Discovering molecular pathways from protein interaction and gene expression data

By: E. Segal, H. Wang and D. Koller
Presented By: Benjamin Cooper, BBSI
Group Members: Emily Germain and Amanda Sitterly

Introduction

- Processes occur via pathways
- Global Picture
- Map Kinase Pathway

1. [Reference]
2. [Reference]
Integration

- Two Types of Data
  - Gene Expression
  - Protein-Protein Interaction

Procedure

- Genes Partitioned into arbitrary or intentional "pathways"
- Algorithm returns a higher score for "pathway" when expression levels match
- Maximizes data using a Expectation Maximization (EM) algorithm
Procedure

Probabilistic Model

- Naive Bayes Model using Markov Networks

\[ P(g.C, g.E_1, \ldots, g.E_m) = P(g.C) \prod_{j=1}^{m} P(g.E_j | g.C). \]
\[ P(g.C, g.E_1, \ldots, g.E_m) = P(g.C) \prod_{j=1}^{m} P(g.E_j \mid g.C). \]

- **G** – Set of Genes \( G = \{g_1, \ldots, g_n\} \)
- Each gene belongs to only one of \( k \) pathways \( g.C \in \{1, \ldots, k\} \)
- \( g.E_j \) – Represents the mRNA expression levels
- Conditional Probability Distribution \( P(g.E_j \mid g.C = p) \)
- uses a Gaussian Distribution

**Naive Bayes Model**
Protein-Protein Interactions

- If they interact they’re in the same pathway
- Markov networks

\[ P(V_1, \ldots, V_n) = \frac{1}{Z} \prod_{i=1}^{n} \phi_i(V_i) \prod_{[V_i - V_j] \in \mathcal{E}} \phi_{i,j}(V_i, V_j) \]

- Binary Markov Network
- Compatibility potential - \( \phi_{i,j}(v_i, v_j) \)
- \( \mathcal{V} \) - pathway assignments

\[ \phi_2(g_i.C = p, g_j.C = q) = \begin{cases} \alpha & p = q \\ 1 & \text{otherwise} \end{cases} \]

When \( \alpha > 1 \) there is said to be a correlation in pathways
Protein Interaction Model

Unified Model

- Uses the term $P(g_f, C)$ which is present in both models
Maximization

- Expectation Maximization of the probability
- Two Steps
  - E-Step
  - M-Step
- These steps are then iterated until convergence

Results

- Two yeast gene expression data sets
  - 173 Microarrays under various stresses
  - 77 Microarrays during the cell cycle
- Protein-Protein Interaction data set from DIP with 10705 interactions
- Used an algorithm that partitioned the genes into 60 groups initially
Evaluation

- Prediction of held-out interactions
- Coherence of pathways according to functional annotation
- Convergence of protein complexes

Overall the results obtained by this combined method were better than standard clustering methods.
Proteins that participate in:
- Translation
- Protein Degradation
- Transcription
- DNA Replication
Conclusions

- Very useful model
- Best available now
- Has a very serious limitation of only allowing a gene to be in one pathway

References


2) http://www.bio.davidson.edu/courses/immunology/Flash/MAPK.html