Network-assisted data analysis

Bing Zhang
Department of Biomedical Informatics
Vanderbilt University
bing.zhang@vanderbilt.edu
Protein identification in shotgun proteomics

Protein digestion

LC-MS/MS

Database search

Protein assembly

VIQQLEGAFALVF
IQALAEETAQNLK
......
YPIEHGIITNWDDMEK
Protein assembly and classification

Current protein assembly pipelines treat proteins as individual entities. Biologically interesting proteins may be eliminated due to insufficient experimental evidence.

Most biological functions arise from interactions among proteins. Can we use protein interaction network information to improve protein identification?

Hypothesis: an eliminated protein is more likely to be present in the original sample if it lies close to other confidently identified proteins in the protein interaction network.
Do confidently identified proteins cluster together?

Li et al. Mol Syst Biol, 5:303, 2009
Neighborhood majority voting

- Direct neighbors (local)
- Iterative majority voting (global)
Module-based analysis

Direct neighbor voting

Module-based

Peptide identification

Protein assembly

Network mapping

Confident

Non-confident
Method evaluation:
Cross validation for proteins with known labels

- **Yeast**
  - Yeast cell culture dataset
  - YPIN, 5,666 nodes; 126,126 edges
  - Results: sensitivity of 56% at the specificity of 90%. CEA is more accurate and robust than other methods

- **Mouse**
  - Mouse organ datasets: brain, lung, placenta
  - MPIN1: 9,776 nodes; 69,470 edges
  - MPIN2: 12,271 nodes; 236,675 edges
  - Results: sensitivity of ~45% at the specificity of 90%

*Li et al. Mol Syst Biol, 5:303, 2009*
Method evaluation: Supporting evidence for rescued proteins

- **Mouse-MPIN1**

- **Proteins rescued**
  - Brain: 171 (12%)
  - Lung: 181 (10%)
  - Placenta: 156 (11%)

- **Independent evidences**
  - Microarray
  - EST library
  - Publication

Application: Breast cancer data set (normal vs tumor)

- **Rescued proteins**
  - Normal: 139 (23%)
  - Tumor: 95 (8%)

- **Rescued cancer-related proteins**
  - Ctnnb1
  - Top1
  - …

- **Cancer specific sub-networks**
  - Wnt signaling pathway
  - Cell adhesion
  - Apoptosis
  - …

Li et al. Mol Syst Biol, 5:303, 2009
Extended hypothesis

- **Hypothesis:** an eliminated protein is more likely to be present in the original sample if it lies close to other confidently identified proteins in the protein interaction network.

- **Extended hypothesis:** Proteins lie close to each other in a protein interaction network are more likely to share common attributes.
  - Protein expression
  - Protein function
  - Disease association
Protein function prediction: motivation

Gene Ontology annotation coverage in different organisms

Network-based protein function prediction

- Proteins that lie closer to one another in a protein interaction network are more likely to have similar function and involve in similar biological process.

- GO semantic similarity

Disease gene prioritization: motivation

- Most common genetic disorders may result from variants in many genes, each contributing only weak effects to the disease.
- Quantitative trait locus mapping and genome-wide association studies identify a large set of candidate disease genes.
- Assessing and selecting genes for validation is nontrivial.
- Causal genes for the same disease are often involved in a few biological processes or molecular pathways.
  - Limb-girdle muscular dystrophy: many disease genes involve in the dystrophy complex.
  - Fanconi anemia: 5 out of the 10 disease genes identified in a genetic study function in a nuclear complex.
Network-based disease causal gene prioritization: method

Network-based disease causal gene prioritization: result

Disease biomarker identification: gene-level approach

- Scoring individual genes to generate a marker set
- Use gene expression matrix to train a classifier
- High variability: two breast cancer biomarker sets of ~70 genes each shared only 3 genes in common
- Doesn’t improve mechanistic understanding of cancer

Disease biomarker identification: pathway-level approach

- Scoring known pathways by the coherency of expression changes among their member genes
- More robust
- More sensitive
- Better interpretation
- The majority of human genes have not yet been assigned to a definitive pathway
Disease biomarker identification: network approach

- Overlay gene expression data on the protein interaction network
- Search for subnetworks that are highly discriminative of metastasis
- Use subnetwork activity matrix to train a classifier

Subnetwork markers are more robust

Subnetwork markers are more informative

- 47.3% and 65.4% subnetwork markers showed functional enrichment
- 66 and 153 subnetworks corresponded to the major events that have been implicated in the progression of cancer
  - Signaling of cell growth and survival
  - Cell proliferation and replication
  - Apoptosis
  - Cell and tissue remodeling
  - Circulation and coagulation
  - Metabolism
- Detecting important non-differentially expressed genes
  - TP53, KARS, HARS, ERBB2, PIK3CA

Summary

- Guilt by association: Proteins lie close to each other in a protein interaction network are more likely to share common attributes.
  - Protein expression
  - Protein function
  - Disease-related proteins

- Network-based disease biomarkers
Key references