statistical systems biology

department of applied physics and applied mathematics
+
center for computational biology and bioinformatics
columbia university
chris.wiggins@columbia.edu
statistical systems biology: agenda

1. challenges to keep in mind
2. microarrays / regulation
3. networks
4. final thoughts
statistical systems biology: challenges

1. statistics
2. modeling
3. validation
4. interpretation
microarrays + transcriptional regulation

1. biological questions
2. history/context
3. methods
   - “unsupervised”: cluster first, ask questions later
   - “supervised”: predicting methods
biology as told by a theorist
Figure 1.24. The effect of Cro. Cro first abolishes synthesis of repressor from $P_{br}$ and then turns off synthesis of its own gene as well.

ptashne’s “a genetic switch”
what is to be measured?

1. “expression” via RNA abundance
what is to be measured?

2. regulatory sequence

>YLR081W       GAL2
AGGTTGCAATTTCTTTTTCTATTAGTAGCTAAAAATGGGTCACGTGATCT
                                              CEN
GAL4
                                                        -451
AGGTTGCAATTTCTTTTTCTATTAGTAGCTAAAAATGGGGTAGTTA
ATATCGAAAGGGCGGGTGGCCTCAGGAAGGCACCGGCGGTCTTTCGTCC
                                                        -401
GTGC
GAL4
GGAGATATCTGCGCCGTTCAGGGGTCCATGTGCCTTGGACGATATT
AAGGCAGAAGGCAGTATCGGGCGGATCACTCCGAACCAGATATTAGTTA
GCCCTTCCCATCTCAAGATGGGAGCAATGGCATTATACCTCCTGCTAGA
AAGTTAACTGTGCACATATCTTAAATTATAAAGGACATTCGAGACCTAT
TGTTCAAAAAACAACATTTTCGCAAGCTAAAATGTGGAGATAGGATAAGT
TTTGTAGACATATATAAACAATCAGTAATGGGATTGAAAAATTTGTTGTTG
TGAATTGCTCTCTATTATGCAAGCTTAATTATATCAGAAAGAATAGTAAA
TAGTTAAGTTAAACACAGATTACATAAATAAAAAATAAAATTTCTTTCTATA
ATGGCAGTTGAGAGAACAATATG CCTGTGTTGTTTACAGCAACCCCAAGC
                                                        +50
Affymetrix’ GeneChip® technology was invented in the late 1980's by a team of scientists led by Stephen P.A. Fodor, Ph.D. The theory behind their work was revolutionary - a notion that semiconductor manufacturing techniques could be united with advances in combinatorial chemistry to build vast amounts of biological data on a small glass chip. This technology became the basis of a new company, Affymetrix, formed as a division of Affymax, N.V. in 1991. Affymetrix began operating independently in 1992.

Circa 1989 - The world's first microarray prototype built using a microscope slide.

Affymetrix has headquarters in Santa Clara, California with offices
cDNA “spot” arrays: 1995

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena (1), Dari Shalon (1), Ronald W. Davis (2), Patrick O. Brown (3)

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 Arabidopsis genes were made by means of simultaneous, two-color fluorescence hybridization.
the hope
other relevant innovation:

shared data.
microarrays + transcriptional regulation

1. biological questions
2. history/context
3. methods
   • “unsupervised”: cluster first, ask questions later
   • “supervised”: predicting methods
descriptive “models” of regulation:


- “unsupervised” (no input-output relation)
descriptive “models” of regulation:

• “unsupervised” (no input-output relation)
microarrays + transcriptional regulation

1. biological questions
2. history/context
3. methods
   - “unsupervised”: cluster first, ask questions later
   - “supervised”: predicting methods
Regulatory element detection using correlation with expression

Harmen J. Bussemaker\textsuperscript{1,2}, Hao Li\textsuperscript{1} & Eric D. Siggia\textsuperscript{1}

Acknowledgments
We thank B. Shraiman for suggesting linear multivariate fits to expression data, and L. Grivell, R. Lascaris and H. de Nobel for discussions and critical reading of the manuscript. Support was received from the NSF under grant number DMR 9732083 and from the Keck foundation to H.L.

Received 23 February 2000; accepted 3 January 2001.
REDUCE: why 7?

ptashne’s “a genetic switch”
learning networks from biology

\[ x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \]
“learning networks”: learn network-shaped $f$

\[ A^t_g = f(\mu_g, \pi^t) \]

\[ x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \]
**GeneClass**: predict expression as class

- complex enough to learn from data
- simple enough
  - to generalize
    (predict on “held out” experiments)
  - and to be interpretable
    (based on biological rules)
- will exploit 3 tricks
trick #1: base on biological rules

parents - “motifs” - children

- 10M-dimensional feature space
- approx 100*6000 examples
trick #2: predict expression as class

build a theory of 3’s?
1-slide summary of classification

• banana or orange?

large deviation theory: “maximum margin”
1-slide summary of classification

- banana or orange?

boosting (1997)
SVMs (1990s)
1-slide summary of classification

• up- or down-regulated?

“cat” & gene 11 up?
“gataca” & gene 37 down?
“tag” & gene 34 up?
“gaga” & gene 3184 up?
“acgt” & gene 45 down?

learn predictive features from data
model framework: \[ A_g^t = f(\mu_g, \pi^t) \]

- Parent states: \( P_{\pi e} \in \{-1, 0, 1\} \)
- Feature vector: motif-parent pairs
  \[ x_{g,e} = \{ M_{\mu g} P_{\pi e} \}_{\mu, \pi} \]
- Label: \( y \in \{+1, -1\} \)
- Motifs: \( M_{\mu g} \in \{0, 1\} \)
1-slide summary of classification

• up- or down-regulated?

“gataca” & gene 37 down?

“gaga” & gene 3184 up?

learn predictive features from data
"boosting"?

• Anachronistic observation:

\[ \langle e^{-\sigma B(\vec{x})} \rangle \text{ minimized by } B(\vec{x}) = \frac{1}{2} \ln \frac{p(+) | \vec{x})}{p(-) | \vec{x})} \]

• Therefore approximate

\[ \langle e^{-\sigma B(\vec{x})} \rangle \approx Z \equiv \sum_s e^{-\sigma_s} \sum_k c_k b_k(\vec{x}_s) \]

• Coordinate descent

• Interpretations:
  - Add weight to hard examples
  - Greedily add 1 rule per iteration
  - learn predictive features from data.
**trick #3:** boosted alternating decision trees

- **One tree:** control logic all genes, all expts

  1 interaction

  quantify regulation

  - play “20 questions”
  - output \( \log(p(+)/p(-)) \)
  - highly interpretable

[ADTs: Freund & Mason 1999]
gene-centric vs. expt-centric vs. integrative

Learn *regulatory program* that makes genome-wide, context-specific predictions for differential (up/down) expression of target genes.
yeast environmental stress response

- Gasch et al. (2000) dataset, 173 microarrays, 13 environmental stresses
- ~5500 target genes, 475 regulators (237 TF+ 250 SM)
- 500bp upstream promoter sequences
- Binning into +1/0/-1 expression levels based on wildtype vs.
basic notions: fitting vs. overfitting

- “10-fold cross-validation” yields test loss of 13.6%

<table>
<thead>
<tr>
<th>True Bins</th>
<th>Down</th>
<th>Baseline</th>
<th>Up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Down</td>
<td>16.5%</td>
<td>8.9%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Baseline</td>
<td>9.3%</td>
<td>32.4%</td>
<td>6.3%</td>
</tr>
<tr>
<td>Up</td>
<td>2.8%</td>
<td>9.9%</td>
<td>12.0%</td>
</tr>
</tbody>
</table>

- Empirical estimate of generalization error
- not chi squared (not training data, and not normal)
basic notions: mining vs. understanding

- **Test Loss** vs. "boosting iteration" = number of edges

- establish a baseline via randomizing
4th trick: learn predictive “f”+ motifs \textit{ab initio}

- Use \textit{boosting} to iteratively combine predictive regulators and motifs into a tree-structure.
- Alternating decision tree = margin-based generalization of decision trees.
- Learn motifs \textit{ab initio} from promoter sequences.
- Lower nodes are conditionally dependent on higher nodes $\Rightarrow$ can possibly reveal \textit{combinatorial interactions}. 
binding sites + “motif discovery”

Learning problems:

• Understand which regulators control which target genes

• Discover motifs representing regulatory elements
MEDUSA: why dimers?

ptashne’s “a genetic switch”
MEDUSA’s individual interactions

...AGCTATGCCATCGACTGCTCCAGTGCAGACACACCAAGATTTGAG
GCTATAGCTACTTTTATAAAGGGGCTACCGGCAAATT...

**k-mers (k≤7)**
- AGCTATG
- GCTATGC
- CTATGCC

**dimers (gapped elements)**
- TTT_AAA
- GCTA_GCTA

**Regulator expression**
- Is AGCTATG present and USV1 up?
- Is AGCTATG present and USV1 down?
- Is GCTATGC present and USV1 up?
- Is GCTATGC present and TPK1 up? ...

**boosting loss**
- minimizes boosting loss

**PSSMs**
- hierarchical sequence agglomeration

**Is**
- AGCTATGCC\_CATCGACTGCTCCAGTGCAGACACACCAAGATTTGAG
- GCTATAGCTACTTTTATAAAGGGGCTACCGGCAAATT...

**Is**
- AGCTATG\_present and USV1 up?
- GCTATGC\_present and USV1 up?
- TCTATGC\_present and USV1 up?
- GCTTTTG\_present and USV1 up?

**Try all motif-regulator pairs as individual interactions**

**minimize boosting loss**
⇒ **selected interaction**
- Is **bold AGCTATGC** present and USV1 up?
- Is **bold GCTTTTG** present and USV1 up? ...

**minimize boosting loss**
⇒ **selected interaction**
hierarchical sequence agglomeration

• Avoids masking of *correlated individual interactions*
• Improves prediction accuracy on test data

\[
PSSM\ p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i),\ x_i \in \{A, C, G, T\}
\]

score \[
S = \sum_{i=1}^{n} \ln \left( \frac{p_i(x_i)}{p^{bg}(x_i)} \right)
\]

\[
d(p, q) \equiv \min_{\text{offsets}} \left[ w_1 D_{KL}(p || w_1 p + w_2 q) + w_2 D_{KL}(q || w_1 p + w_2 q) \right],
\]

GCTATGC
GCAATGC
GGATGTC
CCTAAGC
GCTATT

... 

GGATATGG

PSSMs

...
MEDUSA: summary

1. Integrate sequence+ expression to learn a global regulatory program;
2. Avoid overfitting
3. Learn functional regulators-motif combos
4. Learn binding site motifs, and thresholds, directly from sequence without seeding
reminder: fitting vs. overfitting

- “10-fold cross-validation” yields test loss of 13.6%

<table>
<thead>
<tr>
<th>True Bins</th>
<th>Predicted Bins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Down</td>
<td>Baseline</td>
</tr>
<tr>
<td>Down</td>
<td>16.5%</td>
</tr>
<tr>
<td>Baseline</td>
<td>9.3%</td>
</tr>
<tr>
<td>Up</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

- Empirical estimate of generalization error
- not chi squared (not training data, and not normal)
basic notions: fitting vs. overfitting

- 10-fold cross-validation (held-out experiments), \(\sim 60,000\) (gene,experiment) training examples, 700 iterations
- \((N_{k\text{-mers}}+N_{\text{dimers}}+N_{\text{PSSMs}})\times N_{\text{reg}}\times 2 \approx 10^7\) possible individual interactions at every node
- \textit{MEDUSA}'s motifs give a \textit{better prediction accuracy} on held-out experiments than database motifs

<table>
<thead>
<tr>
<th></th>
<th>test-loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{MEDUSA}</td>
<td>13.4%</td>
</tr>
<tr>
<td>AlignACE (Pilpel et al. 2001)</td>
<td>16.1%</td>
</tr>
<tr>
<td>TRANSFAC</td>
<td>20.8%</td>
</tr>
</tbody>
</table>
**basic notions: fitting vs. overfitting**

- Large-scale results: yeast ESR data set, ~170 microarrays, 5-fold cross-validation (held-out experiments), ~60,000 (gene, experiment) training examples, 700 iterations
- *MEDUSA*’s motifs give a *better prediction accuracy* on *held-out experiments* than database motifs

<table>
<thead>
<tr>
<th></th>
<th>test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRANSFAC motifs +</td>
<td>31.3%</td>
</tr>
<tr>
<td>nearest neighbor</td>
<td></td>
</tr>
<tr>
<td>TRANFAC motifs +</td>
<td>20.8%</td>
</tr>
<tr>
<td>ADT</td>
<td></td>
</tr>
<tr>
<td>AlignACE motifs + ADT</td>
<td>16.1%</td>
</tr>
<tr>
<td>MEDUSA</td>
<td>13.4%</td>
</tr>
</tbody>
</table>
### MEDUSA: ab initio PSSM discovery

<table>
<thead>
<tr>
<th>TF name</th>
<th>MEDUSA logo</th>
<th>Pattern matched</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN2/4</td>
<td><img src="image1" alt="MEDUSA logo" /></td>
<td>AGGGGG</td>
<td>TRANSFAC Sites</td>
</tr>
<tr>
<td>HSF1</td>
<td><img src="image2" alt="MEDUSA logo" /></td>
<td>NGAANNTTCN</td>
<td>YPD</td>
</tr>
<tr>
<td>GIS1</td>
<td><img src="image3" alt="MEDUSA logo" /></td>
<td>AAGGGGAT</td>
<td>YPD</td>
</tr>
<tr>
<td>YAP1</td>
<td><img src="image4" alt="MEDUSA logo" /></td>
<td>AAGCCAC</td>
<td>YPD</td>
</tr>
<tr>
<td>RAP1</td>
<td><img src="image5" alt="MEDUSA logo" /></td>
<td>ATGTACGGGATG</td>
<td>YPD</td>
</tr>
<tr>
<td>RAP1</td>
<td><img src="image6" alt="MEDUSA logo" /></td>
<td>ACACCCCATACAT</td>
<td>YPD</td>
</tr>
</tbody>
</table>
### yeast ESR: biological validation

<table>
<thead>
<tr>
<th>TFNAME</th>
<th>DB-MOTIF</th>
<th>MOTIF</th>
<th>DBNAME</th>
<th>d(p,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBF1</td>
<td>CACGTG</td>
<td></td>
<td>YPD</td>
<td>0.032635</td>
</tr>
<tr>
<td>CGG everted repeat</td>
<td>CGGN*CCG</td>
<td></td>
<td>YPD</td>
<td>0.032821</td>
</tr>
<tr>
<td>MSN2</td>
<td>AGGGG</td>
<td></td>
<td>TRANSFAC</td>
<td>0.085626</td>
</tr>
<tr>
<td>HSF1</td>
<td>TTCCNNNGAA</td>
<td></td>
<td>SCPD</td>
<td>0.102410</td>
</tr>
<tr>
<td>XBP1</td>
<td>TCCCAAG</td>
<td></td>
<td>TRANSFAC</td>
<td>0.140561</td>
</tr>
<tr>
<td>STE12</td>
<td>TGAAAAC</td>
<td></td>
<td>TRANSFAC</td>
<td>0.256750</td>
</tr>
<tr>
<td>GCN4</td>
<td>TGAtGc</td>
<td></td>
<td>SCPD</td>
<td>0.292221</td>
</tr>
<tr>
<td>mPAC</td>
<td>TCTCATC</td>
<td></td>
<td>AlignACE</td>
<td>0.552493</td>
</tr>
<tr>
<td>mRRPE</td>
<td>AAAATTTT</td>
<td></td>
<td>AlignACE</td>
<td>0.630740</td>
</tr>
</tbody>
</table>

- **STRE element**
- **Heat shock factor**
Important regulators identified by MEDUSA

<table>
<thead>
<tr>
<th># of weak rules</th>
<th>regulator</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>96</td>
<td>TPK1</td>
<td>Cellular localization of MSN2/4</td>
</tr>
<tr>
<td>64</td>
<td>USV1</td>
<td>Segal et al. 2003</td>
</tr>
<tr>
<td>57</td>
<td>AFR1</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>XBP1</td>
<td>Universal stress repressor</td>
</tr>
<tr>
<td>19</td>
<td>ATG1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ETR1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>SDS22</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>CIN5</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>PDR3</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>GPA2</td>
<td></td>
</tr>
</tbody>
</table>
conclusions

• motif discovery + learning transcriptional regulation using *large-margin classification*
• learn binding sites *ab initio*
• PSSMs predictive on *test data*
• learn model of transcriptional regulation for *all genes* and *all experiments*
• simultaneous *discovery of important regulators*
• no gene clustering, no initialization
• open source:

http://www.cs.columbia.edu/compbio/medusa
agenda

• **Theme**: a predictive network model
  - predict expression
  - learn binding sites *ab initio*

• **Breakdown**: prediction? $y=f(x)$

• **Variation**: predicting evolution
  - validating models
  - letting the data decide
• **task:** learn *predictive* network from microarray and sequence data, w/o prior sequence annotation

• **tool:** \( y=f(x) \)

• **task:** predict evolution from topology

• **tool:** \( y=f(x) \)
what's so great about $y=f(x)$???!?!?!

1. nothing up my sleeve:
   CV: $y_V = f_L(x_V)$?
   sig. $y_V = f_R(x_V)$? $y_V = f_L(x_R)$?

2. which $x$ matter?
Statistical network physics: definition

Statistical analysis to reveal the mechanism responsible for an observed network topology

What information is in ?
agenda

- statistical network physics
  - pseudohistory
  - the problem

- statistical learning

- biological networks
statistical network physics: pseudohistory

1999-2001:

1. measure $p(k)$ for real networks
2. posit models/mechanisms:
   
   1. Erdos-Renyi $p(\omega) \sim 1$
   
   2. Yule/Simon/Barabasi-Albert $\dot{p}_k = f[k, p(k)]$

3. calculate $p(x) = \int_{\omega \in \Omega} d\omega p(x|\omega)p(\omega)$
4. select model which better agrees
1800s–:

1. measure $p(x)$ (or $\langle x \rangle$)
2. posit models, e.g.:
   
   $$ p(\omega) \sim e^{-E(\omega)/k_B T} $$
3. calculate
   
   $$ p(x) = \int_{\omega \in \Omega} d\omega p(x|\omega)p(\omega) $$
4. select model which best agrees
statistical network physics: measure

$p(k)$:

(c) World-Wide Web

(d) Internet

(f) protein interactions

Newman SIAM Review 2003
the problem:

informative statistics?
statistical network physics: history

1999-2001; 2001-2005

1. measure $p(k)$ for real networks
2. posit models/mechanisms:
   1. Erdos-Renyi $p(\omega) \sim 1$
   2. Yule/Simon/Barabasi-Albert $\dot{p}_k = f[k, p(k)]$
3. calculate $p(x) = \int_{\omega \in \Omega} d\omega p(x|\omega)p(\omega)$
4. mega)
5. select model which better agrees
6. overuniversality: almost all models can agree
proliferation of models (+metrics)

1. DMC  
   (Vazquez, Flammini, Maritan, Vespignani, 2003)
2. DMR  
   (Sole, Pastor-Satorras, Smith, Kepler, 2002)
3. RDS  
   (Erdos, Renyi, 1959)
4. RDG  
   (Callaway, Hopcroft, Kleinberg, Newman, Strogatz, 2001)
5. LPA  
   (Barabasi, Albert 1999)
6. AGV  
   (Klemm, Eguiluz, 2002)
7. SMW  
   (Watts, Strogatz 1998)
“First, power law distributions are neither new nor rare;
second, fitting available data to such distributions is suspiciously easy;
third, even when the fit is robust, it adds little if anything to our knowledge of the actual architecture of the network (many different architectures can give rise to the same power laws)”

- Revisiting “Scale-Free” Networks, E.F. Keller
inferring design in the presence of overuniversality for a target network

**algorithm:**

- forget your favorite design.
- forget your favorite feature.
- forget the target network.
- define a system for feature-generation.
- build a classifier to discriminate proposed designs.
- classify the target network.
1-slide summary of classification

- banana or orange?

boosting (1997)
SVMs (1990s)
1-slide summary of classification

- Watts or Barabasi?
1-slide summary of classification

- Watts or Barabasi?
- learn predictive (not “overrepresented”) features from data;
- no null model assumed;
- no distribution assumed;
calculate discriminative features

(and let the data decide which is best model)
agenda

- statistical network physics
  - the problem
  - probability
  - statistics
- statistical learning
- biological networks: predicting evolution
  - validating models
  - letting the data decide
systematic enumeration of network features

• Subgraph census
  - exploit sparseness ("walks")
  - use a pre-processed hash-table for subgraph isomorphisms
  - 148 subgraphs shown, can easily do 181 subgraphs
NetBoost: 20 questions

- play “20 questions”
- output \( \log(p(\text{model})/p(\text{not})) \)
- highly interpretable

[ADTs: Freund & Mason 1999]
conditionally important subgraphs
high accuracy (fit vs. overfit; test-loss)

Table 1. Prediction accuracy (%) for tested networks using fivefold cross-validation (13)

<table>
<thead>
<tr>
<th>Truth</th>
<th>DMR</th>
<th>DMC</th>
<th>AGV</th>
<th>LPA</th>
<th>SMW</th>
<th>RDS</th>
<th>RDG</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMR</td>
<td>99.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>DMC</td>
<td>0.0</td>
<td>99.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>AGV</td>
<td>0.0</td>
<td>0.1</td>
<td>84.7</td>
<td>13.5</td>
<td>1.2</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>LPA</td>
<td>0.0</td>
<td>0.0</td>
<td>10.3</td>
<td>89.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>SMW</td>
<td>0.0</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>99.0</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>RDS</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.8</td>
<td>99.0</td>
<td>0.0</td>
</tr>
<tr>
<td>RDG</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>99.0</td>
</tr>
</tbody>
</table>

- Empirical estimate of generalization error
- not chi squared (not normal, too many parts=parameters)
now look @ target: robust predictions

Dup/Mut/Comp Erdős
DMC?

(from Rice, Kershenbaum, Stolovitzky’s Commentary)
rank scores
not just for flies: yeast P-P network

data courtesy O. Troyanskaya
why subgraphs?
• Triad Census to test for **transitivity**, Holland and Leinhardt, 1970

‘FFL’ = \begin{align*}
&= '030T' triad
\end{align*}
subgraph census: problems

• Number of isomorphism classes grows rapidly with graph size (Haray, 1955)
  
  3 dyads
  16 triads
  218 tetrads
  9608 pentads

• Census sensitive to density, clustering, degree distributions
• Traditional algorithms limited to n=3 or n=4
• Larger structures require tailored, parameterized algorithms
systematic enumeration of network features

- Subgraph census
  - exploit sparseness ("walks")
  - use a pre-processed hash-table for subgraph isomorphisms
  - 148 subgraphs shown, can easily do 181 subgraphs

or

- Adjacency matrix functionals ("words") (Ziv et al. cond-mat/0306610)
  - more efficient than subgraph census for denser networks
  - up to 4670 features tested
• 030T (FFL) signature

\[ A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}; \text{diag}(A^2 A^T) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \]

Path operators
- \( A \) = adjacency; (walking down the graph)
- \( A^T \) = transpose; (walking up the graph)
- \( D \) = diag; (restriction to closed walks)
- \( U \) = \( I - D \); (restriction to open walks)
In other words ...

\[ \text{Number of FFLs} = \text{"sum } D(A^2A^T)\text{"} \]

Example 1: \( S(D(A^2A^T)) = 40 \) = the number of FFLs in the E. coli network

Example 2: \( \text{nnz}(D(A^2A^T)) = 10 \)
   16 of 40 FFLs associated with gene csgBA

\text{sum} \Rightarrow \text{number of distinct paths between all pairs of endpoints}
\text{nnz} \Rightarrow \text{number of distinct paths between unique pairs of endpoints}
Tunable, preferential-attachment (PA) parameter
• Barabasi and Albert, Science ‘99

Scalars perform better for networks that are dense, clustered, or networks with long-tailed degree distributions.
NetClass: predict mechanism as class

q-bio/0402017; BMC Bioinformatics 2004, 5:181
NetClass: *E. coli* Transcriptional Network

Kumar-C wins (words)

\[ N=324; m=519; d=.3\%; r=1.0 \]
NetClass: *E. coli* Transcriptional Network

Kumar-C wins (walks)
NetClass: *C. elegans* Neural Network

“MZ” wins
(new model)

\[ N=306; \, m=2359; \, d=2.5\%; \, r=.97 \]
what is important? let the data decide

Fig. 1. Discriminating similar networks. Ten graphs of two different mechanisms exhibit similar average geodesic lengths and almost identical degree distribution and clustering coefficients. (a) Cumulative degree distribution $p(k > k_0)$, average clustering coefficient $\langle C \rangle$ and average geodesic length $\langle \ell \rangle$, all quantities averaged over a set of 10 graphs. (b) Prediction scores for all 10 graphs and all five cross-validated (13) ADTs. The two sets of graphs can be perfectly separated by our classifier, even though none of these graphs is used in the classifier training.
statistical systems biology: agenda

1. challenges to keep in mind
2. microarrays / regulation
3. networks
4. final thoughts
things to watch out for:

1. methods / how to read
2. different data, same issues
3. “prediction”
4. validation
how to read/write a comp. sys. bio. paper:

1. background
2. intuition
3. question to be answered, in words
4. question to be answered, in math:
   \[ \vartheta = \underset{\vartheta \in \Omega}{\text{argmin}} \ L(D, \vartheta; \lambda) \]
5. algorithm
6. validation
“prediction”

1. overfitting
2. feature ranking / hypothesis generation ("qualitative predictions")
3. predicting unseen data
validation, closely related to prediction

1. in literature / by friends
2. statistical validation (e.g., CV)
3. experiment
different data, same issues

1. RNAi
2. ChIP-chip
3. PPI
4. image data
5. ...
learning networks from biology

• thanks:
  Freund, Kundaje, Leslie, Middendorf, + Shah

• for more info:
  • RECOMB, ISMB

• funding:
  • NIH NCBC

• open source.
learning biology from networks

• thanks: Middendorf, Ziv
• for more info: • BMC Bioinfo, PNAS
• funding: • NSF/NIH/DOE
• open source: • sourceforge.net