Nir Friedman
April 26, 2004
7.90 / 6.874 Lecture

Goal is Biological Data -> Biological Insight

One approach is Data -> Pattern Discovery
A second approach is Data -> Model

Begin with a real system (such as yeast) that produces (enironment specific) data
- DNA sequence
- Gene expression
- Protein-protein interaction

There is general agreement about how to measure mRNA (gene expression)
There is not a single "gold standard" for protein-protein interactions

What is a model?
- Something that explains our observed data

Model Components
- transcript level
- protein levels
- protein modifications
  - localization
- chromatin
  - promoter region

Model relationships
- Could begin by examining pair-wise correlations in absence of a model (no why)
- Could have a detailed list of equations (perhaps going overboard)
- It is hard to build a detailed model given the kinds of data that we have discussed

Let's look at possible relationships we could model
- Phosphorolation of a specific protein
- TF binding to a promoter

A specific example - TF binding to a promoter
- Binding site model
- Binding motifs in a a promoter
- Expression

Simple idea - a function from motifs -> expression
More complex model - takes into account the state of the cell
- Need to take into account transcription factor levels
- If we assume a linear relation, we can run into problems

How can we inject biological knowledge into our model?

Question - how much does it matter what underlying model you use?
Answer - it is very hard to compare the results as the methodologies are not standardized
- in the future there will be standards that will allow us to more directly compare methods

There is a good deal of hidden state in the systems we study
- True binding motifs
- TF activity levels
- Standard approach is to seed hidden variables with a good guess and adjust to fit the data
- Solution is to do cross-validation

If we learn on 80% and predict expression of 20%, and we do better than random, is this compelling?
- Perhaps, if we do not over fit by tuning the model to do well in cross validation

One approach is to take away a biological hypothesis, and test it explicitly

Protein-protein interactions
- Everything that happens in a cell involves protein-protein interactions

Given a protein-protein interaction
- Question 1 - do we really believe they are interacting?
  - we could make our measurements depend on that (yeast two hybrid, mass spec)
- Question 2 - how can we use other information we know about the proteins
  - e.g. localization - proteins that interact should be in the same neighborhood
  - perhaps we are uncertain about the localization

I(p, q) "interaction"
Loc(p, q) "localization"
Y2H(p, q) "yeast two hybrid assay"

P(Y2H(p,q) | I(p,q)) -- one way to examine the data
P(Y2H(p,q) | I(p,q), Loc(p, nuclear), Loc(q, nuclear)) -- conditioned upon localization

Bayesian network formulation:

P(I(p, q) | {Loc(p, c), Loc(q, c), c in Compartments})

Can rewrite this as a product of potential functions that put constrains on the probability

Another idea is to include weak transitivity into the model
- Could include the desirability to see triplets in the model