq 0 \)
  – \( d(x, y) = 0 \) if and only if \( x = y \)
  – \( d(x, y) = d(y, x) \)
  – Triangle inequality:
    \[ d(x, z) \leq d(x, y) + d(y, z) \]

• The triangle inequality need not hold for a measure of “similarity.”

• Distance ~ Dissimilarity = 1 - similarity
Can we capture the similarity of these patterns?
Pearson Correlation

- $X_{i,j} = \text{Expression of gene } i \text{ in condition } j$
- $Z_i = \text{z-score of gene } i \text{ one experiment:}$

$$Z_A = \frac{X_A - \overline{X_A}}{\sigma} \quad \sigma^2 = \frac{\sum (X - \overline{X})}{N}$$
Pearson Correlation

- $X_{i,j} = $ Expression of gene $i$ in condition $j$
- $Z_i = $ z-score of gene $i$ one experiment:

- Pearson correlation
  - $r_{A,B} = \frac{\sum Z_A Z_B}{N}$ over all experiments
  - from +1 (perfect correlation) to -1 (anti-correlated)

- Distance $= 1 - r_{A,B}$

$$Z_A = \frac{X_A - \bar{X}_A}{\sigma}$$
$$\sigma^2 = \frac{\sum (X - \bar{X})}{N}$$
Expression

Z-score

\[ Z_A = \frac{X_A - \bar{X}_A}{\sigma} \]

\[ r_{A, B} = \frac{\sum Z_A Z_B}{N} \]
$R_{A,B} = -0.01$
$R_{A,C} = 0.999$
$R_{B,C} = -0.03$

$$r_{A,B} = \frac{\sum Z_A Z_B}{N}$$

$$Z_A = \frac{X_A - \bar{X}_A}{\sigma}$$
\[ R_{A,B} = -0.01 \]
\[ R_{A,D} = -1.0 \]
\[ R_{B,D} = 0.007 \]

\[ r_{A,B} = \frac{\sum Z_A Z_B}{N} \]
Distance Metrics

\[ d(X_A, X_B) = \sqrt{\sum_{k=1}^{N} (X_{A,k} - X_{B,k})^2} \]

\[ 1 - r_{A,B} = \frac{\sum Z_A Z_B}{N} \]
Missing Data

• What if a particular data point is missing? (Back in the old days: there was a bubble or a hair on the array)
  – ignore that gene in all samples
  – ignore that sample for all genes
  – replace missing value with a constant
  – “impute” a value

• example: compute the K most similar genes (arrays) using the available data; set the missing value to the mean of that for these K genes (arrays)
Outline

• Bayesian Networks for PPI prediction
• Gene expression
  – Distance metrics
  – Clustering
  – Signatures
  – Modules
    • Bayesian networks
    • Regression
    • Mutual Information
    • Evaluation on real and simulated data
Clustering

• Intuitive idea that we want to find an underlying grouping
• In practice, this can be hard to define and implement.
• An example of unsupervised learning
Unsupervised Learning
Clustering 8600 human genes based on time course of expression following serum stimulation of fibroblasts

Key: Black = little change   Green = down   Red = up
(relative to initial time point)

(A) cholesterol biosynthesis
(B) the cell cycle
(C) the immediate-early response
(D) signaling and angiogenesis
(E) wound healing and tissue remodeling

Iyer et al. Science 1999
Why cluster?

• Cluster genes (rows)
  – Measure expression at multiple time-points, different conditions, etc.

  Similar expression patterns may suggest similar functions of genes

• Cluster samples (columns)
  – e.g., expression levels of thousands of genes for each tumor sample

  Similar expression patterns may suggest biological relationship among samples
Hierarchical clustering

Two types of approaches:
• Agglomerative
• Divisive
Agglomerative Clustering Algorithm

• Initialize: Each data point is in its own cluster
• Repeat until there is only one cluster:
  – Merge the two most similar clusters.
Agglomerative Clustering Algorithm

- Initialize: Each data point is in its own cluster
- Repeat until there is only one cluster:
  - Merge the two most similar clusters.

If distance is defined for a vector, how do I compare clusters?
• Clusters Y, Z with A in Y and B in Z
• Single linkage = \( \min\{d_{A,B}\} \)
• Complete linkage = \( \max\{d_{A,B}\} \)
• UPGMC (Unweighted Pair Group Method using **Centroids**)
  
  \[
  \text{centroid} = \hat{Y} = \frac{1}{N_Y} \sum_{i \in Y} X_{i,j}
  \]

  — Define distance as \( \delta_{Y,Z} = d_{\hat{Y},\hat{Z}} \)

• UPGMA (Unweighted Pair Group Method with Arithmetic **Mean**) average of pairwise distances:
  
  \[
  \delta_{Y,Z} = \frac{1}{N_Y N_Z} \sum_{i \in Y} \sum_{j \in Z} d_{i,j}
  \]
• Single linkage = min\{d_{A,B}\}
• Complete linkage = max\{d_{A,B}\}
• If clusters exist and are compact, it should not matter.
• Single linkage will “chain” together groups with one intermediate point.
• Complete linkage will not combine two groups if even one point is distant.
Interpreting the Dendogram

- This produces a binary tree or *dendrogram*
- The final cluster is the root and each data item is a leaf
- The heights of the bars indicate how close the items are
- Can ‘slice’ the tree at any distance cutoff to produce discrete clusters
- Dendogram represents the results of the *clustering*; its usefulness in representing the *data* is mixed.
- The results will always be hierarchical, even if the data are not.
K-means clustering

• Advantage: gives sharp partitions of the data
• Disadvantage: need to specify the number of clusters (K).
• Goal: find a set of k clusters that minimizes the distances of each point in the cluster to the cluster mean:

\[
\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i
\]

\[
\arg\min_C \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2
\]
K-means clustering algorithm

- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: place each point $X_i$ in the cluster with the closest mean.
  - Update: recalculate the mean for each cluster
round 1 distance 54
What if you choose the wrong K?
Big steps occur when we are dividing data into natural clusters.

Smaller steps occur when we are overclustering.

![Graph showing within cluster distance over number of clusters](image)
What if we choose pathologically bad initial positions?
round 4 distance 136
round 6 distance 67
round 8 distance 54
What if we choose pathologically bad initial positions?

Often, the algorithm gets a reasonable answer, but not always!
Convergence

• K-means always converges.
• The assignment and update steps always either reduce the objective function or leave it unchanged.

$$\text{argmin}_{C} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$
Convergence

• However, it doesn’t always find the same solution.

\[ K=2 \]
Fuzzy K-means

K=2
K-means

- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: place each point $X_i$ in the cluster with the closest mean.
  - Update: recalculate the mean for each cluster

Fuzzy k-means

- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: calculate probability of each point belonging to each cluster.
  - Update: recalculate the mean for each cluster using these probabilities
K-means

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$

$$\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i$$

Fuzzy k-means

$$\arg\min_{\mu, Y} \sum_{i=1}^{k} \sum_{j=1}^{N} \mu_{i,j}^r \left| X_j - \hat{Y}_i \right|^2$$

$$\text{centroid}_j = \hat{Y}_j = \frac{\sum_{i=1}^{N} \mu_{i,j}^r X_i}{\sum_{i=1}^{N} \mu_{i,j}^r}$$

$$\mu_{i,j}^r = \text{membership of point } j \text{ in cluster } i$$

Larger values of r make the clusters more fuzzy.

Relationship to EM and Gaussian mixture models
Example of Fuzzy K-means


Limits of k-means

K-means uses Euclidean distance

\[
\arg\min_C \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2
\]

\[
\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i
\]

- Gives most weight to largest differences
- Can’t be used if data are qualitative
- Centroid usually does not represent any datum
K-means
• Best clustering minimizes within-cluster Euclidean distance of from centroids

$$\text{centroid} = \hat{Y} = \frac{1}{N_Y} \sum_{i \in Y} X_{i,j}$$

K-medoids
• Best clustering minimizes within-cluster dissimilarity from medoids (exemplar)

$$\text{medoid}_k = \arg\min_i \sum_{i' \in C(k)} D(X_i, X_{i'})$$
K-medoids clustering

• Initialize: choose k points as cluster means
• Repeat until convergence:
  – Assignment: place each point $X_i$ in the cluster with the closest medoid.
  – Update: recalculate the medoid for each cluster
Other approaches

• SOM (Text 16.3)
• Affinity Propagation
So What?

• Clusters could reveal underlying biological processes not evident from complete list of differentially expressed genes

• Clusters could be co-regulated. How could we find upstream factors?
Outline

• Bayesian Networks for PPI prediction

• Gene expression
  – Distance metrics
  – Clustering
  – Signatures
  – Modules
    • Bayesian networks
    • Regression
    • Mutual Information
    • Evaluation on real and simulated data
Personalized Medicine

• Can gene expression be used for diagnosis and to determine the best treatment?
Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling
OS = the fraction of patients alive (overall survival)
Hazard Ratio = Death rate vs. control

post-prandial laughter

Courtesy of Venet et al. License: CC-BY.

© FarWorks, Inc. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
OS= the fraction of patients alive (overall survival)
Hazard Ratio= Death rate vs. control

 Courtesy of Venet et al. License: CC-BY.

© FarWorks, Inc. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet\textsuperscript{1}, Jacques E. Dumont\textsuperscript{2}, Vincent Detours\textsuperscript{2,3}\textsuperscript{*}

\textsuperscript{1}IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, \textsuperscript{2}IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, \textsuperscript{3}WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium
OS = the fraction of patients alive (overall survival)
Hazard Ratio = Death rate vs. control

post-prandial laughter

localization of skin fibroblasts

social defeat in mice

Courtesy of Venet et al. License: CC-BY.
Published Signature

Distribution for random signatures

Best 5% of random signatures

Courtesy of Venet et al. License: CC-BY.
Hazard Ratio=
Death rate vs. control

R^2 = 0.9

PCNA metagene = 1% genes the most positively correlated with expression of PCNA (proliferating cell nuclear antigen, a known marker) across 36 tissues

Courtesy of Venet et al. License: CC-BY.
Outline

• Bayesian Networks for PPI prediction
• Gene expression
  – Distance metrics
  – Clustering
  – Signatures
  – Modules
    • Bayesian networks
    • Regression
    • Mutual Information
    • Evaluation on real and simulated data
Reconstructing Regulatory Networks

Courtesy of Elsevier B.V. Used with permission.
Clustering vs. “modules”

• Clusters are purely phenomenological – no claim of causality
• The term “module” is used to imply a more mechanistic connection
Wisdom of crowds for robust gene network inference

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, The DREAM5 Consortium, Manolis Kellis, James J Collins & Gustavo Stolovitzky

Affiliations  |  Contributions  |  Corresponding author

Received 31 October 2011  |  Accepted 22 May 2012  |  Published online 15 July 2012
Wisdom of crowds for robust gene network inference
Wisdom of crowds for robust gene network inference
Outline

• Bayesian Networks for PPI prediction
• Gene expression
  – Distance metrics
  – Clustering
  – Signatures
  – Modules
    • Bayesian networks
    • Regression
    • Mutual Information
    • Evaluation on real and simulated data
Bayesian Networks

A "natural" way to think about biological networks.

Predict unknown variables from observations.
Is the p53 pathway activated?

Courtesy of Looso et al. License: CC-BY.
Is the p53 pathway activated?

**Possible Evidence**

- Known p53 targets are up-regulated

Is the p53 pathway activated?

• Formulate problem probabilistically

• Compute
  – $P(\text{p53 pathway activated} \mid \text{data})$

• How?
  – Relatively easy to compute $p(X \text{ up} \mid \text{TF up})$
  – How?

![Diagram](image-url)
Is the p53 pathway activated?

- Formulate problem probabilistically
- Compute
  - $P(\text{p53 pathway activated} | \text{ data})$
- How?
  - Relatively easy to compute $p(X \text{ up } | \text{ TF up})$
  - Look over lots of experiments and tabulate:
    - $X1 \text{ up } & \text{ TF up}$
    - $X1 \text{ up } & \text{ TF not up}$
    - $X1 \text{ not up } & \text{ TF not up}$
    - $X1 \text{ not up } & \text{ TF up}$
Is the p53 pathway activated?

- Formulate problem probabilistically

- Compute
  - $P(\text{p53 pathway activated} | \text{data})$

- How?
  - Relatively easy to compute $p(X \text{ up } | \text{ TF up})$
  - $P(\text{TF up} | X \text{ up}) = p(X \text{ up } | \text{ TF up}) \frac{p(\text{TF up})}{p(X \text{ up})}$
Is the p53 pathway activated?

- Formulate problem probabilistically
- Compute
  - $P(\text{p53 pathway activated} \mid \text{data})$
- How?
  - Even with $p(\text{TF up} \mid \text{X up})$ how do we compare this explanation of the data to other possible explanations?
  - Can we include upstream data?
Application to Gene Networks

- Which pathway activated this set of genes?
- Either A or B or both would produce similar but not identical results.

- Bayes Nets estimate conditional probability tables from lots of gene expression data.
- How often is TF B2 expressed when TF B1 is expressed, etc.
Does the probability that it’s raining depend on whether the sprinkler is on?

In a causal sense, clearly not.

But in a probabilistic model, the knowledge that it is raining influences our beliefs.
Multi-layer networks are possible, but feedback is not
Learning Models from Data

• Searching for the BN structure: NP-complete
  – Too many possible structures to evaluate all of them, even for very small networks.
  – Many algorithms have been proposed
  – Incorporated some prior knowledge can reduce the search space.
    • Which nodes should regulate transcription?
    • Which should cause changes in phosphorylation?
  – Intervention experiments help
Without interventions, all we can say is that X and Y are correlated

Interventions allow us to determine which is the parent.

If we don’t measure “Y” can we still model the data? The relationship of X and Z,W will be noisy and might be missed.
Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
- Modules
  - Bayesian networks
  - Regression
  - Mutual Information
  - Evaluation on real and simulated data
Regression-based models

Predicted expression: \( Y_g = f_g(X_{Tg}) + \epsilon \)

Assume that expression of gene \( X_g \) is some function of the expression of its transcription factors \( X_{Tg} = \{X_t, t \in T_g\} \)

\( X_i \) = measured expression of i-th gene

\( X_{Ti} \) = measured expression of a set of TFs potentially regulating gene i

\( f_g \) is an arbitrary function

\( \epsilon \) is noise
Regression-based models

\[ f_g (X_{Tg}) = \sum_{t \in T_g} \beta_{t,g} X_t \]

\( f_g \) is frequently assumed to be a linear function
The values of the \( \beta_{t,g} \) reflect the influence of each TF on
gene \( g \)

How do we discover the values of the \( \beta_{t,g} \) ?


Regression-based models

\[ Y_g = \sum_{t \in T_g} \beta_{t,g} X_t + \varepsilon \]

Define an objective function:
Sum over M training data sets and N genes
Find parameters that minimize “residual sum of squares” between observed (X) and predicted (Y) expression levels.

\[ \text{RSS} = \sum_{j=1}^{M} \sum_{i=1}^{N} (X_{i,j} - Y_{i,j})^2 \]
Regesration-based models

\[ Y_g = \sum_{t \in T_g} \beta_{t,g} X_t + \varepsilon \]

\[ RSS = \sum_{j=1}^{M} \sum_{i=1}^{N} (X_{i,j} - Y_{i,j})^2 \]

Problems:
Standard regression will produce many very small values of \( \beta \), which makes interpretation difficult
\( \beta \) values can be unstable to changes in training data

Solutions:
Subset Selection and Coefficient Shrinkage
Outline

• Bayesian Networks for PPI prediction

• Gene expression
  – Distance metrics
  – Clustering
  – Signatures

– Modules
  • Bayesian networks
  • Regression
  • Mutual Information
  • Evaluation on real and simulated data
Quick Review of Information Theory

Information content of an event $E$

$$I(E) = \log_2 \frac{1}{P(E)}$$

Rare letters have higher information content
Quick Review of Information Theory

Information content of an event $E$

$$I (E) = \log_2 \frac{1}{P(E)}$$

Entropy is evaluated over all possible outcomes

$$H (S) = \sum_i p_i I(s_i) = \sum_i p_i \log_2 \frac{1}{p_i}$$

$$H(f) = - \int f(x) \ln f(x) dx.$$
Mutual Information

• Does knowing variable X reduce the uncertainty in variable Y?
• Example:
  – \( P(\text{Rain}) \) depends on \( P(\text{Clouds}) \)
  – \( P(\text{target expressed}) \) depends on \( P(\text{TF expressed}) \)

\[
I(x, y) = H(x) + H(y) - H(x, y)
\]

• \( I(x,y) = 0 \) means variables are independent
• Reveals non-linear relationships that are missed by correlation.
Mutual information detects non-linear relationships

Incoherent feed-forward loop (FFL)

Mutual information $= 1.7343$

Correlation coefficient $= -0.0464$

No correlation, but knowing A reduces the uncertainty in the distribution of B

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Mutual information detects non-linear relationships

- Complex regulatory network structure => complex relationships between protein levels
- Example: incoherent feed-forward loop (FFL)
ARACNe

Reverse engineering of regulatory networks in human B cells

Katia Basso¹, Adam A Margolin², Gustavo Stolovitzky³, Ulf Klein¹, Riccardo Dalla-Favera¹,⁴ & Andrea Califano²
ARACNe

• Find TF-target relationships using mutual information

\[ H(f) = - \int f(x) \ln f(x) \, dx. \]

• How do you recognize a significant value of MI?
  – randomly shuffle expression data
  – compute distribution of Mutual information
ARACNE

- Data processing inequality
  - Eliminate indirect interactions
  - If G2 regulates G1,G3
    \[ I(G1,G3) > 0 \] but adds no insight
  - Remove edge with smallest mutual information in each triple

\[
I(g_1, g_3) \leq \min [I(g_1, g_2); I(g_2, g_3)]
\]
MINDy

- Identify proteins that modulate TF function
  - Other TFs

Genome-wide identification of post-translational modulators of transcription factor activity in human B cells

Kai Wang¹,²,⁵,⁶, Masumichi Saito³,⁵,⁶, Brygida C Bisikiriska², Mariano J Alvarez², Wei Keat Lim¹,²,⁵, Presha Rajbhandari², Qiong Shen³, Ilya Nemenman²,⁵, Katia Basso³, Adam A Margolin¹,²,⁵, Ulf Klein³, Riccardo Dalla-Favera³,⁴ & Andrea Califano¹–³
Model

- Assumes that expression of target T is determined by TF and modulator (M)

\[
[T] = C \cdot [TF]^h \cdot [M]^g
\]

- Modulator present at highest levels
- Modulator present at lowest levels
- \(\rightarrow\) Suggests M is an activator
Filters

1. expression of the modulator and of the TF must be statistically independent
2. the modulator expression must have sufficient range
3. may be filtered by additional criteria—for example, molecular functions.
Estimate conditional mutual information

Courtesy of Macmillan Publishers Limited. Used with permission.
Supplementary Table 12. Inferring the biological activity of a MINDy modulator. MoA:

MINDy mode of action: \( \rho \): Pearson correlation between TF and the target gene \( t \); \( \mu_t^+ \): the mean expression of \( t \) in the most and least expressed condition of the modulator. BA: biological activity. The schematic scatter plots shown in the table demonstrate the relationship between TF and \( t \) when the modulator is most (red dots) and least (blue dots) expressed.

\[
\begin{align*}
\text{activator} & \quad \text{if } \rho (\mu_t^+ - \mu_t^-) > 0 \\
\text{antagonist} & \quad \text{if } \rho (\mu_t^+ - \mu_t^-) < 0 \\
\text{undetermined} & \quad \text{if } \rho (\mu_t^+ - \mu_t^-) \approx 0
\end{align*}
\]

where \( \rho \) is the Pearson correlation between TF and \( t \), and \( \mu_t^+ \) is the mean expression of \( t \) in \( L^+_w \). In practice, however, the difference between \( \mu_t^+ \) has to be assessed statistically.

In this work, we choose to use the two sample Student t-test (two sided) that assess the null hypothesis of \( \mu_t^+ = \mu_t^- \). If the null hypothesis can not be rejected at \( \alpha = 0.1 \), we assign the mode to be undermined; otherwise, \( M_i \) is considered an activator or antagonist (depending on which tail is tested) of the interaction between TF and \( t_i \).

Note than none of these curve saturate

<table>
<thead>
<tr>
<th>MoA</th>
<th>( \rho )</th>
<th>( \mu_t^+ - \mu_t^- )</th>
<th>Plot</th>
<th>BA</th>
<th>Sign(( \rho (\mu_t^+ - \mu_t^-) ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
<td>t</td>
<td>Activator</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>-</td>
<td>t</td>
<td>Antagonist</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>-</td>
<td>t</td>
<td>Activator</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td>+</td>
<td>t</td>
<td>Antagonist</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>-</td>
<td>t</td>
<td>Antagonist</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td>+</td>
<td>t</td>
<td>Activator</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>+</td>
<td>t</td>
<td>Antagonist</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>t</td>
<td>Activator</td>
<td>+</td>
</tr>
</tbody>
</table>

What regulates MYC?

Input:
254 expression profiles in B cells
(normal and tumor)
various sets of candidate regulators

Evaluation:
1. comparison to known modulators
2. experimental tests of four candidates
What regulates MYC?

Limitations

• Need huge expression datasets
• Can’t find:
  – modulator that do not change in expression
  – modulator that are highly correlated with target
  – modulators that both activate and repress
Huge networks!

This is just the nearest neighbors of one node of interest from ARACNe!


Courtesy of Macmillan Publishers Limited. Used with permission.
Huge networks!

Conditional MI network of miR modulators
248,000 interactions


Courtesy of Elsevier B.V. Used with permission.
MINDy modulators

<table>
<thead>
<tr>
<th>Source of targets</th>
<th>Potential Modulators</th>
<th>Database</th>
<th>ARACNe</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Signaling (542)</td>
<td>91</td>
<td>80</td>
<td>[25/296]</td>
</tr>
<tr>
<td></td>
<td>TFs (598)</td>
<td>99</td>
<td>85</td>
<td>[32/296]</td>
</tr>
<tr>
<td></td>
<td>Any (3,131)</td>
<td></td>
<td></td>
<td>296</td>
</tr>
</tbody>
</table>

MINDy selects between 10-20% of candidates!
Outline

• Bayesian Networks for PPI prediction
• Gene expression
  – Distance metrics
  – Clustering
  – Signatures

– Modules
  • Bayesian networks
  • Regression
  • Mutual Information
  • Evaluation on real and simulated data
Wisdom of crowds for robust gene network inference

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, The DREAM5 Consortium, Manolis Kellis, James J Collins & Gustavo Stolovitzky

Affiliations  Contributions  Corresponding author

Received 31 October 2011  Accepted 22 May 2012  Published online 15 July 2012
AUPR = area under precision-recall curve

Wisdom of crowds for robust gene network inference

Courtesy of Macmillan Publishers Limited. Used with permission.
Wisdom of crowds for robust gene network inference
Wisdom of crowds for robust gene network inference
Wisdom of crowds for robust gene network inference
Wisdom of crowds for robust gene network inference

Area under precision-recall curve

AUPR = area under precision-recall curve

Courtesy of Macmillan Publishers Limited. Used with permission.
Wisdom of crowds for robust gene network inference
Wisdom of crowds for robust gene network inference
Thoughts on Gene Expression Data

• Useful for classification and clustering
• Not sufficient for reconstructing regulatory networks in yeast
• Can we infer levels of proteins from gene expression?
Approach

mRNA levels do not predict protein levels

![Scatter plot showing mRNA expression levels vs. protein expression levels. The R² value is 0.22, and Rₛ is 0.46.](image)

© Royal Society of Chemistry. All rights reserved. This content is excluded from our Creative Commons license. For more information, see [http://ocw.mit.edu/help/faq-fair-use/](http://ocw.mit.edu/help/faq-fair-use/).


<table>
<thead>
<tr>
<th>SpectrumMill</th>
<th>msInspect</th>
<th>msBID</th>
<th>NSAF</th>
<th>RPKM</th>
<th>Microarray</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SpectrumMill</strong></td>
<td>-</td>
<td>0.91 (0.92)</td>
<td>0.91 (0.91)</td>
<td>0.90 (0.90)</td>
<td>0.49 (0.51)</td>
</tr>
<tr>
<td><strong>msInspect</strong></td>
<td>0.91 (0.92)</td>
<td>-</td>
<td>0.89 (0.91)</td>
<td>0.87 (0.88)</td>
<td>0.51 (0.53)</td>
</tr>
<tr>
<td><strong>msBID</strong></td>
<td>0.91 (0.91)</td>
<td>0.89 (0.91)</td>
<td>-</td>
<td>0.84 (0.89)</td>
<td>0.54 (0.54)</td>
</tr>
<tr>
<td><strong>NSAF</strong></td>
<td>0.90 (0.90)</td>
<td>0.87 (0.88)</td>
<td>0.84 (0.89)</td>
<td>-</td>
<td>0.51 (0.53)</td>
</tr>
</tbody>
</table>


Global quantification of mammalian gene expression control.
Schwanhäusser B1, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M.

• L12 - Introduction to Protein Structure; Structure Comparison & Classification
• L13 - Predicting protein structure
• L14 - Predicting protein interactions
• L15 - Gene Regulatory Networks
• L16 - Protein Interaction Networks
• L17 - Computable Network Models