PCA
Scatterplot and Variances

FIGURE 17.26: The PCA can also be used to separate the within-experiment variation from the inter-experiment variation. The new axis $P_1$ will be aligned with the direction of the within-experiment variation. The within-experiment variation stems from the fact that genes are expressed at all levels in either sample. The new axis $P_2$ will be perpendicular on $P_1$ and will capture the inter-experiment variation. The inter-experiment variation is the interesting variations if we want to find those genes that differ between the two samples.
**FIGURE 17.25:** In spite of the fact that each point has two coordinates, this data set is essentially one dimensional: most of the variance is along the first eigenvector $p_1$ with the variance along the second direction $p_2$ being probably due to the noise. The PCA will find a new coordinate system such as the first coordinate is the direction on which the data have maximum variance (the first eigenvector), the second coordinate is perpendicular on the first and captures the second largest variance, etc.
PCA

Introduced by Pearson (1901) and Hotelling (1933) to describe variation in a multivariate data in terms of uncorrelated variables.
Dimensionality Reduction - PCA

• Principal component analysis (PCA)
  – Geometric meaning – fit a low-dimensional (hyper)plane to high dimensional data points; transform the data points into the low-dimensional space.
  – Algebraic meaning – optimal (least square error) low rank approximation of a data matrix
  – Statistical meaning – data transformation preserving the direction with the largest variance

Scree Plot

\[ U S V^T = [x_1 - \hat{\mu}, x_2 - \hat{\mu}, \ldots x_N - \hat{\mu}] \in \mathbb{R}^{K \times N} \]

- Eyeballing – “knee point”
- Percentage of “energy” – variances explained by the PCs

\[
\begin{pmatrix}
A_1 & \\
A_2 & \\
\vdots & \\
A_K & \\
\end{pmatrix}
= 
\begin{pmatrix}
U & w_1 & w_2 & w_3 & U^T \\
& & & & \\
& & & & \\
& & & & \\
& & & & \\
\end{pmatrix}
\]

\[ w_1 \geq w_2 \geq w_3 \geq 0 \]

\[ A_{11}^2 + A_{12}^2 + \cdots + A_{53}^2 = w_1^2 + w_2^2 + w_3^2 \]
Dimensionality Reduction - MDS

- Need to do PCA – but don’t know the coordinates of the points
- Only know the distances between every pair of points
- Multi-dimensional scaling (MDS)
- Embedding problem (Whitney, Nash)
PCA - other things to consider

- Numerical balance/data normalization

\[
\begin{bmatrix}
1000 & 1020 & 980 & 1050 & 950 \\
-2 & -1 & 0 & 1 & 2
\end{bmatrix}
\]

- Noisy direction
- Continuous vs. discrete data
- Principal components are linear combinations of original data
- Principal components are orthogonal to each other, however, biological data are not
- Prior knowledge is important
- PCA is not clustering!
Issues with Clustering

Failed PCA – need other methods such as ICA

PCA (orthogonal coordinate)  ICA (non-orthogonal coordinate)
Figure 3
Three independent components of the human normal tissue data (dataset 5). Each gene is mapped to a point based on the value assigned to the gene in the 14th (x-axis), 15th (y-axis) and 55th (z-axis) independent components, which are enriched with liver-specific (red), muscle-specific (orange), and vulva-specific (green) genes, respectively. Genes not annotated as liver-, muscle- or vulva-specific are colored yellow.
Heatmaps
A HeatMap is A Color Table!
Clustered display of data from time course of serum stimulation of primary human fibroblasts.

Clustered display of data from time course of serum stimulation of primary human fibroblasts. Experimental details are described elsewhere (11). Briefly, foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15 min, 30 min, 1 hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr. The final datapoint was from a separate unsynchronized sample. Data were measured by using a cDNA microarray with elements representing approximately 8,600 distinct human genes. All measurements are relative to time 0. Genes were selected for this analysis if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. The dendrogram and colored image were produced as described in the text; the color scale ranges from saturated green for log ratios −3.0 and below to saturated red for log ratios 3.0 and above. Each gene is represented by a single row of colored boxes; each time point is represented by a single column. Five separate clusters are indicated by colored bars and by identical coloring of the corresponding region of the dendrogram. As described in detail in ref. 11, the sequence-verified named genes in these clusters contain multiple genes involved in (A) cholesterol biosynthesis, (B) the cell cycle, (C) the immediate–early response, (D) signaling and angiogenesis, and (E) wound healing and tissue remodeling. These clusters also contain named genes not involved in these processes and numerous uncharacterized genes.
Cluster analysis of combined yeast data sets.

Cluster analysis of combined yeast data sets. Data from separate time courses of gene expression in the yeast *S. cerevisiae* were combined and clustered. Data were drawn from time courses during the following processes: the cell division cycle (9) after synchronization by alpha factor arrest (ALPH; 18 time points); centrifugal elutriation (ELU; 14 time points), and with a temperature-sensitive cdc15 mutant (CDC15; 15 time points); sporulation (10) (SPO, 7 time points plus four additional samples); shock by high temperature (HT, 6 time points); reducing agents (D, 4 time points) and low temperature (C; 4 time points) (P.T.S., J. Cuoczo, C. Kaiser, P.O. B., and D. B., unpublished work); and the diauxic shift (8) (DX, 7 time points). All data were collected by using DNA microarrays with elements representing nearly all of the ORFs from the fully sequenced *S. cerevisiae* genome (8); all measurements were made against a time 0 reference sample except for the cell-cycle experiments, where an unsynchronized sample was used. All genes (2,467) for which functional annotation was available in the Saccharomyces Genome Database were included (12). The contribution to the gene similarity score of each sample from a given process was weighted by the inverse of the square root of the number of samples analyzed from that process. The entire clustered image is shown in A; a larger version of this image, along with dendrogram and gene names, is available at http://rana.stanford.edu/clustering/yeastall.html. Full gene names are shown for representative clusters containing functionally related genes involved in (B) spindle pole body assembly and function, (C) the proteasome, (D) mRNA splicing, (E) glycolysis, (F) the mitochondrial ribosome, (G) ATP synthesis, (H) chromatin structure, (I) the ribosome and translation, (J) DNA replication, and (K) the tricarboxylic acid cycle and respiration. The full-color range represents log ratios of −1.2 to 1.2 for the cell-cycle experiments, −1.5 to 1.5 for the shock experiments, −2.0 to 2.0 for the diauxic shift, and −3.0 to 3.0 for sporulation. Gene name, functional category, and specific function are from the Saccharomyces Genome Database (13). Cluster I contains 112 ribosomal protein genes, seven translation initiation or elongation factors, three tRNA synthetases, and three genes of apparently unrelated function.

Eisen M B et al. PNAS 1998;95:14863-14868
To demonstrate the biological origins of patterns seen in Figs. 1 and 2, data from Fig. 1 were clustered by using methods described here before and after random permutation within rows (random 1), within columns (random 2), and both (random 3).
HeatMaps in R/Bioconductor

```r
> library("ALL")
> Data(“ALL”)
> eset <- ALL[, ALL$mol.biol %in% c("BCR/ABL", "ALL1/AF4")]
> heatmap(exprs(eset[1:100,]))
```
Heat map (produced by the Bioconductor function heatmap()) of the ALL leukemia data.

Gentleman et al.  
R/BioConductor

http://www2.warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/heatmap/

> ann3 = annHeatmap(exprs(exdat2), ann=pData(exdat2),
+               cluster=list(cuth=7500, label=c("Control-like", "Case-like")))
> plot(ann3)

Figure 7: Annotated heatmap with row- and column dendrograms and a legend for 46 genes and 26 samples. The column dendrogram is cut at $h = 7500$, and we add cluster names, based on the annotation.
Distance Measure (Metric?)

- What do we mean by “similar”?
- Euclidean
- Pearson correlation
- Uncentered correlation (cosine)
- Manhattan distance (city block distance)
- …
Distance Metric

- Euclidean (L2)
  \[ x = (x_1, x_2, \cdots, x_n)^T \]
  \[ y = (y_1, y_2, \cdots, y_n)^T \]
  \[ d^E(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \]

- Cityblock (Manhattan) distance (L1)

- Pearson correlation
  \[ r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \]
Gene Co-Expression

ARTICLES

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 Rennert3, 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 Solé7, Pilar Hernández13, Conxi Lázaro13, Katherine L Nathanson14, Barbara L Weber14, Michael E Cusick1,2, David E Hill1,2, Kenneth Offit8, David M Livingstone3, Stephen B Gruber2,6,15, Jeffrey D Parvin1,6 & Marc Vidal1,2

Many cancer-associated genes remain to be identified to clarify the underlying molecular mechanisms of cancer susceptibility and progression. Better understanding is also required of how mutations in cancer genes affect their products in the context of complex cellular networks. Here we have used a network modeling strategy to identify genes potentially associated with higher risk of breast cancer. Starting with four known genes encoding tumor suppressors of breast cancer, we combined gene expression profiling with functional genomic and proteomic (or ‘omic’) data from various species to generate a network containing 118 genes linked by 866 potential functional associations. This network shows higher connectivity than expected by chance, suggesting that its components function in biologically related pathways. One of the components of the network is HMMP, encoding a centrosome subunit, for which we demonstrate previously unknown functional associations with the breast cancer-associated gene BRCA1. Two case-control studies of incident breast cancer indicate that the HMMP locus is associated with higher risk of breast cancer in humans. Our network modeling strategy should be useful for the discovery of additional cancer-associated genes.
Gene Co-Expression

- Correlated gene expression profiles
  - Both positive and negative correlation
- Do these genes form clusters?

![Gene Co-Expression Network](image)
Correlation Coefficients

- Pearson correlation coefficients (PCC, $\rho$)
- Linear model
- Assumption: Gaussian distribution
- Widely used for microarray data – log-normal distribution

$$R_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$
Correlation Coefficients

- Spearman rank correlation coefficients (SCC)
- Nonlinear model
- PCC of the ranks
- Example – RNA-seq data
- Related – Kendall’s $\tau$
Other Metrics

- Challenge 1 – nonlinear relationship
- Challenge 2 – non-Gaussian distribution
- Mutual information
- Canonical correlation
- Distance correlation
- ...

[Graph showing data points on a coordinate plane]
Correlation is NOT causation!!!
Gene Expression Regulation
Gene Expression Regulation

DNA Copy Number Variance

Proteins in the Networks

1. MYOCD
2. MKL1
3. MKL2
4. FLYWCH2
5. CISD3
6. HN1L
7. STUB1
8. SUCLG1

Protein Expression

MKL1
- R square = 0.26018
- PCC = 0.51008

MKL2
- R square = 0.28307
- PCC = 0.53205

STUB1
- R square = 0.33351
- PCC = 0.5775

CISD3
- R square = 0.28342
- PCC = 0.53237

Protein Expression vs. Copy Number Variance
Tissue Mixture

- Samples have different composition of different tissues
- The expression levels of tissue specific genes “co-express”
From Correlations to Gene Co-expression Network (GCN)

<table>
<thead>
<tr>
<th>Gene/Patient</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
<th>P14</th>
<th>P15</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>-1</td>
<td>-2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>-3</td>
<td>2</td>
<td>-3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>-2</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G2</td>
<td>-2</td>
<td>-4</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>-6</td>
<td>4</td>
<td>-6</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>-4</td>
<td>8</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>G3</td>
<td>1</td>
<td>2</td>
<td>-2</td>
<td>-1</td>
<td>-4</td>
<td>3</td>
<td>-2</td>
<td>3</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>2</td>
<td>-4</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>G4</td>
<td>2</td>
<td>4</td>
<td>-4</td>
<td>-2</td>
<td>-8</td>
<td>6</td>
<td>-4</td>
<td>6</td>
<td>4</td>
<td>-2</td>
<td>-4</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>G5</td>
<td>-4</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>G6</td>
<td>8</td>
<td>-6</td>
<td>-8</td>
<td>0</td>
<td>6</td>
<td>-6</td>
<td>-12</td>
<td>-4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>G7</td>
<td>-4</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>-4</td>
<td>-2</td>
<td>1</td>
</tr>
<tr>
<td>G8</td>
<td>-1</td>
<td>-2</td>
<td>2</td>
<td>4</td>
<td>-3</td>
<td>2</td>
<td>3</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>2</td>
<td>-4</td>
<td>0</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>G9</td>
<td>1</td>
<td>2</td>
<td>-2</td>
<td>-1</td>
<td>-4</td>
<td>3</td>
<td>-2</td>
<td>-3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>-2</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G10</td>
<td>-2</td>
<td>-4</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>-2</td>
<td>-4</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

Genes/genes:

<table>
<thead>
<tr>
<th>genes</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5</th>
<th>G6</th>
<th>G7</th>
<th>G8</th>
<th>G9</th>
<th>G10</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>1</td>
<td>1</td>
<td>-0.96</td>
<td>-0.96</td>
<td>-0.31</td>
<td>0.35</td>
<td>-0.33</td>
<td>-0.05</td>
<td>0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>G2</td>
<td>1</td>
<td>1</td>
<td>-0.96</td>
<td>-0.96</td>
<td>-0.31</td>
<td>0.35</td>
<td>-0.33</td>
<td>-0.05</td>
<td>0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>G3</td>
<td>-0.96</td>
<td>-0.96</td>
<td>1</td>
<td>1</td>
<td>0.35</td>
<td>-0.31</td>
<td>0.37</td>
<td>0.08</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>G4</td>
<td>-0.96</td>
<td>-0.96</td>
<td>1</td>
<td>1</td>
<td>0.35</td>
<td>-0.31</td>
<td>0.37</td>
<td>0.08</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>G5</td>
<td>-0.31</td>
<td>-0.31</td>
<td>0.35</td>
<td>0.35</td>
<td>1</td>
<td>-0.96</td>
<td>1</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>G6</td>
<td>0.35</td>
<td>0.35</td>
<td>-0.31</td>
<td>-0.31</td>
<td>-0.96</td>
<td>1</td>
<td>-0.96</td>
<td>-0.19</td>
<td>0.23</td>
<td>-0.19</td>
</tr>
<tr>
<td>G7</td>
<td>-0.33</td>
<td>-0.33</td>
<td>0.37</td>
<td>0.37</td>
<td>1</td>
<td>-0.96</td>
<td>1</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>G8</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.08</td>
<td>0.08</td>
<td>0.23</td>
<td>0.19</td>
<td>-0.19</td>
<td>0.23</td>
<td>1</td>
<td>-0.96</td>
</tr>
<tr>
<td>G9</td>
<td>0.08</td>
<td>0.08</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.19</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
<td>1</td>
<td>-0.96</td>
</tr>
<tr>
<td>G10</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.08</td>
<td>0.08</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
<td>1</td>
<td>-0.96</td>
<td>1</td>
</tr>
</tbody>
</table>

Correlations to Gene Co-expression Network (GCN)
## Unweighted vs Weighted GCN

<table>
<thead>
<tr>
<th>genes/genes</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5</th>
<th>G6</th>
<th>G7</th>
<th>G8</th>
<th>G9</th>
<th>G10</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>1</td>
<td>1</td>
<td>-0.96</td>
<td>-0.96</td>
<td>-0.31</td>
<td>0.35</td>
<td>-0.33</td>
<td>-0.05</td>
<td>0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>G2</td>
<td>1</td>
<td>1</td>
<td>-0.96</td>
<td>-0.96</td>
<td>-0.31</td>
<td>0.35</td>
<td>-0.33</td>
<td>-0.05</td>
<td>0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>G3</td>
<td>-0.96</td>
<td>-0.96</td>
<td>1</td>
<td>1</td>
<td>0.35</td>
<td>-0.31</td>
<td>0.37</td>
<td>0.08</td>
<td>-0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>G4</td>
<td>-0.96</td>
<td>-0.96</td>
<td>1</td>
<td>1</td>
<td>0.35</td>
<td>-0.31</td>
<td>0.37</td>
<td>0.08</td>
<td>-0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>G5</td>
<td>-0.31</td>
<td>-0.31</td>
<td>0.35</td>
<td>0.35</td>
<td>1</td>
<td>-0.96</td>
<td>1</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>G6</td>
<td>0.35</td>
<td>0.35</td>
<td>-0.31</td>
<td>-0.31</td>
<td>-0.96</td>
<td>1</td>
<td>-0.96</td>
<td>-0.19</td>
<td>0.23</td>
<td>-0.19</td>
</tr>
<tr>
<td>G7</td>
<td>-0.33</td>
<td>-0.33</td>
<td>0.37</td>
<td>0.37</td>
<td>1</td>
<td>-0.96</td>
<td>1</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>G8</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.08</td>
<td>0.08</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
<td>1</td>
<td>-0.96</td>
<td>1</td>
</tr>
<tr>
<td>G9</td>
<td>0.08</td>
<td>0.08</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.19</td>
<td>0.23</td>
<td>-0.19</td>
<td>-0.96</td>
<td>1</td>
<td>-0.96</td>
</tr>
<tr>
<td>G10</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.08</td>
<td>0.08</td>
<td>0.23</td>
<td>-0.19</td>
<td>0.23</td>
<td>1</td>
<td>-0.96</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Unweighted** – hard or adaptive threshold
- **Weighted** – no threshold
Weighted GCN

- Unweighted approach – need hard threshold
- Step function as Adjacency function
- Too rigid – if 0.6 is the threshold, how about 0.59?
  Remember the data is always noisy
Weighted GCN

- Weighted is more flexible
- Directly use PCC value (or absolute value)
- Nonlinear transformation can help to “emphasize” preferred values
- Monotonic transformation
- Straightline, power, sigmoidal curves
Weighted GCN

**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.

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)

Weighted GCN – PCC is not enough

• One more problem – unbalanced weights
• The densities of clusters are different
• Similarity can be defined by similar neighborhood
TOM metric in clustering

\[ TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} \]

\[ DistTOM_{ij} = 1 - TOM_{ij} \]

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

\[ TOM(i, j) = \frac{|N_1(i) \cap N_1(j)| + a_{ij}}{\min(|N_1(i)|, |N_1(j)|) + 1 - a_{ij}} \]

\( N_1(i) \) denotes the set of 1-step (i.e. direct) neighbors of node \( i \)

\( | | \) measures the cardinality

Adding \( 1 - a(i,j) \) to the denominator prevents it from becoming 0.
Mining Networks

- Dense modules / network motifs
- Many considerations
  - Clique (fully connected)
  - Quasi-clique (dense)
  - K-core (contain a k-clique)
- Density
  - Total # of edges / Total number of edges
  - Sum of weights / Sum of largest possible weights
Mining Networks

- Numerous algorithms from the data mining community
- CODENSE algorithm
- QCM / eQCM algorithm
- WGCNA package
  - Developed by Horvath’s group at UCLA
  - R language
  - Hierarchical clustering instead of mining
WGCNA

Here modules correspond to branches of the dendrogram

Genes correspond to rows and columns

Hierarchical clustering dendrogram
Module Eigengene

measure of over-expression=average redness

Rows=genes, Columns=microarray

The brown module eigengenes across samples
Module Eigengene

• Average

• Weighted average
  – Singular value decomposition (SVD)
  – First column of V
  – Pay attention to **sign**

\[
A = U \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} V^T
\]
Module Eigengene

- Average
- Weighted average
  - Singular value decomposition (SVD)
Module Eigengene

• Less number of eigengene than genes, increase statistical power
• Directly linked to function/structure
Module Eigengene

- Eigengenes can be used to build their own networks
- Eigengenes are quantitative traits

Comparison of GCN
Summary

• GCN contains rich biological information.
• Building GCN may involve different types of correlation metric and different transformation of numbers.
• There are multiple approaches for mining GCN. One common goal is to extract densely connected modules.
• The modules can be summarized into “eigengenes” and then used as quantitative traits.