



# Limitations of Statistical Learning from Gene Expression Data

---

Tianjiao Chu  
Institute for Human and Machine Cognition  
University of West Florida  
May. 2004



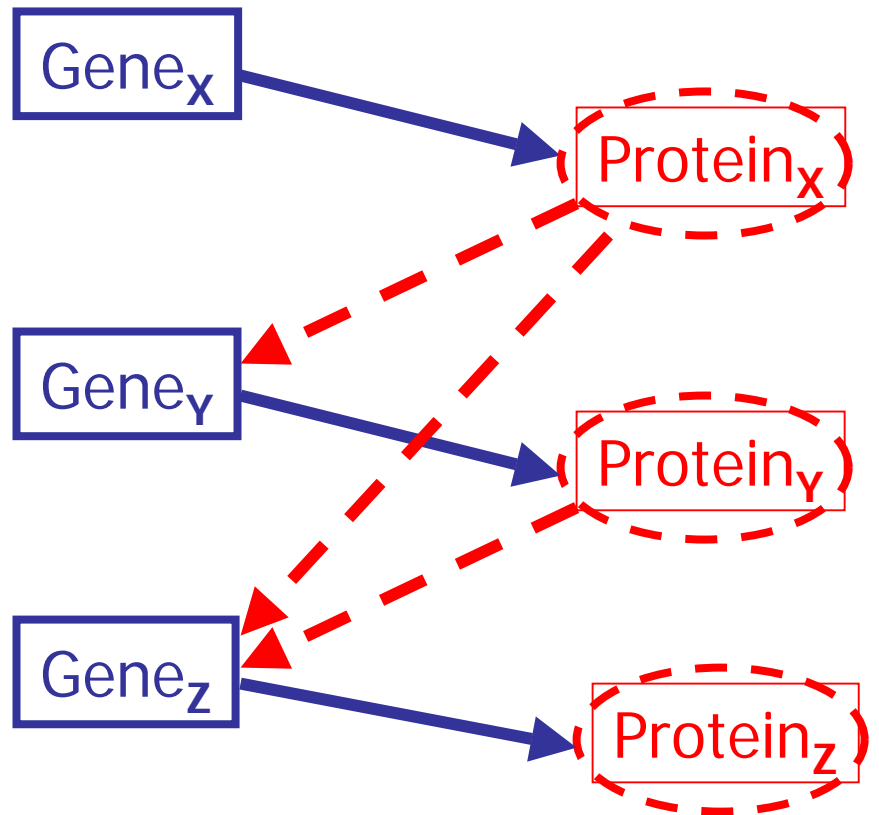
# Outline

---

- Gene Regulation and Gene Expression Data
- Principle of Causal Inference
- Limitations imposed by current technology (Problem of Aggregation)

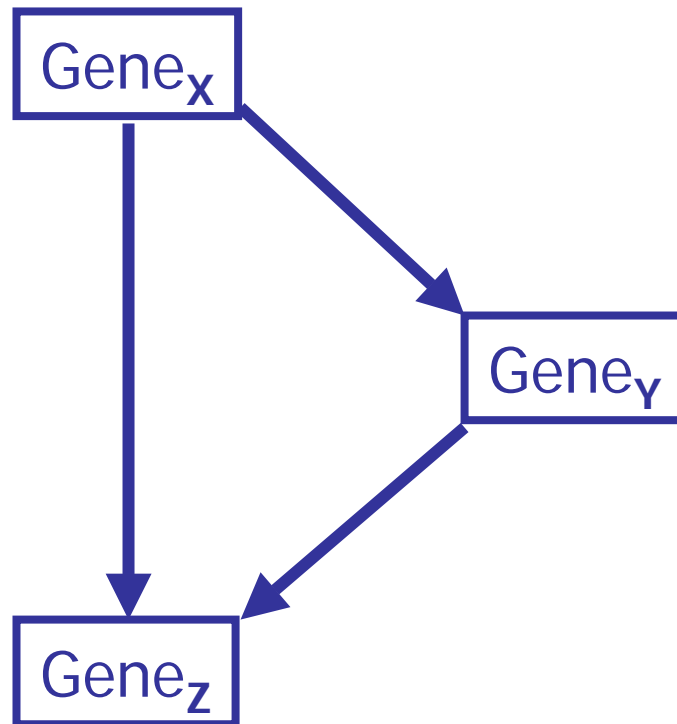
# Gene Regulation in a Cell

- Genes expression determine the production of proteins
- Proteins control the expression of genes



# Gene Regulatory Network

- A network of genes regulating each other
- Technologies are available for measuring expression levels of large number of genes efficiently



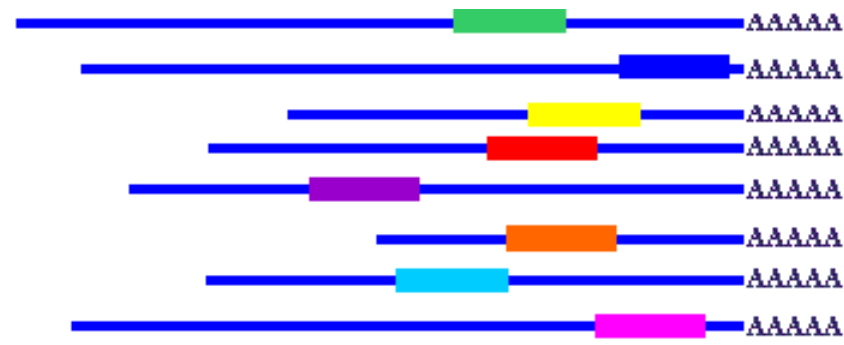


# SAGE Technology

---

- Serial Analysis of Gene Expression
- Cutting 10-base long sequence (tags) from specific position of each gene
- Each SAGE library consists of about 30,000 tags
- Count of tags tell the expression levels of the corresponding genes

# SAGE



↓ Isolate SAGE tags



PCR →

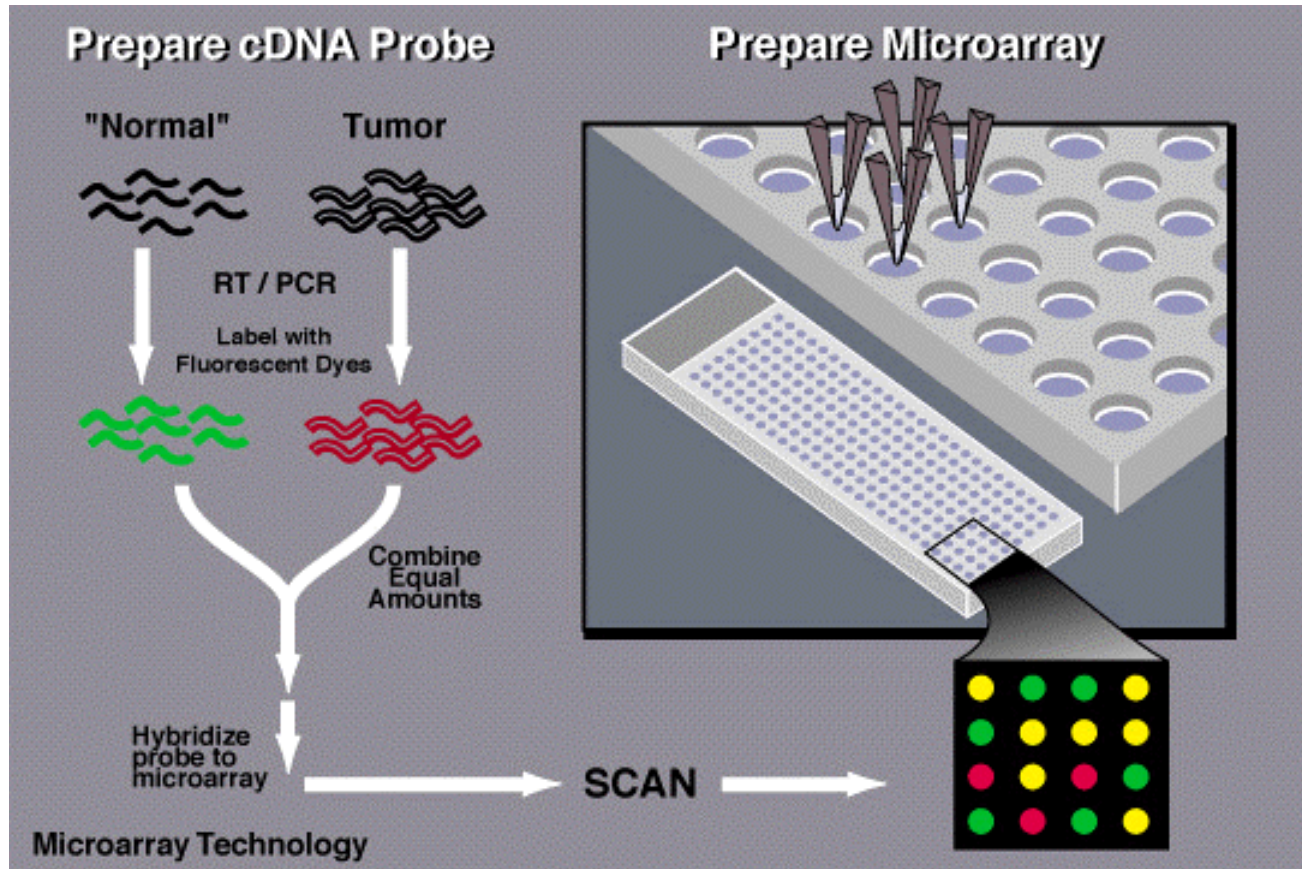
↓ Link tags together



↓ Sequence linked tags



# cDNA MicroArray Technology





# cDNA MicroArray Technology

---

- Luminosity of a spot is proportional to the number of the copies of the genes hybridized to that spot



# Other Technologies

---

- Affymetrix: Each probe contains about 11-20 pairs of short nucleotides
- Codelink: More recent. Each probe contains 30 base long nucleotides.



# Outline

---

- Gene Regulation and Gene expression Data
- Principle of Causal Inference
- Limitations imposed by current technology (Problem of Aggregation)



# What we would like to do

---

with gene expression data from  
MicroArray or SAGE?

*Identify the **gene regulatory network***



# Traditional approach

---

- Controlled Experiments:
  - Manipulate the expression level of a certain gene
  - See what happens to the expression levels of other genes



# Main drawback

---

- Very slow, expensive
  - Only work on a small network with a few genes
  - Thousands of genes
  - Combinatoric problems (manipulating two genes at the same time, three genes at the same time ...)



# Alternative approach

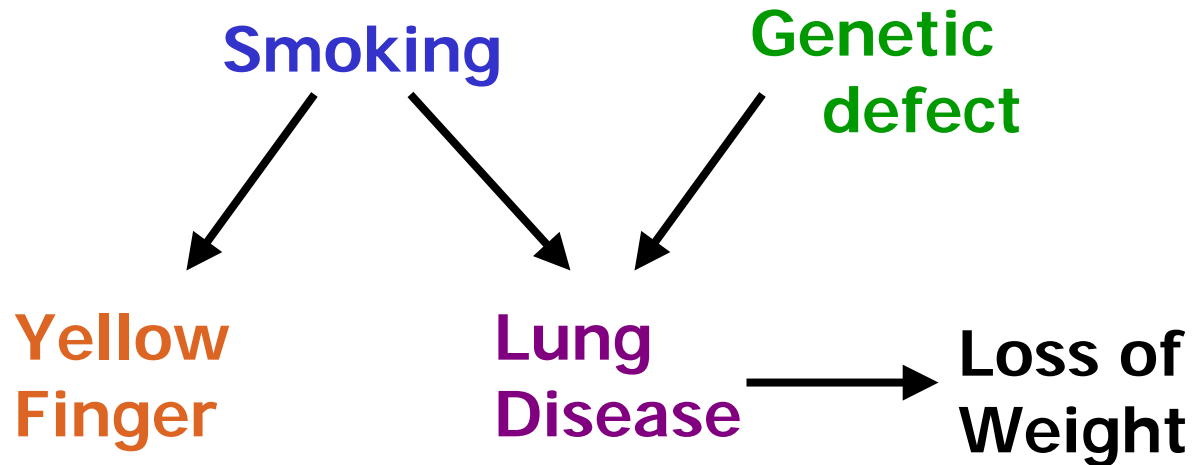
---

- Explore the relation between **causation** and **conditional independence**
- Infer causal relations from observational data
- Reference: Spirtes, Glymour, Scheines 2001, Pearl 2000



# Causal Graph

---





# Markov Condition

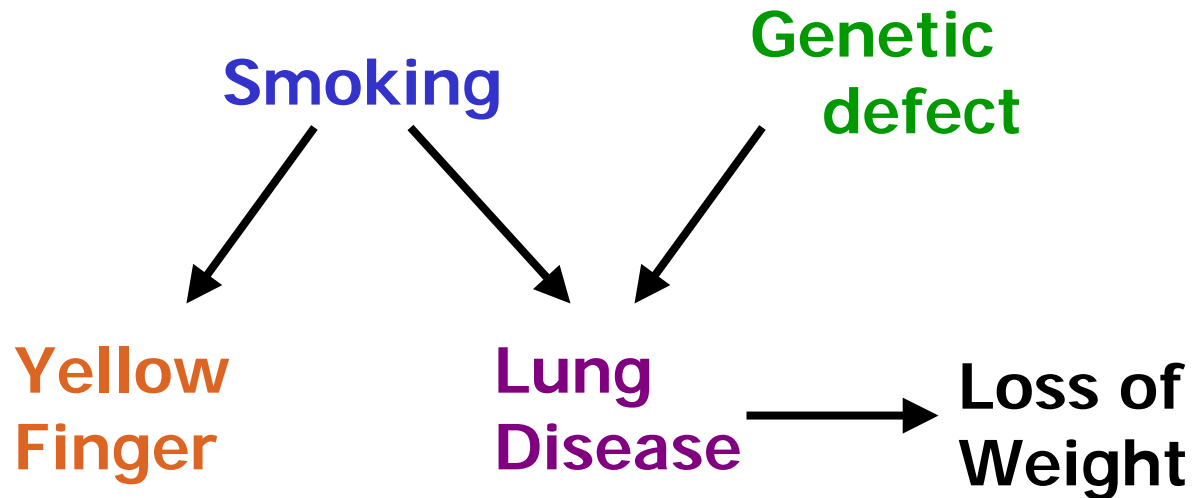
---

- In a causal graph, conditional on the parents of  $X$ ,  $X$  is independent of all the variables that are neither parents nor descendants of  $X$



# Causal Graph

---





# Faithfulness Condition

---

- Faithfulness: All the conditional independencies observed are implied by the causal graph and Markov condition



# Causal inference

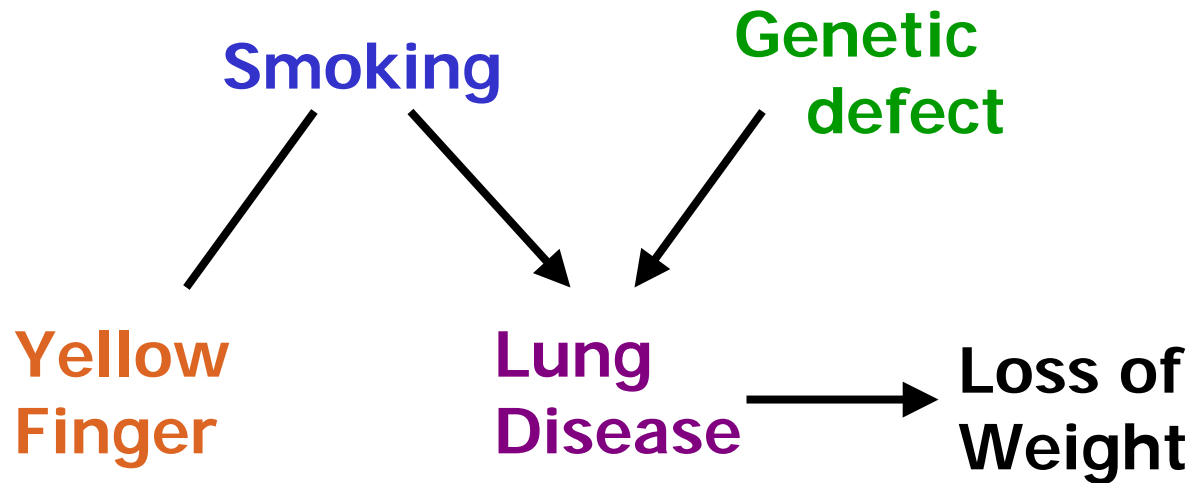
---

- Assuming Markov condition and faithfulness condition, from the conditional independency relations among the observed data, we could identify some of the causal relations among the variables



# Causal Pattern

---





# Outline

---

- Gene Regulation and Gene expression Data
- Principle of Causal Inference
- Limitations imposed by current technology (Problem of Aggregation)



# Aggregation of Cells

---

- Both the SAGE and the MicroArray technologies measure the concentration levels of the genes in a sample tissue
- We are interested in the causal relation among the genes within a single cell

# Conditional Independence Under Aggregation

- $X_i$  and  $Z_i$  are independent given  $Y_i$

$$X_1 \longrightarrow Y_1 \longrightarrow Z_1$$

Cell 1: Latent

$$X_2 \longrightarrow Y_2 \longrightarrow Z_2$$

Cell 1: Latent

⋮

⋮

⋮

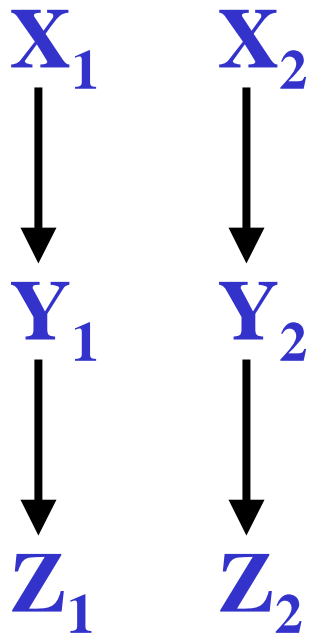
$$X_n \longrightarrow Y_n \longrightarrow Z_n$$

Cell n: Latent

$$\Sigma_i X \quad ? \quad \Sigma_i Y \quad ? \quad \Sigma_i Z$$

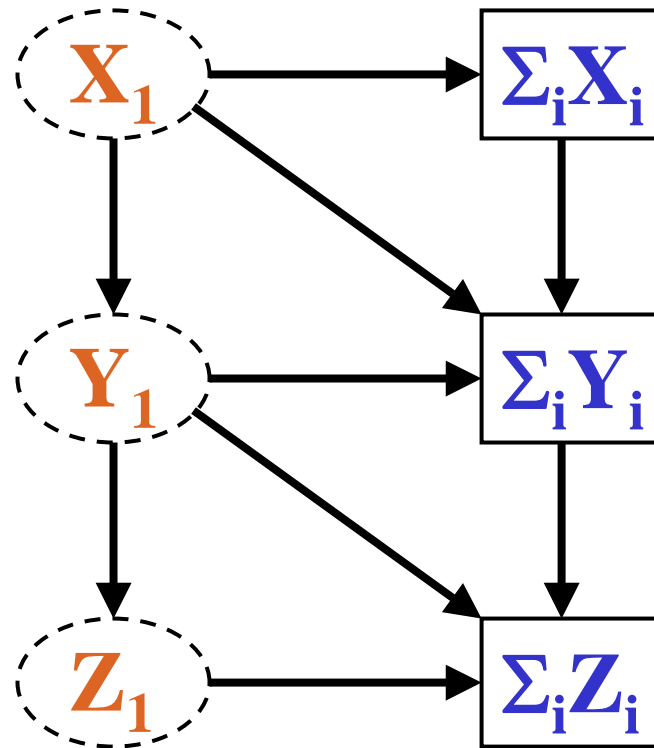
n Cells: Observed

## Before aggregation



$$Y_i = f(X_i, \varepsilon_{1i})$$
$$Z_i = g(Y_i, \varepsilon_{2i})$$

## After aggregation



$$\Sigma_i X_i = X_1 + X_2$$
$$\Sigma_i Y_i = Y_1 + f(\Sigma_i X_i - X_1, \varepsilon_{12})$$
$$\Sigma_i Z_i = Z_1 + g(\Sigma_i Y_i - Y_1, \varepsilon_{22})$$



# The Bad News

---

- In general,  $X_1 + X_2$  is not independent of  $Z_1 + Z_2$  given  $Y_1 + Y_2$ :
- The conditional independence relation does not hold under aggregation



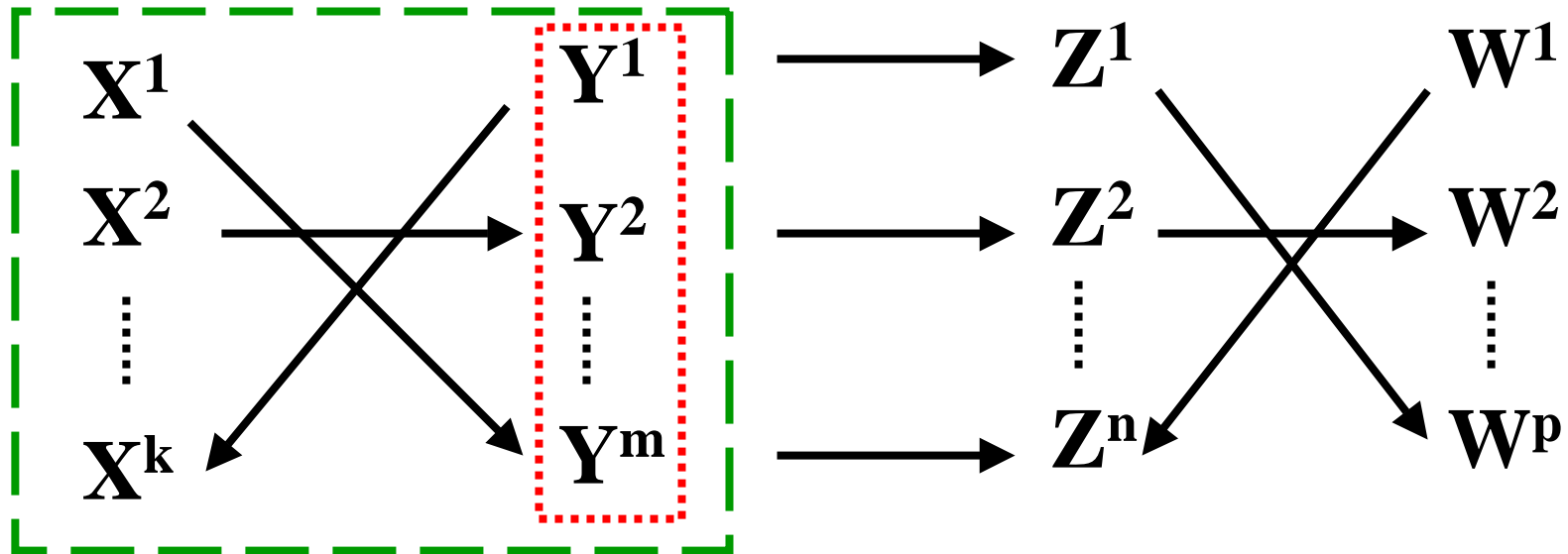
# What We Need to Do

---

- Find general sufficient conditions under which the conditional independencies will hold even after aggregation (Chu, Clark, et al, 2003)
- Find general conditions under which the conditional independencies will NOT hold (This talk)

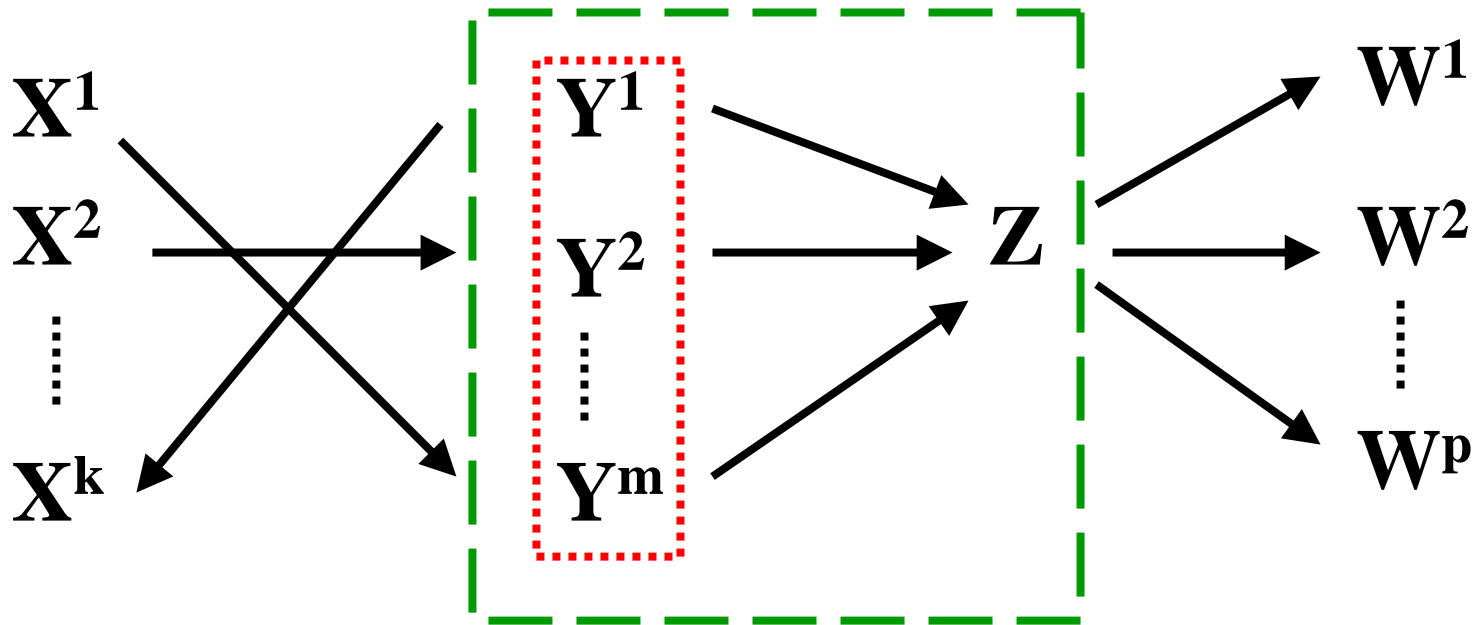
# When the Conditional Independencies will Hold (1)

- $X$ 's and  $Y$ 's are multivariate normal



# When the Conditional Independencies will Hold (2)

- Or:  $Z$  is linear in  $Y$ 's





# When the Conditional Independencies will NOT Hold

---

- Large Sample Aggregation
  - Current data are measurements of summed expression levels from hundreds of thousands of cells



# Main Result

---

- For at least two general classes of distributions, conditional independencies among large sample means/sums are essentially determined by the correlation matrix ...



# Two Classes of Distributions

---

- Continuous distributions with bounded densities
  - Good approximation of the exact distributions of the gene expression levels
- Regular lattice distributions
  - $X = k\hbar + c, k = 0, \pm 1, \pm 2, \dots$
  - Include the exact distribution of the number of mRNA transcripts in a cell



# Conditional Independencies and Correlation Matrix

---

- **$\text{Corr}(X, Y; Z) = 0$ :**
  - Eventually the test for conditional independence will accept the null that  $\Sigma X$  and  $\Sigma Y$  are independent given  $\Sigma Z$
- **$\text{Corr}(X, Y; Z) \neq 0$ :**
  - Eventually the test for conditional independence will reject the null that  $\Sigma X$  and  $\Sigma Y$  are independent given  $\Sigma Z$



# Large Sample Variance Estimation

---

- 100,000 cells (MicroSAGE)
- 15,000 mRNA's per cell (Yeast)
- Consider two genes  $X$  and  $Y$  within a single cell
  - $E[X] = E[Y] = 15$
  - $\text{Var}(X) = \text{Var}(Y) = 225$

# Sample Size for Testing

## $\text{Corr}(x,y) = 0$ Vs $0.5$

---

- Null:  $\text{Corr}(X,Y) = 0$ 
  - Measured correlation in the SAGE library:  
-0.001
- Alternative:  $\text{Corr}(X,Y) = 0.5$ 
  - Measured correlation in the SAGE library:  
-0.00085
- Need  $1.7 \times 10^8$  SAGE libraries for a 15% level test (Fisher's Z transformation)



# Things Are Worse for Microarray Data

---

- Higher noise (chip to chip, dye to dye, cross-hybridization ... )
- Not direct counts (more complicated model)
- ...



# Conclusion

---

- Given the current technology for measuring gene expression levels
  - What we can learn:
    - Expected expression level
  - What we can learn in theory, but unlikely in practice:
    - Correlation matrix
  - What we cannot learn in principle
    - Conditional independencies



# Acknowledgement

---

- Thanks to Clark Glymour for first raising the problem of aggregation and many helpful discussions.