CSE 5559 – Translational Bioinformatics
Pathways, Networks, Graphs
Using Graphs

- **Graph**
  - Nodes
  - Edges
  - Directions
  - Edge labels
  - Weights

- **Graph theory**
  - Connectivity
  - Degree of nodes
  - Modularity

Wexner Medical Center
Examples

- Linear
  - Ligand
  - Receptor
  - Effector (E)
  - Action

- Divergent
  - Ligand
  - Receptor
  - E1: Action A
  - E2: Action B
  - E3: Action C

- Network
  - Ligand
  - Receptor
  - E1
  - E2
  - E3
  - A: Action
  - B: Action
  - C: Action
  - D: Action
  - E: Action

- Network with critical nodes
  - Ligand I
  - Receptor I
  - E1
  - E2a
  - E2b
  - Node 1
  - Action A
  - E3a
  - E3b
  - E3c
  - E3d
  - Node 2
  - E4a
  - E4b
  - E4c
  - Receptor II
  - E6
  - E7
  - E8a
  - E8b
  - Node 3
  - Node 4
  - Receptor IIIa
  - Receptor IIIb
  - Node 5
  - E10
  - E11
  - E12
  - E13
  - Action A
  - B
  - C
  - D
  - E
  - F
  - G
  - H
  - I
  - J

Copyright © 2006 Nature Publishing Group
Nature Reviews | Molecular Cell Biology

Wexner Medical Center
Types
Networks in biology

✓ Protein-protein interaction networks
✓ Gene regulatory networks
  - DNA-protein interaction networks
✓ Gene co-expression networks
  - transcript-transcript association networks
✓ Metabolic networks
✓ Signaling networks
✓ Neuronal networks
✓ Between-species interaction networks
✓ Within-species interaction networks
✓ And more …
Information Flow
Different Network Types

![Network Diagram]

- **Proteins**
- **Metabolites**
- **Metabolism**
- **Gene regulation**
- **Cell signaling**
- **PPIs**

- **A**
- **B**
- **C**
- **D**
- **E**
- **F**

**Nodes:**
- m₁
- m₂
- m₃

**Connections:**
- Metabolism
- Gene regulation
- Cell signaling
- PPIs

*Wexner Medical Center*
Different Network Elements

$X \rightarrow Y$ represents

transcription network

$X \rightarrow Y$

gene $x$  gene $y$

neuron synaptic connection network

ecological food web

Eat $Y$
Protein-Protein Interaction Network
Protein-Protein Interaction Network
A *protein-protein interaction (PPI)* usually refers to a physical interaction, i.e., binding between proteins.
PPI Network

PPIs are very important for structure and function of a cell:

- Participate in signal transduction
  - Play a role in many diseases (e.g., cancer)
- Can be stable interactions forming a functional protein complex
  : ribosome, hemoglobin – illustrated below)
PPI Network

• Interactions
  – Can be *transient*
    • Brief interactions to modify a protein & alter PPIs: protein kinases (add a phosphate group to a target protein)
    • A protein can carry another protein, e.g., *nuclear pore importins* (cytoplasm to nucleus and vice versa)
    • Transient interactions allow dynamic PPIs
      – About 70% of interactions are stable and 30% are dynamic

• Crucial for understanding disease, development of drugs
A network of protein–protein interactions in yeast
Benno Schwikowski, Peter Uetz & Stanley Fields
Gene Regulatory Network
Gene Regulatory Network

Cell cycle gene network

INPUT
signal A
receptor proteins
cascade of
interacting
kinase proteins
or other
molecules
active
transcription
factor A
DNA

e-JUN
NF1
MDM2

e-jun

p53

p53

p53

NF-kB

p21waf
p21WAF
DNA synthesis enzymes

P53

CycA

CYCA

CYC-D

E2F1

c2f-1

RB1

pRB

CDC2

cdc2

Wexner Medical Center
Protein-DNA interactions

▲ Chromatin IP
▼ DNA microarray

Gene levels (up/down)

Protein-protein interactions

▲ Protein coIP
▼ Mass spectrometry

Protein levels (present/absent)

Biochemical reactions

▲ none
▼ Metabolic flux measurements

Biochemical levels
Transcriptional Regulation

Gene regulation

– Gives a cell control over its structure and function, e.g.:

  • *Cellular differentiation* – a process by which a cell turns into a more specialized cell type
  • *Morphogenesis* (a process by which an organism develops its shape)
  • ...
Differentiation

+ Gene A is high in Stem cell
+ Gene B is high in Differentiated B cell.
+ Data shows that A high => B low
  (B high => A low)
+ Gene X associated with Differentiation?

Find all X s.t.
A high => X low and B high => X high

Sahoo et al. PNAS, 2011: **MiDReG**: A method of mining developmentally regulated genes using Boolean implications
Transcriptional Regulation

• Nodes correspond to genes
  – DNA sequences which are transcribed into mRNAs that translate into proteins

• Directed edges correspond to interactions through which the products of one gene affect those of another
  – Protein-protein, protein-DNA and protein-mRNA interactions
Transcriptional Regulation

*Transcription factor X* (protein product of gene X) binds regulatory DNA regions of gene Y to regulate the production rate (i.e., stimulate or repress transcription) of protein Y.

Note: proteins are products of gene expression that play a key role in regulation of gene expression.
E.Coli
Modeling
Example

Figure 7.1. The *E. coli* transcription factors CRP and LacI interact on the lac promoter to control gene expression. (A) Schematic of the various inputs to the lac promoter, the combinations of which determine the transcriptional state of the lac operon. (B) The fundamental elements we need for our modeling.
The classic real-world example of such regulation is the lac operon, whose gene products enable *E. coli* to grow on lactose, and whose expression depends on two transcription factors (Figure 7.1A). *E. coli* prefers to eat glucose, and it won’t metabolize anything else until the glucose is gone. In *E. coli*, this metabolic switching is accomplished with the transcription factor CRP (you encountered CRP in Figure 1.1). CRP binds the promoters of hundreds of genes once it is bound to cAMP, a small molecule whose presence indicates that none of *E. coli*’s favorite sugar sources are available. The CRP-cAMP complex can then bind to operator sites that control the expression of genes that enable the uptake and metabolism of other carbon sources.

One of these carbon sources is lactose, a sugar characterized by a particular link between galactose and glucose. Utilization of lactose depends on enzymes and a transporter, the genes for which appear together on the *E. coli* chromosome as a single transcription unit (an “operon”). The transcription of multiple genes as a single transcript is a more efficient way for bacteria to coordinate gene expression. Operons are common in bacteria, but are absent from more complex organisms, which tend to rely on complex post-transcriptional regulatory processes.

The promoter of the lac operon contains a binding site for the CRP-cAMP complex. Thus, the operon is only fully expressed in the presence of CRP-cAMP, which only appears in the absence of glucose. However, what if there is no lactose in the environment either? It doesn’t make sense to express the lac genes unless both glucose is absent and lactose is present. *E. coli* addresses this problem with another transcription factor: LacI (the “I” stands for “inhibitor”). Free LacI inhibits transcription by binding its own operator in the lac operon promoter. However, when lactose is present in the external environment, one of its metabolic products (allo-lactose) binds LacI, reducing LacI’s binding affinity to the operator and enabling the transcription of the lac operon.
Figure 7.2. Schematic of the complexity associated with eukaryotic transcriptional regulation. Compared to prokaryotic promoters, eukaryotic promoters are typically ten to hundreds of times longer and involve dozens of co-activators and additional transcription factors.
Figure 7.4. Possible instances of regulatory feed-forward loops. In coherent loops, both arms of the motif exert the same type of control (positive or negative). Incoherent loops have one positive arm and one negative arm. The red box highlights the two instances that were specifically overrepresented in both *E. coli* and yeast. [Permission language for Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2007. It’s figure 4.3 on page 47.]
Coherent FFL

Network motifs in the transcriptional regulation network of *Escherichia coli*
Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan & Uri Alon

Coherent FFL

Sign sensitive delay for ON signal
Some Math

Figure 7.5. Detailed schematic (A) and shorter notation (B) for a typical coherent feed-forward loop. Genes, proteins, and signals are defined in the main text. [Permission language for Alon, U. An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman & Hall/CRC, 2007. Figures 4.5 & 4.6 on page 49.]
Coherent FFL

Sign sensitive delay for ON signal
Coherent FFL

Network motifs in the transcriptional regulation network of *Escherichia coli*
Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan & Uri Alon

Nature Genetics 31, 64 - 68 (2002)
Summary

Table 1. Structure and function of the coherent FFL types, with AND- and OR-gates at the Z promoter

<table>
<thead>
<tr>
<th>Coherent type 1</th>
<th>Coherent type 2</th>
<th>Coherent type 3</th>
<th>Coherent type 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Species</strong></td>
<td><strong>Structure</strong></td>
<td><strong>Abundance</strong></td>
<td><strong>Structure</strong></td>
</tr>
<tr>
<td>E. coli</td>
<td><img src="image1" alt="Diagram" /></td>
<td>X 28</td>
<td><img src="image2" alt="Diagram" /></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td><img src="image5" alt="Diagram" /></td>
<td>Z 26</td>
<td><img src="image6" alt="Diagram" /></td>
</tr>
</tbody>
</table>

**Z Logic:---**

- **AND**
- **OR**

**Steady-state**
- Z(x, y)
- Response delay
- Sx on step
- Sx off step
- Sy off step
- Inverted out

**Incoherent FFL types and their abundance:**
- Coherent FFL types and their abundance in transcription databases (E. coli: 8, 11, Z(x, y): Steady-state Z expression of coherent FFLs for the four combinations of Sx and Sy on and off levels (Y = 0, Y = 1). Response delay of coherent FFLs to on and off Sx steps in the presence of Sy = 0, not delayed. Inverted out means that Sx goes off in response to Sy on step.

Table 2. Structure and function of the incoherent FFL types, with AND-gates at the Z promoter

<table>
<thead>
<tr>
<th>Incoherent type 1</th>
<th>Incoherent type 2</th>
<th>Incoherent type 3</th>
<th>Incoherent type 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Species</strong></td>
<td><strong>Structure</strong></td>
<td><strong>Abundance</strong></td>
<td><strong>Structure</strong></td>
</tr>
<tr>
<td>E. coli</td>
<td><img src="image9" alt="Diagram" /></td>
<td>X 5</td>
<td><img src="image10" alt="Diagram" /></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td><img src="image13" alt="Diagram" /></td>
<td>Z 21</td>
<td><img src="image14" alt="Diagram" /></td>
</tr>
</tbody>
</table>

**Z Logic:---**

- **AND**
- **AND**

- **Steady-state**
- Z(x, y)
- Response delay
- Sx on step
- Sx off step
- Sy off step
- Step effect

- **Response acceleration**
  - Sx on step
  - Sx off step

- Incoherent FFL types and their abundance in transcription databases (E. coli: 8, 11, Z(x, y): Steady-state Z expression of incoherent FFL with no basal level of Y (Y = 0, Y = 1). Response delay of incoherent FFLs to on and off Sx steps in the presence of Sy = 0, not delayed. Inverted out means that Sx goes off in response to Sy on step. Response acceleration: Acceleration of response of and steady-state values of incoherent FFL with basal activity to on and off steps in the presence of Sy = 0, not accelerated.
### Summary

<table>
<thead>
<tr>
<th>Function</th>
<th>Circuit class</th>
<th>Circuit types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steady-state logic is sensitive to both Sx and Sy</td>
<td>Coherent and Incoherent*</td>
<td>Types 1, 2 AND</td>
</tr>
<tr>
<td>Sign-sensitive delay upon Sx steps</td>
<td>Coherent</td>
<td>Types 3, 4 OR</td>
</tr>
<tr>
<td>Sy-gated pulse generator upon Sx steps</td>
<td>Incoherent with no basal Y level</td>
<td>Types 1, 2 OR</td>
</tr>
<tr>
<td>Sign-sensitive acceleration upon Sx steps</td>
<td>Incoherent with basal Y level</td>
<td>Types 1, 2, 3, 4</td>
</tr>
<tr>
<td>Cooperativity enhancement for Sx input</td>
<td>Coherent</td>
<td>Type 1 AND</td>
</tr>
</tbody>
</table>

*In incoherent FFL with basal level, Sy modulates Z between two nonzero levels.*
The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks
Complex FFLs
Other Kinds

Cell cycle
- p21
- p27
- p57
  - Cyclin-CDK

Transcription
- p21 → E2F-1
- p21 → STAT-3
- p21 → c-Myc
- p27 → Neurogenin-2
- p57 → MyoD
- p300/CBP

Cytoskeleton
- p27
  - RhoA
- p21
  - ROCK
- p57
  - LIMK
  - cofilin

Apoptosis
- p57
- p21
  - JNK/SAPK
  - ASK1/MEKK5
- p21
  - Procaspase-3
- p21
  - Caspase-3
- Actin remodeling
- Cell movement

Stress-induced apoptosis
- Fas-mediated apoptosis

Graphs and diagrams showing time courses for different processes.
<table>
<thead>
<tr>
<th>Case</th>
<th>Regulation factor ($F_{reg}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Simple repressor</td>
<td>$(1 + r)^{-1}$</td>
</tr>
<tr>
<td>2. Simple activator</td>
<td>$\frac{e_{\text{act}}}{1 + a \cdot \frac{e_{\text{act}}}{K_a \cdot T}}$</td>
</tr>
<tr>
<td>3. Activator recruited by a helper ($H$)</td>
<td>$\frac{1}{1 + a \cdot \frac{h_e}{K_B \cdot T}}$</td>
</tr>
<tr>
<td>4. Repressor recruited by a helper ($H$)</td>
<td>$\left(1 + \frac{1}{1 + \frac{h_e}{K_B \cdot T}}\right)^{-1}$</td>
</tr>
<tr>
<td>5. Dual repressors</td>
<td>$(1 + r_1)(1 + r_2)$</td>
</tr>
<tr>
<td>6. Dual repressors interacting</td>
<td>$\left(1 + r_1 + r_2 + r_1 r_2 e^{-\frac{e_{\text{act}}}{K_B \cdot T}}\right)^{-1}$</td>
</tr>
<tr>
<td>7. Dual activators interacting</td>
<td>$\frac{e_{\text{act}}}{1 + a_1 e^{-\frac{e_{\text{act}}}{K_B \cdot T}} + a_2 e^{-\frac{e_{\text{act}}}{K_B \cdot T}} + a_1 a_2 e^{-\frac{e_{\text{act}}}{K_B \cdot T}}}$</td>
</tr>
<tr>
<td>8. Dual activators cooperating via looping</td>
<td>$\left(1 + \frac{e_{\text{act}}}{1 + a_1 e^{-\frac{e_{\text{act}}}{K_B \cdot T}} + a_2 e^{-\frac{e_{\text{act}}}{K_B \cdot T}} + a_1 a_2 e^{-\frac{e_{\text{act}}}{K_B \cdot T}}}</td>
</tr>
</tbody>
</table><p>ight)^{-1}$ |
| 9. Repressor with two DNA binding units and DNA looping | $\left(1 + \frac{r_m + \frac{\Delta e_{\text{act}}}{1 + e^{-\frac{e_{\text{act}}}{K_B \cdot T}}}}{1 + \frac{r_s}{1 + r_s}}\right)^{-1}$ |</p>
TF Business

• Available for *model organisms*
  – Non-human species manipulated and studied to get insights into workings of other organisms, e.g.:
    • Baker's yeast, *S. cerevisiae* (Milo et al., 2002)
    • *E. coli* (Shen-Orr et al., 2002)
    • Sea urchin (Davidson et al., 2002)
    • Fruitfly, *D. melanogaster*
  – Available from: EcoCyc, GeneNet, KEGG, RegulonDB, Reactom, TRANSPATH, TRANSFAC
Representation of the *E. coli* transcriptional regulatory network. (a) Representation of the transcription-factor gene regulatory network of *E. coli*. Green circles represent transcription factors, brown circles denote regulated genes, and those with both functions are coloured in red. Projections of the network onto (b) transcription factor and onto (c) regulated gene nodes are also shown.

Metabolic Network Model

Escherichia coli
Metabolic Reactions

a

ATP
ADP
UDP
ATP
ADP
UTP
ATP
ADP
CTP

Mg²⁺

Mg²⁺

NH₂⁺

Orthophosphate

b

UMP
UDP
ADP
ATP
UTP
CTP

NH₂⁺

orthophosphate

c

UMP
UDP
UTP
CTP
Other Crazy Networks
### Table I  Sickle cell disease

<table>
<thead>
<tr>
<th>Primary molecular abnormality (disease genome or proteome)</th>
<th>Disease-modifying genes or proteins (secondary disease genome or proteome)</th>
<th>Intermediate phenotype (response genome or proteome and pathological manifestations)</th>
<th>Environmental determinant</th>
<th>Pathophenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hb A Val6Glu</td>
<td>Hb F</td>
<td>Thrombosis</td>
<td>Hypoxia</td>
<td>Hemolytic anemia</td>
</tr>
<tr>
<td></td>
<td>Hb C</td>
<td>Inflammation</td>
<td>Dehydration</td>
<td>Aplastic anemia</td>
</tr>
<tr>
<td></td>
<td>β-Thalassemia</td>
<td>Immune response</td>
<td>Infective agent</td>
<td>Stroke</td>
</tr>
<tr>
<td></td>
<td>G6PD</td>
<td>Fibrosis</td>
<td></td>
<td>Bone infarction</td>
</tr>
<tr>
<td></td>
<td>TGF-β</td>
<td>Apoptosis/necrosis</td>
<td></td>
<td>Painful crisis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acute chest syndrome</td>
</tr>
</tbody>
</table>

Abbreviation: Hb, hemoglobin.

### Table II  Pulmonary arterial hypertension

<table>
<thead>
<tr>
<th>Primary molecular abnormality (disease genome or proteome)</th>
<th>Disease-modifying genes or proteins (secondary disease genome or proteome)</th>
<th>Intermediate phenotype (response genome or proteome and pathological manifestations)</th>
<th>Environmental determinant</th>
<th>Pathophenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMPR-2 mutations</td>
<td>5-HT2B</td>
<td>Thrombosis</td>
<td>Hypoxia</td>
<td>Pulmonary hypertension</td>
</tr>
<tr>
<td>Alk-1 mutations</td>
<td>5-HTT</td>
<td>Vasospasm</td>
<td>Infective agent (HIV, HHV-8)</td>
<td>Cor pulmonale</td>
</tr>
<tr>
<td>Endoglin mutations</td>
<td>Thromboxane synthetase</td>
<td>Inflammation</td>
<td>Crotolaria sp./toxin</td>
<td>Pulmonary thromboembolism</td>
</tr>
<tr>
<td></td>
<td>Prostacyclin synthetase</td>
<td>Fibrosis</td>
<td>Cocaine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5-Lipoxygenase</td>
<td>Proliferation</td>
<td>Anorexigens</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NADPH oxidase</td>
<td>Immune response</td>
<td>Alcoholic cirrhosis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Endothelin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hemoglobinopathies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hereditary spherocytosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HHT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thrombocytosis</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: 5-HTT, serotonin transporter; 5-HT2b, serotonin 2b receptor; HHT hereditary hemorrhagic telangiectasias.
Figure 2 (A) Theoretical human disease network illustrating the relationships among genetic and environmental determinants of the pathophenotypes. Key: G, primary disease genome or proteome; D, secondary disease genome or proteome; I, intermediate phenotype; E, environmental determinants; PS, pathophysiological states leading to P, pathophenotype. (B) Example of this theoretical construct applied to sickle cell disease. Key: red, primary molecular abnormality; gray, disease-modifying genes; yellow, intermediate phenotypes; green, environmental determinants; blue, pathophenotypes.
Figure 2  Drug–target network (DT network). The DT network is generated by using the known associations between FDA-approved drugs and their target proteins. Circles and rectangles correspond to drugs and target proteins, respectively. A link is placed between a drug node and a target node if the protein is a known target of that drug. The area of the drug (protein) node is proportional to the number of targets that the drug has (the number of drugs targeting the protein). Color codes are given in the legend. Drug nodes (circles) are colored according to their Anatomical Therapeutic Chemical Classification, and the target proteins (rectangular boxes) are colored according to their cellular component obtained from the Gene Ontology database.
An Important One
Gene Co-Expression Network

Nature Communications 4, Article number: 1701 doi:10.1038/ncomms2660
Gene Co-Expression Network


CC: extracellular matrix
BP: immune system process
BP: cell cycle process
BP: translational termination
CC: cytosolic ribosome
Networks and Network Topology
Network Representation

- gene A regulates gene B
- gene A binds gene B
- reaction product is a substrate for gene B

regulatory interactions (protein-DNA)
functional complex B is a substrate of A (protein-protein)
metabolic pathways
Different Network Elements

\[ X \rightarrow Y \text{ represents} \]

**transcription network**

\[ \text{gene } x \quad \text{gene } y \]

**neuron synaptic connection network**

**ecological food web**

\[ \text{Eat} \quad Y \]
Formal Definition

A *network* (or *graph*)?

A set of nodes (vertices) and edges (links)

Edges describe a relationship between the nodes
Network Measures

- Degree $k_i$
- Degree distribution $P(k)$
- Mean path length
- Network Diameter
- Clustering Coefficient
Network Analysis

Paths: metabolic, signaling pathways

Cliques: protein complexes

Hubs: regulatory modules

Subgraphs: motifs maximally weighted
What to learn from networks?

- Relationships
- Community information
- Logical or dynamic information
- (Sometimes) $1 + 1 > 2$ in networks
- Need new theories from physics, mathematics, statistics, ...

Wexner Medical Center
Graphs

- Graph $G = (V,E)$ is a set of vertices $V$ and edges $E$

- A subgraph $G'$ of $G$ is induced by some $V' \subset V$ and $E' \subset E$

- Graph properties:
  - Connectivity (node degree, paths)
  - Cyclic vs. acyclic
  - Directed vs. undirected
Subgraphs

- Subgraph: a connected graph consisting of a subset of the nodes and links of a network
- Subgraph properties:
  
  n: number of nodes

  m: number of links

(n=3, m=3)  

(n=3, m=2)  

(n=4, m=4)  

(n=4, m=5)
<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>( N_{\text{real}} )</th>
<th>( N_{\text{rand}} \pm SD ) Z score</th>
<th>( N_{\text{real}} )</th>
<th>( N_{\text{rand}} \pm SD ) Z score</th>
<th>( N_{\text{real}} )</th>
<th>( N_{\text{rand}} \pm SD ) Z score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene regulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(transcription)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>424</td>
<td>519</td>
<td>40 7 ± 3</td>
<td>10</td>
<td>203 47 ± 12</td>
<td>13</td>
<td>1812 360 ± 40</td>
<td>41</td>
</tr>
<tr>
<td>S. cerevisiae*</td>
<td>685</td>
<td>1,052</td>
<td>70 11 ± 4</td>
<td>14</td>
<td>1812 360 ± 40</td>
<td>41</td>
<td>1812 360 ± 40</td>
<td>41</td>
</tr>
<tr>
<td>Neurotransmitters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. elegans†</td>
<td>252</td>
<td>509</td>
<td>128 90 ± 10</td>
<td>3.7</td>
<td>127 55 ± 13</td>
<td>5.3</td>
<td>227 35 ± 10</td>
<td>20</td>
</tr>
<tr>
<td>Food webs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little Rock</td>
<td>92</td>
<td>984</td>
<td>3219 3120 ± 50</td>
<td>2.1</td>
<td>7295 2220 ± 210</td>
<td>25</td>
<td>1337 230 ± 50</td>
<td>23</td>
</tr>
<tr>
<td>Yihan</td>
<td>92</td>
<td>984</td>
<td>1182 1020 ± 20</td>
<td>7.2</td>
<td>1337 230 ± 50</td>
<td>23</td>
<td>1337 230 ± 50</td>
<td>23</td>
</tr>
<tr>
<td>St. Martin</td>
<td>42</td>
<td>205</td>
<td>469 450 ± 10</td>
<td>NS</td>
<td>382 130 ± 20</td>
<td>12</td>
<td>382 130 ± 20</td>
<td>12</td>
</tr>
<tr>
<td>Chesapeake</td>
<td>31</td>
<td>67</td>
<td>80 82 ± 4</td>
<td>NS</td>
<td>26 5 ± 2</td>
<td>8</td>
<td>26 5 ± 2</td>
<td>8</td>
</tr>
<tr>
<td>Coachella</td>
<td>29</td>
<td>243</td>
<td>279 235 ± 12</td>
<td>3.6</td>
<td>181 80 ± 20</td>
<td>5</td>
<td>181 80 ± 20</td>
<td>5</td>
</tr>
<tr>
<td>Skidaway</td>
<td>25</td>
<td>190</td>
<td>184 150 ± 7</td>
<td>5.5</td>
<td>377 80 ± 25</td>
<td>13</td>
<td>377 80 ± 25</td>
<td>13</td>
</tr>
<tr>
<td>B. Brook</td>
<td>25</td>
<td>104</td>
<td>181 130 ± 7</td>
<td>7.4</td>
<td>267 30 ± 7</td>
<td>32</td>
<td>267 30 ± 7</td>
<td>32</td>
</tr>
<tr>
<td>Electronic circuits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(forward logic chips)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s15850</td>
<td>10,383</td>
<td>14,240</td>
<td>424 2 ± 2</td>
<td>285</td>
<td>1040 1 ± 1</td>
<td>1200</td>
<td>480 2 ± 1</td>
<td>335</td>
</tr>
<tr>
<td>s38584</td>
<td>20,717</td>
<td>34,204</td>
<td>413 10 ± 3</td>
<td>120</td>
<td>1739 6 ± 2</td>
<td>800</td>
<td>711 9 ± 2</td>
<td>320</td>
</tr>
<tr>
<td>s38417</td>
<td>23,843</td>
<td>33,661</td>
<td>612 3 ± 2</td>
<td>400</td>
<td>2404 1 ± 1</td>
<td>2550</td>
<td>531 2 ± 2</td>
<td>340</td>
</tr>
<tr>
<td>s5234</td>
<td>5,844</td>
<td>8,197</td>
<td>211 2 ± 1</td>
<td>140</td>
<td>794 1 ± 1</td>
<td>1050</td>
<td>209 1 ± 1</td>
<td>230</td>
</tr>
<tr>
<td>s13207</td>
<td>8,651</td>
<td>11,831</td>
<td>403 2 ± 1</td>
<td>225</td>
<td>4445 1 ± 1</td>
<td>4950</td>
<td>264 2 ± 1</td>
<td>200</td>
</tr>
<tr>
<td>Electronic circuits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(digital fractional multipliers)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s208</td>
<td>122</td>
<td>189</td>
<td>10 1 ± 1</td>
<td>9</td>
<td>4 1 ± 1</td>
<td>3.8</td>
<td>5 1 ± 1</td>
<td>5</td>
</tr>
<tr>
<td>s420</td>
<td>252</td>
<td>399</td>
<td>20 1 ± 1</td>
<td>18</td>
<td>10 1 ± 1</td>
<td>10</td>
<td>11 1 ± 1</td>
<td>11</td>
</tr>
<tr>
<td>s838†</td>
<td>512</td>
<td>819</td>
<td>40 1 ± 1</td>
<td>38</td>
<td>22 1 ± 1</td>
<td>20</td>
<td>23 1 ± 1</td>
<td>25</td>
</tr>
<tr>
<td>World Wide Web</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nd.edu§</td>
<td>325,729</td>
<td>1.46e6</td>
<td>1.1e5 2e3 ± 1e2</td>
<td>800</td>
<td>6.8e6 5e4±4e2</td>
<td>15,000</td>
<td>1.2e6 1e4 ± 2e2</td>
<td>5000</td>
</tr>
</tbody>
</table>


Wexner Medical Center
Motif Topology

Each edge has 4 choices (why?). Three edges $4 \times 4 \times 4 = 64$ choices. There are symmetry redundancy. Despite the choices of activation and repression, there are 13 types.
Sparse vs Dense

- $G(V, E)$ where $|V|=n$, $|E|=m$ the number of vertices and edges

- Graph is **sparse** if $m \sim n$

- Graph is **dense** if $m \sim n^2$

- **Complete graph** when $m=n^2$
Connected Components

• $G(V,E)$
• $|V| = 69$
• $|E| = 71$
Connected Components

- \( G(V,E) \)
- \(|V| = 69\)
- \(|E| = 71\)
- 6 connected components
Paths

A path is a sequence \( \{x_1, x_2, \ldots, x_n\} \) such that \((x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n)\) are edges of the graph.

A closed path \(x_n = x_1\) on a graph is called a graph cycle or circuit.
Shortest-Path between nodes
Shortest-Path between nodes
Longest Shortest-Path
Scale Free Networks
Small-world Network

• Every node can be reached from every other by a small number of hops or steps

• High clustering coefficient and low mean-shortest path length
  – Random graphs don’t necessarily have high clustering coefficients

• Social networks, the Internet, and biological networks all exhibit small-world network characteristics
Small world phenomena

\[ P(k) \sim k^{-\gamma} \]


Wexner Medical Center
Network Measures: Degree

a  Undirected network

B

A

C

D

E

F

G

k = 5

b  Directed network

B

E

F

G

A

D

C

k_{in} = 4

k_{out} = 1
Degree Distribution

\( P(k) \) is probability of each degree \( k \), i.e. fraction of nodes having that degree.

For random networks, \( P(k) \) is normally distributed.

For real networks the distribution is often a power-law:

\[
P(k) \sim k^{-\gamma}
\]

Such networks are said to be scale-free.
Metabolic network

Organisms from all three domains of life are scale-free networks!
Scale-Free vs Random

(a) Random network  (b) Scale-free network
Hubs & Power Law

Small World - Rich Get Richer (preferential attachment) Self-similarity

HUBS!
The huba-Hub


Surfing the p53 network

Bert Vogelstein, David Lane and Arnold J. Levine

The p53 tumour-suppressor gene integrates numerous signals that control cell life and death. As when a highly connected node in the Internet breaks down, the disruption of p53 has severe consequences.

“One way to understand the p53 network is to compare it to the Internet. The cell, like the Internet, appears to be a ‘scale-free network’.”
All About Hubs

• Cellular networks are assortative, hubs tend not to interact directly with other hubs.

• Hubs tend to be “older” proteins (so far claimed for protein-protein interaction networks only)

• Hubs also seem to have more evolutionary pressure—their protein sequences are more conserved than average between species (shown in yeast vs. worm)

• Experimentally determined protein complexes tend to contain solely essential or non-essential proteins—further evidence for modularity.
Hubs and Robustness

Complex systems maintain their basic functions even under errors/ failures (cell $\rightarrow$ mutations; Internet $\rightarrow$ router breakdowns)

Hubs and Robustness

Complex systems maintain their basic functions even under errors and failures (cell → mutations; Internet → router breakdowns)

Knock-out Lethality/Connectivity

\[ y = 1.2x^{-1.91} \]

![Graph showing the relationship between Knock-out Lethality/Connectivity and Degree k. The graph includes a linear regression line.](image)
Target the hubs to have an efficient safe sex education campaign

Lewin Bo, et al., Sex i Sverige; Om sexuallivet i Sverige 1996, Folkhälsoinstitutet, 1998
Yeast protein network

- lethality and topological position

Highly connected proteins are more essential (lethal)...


Wexner Medical Center
Achilles’ Heel

The density of the network surrounding node $I$, characterized as the number of triangles through $I$. Related to network modularity.

\[ C_I = \frac{n_I}{\binom{k}{2}} = \frac{2n_I}{k \cdot (k-1)} \]

- $k$: neighbors of $I$
- $n_I$: edges between node $I$'s neighbors

The center node has 8 (grey) neighbors
There are 4 edges between the neighbors

\[ C = \frac{2 \cdot 4}{(8 \cdot (8-1))} = \frac{8}{56} = \frac{1}{7} \]
Hierarchical Networks

B  Scale-free network

Ba

C  Hierarchical network

Ca
Detecting Hierarchical Org.
Summary

• **Degree** $k_i$
  The number of edges involving node $i$

• **Degree distribution** $P(k)$
  The probability (frequency) of nodes of degree $k$

• **Mean path length**
  The avg. shortest path between all node pairs

• **Network Diameter**
  – i.e. the longest shortest path

• **Clustering Coefficient**
  – A high CC is found for modules

• **Motifs**
  – Finding them is hard but useful
Revisit your old ideas with new eyes.

This is an announcement from Jereneville Community Service.
Gene Co-Expression

Co-expressed!

Q: What could be the mechanisms and implications for co-expression?
A Noteworthy Example

Network modeling links breast cancer susceptibility and centrosome dysfunction

Miguel Angel Pujana1,2,16,17, Jing-Dong J Han1,2,16,17, Lea M Starita3,16,17, Kristen N Stevens4,17, Muneesh Tewari1,2,16, Jin Sook Ahn1,2, Gad Rennert5, Victor Moreno6,7, Tomas Kirchhoff8, Bert Gold9, Volker Assmann10, Wael M ElShamy2, Jean-François Rual1,2, Douglas Levine8, Laura S Rozek6, Rebecca S Gelman11, Kristin C Gunsalus12, Roger A Greenberg2, Bijan Sobhian2, Nicolas Bertin1,2, Kavitha Venkatesan1,2, Nono Ayivi-Guedehoussou1,2,16, Xavier Sole7, Pilar Hernández13, Conxi Lázaro13, Katherine L Nathanson14, Barbara I Weber14, Michael E Cusick1,2, David E Hill1,2, Kenneth Offit8, David M Livingston2, Stephen B Gruber4,6,15, Jeffrey D Parvin5,16 & Marc Vidal1,2

Gene Co-Expression Network to Infer Disease Gene
A Noteworthy Example

Figure 5. Identification of HMMR siRNA
A. Proteins known to be associated with HPRD database of microarray datasets can be identified using the a large scale of PCC > 0.4, a high correlation.
B. Using PCC > 0.4 as a low limit, genes co-expressed with BRCA1, ATM, BRCA2, and CHEK2 were identified. The intersection should have a low false positive rate and was defined as the BCN.
Inferring Regulatory Networks
Common methods

- Mutual information based methods – e.g., ARACNE
- Bayesian network based methods
- Network hierarchy
- Integrate transcription factor binding data – ChIP-chip, ChIP-seq, ENCODE
- Integrate gene expression data
- Time course data
- DREAM Contests
2015

Alzheimer’s Disease Big Data DREAM Challenge #1
Predict the best biomarkers for early AD-related cognitive decline and for the mismatch between high amyloic levels and cognitive decline.

The Broad-DREAM Gene Essentiality Prediction Challenge
Develop predictive models to infer genes that are essential to cancer cell viability using gene expression and/or gene copy number features.

The DREAM9 Acute Myeloid Leukemia (AML) Outcome Prediction Challenge
Predict the outcome of treatment of AML patients (resistant or remission), their remission duration and overall survival based on clinical cytogenetics, known genetics markers and phosphoproteomic data.
2016 – Somatic Mutation Calling

https://www.synapse.org/#!Synapse:syn2813581/wiki/303137

Challenges
The ICGC-TCGADREAM Somatic Mutation Calling - Tumour Heterogeneity Challenge (SMC-Het) is an international effort to improve standard methods for subclonal reconstruction: to quantify and genotype each individual cell population present within a tumor. Leaders of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) cancer genomics projects are joining with Sage Bionetworks and IBM-DREAM to initiate this innovative open crowd-sourced Challenge [1-3].

The goal of this challenge is to identify the best subclonal reconstruction algorithms and to identify the conditions that affect their performance. By running an unbiased, comprehensive evaluation, the most appropriate method(s) will be identified.

**Subchallenge 1:** Identify the purity of the tumour sample, the number of subclones present in the sample, and for each subclone the cellular prevalence and number of mutations in that subclone.

**Subchallenge 2:** Identify which mutations define each subpopulation.

**Subchallenge 3:** Identify the evolutionary relationship between the subpopulations.
BioVis 2016 + DREAM

Subchallenge 4: Visualize
Gene co-expression network

- Horvath’s framework
- http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/
  - Overview
  - Paper 1
  - Paper 6

![Gene expression network graph](image)
An Overview of Weighted Gene Co-Expression Network Analysis

Steve Horvath@UCLA
Software and Data Availability

Sample data and R software tutorials
http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

An R package and accompanying tutorials
http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/

Tutorial for WGCNA R package
http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA/TutorialWGCNApackage.doc
Contents

• How to construct a weighted gene co-expression network?
• Why use soft thresholding?
• How to detect network modules?
• How to relate modules to an external clinical trait?
• What is intramodular connectivity?
• How to use networks for gene screening?
• How to integrate networks with genetic marker data?
• What is weighted gene co-expression network analysis (WGCNA)?
• What is neighborhood analysis?
Philosophy of WGCNA

• Understand the “system” instead of reporting a list of individual parts
  – Describe the functioning of the engine instead of enumerating individual nuts and bolts
• Focus on modules as opposed to individual genes
  – this greatly alleviates multiple testing problem
• Network terminology is intuitive to biologists
How to construct a weighted gene co-expression network?

A network can be represented by an adjacency matrix, $A = [a_{ij}]$, that encodes whether/how a pair of nodes is connected:

- $A$ is a symmetric matrix with entries in $[0,1]$.
- For unweighted network, entries are 1 or 0 depending on whether or not 2 nodes are adjacent (connected).
- For weighted networks, the adjacency matrix reports the connection strength between gene pairs.
Co-expression Network

A) Microarray gene expression data

B) Measure concordance of gene expression with a Pearson correlation

C) The Pearson correlation matrix is either dichotomized to arrive at an adjacency matrix \( \rightarrow \) unweighted network

Or transformed continuously with the power adjacency function \( \rightarrow \) weighted network
Power adjacency function

\[ a_{ij} = | \text{cor}(x_i, x_j) |^\beta \]

Often choosing beta=6 works well but in general we use the “scale free topology criterion” described in Zhang and Horvath 2005.
Comparing Adjacency Functions

Power Adjacency vs Step Function
Comparing Adjacency Functions

• While network analysis results are highly robust with respect to the network construction method there are several reasons for preferring the power adjacency function.

  – Empirical finding: Network results are highly robust with respect to the choice of the power beta
    • Zhang B and Horvath S (2005)

  – Theoretical finding: Network Concepts make more sense in terms of the module eigengene.
How to detect network modules?
Module Definition

• Numerous methods exist

• Hovrath uses average linkage hierarchical clustering coupled with topological overlap dissimilarity measure.

• Once a dendrogram is obtained from a hierarchical clustering method, we choose a height cutoff to arrive at a clustering.

• Modules correspond to branches of the dendrogram
TOM metric in clustering

\[
TOM_{ij} = \sum_u a_{iu} a_{uj} + a_{ij} \\
\text{DistTOM}_{ij} = 1 - TOM_{ij}
\]

- Generalized in Zhang and Horvath (2005) to the case of weighted networks
- Generalized in Yip and Horvath (2006) to higher order interactions
TOM to Cluster Genes

Here modules correspond to branches of the dendrogram

**TOM plot**

Genes correspond to rows and columns

Hierarchical clustering dendrogram
Genetics of gene expression and its effect on disease

Valur Emilsson1,2, Gudmar Thorleifsson1, Bin Zhang1, Amy S. Leonardson2, Florian Zink1, Jun Zhu2, Sonia Carlson2, Agnar Helgason1, G. Bragi Walters1, Steinunn Gunnarsdottir2, Magali Moxy1, Valgerdur Steinhorsdottir1, Gudrun H. Einarsdottir2, Gyda Bjornsdottir1, Inga Reynisdottir1, Daníel Guðbjartsson1, Anna Helgadottir2, Aslaug Jonasdottir1, Adalbjorg Jonasdottir1, Unnur Stykarsdottir1, Solveig Gretarsdottir1, Kristen P. Magnusson1, Hreinn Stefansson1, Ragnheidur Fosdahl1, Kristleifur Kristjansson1, Hjortur G. Gislason1, Tryggvi Stefansson2, Bjorn G. Leifsson1, Unnur Thorsteinsdottir1, John R. Lamb2, Jeffrey R. Gulcher1, Marc L. Reibman3, Augustine Kong1, Eric E. Schadt1# & Kari Stefansson1#
Depicting Gene Modules

1) Rows and columns correspond to genes
2) Red boxes along diagonal are modules
3) Color bands = modules

Idea:
Use network distance in MDS
Heatmap

Columns= tissue samples

Rows=Genes Color band indicates module membership

Message: characteristic vertical bands indicate tight co-expression of module genes
Module Eigengene

measure of over-expression=average redness

Rows=genes, Columns=microarray

The brown module eigengenes across samples
Mantra

Module eigengenes can be used to determine whether 2 modules are correlated. If correlation of MEs is high-> consider merging.
Eigengenes can be used to build separate networks…
Consensus eigengene networks in male and female mouse liver data and their relationship to physiological traits

How to relate modules to external data?
Clinical Trait

- Abstract definition of a gene significance measure
  - GS(i) is non-negative,
  - the bigger, the more *biologically* significant for the i-th gene

Equivalent definitions
- GS.ClinicalTrait(i) = |cor(x(i),ClinicalTrait)| where x(i) is the gene expression profile of the i-th gene
- GS(i)=|T-test(i)| of differential expression between groups defined by the trait
- GS(i)=-log(p-value)
Significance

A SNP marker naturally gives rise to a measure of gene significance

\[ GS.SNP(i) = |\text{cor}(x(i), \text{SNP})| \]

- Additive SNP marker coding: AA->2, AB->1, BB->0
- Absolute value of the correlation ensures that this is equivalent to AA->0, AB->1, BB->2
  - Dominant or recessive coding may be more appropriate in some situations
  - Conceptually related to a LOD score at the SNP marker for the i-th gene expression trait
Significance

A gene significance naturally gives rise to a module significance measure

- Define module significance as mean gene significance
- Often highly related to the correlation between module eigengene and trait
**Important Task in Many Genomic Applications:**

Given a network (pathway) of interacting genes how to find the central players?
Hubs

The nodes with the largest number of links (connections) are most important!

**Slide courtesy of A Barabasi**
What is intramodular connectivity?
Generalized Connectivity

- Gene connectivity = row sum of the adjacency matrix
  - For unweighted networks = number of direct neighbors
  - For weighted networks = sum of connection strengths to other nodes

\[ k_i = \sum_j a_{ij} \]
Comparison

Gene significance vs. intramodular connectivity

- $\text{cor} = -0.34, p = 2.6\times10^{-5}$
- $\text{cor} = 0.069, p = 0.14$
- $\text{cor} = 0.19, p = 7.5\times10^{-5}$
- $\text{cor} = 0.31, p = 4.1\times10^{-6}$
- $\text{cor} = -0.28, p < 10^{-20}$
- $\text{cor} = -0.16, p = 0.096$
- $\text{cor} = 0.44, p = 6.8\times10^{-9}$
- $\text{cor} = 0.3, p = 9.3\times10^{-16}$
- $\text{cor} = 0.27, p = 2.7\times10^{-7}$
How to use networks for gene screening?
Comparison

Gene significance vs. intramodular connectivity

• Note the relatively high correlation between gene significance and intramodular connectivity in some modules
• In general, kIN is a more reliable measure than GS
• In practice, a combination of GS and k should be used
• Module eigengene turns out to be the most highly connected gene (under mild assumptions)
What is weighted gene co-expression network analysis?
Workflow

Construct a network
Rationale: make use of interaction patterns between genes

Identify modules
Rationale: module (pathway) based analysis

Relate modules to external information
Array Information: Clinical data, SNPs, proteomics
Gene Information: gene ontology, EASE, IPA
Rationale: find biologically interesting modules

Study Module Preservation across different data
Rationale:
• Same data: to check robustness of module definition
• Different data: to find interesting modules
Workflow (contd.)

Find the key drivers in interesting modules
Tools: intramodular connectivity, causality testing
Rationale: experimental validation, therapeutics, biomarkers
Difference with Other Methods

- Emphasis on modules (pathways) instead of individual genes
  - Greatly alleviates the problem of multiple comparisons
    - Less than 20 comparisons versus 20000 comparisons
- Use of intramodular connectivity to find key drivers
  - Quantifies module membership (centrality)
  - Highly connected genes have an increased chance of validation
- Module definition is based on gene expression data
  - No prior pathway information is used for module definition
  - Two module (eigengenes) can be highly correlated
- Emphasis on a unified approach for relating variables
  - Default: power of a correlation
  - Rationale:
    - puts different data sets on the same mathematical footing
    - Considers effect size estimates (cor) and significance level
    - $p$-values are highly affected by sample sizes ($\text{cor}=0.01$ is highly significant when dealing with 100000 observations)
- Technical Details: soft thresholding with the power adjacency function, topological overlap matrix to measure interconnectedness
Case Study 1: Finding brain cancer genes

Modules Cancer & Normal

55 Brain Tumors
VALIDATION DATA: 65 Brain Tumors

Messages:
1) Cancer modules can be independently validated
2) Modules in brain cancer tissue can also be found in normal, non-brain tissue.

--> Insights into the biology of cancer
Mean Prognostic Significance

Message: Focus the attention on the brown module genes
Hub Genes – Cancer Survival

1. Cox model to regress survival on gene expression levels
2. Defined prognostic significance as $-\log_{10}(\text{Cox-p-value})$ the survival association between each gene and glioblastoma patient survival
3. A module-based measure of gene connectivity significantly and reproducibly identifies the genes that most strongly predict patient survival

Test set – 55 gbms
$r = 0.56; p = 2.2 \times 10^{-16}$

Validation set – 65 gbms
$r = 0.55; p = 2.2 \times 10^{-16}$
Hub Genes – Cancer Survival

The fact that genes with high intramodular connectivity are more likely to be prognostically significant facilitates a novel screening strategy for finding prognostic genes

- Focus on those genes with significant Cox regression p-value AND high intramodular connectivity.
  - It is essential to take a module centric view: focus on intramodular connectivity of disease related module

- Validation success rate = proportion of genes with independent test set Cox regression p-value < 0.05.
- Validation success rate of network based screening approach (68%)
- Standard approach involving top 300 most significant genes: 26%
Validation

300 most significant genes (Cox p-value < $1.3 \times 10^{-3}$)

Network based screening
p < 0.05 and high intramodular connectivity

26%

67%
The network-based approach uncovers novel therapeutic targets

Five of the top six hub genes in the mitosis module are already known cancer targets: topoisomerase II, Rac1, TPX2, EZH2 and KIF14. We hypothesized that the 6-th gene ASPM gene is novel therapeutic target. ASPM encodes the human ortholog of a drosophila mitotic spindle protein.

Biological validation: siRNA mediated inhibition of ASPM
Questions

- Expansion
  - Negative correlation
  - Multiple breast cancer datasets
  - More anchor genes
  - ...

- Is there a way to find all highly correlated genes in multiple datasets?

- Do these genes form clusters?
One Can Mine Too
Frequent network mining

CODENSE

- Originally applied to yeast microarray data, later expanded to cancers
- Used for functional annotation

BIOINFORMATICS

Mining coherent dense subgraphs across massive biological networks for functional discovery
Haiyan Hu¹, Xifeng Yan², Yu Huang¹, Jiawei Han² and Xianghong Jasmine Zhou¹,*
Data Selection & Correlation

• 23 datasets from Gene Expression Omnibus (GEO)
  – Search term “human metastatic cancer”
  – Contain both control and tumor, # sample > 15
  – Only primary biopsy

• Correlation – PCC > 0.75 (really high similarity)

• For CODENSE
  – Edge support in at least 4 datasets
  – Connectivity ratio $r > 40\%$ ($L > r \cdot n(n-1)/2$)
  – # of nodes > 20
Results from CODENSE

- 44 networks are identified
- # of nodes: 21 ~ 74 (average 44)
- Connectivity: 0.41 ~ 0.78
GO Enrichment Analysis

- Networks with at least 1/3 of the genes
  - Immune response/system – 15
  - Protein translation (ribosome) – 5
  - Development – 4
  - Metabolism and energy (oxidative phosphorylation or monocarboxylic acid metabolism) – 3
  - Cell cycle – 2
  - Muscle contraction – 1

- 14 networks do NOT satisfy the above criterion
  - Potential new functions
  - New interactions
Cluster 2

- Predict survival outcome
- NKI-295 dataset
- Supervised clustering: k-means, k=2, 100 random initialization
- Kaplan Meier curve and log-rank test for survival analysis and comparison
- Test for different patient groups
Predict Survival Outcome

---

**Graphs:**

1. **List 1:** All 295 Samples, p = 6.7302e-011, N₁ = 115, N₂ = 180
2. **List 2:** All 295 Samples, p = 4.4656e-010, N₁ = 149, N₂ = 145
3. **List 3:** All 295 Samples, p = 1.5317e-008, N₁ = 155, N₂ = 139

---

**Survival Ratio vs. Survival Time (Year):**

- **Graph 1:**
  - Survival Ratio: 1, 0.5, 0.7, 0.8, 0.9, 1
  - Survival Time (Year): 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20

- **Graph 2:**
  - Survival Ratio range: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
  - Survival Time (Year) range: 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20

---

**Institution:**

Wexner Medical Center
Predict Survival Outcome

Vant Veer 70 Genes – Lymph Node Positive Samples p=0.00026302 \( N_1 = 55 \), \( N_2 = 59 \)

List2 – Lymph Node Positive Samples p=1.4822e-006 \( N_1 = 83 \), \( N_2 = 60 \)

List3 – Lymph Node Positive Samples p=1.4877e-005 \( N_1 = 78 \), \( N_2 = 67 \)
Relation to BRCA1
Co-expression Networks

Cancer tissue merged networks

Normal tissue merged networks

Compare Frequent Ones

Predict New Gene Functions

siRNA: PPAF
KIAA0101

siRNA: KIF23

siRNA: KIF14
CLL Prognostic Biomarker

- CLL is the most common adult leukemia in the western world. It is highly heterogeneous, can be indolent or progressive.
- Prognosis at early stage is crucial to progressive patient survival as well as to indolent patients to avoid unnecessary adverse treatment.
- Biological prognostic markers:
  - Serum markers (TK, B2M, sCD23)
  - FISH cytogenetics
  - IgV\textsubscript{\textsc{h}} mutational status - Determination is time consuming and expansive
  - CD\textsc{38} expression - Actually independent of IgV\textsubscript{\textsc{h}} mutational status
  - ZAP-70 expression - Not 100% correlated to IgV\textsubscript{\textsc{h}} mutational status, only accurate when patients in the progressive stage
Network 17

- 51 genes, including ZAP-70 and CD38
- $r = 0.4142$
- Including known ZAP-70 interacting genes - CD8A, CD3G, CD3D, CD247
Immune Response
Workflow of CLL Prognostic Biomarker Selection

Further select prognostic biomarkers by testing on separate CLL dataset

mRMR

Select a group of feature genes that can differentiate IgV_H mu +/- groups

40

Compute gene exp level difference on IgV_H mu +/- groups

40

Genes with exp fold change > 1.5 p <0.05

11

Cross validation

Test the prediction accuracy of each gene on IgV_H mutation status

6

Identify potential prognostic biomarkers

5
Biomarkers for CLL

Predicting IgVH mutation status using GDS1454

<table>
<thead>
<tr>
<th>Genes</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SH2D1A</td>
<td>57.32%</td>
</tr>
<tr>
<td>IL2RB</td>
<td>68.84%</td>
</tr>
<tr>
<td>KLRK1</td>
<td>63.67%</td>
</tr>
<tr>
<td>CD247</td>
<td>66.03%</td>
</tr>
<tr>
<td>GZMB</td>
<td>57.13%</td>
</tr>
<tr>
<td>CD3G</td>
<td>62.52%</td>
</tr>
<tr>
<td>CD3D</td>
<td>64.27%</td>
</tr>
<tr>
<td>GZMK</td>
<td>57.58%</td>
</tr>
<tr>
<td>CD8A</td>
<td>68.31%</td>
</tr>
<tr>
<td>NKG7</td>
<td>64.94%</td>
</tr>
<tr>
<td>ZAP70</td>
<td>68.46%</td>
</tr>
<tr>
<td>LAG3</td>
<td>59.53%</td>
</tr>
<tr>
<td>ZAP70+IL2RB</td>
<td>73.22%</td>
</tr>
<tr>
<td>ZAP70+IL2RB+CD8A</td>
<td>74.62%</td>
</tr>
</tbody>
</table>
Outcome Data

- **LAG3**: involved in T-cell-dependent B-cell activation, reported recently to be highly correlated to \( \text{IgV}_H \) mutational status.

- **IL2RB**: involved in endocytosis and transduction of mitogenic signal of IL2, expression on B-cells was linked to CLL.

- **CD8A**: expression of CD8A on B-cells has been linked to CLL.

- **KLRK1**: involved in immune surveillance exerted by T/B-cells.

Using GSE10138

Kaplan-Meier Curves, Log-rank test \( p = 0.033418 \)

[Graph showing Kaplan-Meier curves with time to treatment in years on the x-axis and ratio on the y-axis.]
Condition Specific Co-Expression

Potential Mechanism

Network 1 (43) - TF enrichment analysis

LEF1 (V$LEF1_Q2)
FOXO4 (V$FOXO4_01)
E12 (V$E12_Q6)
FOXO4, LEF1 (V$FOXO4_01,V$LEF1_Q2)
NFAT (V$NFAT_Q4_01)
MAZ (V$MAZ_Q6)
(V$TATA_01)
AP4 (V$AP4_Q5)
LEF1, NFAT (V$LEF1_Q2,V$NFAT_Q4_01)
Signal transduction (BP)
SP1 (V$SP1_Q6)
ATF3 (V$ATF3_Q6)
hsa-miR-190
has-miR-616
LEF1, E12 (V$LEF1_Q2,V$E12_Q6)

LEF1 is highly enriched (p = 3.34317 × 10^{-10})
Expression levels of LEF1 between non-basal & basal decreased (fold change = 2.48)

Wexner Medical Center
Visualization
BoxPlots

> set.seed(10)
> x=rnorm(2000)
> boxplot(x,col="yellow")

FIGURE 17.1: A box plot in R. The central line in the box shows the position of the median (the value located halfway between the largest and smallest data value). For these settings, the upper and lower boundaries of the box show the location of the upper quartile (UQ) and lower quartile (LQ), respectively. The upper and lower quartiles are the 75th and 25th percentiles, respectively. The data outside the ends of the tails are considered outliers.
BoxPlots

- **UQ, LQ**
- **Inter Quartile Distance, IQD** - length between quartiles
- **Length of tails** = 1.5 * IQD
- **Outliers:**
  - > UQ + 1.5 * IQD
  - < LQ + 1.5 * IQD
Boxplots

**FIGURE 17.2**: Multiple box plots in R. Drawing box plots of several data sets next to each other allows a quick visual comparison between these data sets.
BoxPlots

**Figure 17.3**: Box plot of a data frame showing the data grouped by the various levels of a given factor.

```r
> mydat = data.frame(y = c(100, 90, 105, 83, 78, 95, 93, 79, 85,
                        + 90, 70, 72, 81, 74, 75), Group = factor(rep(c("mRNA1", "mRNA2",
                        + "mRNA3"), each = 5)))
> boxplot(y ~ Group, data = mydat, col = c("darksalmon", "deepskyblue1",
                        + "darkseagreen"))
```
**FIGURE 17.4:** A gene pie plot. Gene pies convey information about both ratios and intensities. The maximum intensity is encoded in the diameter of the pie chart while the ratio is represented by the relative proportion of the two colors within any pie chart. The net effect is that the genes that might be biologically relevant are made more conspicuous independently on their ratio (which may not always be meaningful).
A Workhorse
Scatter Plot

**FIGURE 17.5:** Expression levels in two experiments visualized as a scatter plot (synthetic data). The control is plotted on the horizontal axis, while the experiment is plotted on the vertical axis. Points above the diagonal $y = x$ represent genes with expression levels higher in the experiment (e.g., gene B), whereas points below the diagonal represent genes with expression levels higher in control (gene A).
Scatter Plot

+ Banana Blob

+ All genes seem down regulated

+ Or is it a dye problem?

FIGURE 17.6: Typical graph of the cy5 (vertical axis) versus cy3 (horizontal axis) scatter plot of the raw data obtained from a cDNA experiment (real data). The cy3 was used to label the control sample; the cy5 was used to label the experiment sample. The straight line represents the first diagonal cy3=cy5.
M-A Plot

FIGURE 17.10: An MA plot in R. The plot shows the log of the ratio $\log_2(\frac{\text{sample1}}{\text{sample5}})$, as a function of the average log intensity $\log_2(\frac{\text{sample1} \ast \text{sample5}}{2})$. 
MA-Plot(2)

A: \( \log(cy3) + \log(cy5) = \log(cy3 \cdot cy5) \)

M: \( \log(cy5/cy3) = \log(cy5) - \log(cy3) \)

\( Y=0 \) — line ?

**FIGURE 17.7:** In this ratio versus intensity plot, the horizontal axis is used to plot the log of the product of the intensities (which is the sum of the log intensities) of the two channels. Sometimes, the mean of the log intensities is used instead of the sum on the horizontal axis. The vertical axis plots the ratio between the channels. The typical raw data set will exhibit the same banana shape visible in the cy3 versus cy5 plot.
M-A Plot(3)

- High variance at low intensities
- Low variance at high intensities

FIGURE 17.8: Typically, the data will exhibit high variance at low intensities and lower variance at higher intensities. This will produce a funnel shape. This funnel shape is more accentuated in the ratio plots such as this one. Note that here the data have been normalized and most of the points lie near the y = 0 reference.
Scatter Plots Have Problems
17.4.3 Scatter plot summary

1. The scatter plot is a two- or three-dimensional plot in which a vector is plotted as a point having the coordinates equal to the components of the vector.

2. Scatter plots are highly intuitive and easy to understand. Limitations include the reduced number of dimensions in which data can be plotted.

3. In a scatter plot of two expression experiments, including many genes or whole genomes, in which each experiment is represented on an axis, most data points are expected to lie around the line \( y = x \) (or \( z = y = x \)). The assumption here is that most genes will not change. Points far from the diagonal are differentially expressed in the experiments plotted. The assumption may not hold if the set of genes has been preselected in some relevant way.

4. In a scatter plot of two expression experiments, including many genes or whole genomes in which the ratio is plotted against an intensity, most data points are expected to lie around the line \( y = 0 \).

5. A consistent departure from the reference lines above may indicate a systematic trend in the data and the need for a normalization procedure.

6. Many scatter plots of microarray data will exhibit a funnel shape with the wider variance towards low intensities.
Eine Kleine Plug

Visualizing Multidimensional Data with Glyph SPLOMs

A. Yains*, A. Webb*, H. Chamberlin*, K. Huang*, and R. Machtanju*
* The Ohio State University, USA

SPLOM
+ Summarization

HEATMAP
+ Details
Scatter Plot Matrix
Scatter Plot

> intens=2^eset
> plot(intens[,1],intens[,5],cex=1,pch=19, + col="mediumaquamarine")

> A=1/2*log2(intens[,1]*intens[,5])
> M=log2(intens[,1]/intens[,5])
> plot(A,M,cex=0.75,pch=19,col="steelblue")
Classifying Scatter Plots

% P. access to drinking water

Life exp. at birth years: female

Income per person

Deaths from AIDS per 100k

Infant Mortality Rate per 1k

% P. under 15 years old

Wexner Medical Center
Classify Relationships

elt-3

elt-1
Glyph SPLOM

Explore complex dependencies visually

Top left is 4 gene expressions in human brain. Bottom right is synthetic data in different patterns.
Strength of Relationship

Pearson \((X,Y)\)

Spearman\((X,Y)\)

\[ Y \]

\[ X \]
Distance Correlation
Distance Correlation – All pairs

Let $X$ and $Y$ be random variables of length $n$ $(X_k, Y_k)$, and let $A$ and $B$ be matrices of all centered $n$ by $n$ pairwise distances between elements of $X$ and $Y$. That is,

- $a_{k,\ell} := \|X_k - X_\ell\|$ and $b_{k,\ell} := \|Y_k - Y_\ell\|$ for $k, \ell = 1, 2, \ldots, n$.
- $A_{k,\ell} := a_{k,\ell} - \bar{a}_k - \bar{a}_\ell + \bar{a}$.
- $B_{k,\ell} := b_{k,\ell} - \bar{b}_k - \bar{b}_\ell + \bar{b}$.

The squared distance covariance is the arithmetic average of the element-wise products of $A$ and $B$.

$$dCov_n^2(X, Y) := \frac{1}{n^2} \sum_{k,\ell} A_{k,\ell} B_{k,\ell} \quad \text{for } k, \ell = 1, 2, \ldots, n \tag{1}$$

Likewise, the squared distance variance is the squared sample distance covariance of a variable with itself.

$$dVar_n^2(X) := dCov_n^2(X, X) = \frac{1}{n^2} \sum_{k,\ell} A_{k,\ell}^2 \tag{2}$$

Then the distance correlation between $X$ and $Y$ is the distance covariance of $X$ and $Y$ divided by the square root of the products of the distance variances of $X$ and $Y$.

$$dCor(X, Y) = \frac{dCov(X, Y)}{\sqrt{dVar(X) \cdot dVar(Y)}}$$ \tag{3}
Cluster & Visualize
The Prize – Necessary Network

Wexner Medical Center
Eine Kleine Plug

Visualizing Multidimensional Data with Glyph SPLOMs

A. Yains*, A. Webb*, H. Chamberlin*, K. Huang*, and K. Machiraju*
* The Ohio State University, USA

SPLOM + Summarization

HEATMAP + Details
The p-value vs. Fold Change
**Volcano Plot**

**Why V-shape?**

*FIGURE 17.11:* A volcano plot shows the fold changes on the horizontal axis, and the p-values on the vertical axis, for every probe or gene. Thresholds for both fold changes and p-values can be applied simultaneously. In this plot, the red probes meet both criteria: a fold change more than two-fold (±1 in log2 scale), and a p-value more significant than 0.001 (− log2(p) > 3).
Where do we look?
Volcano Plots

```r
> myvolcano = function(myx, myy, vt1, vt2, ht, myxlabel, myylabel,
+                  + cex.lab=1.2,...){
+    plot(myx, myy, pch=18, col=colors()[552], xlab=myxlabel,
+         ylab=myylabel, cex.lab=cex.lab,...)
+    points(myx[myx<vt2 & myx>vt1 & myy<ht], myy[myx<vt2 &
+         myx>vt1 & myy<ht], pch=18, cex=1.05)
+    abline(v=vt1, lty=2)
+    abline(v=vt2, lty=2)
+    abline(h=ht, lty=2)
+ }

> myvolcano(log2(peDat[, "FC"]), -log10(peDat[, "p.value"]), -1, 1, 3,
+          "log2(fold changes)", "-log10(p-values)"
```
Visualizing Distributions
Histogram
Histograms in R

> par(mfrow=c(2,2))
> hist(eset[,1], breaks=5, col="peachpuff")
Histograms in R

```r
> hist(eset[,1], breaks=10, col="seagreen")
> hist(eset[,1], breaks=20, col="yellow2")
> hist(eset[,1], breaks=50, col="sienna")
> par(mfrow=c(1,1))
```

**FIGURE 17.16:** Histograms in R. The same data are plotted in histograms with 5 (top left), 10 (top right), 20 (bottom left), and 50 (bottom right) equally spaced bins.
Adaptive Binning

```r
> par(mfrow=c(2,2))
> hist(eset[,1], breaks="Scott", col="violetred")
> hist(eset[,1], breaks=c(min(eset[,1]), 6, 6.5, 7, max(eset[,1])),
   + col="yellowgreen")
> hist(eset[,1], breaks=function(x){(max(x)-min(x))/2},
   + col="turquoise")
> hist(eset[,1], breaks="FD", col="wheat")
> par(mfrow=c(1,1))
```

**FIGURE 17.17:** Histograms in R. The same data are plotted in histograms in which the number of bins was calculated with the “Scott” algorithm (top left), arbitrary custom bins defined by the user (top right), calculated with a user-defined function (bottom left), and calculated with the “FD” built-in algorithm (bottom right).
Histogram vs pdf

FIGURE 17.18: The same function `hist` can be used to plot either the raw frequency (left) or the probability density (right) for each bin. The shapes of the two histograms are the same, but note the units on the vertical axis and the meaning of the respective heights of each bin. In the frequency plot, the total area of the bins will sum up to the number of values; in the probability density plot, the total area will be 1.
FIGURE 17.19: The histogram of the normalized log ratios can be used to select the genes that have a minimum desired fold change. In this histogram, the tails beyond ±0.5 contain the genes with a fold change of more than 1.4 \((\log_2 x = 0.5 \Rightarrow x = 2^{0.5} = 1.41)\).
Is there really a problem?
Choose the right bin size

FIGURE 17.15: The histogram of the first experiment in the yeast sporulation data constructed using 88 bins. The data are the same as in Fig. 17.12, but this histogram does not have the artifactual gap near the peak.
Artifact of Binning

**FIGURE 17.14:** An artifact of the binning process. From left to right, the same data are shown with bin sizes of 0.25, 0.33, and 0.5, respectively. Note that all three histograms suggest a clear tendency for higher frequencies at lower values. The data are in fact uniformly distributed, and this tendency is entirely an artifact of the binning process and the relatively small sample size (26 data points).
Sample size

**FIGURE 17.13**: The effect of the bin size on the shape of the histogram. From top to bottom and left to right, the data are exactly the same, but there are 100, 40, 20 and 10 bins, respectively. If bins are very narrow, few values fall into each bin; as bins get wider, more values fall into each bin, and the shape of the histogram changes.
17.6.1 Histograms summary

- Histograms are plots of data counts as a function of the data values. Usually, they are drawn using bars. A value $x$ that occurs $y$ times in the data set will be represented by a bar of height $y$ plotted at location $x$.

- The histogram provides information about the distribution of the data and can be used in certain situations as an empirical approximation of a probability density function (pdf).

- The exact shape of a histogram depends on the number of data collected and the size of the bins used. A small sample size and/or the binning process may create gross artifacts distorting the nature of the data distribution. Such artifacts may be detected by comparing the shapes of several histograms constructed using different number of bins.

- Two experiments can be compared by constructing the histogram of the ratios of the corresponding values. If the experiments involve a large number of genes and the data are suitably preprocessed and normalized, the histogram of the ratios is expected to be centered on either zero (if logs are used) or 1 (no logs are used) and be approximatively symmetrical.

- Differentially expressed genes will be found in the tails of such a histogram.
FIGURE 17.24: The data used for the PCA example. These genes were constructed by adding random noise to each prototype in Fig. 17.22 as well as by adding 150 completely random genes. Each prototype was the base for constructing 50 genes of similar behavior. The set shown here includes a total of 1,050 genes.
Visualization of Graphs
The authorship network of genome-wide association studies

Brendan K. Bulik-Sullivan & Patrick F. Sullivan

Affiliations | Corresponding author

Nature Genetics 44, 113 (2012) | doi:10.1038/ng.1052
Published online 27 January 2012

To the Editor:

Genome-wide association studies (GWAS) of complex human traits have become an important approach in human genetics. Taken together, GWAS are arguably the largest biological investigations of humans ever conducted. The total number of people genotyped to date with a GWAS array is difficult to know but probably exceeds 1,000,000. Major findings from these studies are that many common diseases have a polygenic architecture, the genetic effect sizes of common SNP variants are small, the identification of the involvement of genes and biological processes not previously suspected, and the association of some loci with different diseases. Critically, the sample sizes necessary to identify robust and replicable findings are beyond those achievable by single groups, and collaborations have rapidly evolved to augment statistical power.
Data visualization

- Example

Network visualization

Cytoscape, Gephi, NodeXL, etc.

Graph layout is computationally challenging (demo using NodeXL)
Eine Grosse Problem
E. Coli metabolism

KEGG: Kyoto Encyclopedia of Genes and Genomes
(http://www.genome.ad.jp/kegg/kegg.html)
HairBalls Everywhere
Combing Hairballs
Graph Solutions (Layouts)
Matrices
Layouts

Figure 7.24: Eight visual encodings of the same tree dataset, using different combinations of visual channels. a) Rectilinear vertical node-link, using connection, with vertical spatial position showing tree depth. b) Rectilinear horizontal layered node-link, using connection, with horizontal spatial position showing tree depth. c) icicle, with vertical spatial position and size showing tree depth, and horizontal spatial position showing link relationships. d) Radial node-link, using connection, with radial spatial position showing tree depth. e) Concentric circles, with spatial position and size showing tree depth and radial spatial position showing link relationships. f) Nested circles, using radial containment, with nesting level and size showing tree depth. g) Treemap, using rectilinear containment, with nesting level and size showing tree depth. h) Indented outline, using spatial position channels, with horizontal spatial position showing tree depth and (hmm). From [McGuffin and Robert 10], Figure 1. (Permission granted.)
Layouts

- Maximize space for color coding
- High Information density
- Cannot use white space

Space Filling

- Containment marks
- Area Marks and spatial position channels
Graphviz

http://graphviz.org/
Node Link Diagrams
Tidier Drawings of Trees

EDWARD M. REINGOLD AND JOHN S. TILFORD

Abstract—Various algorithms have been proposed for producing tidy drawings of trees—drawings that are aesthetically pleasing and use minimum drawing space. We show that these algorithms contain some difficulties that lead to aesthetically unpleasing, wider than necessary drawings. We then present a new algorithm with comparable time and storage requirements that produces tidier drawings. Generalizations to forests and m-ary trees are discussed, as are some problems in discretization when alphanumeric output devices are used.

Index Terms—Data structures, trees, tree structures.

INTRODUCTION

In a recent article [6], Wetherell and Shannon presented algorithms for producing “tidy” drawings of trees—drawings that use as little space as possible while satisfying certain aesthetics. The basic task is the assignment of x and y coordinates to each node of a tree after which a straightforward
Tillford-Reingold

Fig. 2. Example tree as drawn by Algorithm TR.
Tillford-Reingold

Criteria:
- Nodes layered by depth in tree
- No edge crossings
- Similar subtrees drawn in similar ways
- Compact representation

Approach:
- Top down recursive approach
- For each parent make sure every subtree is drawn
- Pack subtrees as closely as possible
- Center parent over subtrees
Method

Sample
Compare right and left trees and compress
Method

Center parent over subtrees and update
Example - PPI
Methods for Visual Understanding of Hierarchical System Structures

KOZO SUGIYAMA, MEMBER, IEEE, SHOJIRO TAGAWA, AND MITSUHIKO TODA, MEMBER, IEEE

Abstract—Two kinds of new methods are developed to obtain effective representations of hierarchies automatically: theoretical and heuristic methods. The methods determine the positions of vertices in two steps. First the order of the vertices in each level is determined to reduce the number of crossings of edges. Then horizontal positions of the vertices are determined to improve further the readability of drawings. The theoretical methods are useful in recognizing the nature of the problem, and the heuristic methods make it possible to enlarge the size of hierarchies with which we can deal. Performance tests of the heuristic methods and several applications are presented.

DIGRAPHS are widely utilized in modeling structures of complex systems in various fields where vertices correspond to elements of the systems and edges correspond to relations among the elements. It is empirically recognized that drawings of the digraphs are useful as a visual aid to understand overall images of the structures of the complex systems. For example, block diagrams and flowcharts are commonly used by engineers in performing tasks such as structural modeling, project scheduling, computer programming, etc.

Multilevel digraphs, called hierarchies, constitute an important subclass of digraphs. Interpretive Structural Modeling (ISM) [1] and Program Evaluation and Review Technique (PERT) [2] are the well-known techniques in which hierarchies are utilized for modeling structures of systems.

0018-9472/81/0200-0109$00.75 ©1981 IEEE
Sujiyama

- Great for graphs that have an intrinsic ordering
- ‘Depth’ in graph mapped to one axis
Fig. 2. An example of the improvements of a drawing. Step II: (a) $\rightarrow$ (b). Step III: (b) $\rightarrow$ (c).
## Steps

<table>
<thead>
<tr>
<th>Basic rules of drawing</th>
<th>Functions</th>
<th>Readability elements</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constraints for algorithms</td>
<td>Elements (A), (C₁)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### STEP I
- **Transformation into proper hierarchy**

### STEP II
- **Reduction of the number of crossings**
  - Element (B)
  - Penalty Minimization (PM) method
  - Barycentric (BC) method

### STEP III
- **Improvement of horizontal positions of vertices**
  - Elements (C₂), (D), (E)
  - Quadratic Programming (QP) method
  - Priority (PR) method

### STEP IV
- **Display on graphic terminals**

---

**Element A**: “Hierarchical” layout of vertices.
- The latter, traceability of paths, is broken down into the following four readability elements.

**Element B**: “Less-crossings” of lines (edges). The greatest difficulty in tracing paths is line crossings.

**Element C**: “Straightness” of lines. It is easy to trace straight lines. This element is further broken down into “straightness” of one-span edges (Element C₁) and “straightness” of long span edges (Element C₂).

**Element D**: “Close” layout of vertices connected to each other. It is desirable that paths are short.

**Element E**: “Balanced” layout of lines coming into or going from a vertex. This means that the structural information on branching and joining of paths is drawn clearly.
Step I

- Create layering of graph
  - From domain specific knowledge
  - No cycles
  - Longest path from root
  - Algorithmically determine best layering (NP-Hard)
- Dummy nodes for long edges
Step II

- Layer by layer crossing minimization (NP)
- Number of heuristics available
Step III

- Final assignment of x-coordinates
- Routing of edges
Figure 20.7  Visualizations of the metabolic pathway shown in Figure 20.5 using (a) a force-directed algorithm [KK89] and (b) a hierarchical approach [STT81].
Example
Force Directed Layouts

- Proximity does not mean actual grouping
- Avoid edge crossings and node overlaps
- Spatial position does not directly encode attributes
- Non-deterministic

<table>
<thead>
<tr>
<th>Technique</th>
<th>force-directed placement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Types</td>
<td>network</td>
</tr>
<tr>
<td>View Comp.</td>
<td>connection</td>
</tr>
<tr>
<td>Abstract Tasks</td>
<td>understanding topological structure following paths</td>
</tr>
<tr>
<td>Scalability</td>
<td>nodes: dozens/hundreds</td>
</tr>
<tr>
<td></td>
<td>edges: hundreds</td>
</tr>
<tr>
<td></td>
<td>node/edge density: $E &lt; 4N$</td>
</tr>
</tbody>
</table>
4. Drawing on Physical Analogies

Ulrik Brandes

FDP

- No intrinsic layering, now what?
- Physics model, edges = springs, nodes = repulsive magnets
Eye Candy
Spring Analogy

Fig. 4.1. The spring analogy.
Ulrike's Forces

Given a connected undirected graph $G = (V, E)$, let $p = (p_v)_{v \in V}$ be a vector of vertex positions $p_v = (x_v, y_v)$ in the plane. We denote by $\|p_v - p_u\|$ the length of the difference vector $p_v - p_u$, which is the Euclidean distance between positions $p_u$ and $p_v$. Furthermore, we denote by $\frac{p_v - p_u}{\|p_v - p_u\|}$ the unit length vector pointing from $p_u$ to $p_v$. The model of Eades (1984), now known as the

$$f_{\text{rep}}(p_u, p_v) = \frac{c_e}{\|p_v - p_u\|^2} \cdot \frac{p_u p_v}{\|p_v - p_u\|}$$

$$f_{\text{spring}}(p_u, p_v) = c_\sigma \cdot \log \frac{\|p_u - p_v\|}{l} \cdot \mathbf{j}$$

**Algorithm 6: Spring embedder**

**Input**: connected undirected graph $G = (V, E)$
initial placement $p = (p_v)_{v \in V}$

**Output**: placement $p$ with low internal stress

for $t \leftarrow 1$ to $\text{ITERATIONS}$ do
  for $v \in V$ do
    $F_v(t) \leftarrow \sum_{u : \{u, v\} \not\in E} f_{\text{rep}}(p_u, p_v) + \sum_{u : \{u, v\} \in E} f_{\text{spring}}(p_u, p_v)$
  for $v \in V$ do
    $p_v \leftarrow p_v + \delta \cdot F_v(t)$

**Fig. 4.2.** Magnitude of spring embedder forces.
Method

- Start from random layout
- Loop:
  - For every node pair compute repulsive force
  - For every edge compute attractive force
  - Accumulate forces per node
  - Update node position in direction of accumulated force
- Stop when layout is ‘good enough’
Phenotype Layouts
Figure 20.3  A force-directed 2D layout of protein-protein interactions in yeast (redrawn from [FS03]).
Radial Layout

http://www.cise.ufl.edu/research/sparse/matrices/synopsis/

Wexner Medical Center
Circos Style
The HOS-Tree

Level 00

Level 01

Level 02

Level 03

Increasing Levels
Other
THE END