Lecture 27: Systems Biology and Bayesian Networks

§ Systems Biology and Regulatory Networks

• Definitions
• Network motifs
• Examples

§ Bayesian Networks

• Bayesian Statistics
• Bayesian Networks
System Biology: Definitions

§ Nodes (circles): Genes/Proteins

§ Edges (lines): interactions

§ Edges can represent:

  • protein-protein interactions
  • protein-DNA interactions (e.g. TF binding to promoter)
  • genetic interactions
  • regulatory or functional interactions

§ Edges can have directionality (gene X regulates gene Y)

§ Edges can represent positive or negative regulation

Protein-protein interaction

Activator

Repressor

Target
Transcription Factor Autoregulation

§ Transcription factors (TF) frequently bind to their own promoters and regulate their own transcription

§ Which network would respond more rapidly (by increasing the expression of the TF protein) in response to a signal?

Positive Autoregulation  Negative Autoregulation  No Autoregulation

Transcription Factor Autoregulation

- Negative autoregulation facilitates a more rapid response to a signal.
- Positive autoregulation delays gene expression in response to a signal.

Transcription Factor Autoregulation: Summary

§ Negative autoregulation facilitates a more rapid response to an environmental signal

§ Positive autoregulation delays gene expression in response to an environmental signal

§ Negative autoregulation helps maintain the expression of the TF at a constant level

§ Positive autoregulation can cause bistability
Complex Network Motifs: Feed-Forward Loops

§ Some network motifs are very common in biological networks (over-represented)

§ For example, the coherent Feed-Forward loop, pictured to the right, is common in biological networks

§ In a coherent Feed-Forward loop, Gene X and Gene Y must be activated to activate Gene Z
  • Note that Gene Y activation is dependent on Gene X

§ A coherent Feed-Forward loop requires a sustained input in order to activate the target gene Z

§ Hence, the coherent Feed-Forward loop is a “persistence detector”.

Input

⊙ Gene X

⊙ Gene Y

AND

⊙ Gene Z
Response Delay in Coherent Feed-Forward Loop

§ Note, however, that there is no delay in turning off the network signal

Incoherent Feed-Forward Loops

§ The incoherent Feed-Forward loop is also common in biological networks

§ In an Incoherent Feed-Forward loop, Gene X and must be activated, and Gene Y must be inactive to activate Gene Z.
  • Note again that Gene Y activation is dependent on Gene X

§ An incoherent Feed-Forward loop generates a rapid response that is eventually suppressed as Gene Y accumulates

§ Hence, the incoherent Feed-Forward loop is a “pulse generator”.
Pulse Generation with Incoherent Feed-Forward Loops

Image from:
Bayesian Statistics

Bayes’ Theorem:

\[ P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \]

- \( A \) = Experimental data
- \( B \) = Model parameters
- \( P(A \mid B) \) = Likelihood
- \( P(A) \) is constant

Bayes’ Postulate = \( P(B) \) is constant

§ How do we find the model parameters that best fit our data?

• From Bayes’ theorem we know: \( P(B \mid A) \propto P(A \mid B) \)

• To find the best fitting parameters [maximize value of \( P(B \mid A) \)], then we need to maximize the likelihood of the data given the model parameters [\( P(A \mid B) \)]
Maximum Likelihood Ratio Test

§ Method for testing hypothesis using maximum likelihood estimation (MLE)

§ Compute MLE for null hypothesis ($H_0$) and alternative hypothesis ($H_1$) from experimental data ($x$), as follows:

$$ l = \frac{L(x \mid H_0)}{L(x \mid H_1)} $$

§ The ratio of the MLE for $H_0$ over the MLE for $H_1$ is related to the chi-square distribution via the following equation:

$$ \chi^2 = -2 \ln l = -2 \ln \left[ \frac{L(x \mid H_0)}{L(x \mid H_1)} \right] $$
Bayesian Networks

§ A Bayesian network is a graphical representation of a probability distribution

§ Can use Bayesian networks to model relationships between genes and genetic regulatory networks

§ Advantages of using Bayesian network:
  • Compact and intuitive representation of gene relationships
  • Captures causal relationships between genes
  • Integration of prior knowledge into network
  • Works well with noisy data, a characteristic of gene expression data
  • Efficiently learn new models of gene relationships

Some material adapted from Hartemink, et al. (2001) *Pac Symp Biocomput.* 422-33
Bayesian Networks (continued)

§ Bayesian networks consist two parts: Qualitative and Quantitative

§ Qualitative part of network is a directed acyclic graph (DAG)
  • Nodes represent random variables (gene mRNA levels)
  • Edges represent direct (causal) influences between nodes (genes)
  • A positive edge from X to Y indicates that higher values of X will result in higher values of Y (bias distribution of Y higher)

Gene X ➔ Gene Y

  • A negative edge from X to Y indicates that higher values of X will result in lower values of Y (bias distribution of Y lower)

Gene X ➔ Gene Y

§ Quantitative part of network consists of a set of conditional probability distributions
Example of a Bayesian Network

§ Qualitative part:

Gene A → Gene B → Gene C

positive edge

negative edge

§ Quantitative part:

| Gene A | Gene B | P(C+ | AB) | P(C− | AB) |
|--------|--------|----------|-----------|
| +      | +      | 0.6      | 0.4       |
| −      | +      | 0.01     | 0.99      |
| +      | −      | 0.99     | 0.01      |
| −      | −      | 0.4      | 0.6       |

This row indicates that when Gene A and Gene B are up-regulated, then Gene C has a 60% probability to be up-regulated and a 40% probability to be down-regulated.
Testing Bayesian Networks Models

Microarray data can be used to test the validity of genetic regulatory networks written in the form of Bayesian networks.

Typically, two or more alternate network models will be constructed and tested using the microarray data.

A Bayesian scoring metric is used to identify the network model that best fits the microarray data.

\[
\text{BayesianScore}(M) = \log[P(M | D)] \\
= \log[P(M)] + \log[P(D | M)] + c
\]

- Where \( M \) = model, \( D \) = microarray data, \( c \) = constant

The term \( P(D|M) \) is the likelihood of the data \( D \), given the model \( M \).

We can calculate this term using the Maximum Likelihood method.
Hartemink et al. (2001) used 52 microarray data sets to score the competing models. Using the Bayesian scoring metric, Model 1 received a score of -44.0 while Model 2 received a score of -34.5. These scores indicate that the microarray data was over 13,000 times more likely to be observed under Model 2.
Learning Bayesian Networks from Microarray Data

§ Can also use microarray data as a basis for learning new Bayesian networks

§ Issues in using microarray data in learning Bayesian networks
  • Large number of variables (genes)
  • Relatively small number of microarray data sets
  • Sparse networks (few genes directly affect one another)

§ The Sparse Candidate algorithm (Friedman et al.) is commonly used in Bayesian Network learning
Example of Bayesian Network Learning

Image from Gifford, DK (2001) Science 293:2049-2051
Appendix: Maximum Likelihood Estimation

§ Want to choose the most likely model parameter of the mean $\mu$ given the experimental data $x$.

§ Following Bayes’ theorem and postulate: $P(B \mid A) \propto P(A \mid B)$

§ If the data $x$ follows a Gaussian, then the Maximum Likelihood Estimate (MLE) of $\mu$ is given by:

$$L = \frac{1}{(2\pi)^{n/2} \sigma^n} \prod_{i=0}^{n} e^{-(x_i - \mu)^2 / 2\sigma^2}$$

$$\frac{\partial \ln L}{\partial \mu} = 0 = \frac{1}{\sigma^2} \sum_{i=0}^{n} (x_i - \mu)$$

$$\mu = \frac{1}{n} \sum_{i=0}^{n} x_i$$