Omics

Genomic ← gene/signals/etc.

Transcriptomics → RNA transcripts/copies

Proteomics ← protein content of a cell

Quantify

Over expressed
Under expressed

Normal/Control
Cancer/Treatment

High throughput microarrays

Differential Expression

Translate
**Time course**

1. Repeated study
   - within 3/4 time

2. Different "environmental" conditions
   - temporal
   - spatial

**Expression Matrix**

- mRNA
- Protein

**Gene**

- mRNA fragments

**Interactions between 2 proteins**

**Protein-Protein Interaction Network**

**Gene regulation networks / transcription network**

**Signalling / transduction network**

**Interactions**

- PPI
Microarrays

- CDNA

- In situ

- Sample

- Extract RNA

- Transcribe

- Probe

- ACGCC

- TCGCC

- Tagged with dye

- Tagged

- "tagged"

- Control

- Treatment

- Cy3 green dye

- Cy5 red dye

- Image quantification

- Green versus red

- Treatment

- Raw data

- $c_{ij}$

- $g_{in}$

- $m 	imes m$

- $\log(R/G)$

- True under expressed

- True over expressed

- $0 \Rightarrow$ equal

- Processed
gene expression matrix (flipped/transposed) \( J \leftarrow \text{dimension} \)
\( n \leftarrow \text{points} \)
\( d >> n \)

\[ S_i \rightarrow \text{mean expression} \quad \mu_i = \frac{1}{n} \sum_{j=1}^{n} e_{ji} \]

\[ \sigma_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (e_{ji} - \mu_i)^2 \]

\( e_{ji} = \text{expression value of gene } i \text{ in sample } j \)

**Differential Expression**

Cancer

\( n_1 \) # of samples

\( n_2 \) # of normal samples

\( Q_i \): is gene \( S_i \) differentially expressed?

Compare means

Cancer \( \mu_i \)

\( \mu_i \) Normal

Is there a statistically significant difference?
t-test for significant difference between means

\[
\text{Gene i} \quad \rightarrow \quad \begin{array}{c|c|c}
\text{Cancer} & \text{Normal} \\
\hline
n_1 & n_2 \\
\bar{x}_1 & \bar{x}_2 \\
\sigma_1 & \sigma_2 \\
\text{std.-dev} & \text{std.-dev}
\end{array}
\]

\[ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \]

\[ \text{dof} = \frac{(a+1)^2}{a^2/(n_1-1) + b^2/(n_2-1)} \]

\[ \text{degrees of freedom} \]

\[ \text{p-value of } t \]

\[ \text{if } p\text{-value (of } t) \leq 0.05 \]

Individual → pairwise dependence → groups clustering

\[ \text{cov}(g_a, g_b) = \frac{1}{n-1} \sum_{j=1}^{n} (e_{j_a} - \bar{m}_a)(e_{j_b} - \bar{m}_b) \]

- linear trend
- +ve: correlation
- 0: no dependence
- -ve: anti-correlation
Correlation
\[
\text{cor}(g_a, g_b) = \frac{\text{cor}(g_a, g_b)}{\sigma_a \sigma_b} \in [-1, 1]
\]

Find groups of similar gene \(\rightarrow\) Clustering

Definition of similarity
\[
\text{dist}(g_a, g_b) = \sqrt{\sum_{j=1}^{n} (g_{ja} - g_{jb})^2}
\]

Euclidean distance

\(1 - \text{cor}(g_a, g_b)\)
\(\in [0, 2]\)

Distance matrix

Pairwise gene distance

Dendogram
Hierarchical Clustering of genes
merge the closest pair of clusters.

a) \( d(c_i, c_j) = \) avg pairwise distance

b) single links
\[
d(c_i, c_j) = \min \text{ over the pairs}
\]

c) complete links
\[
d(c_i, c_j) = \max \text{ over gene pair distance}
\]