SISG
2014

SISG Module 22: Pathway & Network Analysis for Omics Data

19th Summer Institute in Statistical Genetics

UNIVERSITY of WASHINGTON

(This page left intentionally blank.)

Introduction to Pathway and Network Analysis

Alison Motsinger-Reif, PhD
Associate Professor
Bioinformatics Research Center
Department of Statistics
North Carolina State University

Pathway and Network Analysis

- High-throughput genetic/genomic technologies enable comprehensive monitoring of a biological system
- Analysis of high-throughput data typically yields a list of differentially expressed genes, proteins, metabolites…
  - Typically provides lists of single genes, etc.
  - Will use "genes" throughout, but using interchangeably mostly
- This list often fails to provide mechanistic insights into the underlying biology of the condition being studied
- How to extract meaning from a long list of differentially expressed genes → pathway/network analysis

What makes an airplane fly?

Chas' Stainless Steel, Mark Thompson's Airplane Parts, About 1000 Pounds of Stainless Steel Wire, and Gagosian's Beverly Hills Space

From components to networks

A biological function is a result of many interacting molecules and cannot be attributed to just a single molecule.

Pathway and Network Analysis

- One approach: simplify analysis by grouping long lists of individual genes into smaller sets of related genesreduces the complexity of analysis.
  - a large number of knowledge bases developed to help with this task
- Knowledge bases
  - describe biological processes, components, or structures in which individual genes \are known to be involved in
  - how and where gene products interact with each other

Pathway and Network Analysis

- Analysis at the functional level is appealing for two reasons:
  - First, grouping thousands of genes by the pathways they are involved in reduces the complexity to just several hundred pathways for the experiment
  - Second, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of genes

Pathway and Network Analysis

- What kinds of data is used for such analysis?
  - Gene expression data
    - Microarrays
    - RNA-seq
  - Proteomic data
  - Metabolomics data
  - Single nucleotide polymorphisms (SNPs)
  - ….

Pathway and Network Analysis

- What kinds of questions can we ask/answer with these approaches?

Pathway and Network Analysis

- The term "pathway analysis" gets used often, and often in different ways
  - applied to the analysis of Gene Ontology (GO) terms (also referred to as a "gene set")
  - physical interaction networks (e.g., protein–protein interactions)
  - kinetic simulation of pathways
  - steady-state pathway analysis (e.g., flux-balance analysis)
  - inference of pathways from expression and sequence data
- May or may not actually describe biological pathways

Pathway and Network Analysis

- For the first part of this module, we will focus on methods that exploit pathway knowledge in public repositories rather than on methods that infer pathways from molecular measurements
  - Use repositories such as GO or Kyoto Encyclopedia of Genes and Genomes (KEGG)

→ *knowledge base–driven pathway analysis*

A History of Pathway Analysis Approaches

- Over a decade of development of pathway analysis approaches
- Can be *roughly* divided into three generations:
  - 1st: Over-Representation Analysis (ORA) Approaches
  - 2nd : Functional Class Scoring (FCS) Approaches
  - 3rd : Pathway Topology (PT)-Based Approaches

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

- The data generated by an experiment using a high-throughput technology (e.g., microarray, proteomics, metabolomics), along with functional annotations (pathway database) of the corresponding genome, are input to virtually all pathway analysis methods.
- ORA methods require that the input is a list of differentially expressed genes
- FCS methods use the entire data matrix as input
- PT-based methods additionally utilize the number and type of interactions between gene products, which may or may not be a part of a pathway database.
- The result of every pathway analysis method is a list of significant pathways in the condition under study.

Over-Representation Analysis (ORA) Approaches

- Earliest methods → over-representation analysis (ORA)
- Statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression
- It is also referred to as "2×2 table method" in the literature

Over-Representation Analysis (ORA)

- Uses one or more variations of the following strategy:
  - First, an input list is created using a certain threshold or criteria
    - For example, may choose genes that are differentially over- or under-expressed in a given condition at a false discovery rate (FDR) of 5%
  - Then, for each pathway, input genes that are part of the pathway are counted
  - This process is repeated for an appropriate background list of genes
    - (e.g., all genes measured on a microarray)
  - Next, every pathway is tested for over- or under-representation in the list of input genes
    - The most commonly used tests are based on the hypergeometric, chi-square, or binomial distribution

ORA tools	
Onto-Express	Web (http://vortex.cs.wayne.edu)
GenMAPP	Standalone (http://www.genmapp.org)
GoMiner	Standalone, Web (http://discover.nci.nih.gov/gominer)
FatiGO	Web (http://babelomics.bioinfo.cipf.es)
GOstat	Web (http://gostat.wehi.edu.au)
FuncAssociate	Web (http://llama.mshri.on.ca/funcassociate/)
GOToolBox	Web (http://genome.crg.es/GOToolBox/)
GeneMerge	Standalone, Web (http://genemerge.cbcb.umd.edu/)
GOEAST	Web (http://omicslab.genetics.ac.cn/GOEAST/)
ClueGO	Standalone (http://www.ici.upmc.fr/cluego/)
FunSpec	Web (http://funspec.med.utoronto.ca/)
GARBAN	Web
GO:TermFinder	Standalone (http://search.cpan.org/dist/GO-TermFinder/)
WebGestalt	Web (http://bioinfo.vanderbilt.edu/webgestalt/)
agriGO	Web (http://bioinfo.cau.edu.cn/agriGO/)
GOFFA	Standalone, Web (http://edkb.fda.gov/webstart/arraytrack/)
WEGO	Web (http://wego.genomics.org.cn/cgi-bin/wego/index.pl)

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Limitations of ORA Approaches

- First, the different statistics used by ORA are independent of the measured changes
  - (e.g., hypergeometric distribution, binomial distribution, chi-square distribution, etc.)
- Tests consider the number of genes alone but ignore any values associated with them
  - such as probe intensities
- By discarding this data, ORA treats each gene equally
  - Information about the extent of regulation (e.g., fold-changes, significance of a change, etc.) can be useful in assigning different weights to input genes/pathways
  - This can provide more information

Limitations of ORA Approaches

- Second, ORA typically uses only the most significant genes and discards the others
  - input list of genes is usually obtained using an arbitrary threshold (e.g., genes with fold-change and/or p-values)
- Marginally less significant genes are missed, resulting in information loss
  - (e.g., fold-change = 1.999 or p-value = 0.051)
  - A few methods avoiding thresholds
    - They use an iterative approach that adds one gene at a time to find a set of genes for which a pathway is most significant

Limitations of ORA Approaches

- Third, ORA assumes that each gene is independent of the other genes
- However, biology is a complex web of interactions between gene products that constitute different pathways
  - One goal might be to gain insights into how interactions between gene products are manifested as changes in expression
  - A strategy that assumes the genes are independent is significantly limited in its ability to provide insights
- Furthermore, assuming independence between genes amounts to "competitive null hypothesis" testing (more later), which ignores the correlation structure between genes
  - the estimated significance of a pathway may be biased or incorrect

Limitations of ORA Approaches

- Fourth, ORA assumes that each pathway is independent of other pathways → NOT TRUE!
- Examples of dependence:
  - GO defines a biological process as a series of events accomplished by one or more ordered assemblies of molecular functions
  - The cell cycle pathway in KEGG where the presence of a growth factor activates the MAPK signaling pathway
    - This, in turn, activates the cell cycle pathway
- No ORA methods account for this dependence between molecular functions in GO and signaling pathways in KEGG

Functional Class Scoring (FCS) Approaches

- *The hypothesis of functional class scoring (FCS) is that although large changes in individual genes can have significant effects on pathways, weaker but coordinated changes in sets of functionally related genes (i.e., pathways) can also have significant effects*
- With few exceptions, all FCS methods use a variation of a general framework that consists of the following three steps.

Step 1

- First, a gene-level statistic is computed using the molecular measurements from an experiment
  - Involves computing differential expression of individual genes or proteins
- Statistics currently used at gene-level include correlation of molecular measurements with phenotype
  - ANOVA
  - Q-statistic
  - signal-to-noise ratio
  - t-test
  - Z-score

Step 1

- Choice of a gene-level statistic generally has a negligible effect on the identification of significantly enriched gene sets
  - However, when there are few biological replicates, a regularized statistic may be better
- Untransformed gene-level statistics can fail to identify pathways with up- and down-regulated genes
  - In this case, transformation of gene-level statistics (e.g., absolute values, squared values, ranks, etc.) is better

Step 2

- Second, the gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic
  - can be multivariate and account for interdependencies among genes
  - can be univariate and disregard interdependencies among genes
- The pathway-level statistics used include:
  - Kolmogorov-Smirnov statistic
  - sum, mean, or median of gene-level statistic
  - Wilcoxon rank sum
  - maxmean statistic

Step 2

- Irrespective of its type, the power of a pathway-level statistic depends on
  - the proportion of differentially expressed genes in a pathway
  - the size of the pathway
  - the amount of correlation between genes in the pathway
- Univariate statistics show more power at stringent cutoffs when applied to real biological data, and equal power as multivariate statistics at less stringent cutoffs

Step 3

- Assessing the statistical significance of the pathway-level statistic
- When computing statistical significance, the null hypothesis tested by current pathway analysis approaches can be broadly divided into two categories:
  - i) competitive null hypothesis
  - ii) self-contained null hypothesis
- A self-contained null hypothesis permutes class labels (i.e., phenotypes) for each sample and compares the set of genes in a given pathway with itself, while ignoring the genes that are not in the pathway
- A competitive null hypothesis permutes gene labels for each pathway, and compares the set of genes in the pathway with a set of genes that are not in the pathway

FCS tools	
GSEA	Standalone (http://www.broadinstitute.org/gsea/)
sigPathway	Standalone (BioConductor)
Category	Standalone (BioConductor)
SAFE	Standalone (BioConductor)
GlobalTest	Standalone (BioConductor)
PCOT2	Standalone (BioConductor)
SAM-GS	Standalone (http://www.ualberta.ca/~yyasui/software.html)
Catmap	Standalone (http://bioinfo.thep.lu.se/catmap.html)
T-profiler	Web (http://www.t-profiler.org)
FunCluster	Standalone (http://corneliu.henegar.info/FunCluster.htm)
GeneTrail	Web (http://genetrail.bioinf.uni-sb.de)
GAzer	Web

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Advantages of FCS Methods

FCS methods address three limitations of ORA

1. Don't require an arbitrary threshold for dividing expression data into significant and non-significant pools.
   Rather, FCS methods use all available molecular measurements for pathway analysis.

2. While ORA completely ignores molecular measurements when identifying significant pathways, FCS methods use this information in order to detect coordinated changes in the expression of genes in the same pathway

3. By considering the coordinated changes in gene expression, FCS methods account for dependence between genes in a pathway

Limitations of FCS Methods

- First, similar to ORA, FCS analyzes each pathway independently
  - Because a gene can function in more than one pathway, meaning that pathways can cross and overlap
  - Consequently, in an experiment, while one pathway may be affected in an experiment, one may observe other pathways being significantly affected due to the set of overlapping genes

- Such a phenomenon is very common when using the GO terms to define pathways due to the hierarchical nature of the GO

Limitations of FCS Methods

- Second, many FCS methods use changes in gene expression to rank genes in a given pathway, and discard the changes from further analysis
  - For instance, assume that two genes in a pathway, A and B, are changing by 2-fold and 20-fold, respectively
  - As long as they both have the same respective ranks in comparison with other genes in the pathway, most FCS methods will treat them equally, although the gene with the higher fold-change should probably get more weight
- Importantly, however, considering only the ranks of genes is also advantageous, as it is more robust to outliers.
  - A notable exception to this scenario is approaches that use gene-level statistics (e.g., t-statistic) to compute pathway-level scores.
  - For example, an FCS method that computes a pathway-level statistic as a sum or mean of the gene-level statistic accounts for a relative difference in measurements (e.g., Category, SAFE).

Pathway Topology (PT)-Based Approaches

- A large number of publicly available pathway knowledge bases provide information beyond simple lists of genes for each pathway
  - KEGG
  - MetaCyc
  - Reactome
  - RegulonDB
  - STKE
  - BioCarta
  - PantherDB
  - ….
- Unlike GO and MSigDB, these knowledge bases also provide information about gene products that interact with each other in a given pathway, how they interact (e.g., activation, inhibition, etc.), and where they interact (e.g., cytoplasm, nucleus, etc.)

Pathway Topology (PT)-Based Approaches

- ORA and FCS methods consider only the number of genes in a pathway or gene coexpression to identify significant pathways, and ignore the additional information available from these knowledge bases
  - Even if the pathways are completely redrawn with new links between the genes, as long as they contain the same set of genes, ORA and FCS will produce the same results
- Pathway topology (PT)-based methods have been developed to use the additional information
  - PT-based methods are essentially the same as FCS methods in that they perform the same three steps as FCS methods
  - The key difference between the two is the use of pathway topology to compute gene-level statistics

Pathway Topology (PT)-Based Approaches

- Rahnenfuhrer et al. proposed ScorePAGE, which computes similarity between each pair of genes in a pathway (e.g., correlation, covariance, etc.)
  - similarity measurement between each pair of genes is analogous to gene-level statistics in FCS methods
  - averaged to compute a pathway-level score
- Instead of giving equal weight to all pairwise similarities, ScorePAGE divides the pairwise similarities by the number of reactions needed to connect two genes in a given pathway

Pathway Topology (PT)-Based Approaches

- Impact factor (IF) analysis
  - IF considers the structure and dynamics of an entire pathway by incorporating a number of important biological factors, including changes in gene expression, types of interactions, and the positions of genes in a pathway

Ali will talk more about these approaches in detail!!!

IF Analysis

- Briefly…
  - Models a signaling pathway as a graph, where nodes represent genes and edges represent interactions between them
  - Defines a gene-level statistic, called perturbation factor (PF) of a gene, as a sum of its measured change in expression and a linear function of the perturbation factors of all genes in a pathway
  - Because the PF of each gene is defined by a linear equation, the entire pathway is defined as a linear system
    - addresses loops in the pathways
  - The IF of a pathway (pathway-level statistic) is defined as a sum of PF of all genes in a pathway

Pathway Topology (PT)-Based Approaches

- FCS methods that use correlations among genes implicitly assume that the underlying network, as defined by the correlation structure, does not change as the experimental conditions change
- This assumption may be inaccurate → PT approaches improve on this

Pathway Topology (PT)-Based Approaches

- NetGSA accounts for the the change in correlation as well as the change in network structure as experimental conditions change
  - like IF analysis, models gene expression as a linear function of other genes in the network
- it differs from IF in two aspects
  - First, it accounts for a gene's baseline expression by representing it as a latent variable in the model
  - Second, it requires that the pathways be represented as directed acyclic graphs DAGs
    - If a pathway contains cycles, NetGSA requires additional latent variables affecting the nodes in the cycle.
    - In contrast, IF analysis does not impose any constraint on the structure of a pathway

Limitations of PT-based Approaches

- True pathway topology is dependent on the type of cell due to cell-specific gene expression profiles and condition being studied
  - information is rarely available
  - fragmented in knowledge bases if available
  - As annotations improve, these approaches are expected to become more useful
- Inability to model dynamic states of a system
- Inability to consider interactions between pathways due to weak inter-pathway links to account for interdependence between pathways

PT-based tools	
ScorePAGE	No implementation available
Pathway-Express	Web (http://vortex.cs.wayne.edu)
SPIA	Standalone (BioConductor)
NetGSA	No implementation available

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Outstanding Challenges

- Broad Categories:

1. annotation challenges
2. methodological challenges

Outstanding Challenges

- Next generation approaches will require improvement of the existing annotations
  - necessary to create accurate, high resolution knowledge bases with detailed condition-, tissue-, and cell-specific functions of each gene
    - PharmGKB ….
  - these knowledge bases will allow investigators to model an organism's biology as a dynamic system, and will help predict changes in the system due to factors such as mutations or environmental changes

Annotation Challenges

- Low resolution knowledge bases
- Incomplete and inaccurate annotations
- Missing condition- and cell-specific information

Green arrows represent abundantly available information, and red arrows represent missing and/or incomplete information. The ultimate goal of pathway analysis is to analyze a biological system as a large, single network. However, the links between smaller individual pathways are not yet well known. Furthermore, the effects of a SNP on a given pathway are also missing from current knowledge bases. While some pathways are known to be related to a few diseases, it is not clear whether the changes in pathways are the cause for those diseases or the downstream effects of the diseases.

Low Resolution Knowledge Bases

- Knowledge bases not as high resolution as technologies
  - using RNA-seq, more than 90% of the human genome is estimated to be alternatively spliced
  - multiple transcripts from the same gene may have related, distinct, or even opposing functions
  - GWAS have identified a large number of SNPs that may be involved in different conditions and diseases.
  - However, current knowledge bases only specify which genes are active in a given pathway
  - Essential that they also begin specifying other information, such as transcripts that are active in a given pathway or how a given SNP affects a pathway

Low Resolution Knowledge Bases

- Because of these low resolution knowledge bases, every available pathway analysis tool first maps the input to a non-redundant namespace, typically an Entrez Gene ID
  - this type of mapping is advantageous, although it can be non-trivial, as it allows the existing pathway analysis approaches to be independent of the technology used in the experiment
  - However, mapping in this way also results in the loss of important information that may have been provided because a specific technology was used
    - XRN2a, a variant of gene XRN2, is expressed in several human tissues, whereas another variant of the same gene, XRN2b, is mainly expressed in blood leukocytes
    - Although RNA-seq can quantify expression of both variants, mapping both transcripts to a single gene causes loss of tissue-specific information, and possibly even condition-specific information

Low Resolution Knowledge Bases

- Therefore, before pathway analysis can exploit current and future technological advances in biotechnology, it is critically important to annotate exact transcripts and SNPs that participate in a given pathway
- While new approaches are being developed in this regard, they may not yet be adequate
  - Braun et al. proposed a method for analyzing SNP data from a GWAS
  - Still relies on mapping multiple SNPs to a single gene, followed by gene-to-pathway mapping

Incomplete and Inaccurate Annotation

- A surprisingly large number of genes are still not annotated
- Many of the genes are hypothetical, predicted, or pseudogenes
  - Although the number of protein-coding genes in the human genome is estimated to be between 20,000 and 25,000, according Entrez Gene, there are 45,283 human genes, of which 14,162 are pseudogenes
  - One could argue that the pseudogenes should not be included when evaluating functional annotation coverage
  - pseudogene-derived small interfering RNAs have been shown to regulate gene expression in mouse oocytes
  - GO provides annotations for 271 pseudogenes
  - A widely used DNA microarray, Affymetrix HG U133 plus 2.0, contains 1,026 probe sets that correspond to 823 pseudogenes
  - Should pseudogenes be included in the count when estimating annotation coverage for the human genome?

Incomplete and Inaccurate Annotation

Number of GO-annotated genes (left panel) and number of GO annotations (right panel) for human from January 2003 to November 2009.As the estimated number of known genes in the human genome is adjusted (between January 2003 and December 2003) and annotation practices are modified (between December 2004 and December 2005, and between October 2008 and November 2009), one can argue that, although the number of annotated genes and the annotations are decreasing (which is mainly due to the adjusted number of genes in the human genome and changes in the annotation process), the quality of annotations is improving, as demonstrated by the steady increase in non-IEA annotations and the number of genes with non-IEA annotations. However, the increase in the number of genes with non-IEA annotations is very slow. In almost 7 years, between January 2003 and November 2009, only 2,039 new genes received non-IEA annotations. At the same time, the number of non-IEA annotations increased from 35,925 to 65,741, indicating a strong research bias for a small number of genes.doi:10.1371/journal.pcbi.1002375.g003

Incomplete and Inaccurate Annotation

- Additionally, many of the existing annotations are of low quality and may be inaccurate
  - >95% of the annotations in the October 2007 release of GO had the evidence code "inferred from electronic annotations (IEA)"
  - the only ones in GO that are not curated manually
  - Annotations inferred from indirect evidence are considered to be of lower quality than those derived from direct experimental evidence
  - If the annotations with IEA code are removed, the number of genes with good quality annotations in the November 2009 release of human GO annotations is reduced from 18,587 to 11,890

Incomplete and Inaccurate Annotation

- It is very likely that the reduced number of annotations and annotated genes since January 2003 is an indicator of improving quality
- This is due in part to the fact that the number of genes in a genome are continuously being adjusted and the functional annotation algorithms are being improved
  - the number of non-IEA annotations is continuously increasing
- However, the rate of increase for non-IEA annotations is very slow (approximately 2,000 genes annotated in 7 years)

Incomplete and Inaccurate Annotation

- Manual curation of the entire genome is expected to take a very long time (~13–25 years)
- Entire research community could participate in the curation process
- One approach to facilitate participation of a large number of researchers is to adopt a standard annotation format similar to Minimum Information About a Microarray Experiment (MIAME)
  - should this be required like GEO?
- A format for functional annotation can be designed or adopted from the existing formats (e.g., BioPAX, SBML)
  - Such a format could allow researchers to specify an experimentally confirmed role of a specific transcript or a SNP in a pathway along with experimental and biological conditions

Missing Condition and cell-specific information

- Most pathway knowledge bases are built by curating experiments performed in different cell types at different time points under different conditions
- These details are typically not available in the knowledge bases!
- One effect of this omission is that multiple independent genes are annotated to participate in the same interaction in a pathway
- This effect is so widespread that many pathway knowledge bases represent a set of distinct genes as a single node in a pathway

Missing Condition and cell-specific information

- Example: *Wnt/beta-catenin pathway in STKE*
  - the node labeled "Genes" represents 19 genes directly targeted by Wnt in different organisms (Xenopus and human) in different cells and tissues (colon carcinoma cells and epithelial cells
  - these non-specific genes introduce bias for these pathways in all existing analysis approaches
  - For instance, any ORA method will assign higher significance (typically an order of magnitude lower p-value) to a pathway with more genes
  - Similarly, more genes in a pathway also increase the probability of a higher pathway-level statistic in FCS approaches, yielding higher significance for a given pathway.

Missing Condition and cell-specific information

- This contextual information is typically not available from most of the existing knowledge bases
- A standard functional annotation format discussed above would make this information available to curators and developers
  - For instance, the recently proposed Biological Connection Markup Language (BCML) allows pathway representation to specify the cell or organism in which each pathway interaction occurs.
  - BCML can generate cell-, condition-, or organism-specific pathways based on user-defined query criteria, which in turn can be used for targeted analysis

Missing Condition and cell-specific information

- Existing knowledge bases do not describe the effects of an abnormal condition on a pathway
  - For example, it is not clear how the Alzheimer's disease pathway in KEGG differs from a normal pathway
  - Nor it is clear which set of interactions leads to Alzheimer's disease
- We are now understanding that context plays an important role in pathway interactions
- Information about how cell and tissue type, age, and environmental exposures affect pathway interactions will add complexity that is currently lacking

Methodological Challenges

- Benchmark data sets for comparing different methods
- Inability to model and analyze dynamic response
- Inability to model effects of an external stimuli

Comparing Different Methods

- How do we compare different pathway analysis methods?
- Simulated data
  - Advantages:
    - Real signal is simulated, so "true" answer is known
  - Disadvantages
    - Cannot contain all the complexity of real data
    - The success of the methods can reflect the similarity of how well the simulation matches the knowledgebase structure used

Comparing Different Methods

- Benchmark data
  - Advantages:
    - Can compare sensitivity and specificity
    - Several datasets have been consistently used in the literature
    - Includes all the complexity of real biological data
  - Disadvantages
    - Affected by confounding factors
      - absence of a pure division into classes
      - presence of outliers
      - ….
    - No true answer known for grounded comparisons – actual biology isnt known

Comparing Different Methods

- A general challenge: *Different definitions of the same pathway in different knowledge bases can affect performance assessment*
  - GO defines different pathways for apoptosis in different cells
    - (e.g., cardiac muscle cell apoptosis, B cell apoptosis, T cell apoptosis)
    - Further distinguishes between induction and regulation of apoptosis
  - KEGG defines a single signaling pathway for apoptosis
    - does not distinguish between induction and regulation
  - An approach using KEGG would identify a single pathway as significant, whereas GO could identify multiple pathways, and/or specific aspects of a single apoptosis pathway

Inability to model and analyze dynamic response

- No existing approach can collectively model and analyze high-throughput data as a single dynamic system
- Current approaches analyze a snapshot assuming that each pathway is independent of the others at a given time
  - measure expression changes at multiple time points, and analyze each time point individually
  - Implicitly assumes that pathways at different time points are independent
- Need models that accounts for dependence among pathways at different time points
  - Much of this limitation is due to technology/experimental design → not all bioinformatics limitations

Inability to model effects of an external stimuli

- Gene set–based approaches often only consider genes and their products
- Completely ignore the effects of other molecules participating in a pathway
  - such as the rate limiting step of a multi-step pathway.
- Example:
  - The amount/strength of Ca^{2+} causes different transcription factors to be activated
  - This information is usually not available.

Summary

- In the last decade, pathway analysis has matured, and become the standard for trying to dissect the biology of high throughput experiments.
- Many similarities across the three main generations of pathway analysis tools.
- Will discuss more details of some of these choices, knowledge bases, and specific approaches next.
- Many open methods development challenges!

Overview of Module

- First Half:
  - Overview of gene set and pathway analysis
    - Commonly used databases and annotation issues
    - 1st and 2nd generation tools
      - Basic differences in methods
      - Details on very popular methods
    - Issues with different "omics" data types
- Second Half
  - Network inference methods
  - "3rd generation" methods for pathway enrichment

Questions?

motsinger@stat.ncsu.edu

Pathway and Gene Set Analysis Part 1

Alison Motsinger-Reif, PhD
Bioinformatics Research Center
Department of Statistics
North Carolina State University
motsinger@stat.ncsu.edu

The early steps of a microarray study

- Scientific Question (biological)
- Study design (biological/statistical)
- Conducting Experiment (biological)
- Preprocessing/Normalizing Data (statistical)
- Finding differentially expressed genes (statistical)

A data example

- Lee et al (2005) compared adipose tissue (abdominal subcutaenous adipocytes) between obese and lean Pima Indians
- Samples were hybridised on HGu95e-Affymetrix arrays (12639 genes/probe sets)
- Available as GDS1498 on the GEO database
- We selected the male samples only
  - 10 obese vs 9 lean

Diabetologia (2005) 48: 1776–1783
DOI 10.1007/s00125-005-1867-3

ARTICLE

Y. H. Lee · S. Nair · E. Rousseau · D. B. Allison ·
G. P. Page · P. A. Tataranni · C. Bogardus ·
P. A. Permana

Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese Pima Indians: increased expression of inflammation-related genes

Received: 10 December 2004 / Accepted: 28 April 2005 / Published online: 30 July 2005
© Springer-Verlag 2005

Abstract *Aims/hypothesis:* Obesity increases the risk of developing major diseases such as diabetes and cardiovascular disease. Adipose tissue, particularly adipocytes, may play a major role in the development of obesity and its comorbidities. The aim of this study was to characterise, in adipocytes from obese people, the most differentially expressed genes that might be relevant to the development of obesity. *Methods:* We carried out microarray gene profiling of isolated abdominal subcutaneous adipocytes from 20 non-obese (BMI 25±3 kg/m^2) and 19 obese (BMI 55±8 kg/m^2) non-diabetic Pima Indians using Affymetrix HG-U95 GeneChip arrays. After data analyses, we measured the transcript levels of selected genes based on their biological functions and chromosomal positions using quantitative real-time PCR. *Results:* The most differentially expressed genes in adipocytes of obese individuals consisted of 433 upregulated and 244 downregulated genes. Of these, 410 genes could be classified into 20 functional Gene Ontology categories. The analyses indicated that the inflammation/immune response category was over-represented, and that most inflammation-related genes were upregulated in adipocytes of obese subjects. Quantitative real-time PCR confirmed the transcriptional upregulation of representative inflammation-related genes (*CCL2* and *CCL3*) encoding the chemokines monocyte chemoattractant protein-1 and macrophage inflammatory protein 1α. The differential expression levels of eight positional candidate genes, including inflammation-related *THY1* and *C1QTNF5*, were also confirmed. These genes are located on chromosome 11q22–q24, a region with linkage to obesity in the Pima Indians. *Conclusions/interpretation:* This study provides evidence supporting the active role of mature adipocytes in obesity-related inflammation. It also provides potential candidate genes for susceptibility to obesity.

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00125-005-1867-3.

The "Result"

Probe Set ID	log.ratio	pvalue	adj.p
73554_at	1.4971	0.0000	0.0004
91279_at	0.8667	0.0000	0.0017
74099_at	1.0787	0.0000	0.0104
83118_at	-1.2142	0.0000	0.0139
81647_at	1.0362	0.0000	0.0139
84412_at	1.3124	0.0000	0.0222
90585_at	1.9859	0.0000	0.0258
84618_at	-1.6713	0.0000	0.0258
91790_at	1.7293	0.0000	0.0350
80755_at	1.5238	0.0000	0.0351
85539_at	0.9303	0.0000	0.0351
90749_at	1.7093	0.0000	0.0351
74038_at	-1.6451	0.0000	0.0351
79299_at	1.7156	0.0000	0.0351
72962_at	2.1059	0.0000	0.0351
88719_at	-3.1829	0.0000	0.0351
72943_at	-2.0520	0.0000	0.0351
91797_at	1.4676	0.0000	0.0351
78356_at	2.1140	0.0001	0.0359
90268_at	1.6552	0.0001	0.0421

What happened to the Biology???

Slightly more informative results

Probe Set ID	Gene Symbo	Gene Title	go biological process term	go molecular function terr	log.ratio	pvalue	adj.p
73554_at	CCDC80	coiled-coil domain contair	---	---	1.4971	0.0000	0.0004
91279_at	C1QTNF5 ///	C1q and tumor necrosis f	visual perception /// embr	---	0.8667	0.0000	0.0017
74099_at	---	---	---	---	1.0787	0.0000	0.0104
83118_at	RNF125	ring finger protein 125	immune response /// mod	protein binding /// zinc ion	-1.2142	0.0000	0.0139
81647_at	---	---	---	---	1.0362	0.0000	0.0139
84412_at	SYNPO2	synaptopodin 2	---	actin binding /// protein bi	1.3124	0.0000	0.0222
90585_at	C15orf59	chromosome 15 open rea	---	---	1.9859	0.0000	0.0258
84618_at	C12orf39	chromosome 12 open rea	---	---	-1.6713	0.0000	0.0258
91790_at	MYEOV	myeloma overexpressed	---	---	1.7293	0.0000	0.0350
80755_at	MYOF	myoferlin	muscle contraction /// blo	protein binding	1.5238	0.0000	0.0351
85539_at	PLEKHH1	pleckstrin homology dom	---	binding	0.9303	0.0000	0.0351
90749_at	SERPINB9	serpin peptidase inhibitor,	anti-apoptosis /// signal tr	endopeptidase inhibitor a	1.7093	0.0000	0.0351
74038_at	---	---	---	---	-1.6451	0.0000	0.0351
79299_at	---	---	---	---	1.7156	0.0000	0.0351
72962_at	BCAT1	branched chain aminotrar	G1/S transition of mitotic	catalytic activity /// branch	2.1059	0.0000	0.0351
88719_at	C12orf39	chromosome 12 open rea	---	---	-3.1829	0.0000	0.0351
72943_at	---	---	---	---	-2.0520	0.0000	0.0351
91797_at	LRRC16A	leucine rich repeat contair	---	---	1.4676	0.0000	0.0351
78356_at	TRDN	triadin	muscle contraction	receptor binding	2.1140	0.0001	0.0359
90268_at	C5orf23	chromosome 5 open read	---	---	1.6552	0.0001	0.0421

If we are lucky, some of the top genes mean something to us

But what if they don't?

And how what are the results for other genes with similar biological functions

How to incorporate biological knowledge

- The type of knowledge we deal with is rather simple:

  We know groups/sets of genes that for example
  - Belong to the same pathway
  - Have a similar function
  - Are located on the same chromosome, etc…

- We will assume these groupings to be given, i.e. we will not yet discuss methods used to detect pathways, networks, gene clusters
  - We will later!

What is a pathway?

- No clear definition
  - Wikipedia: "In biochemistry, **metabolic pathways** are series of chemical reactions occurring within a cell. In each pathway, a principal chemical is modified by chemical reactions."
  - These pathways describe enzymes and metabolites
- But often the word "pathway" is also used to describe gene regulatory networks or protein interaction networks
- In all cases a pathway describes a biological function very specifically

What is a Gene Set?

- Just what it says: a set of genes!
  - All genes involved in a pathway are an example of a Gene Set
  - All genes corresponding to a Gene Ontology term are a Gene Set
  - All genes mentioned in a paper of Smith et al might form a Gene Set
- A Gene Set is a much more general and less specific concept than a pathway
- Still: we will sometimes use two words interchangeably, as the analysis methods are mainly the same

Where Do Gene Sets/Lists Come From?

- Molecular profiling e.g. mRNA, protein
  - Identification → Gene list
  - Quantification → Gene list + values
  - Ranking, Clustering (biostatistics)
- Interactions: Protein interactions, Transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
  - Single nucleotide polymorphisms (SNPs)
  - Copy number variants (CNVs)
  - ……..

What is Gene Set/Pathway analysis?

- The aim is to give one number (score, p-value) to a Gene Set/Pathway
  - Are many genes in the pathway differentially expressed (up-regulated/downregulated)
  - Can we give a number (p-value) to the probability of observing these changes just by chance?

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping
- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Gene Attributes

- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3' UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Gene Attributes

- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3' UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Database Resources

- Use functional annotation to aggregate genes into pathways/gene sets
- A number of databases are available
  - Different analysis tools link to different databases
  - Too many databases to go into detail on every one
  - Commonly used resources:
    - GO
    - KeGG
    - MsigDB
    - WikiPathways

Pathway and Gene Set data resources

- The Gene Ontology (GO) database
  - http://www.geneontology.org/
  - GO offers a relational/hierarchical database
  - Parent nodes: more general terms
  - Child nodes: more specific terms
  - At the end of the hierarchy there are genes/proteins
  - At the top there are 3 parent nodes: biological process, molecular function and cellular component
- Example: we search the database for the term "inflammation"

The genes on our array that code for one of the 44 gene products would form the corresponding "inflammation" gene set

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
  - protein kinase
  - apoptosis
  - membrane
- Ontology: A formal system for describing knowledge

GO Structure

- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child

GO Structure

Species independent. Some lower-level terms are specific to a group, but higher level terms are not

What GO Covers?

- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process

Cell division

glucose-6-phosphate isomerase activity

Terms

- Where do GO terms come from?
  - GO terms are added by editors at EBI and gene annotation database groups
  - Terms added by request
  - Experts help with major development
  - 27734 terms, 98.9% with definitions.
    - 16731 biological_process
    - 2385 cellular_component
    - 8618 molecular_function

Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - Known as 'gene associations' or GO annotations
  - Multiple annotations per gene
- Some GO annotations created automatically

Annotation Sources

- Manual annotation
  - Created by scientific curators
    - High quality
    - Small number (time-consuming to create)
- Electronic annotation
  - Annotation derived without human validation
    - Computational predictions (accuracy varies)
    - Lower 'quality' than manual codes
- Key point: be aware of annotation origin

Evidence Types

- **ISS:** Inferred from Sequence/Structural Similarity
- **IDA:** Inferred from Direct Assay
- **IPI:** Inferred from Physical Interaction
- **IMP:** Inferred from Mutant Phenotype
- **IGI:** Inferred from Genetic Interaction
- **IEP:** Inferred from Expression Pattern
- **TAS:** Traceable Author Statement
- **NAS:** Non-traceable Author Statement
- **IC:** Inferred by Curator
- **ND:** No Data available

- **IEA:** Inferred from electronic annotation

Species Coverage

- All major eukaryotic model organism species
- Human via GOA group at UniProt
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development

Variable Coverage

Lomax J. Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinform. 2005 Sep;6(3):298-304.

Contributing Databases

- Berkeley *Drosophila* Genome Project (BDGP)
- dictyBase (*Dictyostelium discoideum)*
- FlyBase (*Drosophila melanogaster)*
- GeneDB (*Schizosaccharomyces pombe, Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei)*
- UniProt Knowledgebase (Swiss-Prot/TrEMBL/PIR-PSD) and InterPro databases
- Gramene (grains, including rice, *Oryza*)
- Mouse Genome Database (MGD) and Gene Expression Database (GXD) (*Mus musculus)*
- Rat Genome Database (RGD) (*Rattus norvegicus)*
- Reactome
- *Saccharomyces* Genome Database (SGD) (*Saccharomyces cerevisiae)*
- The *Arabidopsis* Information Resource (TAIR) (*Arabidopsis thaliana)*
- The Institute for Genomic Research (TIGR): databases on several bacterial species
- WormBase (*Caenorhabditis elegans)*
- Zebrafish Information Network (ZFIN): (*Danio rerio)*

GO Slim Sets

- GO has too many terms for some uses
  - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
  - Generic, plant, yeast

GO Software Tools

- GO resources are freely available to anyone without restriction
  - Includes the ontologies, gene associations and tools developed by GO
- Other groups have used GO to create tools for many purposes
  - http://www.geneontology.org/GO.tools

Accessing GO: QuickGO

http://www.ebi.ac.uk/ego/

Other Ontologies

http://www.ebi.ac.uk/ontology-lookup

KEGG pathway database

- KEGG = Kyoto Encyclopedia of Genes and Genomes
  - http://www.genome.jp/kegg/pathway.html
  - The pathway database gives far more detailed information than GO
    - Relationships between genes and gene products
  - But: this detailed information is only available for selected organisms and processes
  - Example: Adipocytokine signaling pathway

KEGG pathway database

- Clicking on the nodes in the pathway leads to more information on genes/proteins
  - Other pathways the node is involved with
  - Entries in Gene/Protein databases
  - References
  - Sequence information
- Ultimately this allows to find corresponding genes on the microarray and define a Gene Set for the pathway

Wikipathways

- http://www.wikipathways.org
- A wikipedia for pathways
  - One can see and download pathways
  - But also edit and contribute pathways
- The project is linked to the GenMAPP and Pathvisio analysis/visualisation tools

MSigDB

- MSigDB = Molecular Signature Database

http://www.broadinstitute.org/gsea/msigdb

- Related to the the analysis program GSEA
- MSigDB offers gene sets based on various groupings
  - Pathways
  - GO terms
  - Chromosomal position,…

Some Warnings

- In many cases the definition of a pathway/gene set in a database might differ from that of a scientist
- The nodes in pathways are often proteins or metabolites; the activity of the corresponding gene set is not necessarily a good measurement of the activity of the pathway
- There are many more resources out there (BioCarta, BioPax)
- Commercial packages often use their own pathway/gene set definitions (Ingenuity, Metacore, Genomatix,…)
- Genes in a gene set are usually not given by a Probe Set ID, but refer to some gene data base (Entrez IDs, Unigene IDs)
    - Conversion can lead to errors!

Some Warnings

- In many cases the definition of a pathway/gene set in a database might differ from that of a scientist
- The nodes in pathways are often proteins or metabolites; the activity of the corresponding gene set is not necessarily a good measurement of the activity of the pathway
- There are many more resources out there (BioCarta, BioPax)
- Commercial packages often use their own pathway/gene set definitions (Ingenuity, Metacore, Genomatix,…)
- Genes in a gene set are usually not given by a Probe Set ID, but refer to some gene data base (Entrez IDs, Unigene IDs)
    - Conversion can lead to errors!

Gene Attributes

- Functional annotation
  - Biological process, molecular function, cell location
- Chromosome position
- Disease association
- DNA properties
  - TF binding sites, gene structure (intron/exon), SNPs
- Transcript properties
  - Splicing, 3' UTR, microRNA binding sites
- Protein properties
  - Domains, secondary and tertiary structure, PTM sites
- Interactions with other genes

Sources of Gene Attributes

- Ensembl BioMart (eukaryotes)
  - http://www.ensembl.org
- Entrez Gene (general)
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
- Model organism databases
  - E.g. SGD: http://www.yeastgenome.org/
- Many others…..

Ensembl BioMart

- Convenient access to gene list annotation

Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
  - E.g. Social Insurance Number, Entrez Gene ID 41232
- Gene and protein information stored in many databases
  - → Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
  - Important to recognize the correct record type
  - E.g. Entrez Gene records don't store sequence. They link to DNA regions, RNA transcripts and proteins.

NCBI Database Links

NCBI:
U.S. National Center for Biotechnology Information

Part of National Library of Medicine (NLM)

http://www.ncbi.nlm.nih.gov/Database/datamodel/data_nodes.swf

Common Identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

Annotations
InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757

Experimental Platform
Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

Red = Recommended

Identifier Mapping

- So many IDs!
  - Mapping (conversion) is a headache
- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Genes to proteins, Entrez Gene to Affy
  - Unification during dataset merging
    - Equivalent records

ID Mapping Services

*	entrezgene
YIL062C	854748
YLR370C	851085
YKL013C	853856
YNR035C	855771
YBR234C	852536

- Synergizer
  - http://llama.med.harvard.edu/synergizer/translate/
- Ensembl BioMart
  - http://www.ensembl.org
- UniProt
  - http://www.uniprot.org/

ID Mapping Challenges

- Avoid errors: map IDs correctly
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping

- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Goals

- Pathway and gene set data resources
  - Gene attributes
  - Database resources
    - GO, KeGG, Wikipathways, MsigDB
  - Gene identifiers and issues with mapping
- Differences between pathway analysis tools
  - Self contained vs. competitive tests
  - Cut-off methods vs. global methods
  - Issues with multiple testing

Aims of Analysis

- Reminder: The aim is to give one number (score, p-value) to a Gene Set/Pathway
  - Are many genes in the pathway differentially expressed (up-regulated/downregulated)?
  - Can we give a number (p-value) to the probability of observing these changes just by chance?
  - Similar to single gene analysis statistical hypothesis testing plays an important role

General differences between analysis tools

- Self contained vs competitive test
  - The distinction between “self-contained” and “competitive” methods goes back to Goeman and Buehlman (2007)
  - A self-contained method only uses the values for the genes of a gene set
    - The null hypothesis here is: H = {“No genes in the Gene Set are differentially expressed”}
  - A competitive method compares the genes within the gene set with the other genes on the arrays
    - Here we test against H: {“The genes in the Gene Set are not more differentially expressed than other genes”}

Example: Analysis for the GO-Term “inflammatory response” (GO:0006954)

Back to the Real Data Example

- Using Bioconductor software we can find 96 probesets on the array corresponding to this term
- 8 out of these have a p-value < 5%
- How many significant genes would we expect by chance?
- Depends on how we define "by chance"

The "self-contained" version

- By chance (i.e. if it is NOT differentially expressed) a gene should be significant with a probability of 5%
- We would expect 96 x 5% = 4.8 significant genes
- Using the binomial distribution we can calculate the probability of observing 8 or more significant genes as p = 10.8%, i.e. not quite significant

The "competitive" version

- Overall 1272 out of 12639 genes are significant in this data set (10.1%)
- If we randomly pick 96 genes we would expect 96 x 10.1% = 9.7 genes to be significant "by chance"
- A p-value can be calculated based on the 2x2 table
- Tests for association: Chi-Square-Test or Fisher's exact test

	In GS	Not in GS
sig	8	1264
non-sig	88	11 279

P-value from Fisher's exact test (one-sided): 73.3%, i.e very far from being significant

Competitive Tests

- Competitive results depend highly on how many genes are on the array and previous filtering
  - On a small targeted array where all genes are changed, a competitive method might detect no differential Gene Sets at all
- Competitive tests can also be used with small sample sizes, even for n=1
  - BUT: The result gives no indication of whether it holds for a wider population of subjects, the p-value concerns a population of genes!
- Competitive tests typically give less significant results than self-contained (as seen with the example)
- Fisher's exact test (competitive) is probably the most widely used method!

Cut-off methods vs whole gene list methods

- A problem with both tests discussed so far is, that they rely on an arbitrary cut-off
- If we call a gene significant for 10% p-value threshold the results will change
  - In our example the binomial test yields p= 2.2%, i.e. for this cut-off the result is significant!
- We also lose information by reducing a p-value to a binary ("significant", "non-significant") variable
  - It should make a difference, whether the non-significant genes in the set are nearly significant or completely unsignificant

- We can study the distribution of the p-values in the gene set
- If no genes are differentially expressed this should be a uniform distribution
- A peak on the left indicates, that some genes are differentially expressed
- We can test this for example by using the Kolmogorov-Smirnov-Test
- Here p = 8.2%, i.e. not quite significant
- This would be a "self-contained" test, as only the genes in the gene set are being used

Kolmogorov-Smirnov Test

- The KS-test compares an observed with an expected cumulative distribution
- The KS-statistic is given by the maximum deviation between the two

- Alternatively we could look at the distribution of the RANKS of the p-values in our gene set
- This would be a competitive method, i.e we compare our gene set with the other genes
- Again one can use the Kolmogorov-Smirnov test to test for uniformity
- Here: p= 85.1%, i.e. very far from significance

Other general issues

- Direction of change
  - In our example we didn't differentiate between up or down-regulated genes
  - That can be achieved by repeating the analysis for p-values from one-sided test
    - Eg. we could find GO-Terms that are significantly up-regulated
  - With most software both approaches are possible
- Multiple Testing
  - As we are testing many Gene Sets, we expect some significant findings "by chance" (false positives)
  - Controlling the false discovery rate is tricky: The gene sets do overlap, so they will not be independent!
    - Even more tricky in GO analysis where certain GO terms are subset of others
  - The Bonferroni-Method is most conservative, but always works!

Multiple Testing for Pathways

- Resampling strategies (dependence between genes)
  - The methods we used so far in our example assume that genes are independent of each other…if this is violated the p-values are incorrect
  - Resampling of group/phenotype labels can correct for this
  - We give an example for our data set

Example Resampling Approach

1. Calculate the test statistic, e.g. the percentage of significant genes in the Gene Set
2. Randomly re-shuffle the group labels (lean, obese) between the samples
3. Repeat the analysis for the re-shuffled data set and calculate a re-shuffled version of the test statistic
4. Repeat 2 and 3 many times (thousands…)
5. We obtain a distribution of re-shuffled % of significant genes: the percentage of re-shuffled values that are larger than the one observed in 1 is our p-value

Resampling Approach

- The reshuffling takes gene to gene correlations into account
- Many programs also offer to resample the genes: This does NOT take correlations into account
- Roughly speaking:
  - Resampling phenotypes: corresponds to self-contained test
  - Resampling genes: corresponds to competitive test

Resampling Approaches

- Genes being present more than once
  - Common approaches
    - Combine duplicates (average, median, maximum,…)
    - Ignore (i.e treat duplicates like different genes)

SAFE package

- Using summary statistics vs using all data
  - Our examples used p-values as data summaries
  - Other approaches use fold-changes, signal to noise ratios, etc…
  - Some methods are based on the original data for the genes in the gene set rather than on a summary statistic

Resampling Approaches

- The resampling approaches are highly computationally intensive
- New methods are being developed to speed this up
  - Empirical approximations of permutations
  - Empirical pathway analysis, without permutation.
    - Zhou YH, Barry WT, Wright FA.Biostatistics. 2013 Jul; 14(3):573-85. doi: 10.1093/biostatistics/kxt004. Epub 2013 Feb 20.

Summary

- Databases
- Choice makes a difference
- Not all use the same IDs – watch out ☺
- Major differences between methods
- Issues with multiple testing

- Next lecture, will go into more detail on a few methods

Questions?

Pathway and Gene Set Analysis Part 2

Alison Motsinger-Reif, PhD
Bioinformatics Research Center
Department of Statistics
North Carolina State University
motsinger@stat.ncsu.edu

Goals

Some methods in more detail

- TopGO
- Global Ancova
- Pathvisio/Genmapp
- Impact Factor Analysis
- GSEA

Some methods in detail

- There are far too many methods to give a comprehensive overview

BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 3. 189–197
Advance Access publication January 17, 2008
doi:10.1093/bib/bbn001

Gene-set approach for expression pattern analysis

Dougu Nam and Seon-Young Kim

Submitted: 7th November 2007; Received (in revised form): 28th December 2007

Abstract

Recently developed gene set analysis methods evaluate differential expression patterns of gene groups instead of those of individual genes. This approach especially targets gene groups whose constituents show subtle but coordinated expression changes, which might not be detected by the usual individual gene analysis. The approach has been quite successful in deriving new information from expression data, and a number of methods and tools have been developed intensively in recent years. We review those methods and currently available tools, classify them according to the statistical methods employed, and discuss their pros and cons. We also discuss several interesting extensions to the methods.

Keywords: *gene set analysis; DNA microarray; differential expression of genes*

Table of methods (from Nam & Kim)

Table 1: Cutoff-free gene set analysis methods

Authors	Year	Name	Statistical test	Self-contained versus competitive	Gene versus ample randomization	Reference
Virtaneva *et al.*	2001		sample randomization	self-contained	sample	[8]
Pavlidis *et al.*	2002		gene randomization	competitive	gene	[9]
Mootha *et al.*	2003	GSEA	sample randomization	mixed	sample	[7]
Breslin *et al.*	2004	Catmap	gene randomization	competitive	gene	[3]
Goeman *et al.*	2004	globaltest	sample randomization	self-contained	sample	[17]
Smid *et al.*	2004	GO-Mapper	z-test	competitive	gene	[38]
Volinia *et al.*	2004	GOAL	gene randomization	competitive	gene	[39]
Barry *et al.*	2005	SAFE	sample randomization	competitive	sample	[19]
Beh-Shaul *et al.*	2005		Kolmogorov–Smirnov test	competitive	gene	[5]
Boorsma e al.	2005	T-profiler	*t*-test	competitive	gene	[15]
Kim *et al.*	2005	PAGE	z-test	competitive	gene	[14]
Lee *et al.*	2005	ErmineJ	sample randomization	competitive	gene	[16]
Subramanian *et al.*	2005	GSEA	sample randomization	mixed	gene	[25]
Tian *et al.*	2005	QI, Q2	gene or sample randomization	competitive or self-contained	gene or sample	[10]
Tomfohr *et al.*	2005	PLAGE	sample randomization	self-contained	sample	[20]
Edelman *et al.*	2006	ASSESS	sample randomization	competitive	sample	[28]
Kong *et al.*	2006		Hotelling's T squared	self-contained	sample	[21]
Nam *et al.*	2006	ADGO	z-test	competitive	gene	[29]
Saxena *et al.*	2006	AE	sample randomization	competitive	sample	[31]
Scheer *et al.*	2006	JProGO	Fisher's exact test, Kolmogorov–Smirnov test, *t*-test, unpaired Wilcoxon's test	competitive	gene	[40]
Al-Shahrour *et al.*	2007	Fatiscan	Fisher's exact test, hypergeometric test	competitive	gene	[41]
Backes *et al.*	2007	GeneTrail	Fisher's exact test, hypergeometric test, sample randomization	competitive	gene or sample	[42]
Cavalieri *et al.*	2007	Eu.Gene Analyzer	Fisher's exact test, sample randomization	competitive	gene or sample	[43]
Dinu *et al.*	2007	SAM-GS	sample randomization	self-contained	sample	[22]
Efron *et al.*	2007	GSA	sample randomization	mixed	sample	[26]
Newton *et al.*	2007	Random set	z-test	competitive	gene	[44]

Table of software (from Nam & Kim)

Table 2: Gene set analysis tools

Name	Organism[a]	Application Type	URL	Reference
ADGO	H, M, R, Y	Web server	http://array.kobic.re.kr/ADGO	[29]
ASSESS	H, M, R	Octave/Java standalone	http://people.genome.duke.edu/~jhg9/assess/	[28]
Babelomics	H, M, R, DM, S, C	Web server	http://www.babelomics.org	[45]
Catmap	H	Perl script	http://bioinfo.thep.lu.se/catmap.html	[3]
ErmineJ	H, M, R	Java standalone	http://www.bioinformatics.ubc.ca/ermineJ/	[16]
EuGene Analyzer	H, M, R, Y	Windows/Unix standalone	http://www.ducciocavalieri.org/bio/Eugene.htm	[43]
FatiScan	H, M, R, Y, B, D, G, C, A, S, DM	Web server	http://fatiscan.bioinfo.cipf.es/	[41]
GAZER	H, M, R, Y	Web server	http://integromics.kobic.re.kr/GAzer/index.faces;	[13]
GeneTrail	H, M, R, Y, SA, CG, AT	Web server	http://genetrail.bioinf.uni-sb.de/	[42]
Global test	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/globaltest.html	[17]
GOAL	H, M	Web server	http://microarrays.unife.it	[39]
GO-Mapper	H, M, R, Z, DM, Y	Windows standalone, Perl script	http://www.gatcplatform.nl/	[38]
GSA	H	R package	http://www-stat.stanford.edu/~tibs/GSA/	[26]
GSEA	H	Java standalone, R package	http://www.broad.mit.edu/gsea/	[25]
JProGO	Various prokaryotes	Web server	http://www.jprogo.de/	[40]
MEGO	H	Windows standalone	http://www.dxy.cn/mego/	[46]
PAGE	H, M, R, Y	Python script	From the author (kimsy@kribb.re.kr)	[14]
PLAGE	H, M	Web server	http://dulci.biostat.duke.edu/pathways/	[20]
SAFE	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/safe.html	[19]
SAM-GS	NA	Windows Excel Add-In	http://www.ualberta.ca/~yyasui/homepage.html	[22]
T-profiler	Y, CA	Web server	http://www.t-profiler.org/	[15]

[a]H: *Homo sapiens*; M: *Mus musculus*; R: *Rattus norvegicus*; Y: *Sacchaomyces cerevisiae*; B: *BosTaurus*; D: *Daniel rerio*; G: *Gallus gallus*; C: *Caenorhabditis elegans*; A: *Arabidopsis thaliana*; DM: *Drosophila melanogaster*; Z: *Zebra fish*; CA: *Candida albicans*; SA: *Staphylococcus aureus*; CG: *Corynebacterium glutamicum*; AT: *Arabidopsis thaliana*.

TopGO

- TopGO is a GO term analysis program available from Bioconductor
- It takes the GO hierarchy into account when scoring terms
- If a parent term is only significant because of child term, it will receive a lower score
- TopGO uses the Fisher-test or the KS-test (both competitive)
- TopGO also gives a graphical representation of the results in form of a tree

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1600–1607
doi:10.1093/bioinformatics/btl140

Gene expression

Improved scoring of functional groups from gene expression data by decorrelating GO graph structure

Adrian Alexa*, Jörg Rahnenführer and Thomas Lengauer
Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

Received on September 28, 2005; revised on March 30, 2006; accepted on April 4, 2006
Advance Access publication April 10, 2006
Associate Editor: Martin Bishop

Tree showing the 15 most significant GO terms

Zooming in

Global Ancova

Testing Differential Gene Expression in Functional Groups

Goeman's Global Test versus an ANCOVA Approach

U. Mansmann[1], R. Meister[2]
[1]IBE, Biometry and Bioinformatics, University of Munich, Munich, Germany
[2]Fachbereich II, University of Applied Sciences, Berlin, Germany

- Uses all data (instead of summary statistics)
- NOT a multivariate method (MANOVA)
- One linear model for all genes within the gene set
  - Gene is a factor in the model that interacts with other factors
- Full model (e.g. including difference between lean and obsese) is compared with restricted model (no difference)
- P-values are calculated by group label resampling
- Algorithm allows for complex linear models including covariates
- Related to Goeman's Globaltest, which reverses roles of gene expression and groups: Goeman uses gene expression to explain groups (logistic regression)

10 most significant KEGG pathways according to Global Ancova

Pathway Name	path.size	sig.genes	perc.sig	p.gs	p.fisher	p.globaltest	p.globalAncova
Pantothenate and CoA biosynthesis	11	3	27.27%	7.05%	9.08%	0.55%	0.01%
Valine, leucine and isoleucine biosynthesis	4	2	50.00%	4.10%	5.29%	0.22%	0.02%
Cell Communication	60	10	16.67%	8.77%	7.51%	1.02%	0.03%
PPAR signaling pathway	37	10	27.03%	11.01%	0.28%	1.64%	0.07%
Inositol metabolism	1	1	100.00%	8.46%	10.06%	0.19%	0.10%
Valine, leucine and isoleucine degradation	35	7	20.00%	49.56%	5.65%	1.42%	0.11%
Fatty acid metabolism	27	6	22.22%	49.59%	4.81%	1.54%	0.31%
ECM-receptor interaction	49	8	16.33%	4.91%	11.45%	1.47%	0.83%
Focal adhesion	122	16	13.11%	76.63%	16.40%	2.59%	0.87%
Purine metabolism	78	14	17.95%	26.82%	2.26%	3.42%	1.21%

p.gs = A GSEA related competitive method (available in Limma)

p.fisher = Fisher-Test (competitive)

Genmapp/Pathvisio

- These are two pathway visualisation tools that collaborate
  - http://www.genmapp.org
  - http://www.pathvisio.org
- Both do some basic statistical analysis too (Fisher-Test with normal approximation)
- Main focus is on visually displaying pathways
  - Genes/nodes can be color-coded according to the data
  - Results (p-values, fold changes) can be displayed next to genes/nodes

Impact Factor Analysis

- Impact Factor (IF) analysis combines both ORA and FCS approach, while accounting for the topology of the pathway

- IF analysis computes Perturbation Factor (PF) for each gene in each pathway, which is a gene-level statistic, as follows:

$$PF(g_i) = \Delta F(g_i) + \sum_{j=1}^{n} \beta_{ji} \cdot \frac{PF(g_j)}{N_{ds}(g_j)}$$

- The first term, $\Delta F(g_i)$, represents the signed normalized measured expression change (i.e., fold change) of the gene g_i
- The second term accounts for the topology of the pathway, where gene g_j is upstream of gene g_i
- In the second term, β_{ji} represents the type and strength of interaction between g_j and g_i
- If g_j activates g_i, $\beta_{ji} = 1$, and if g_j inhibits g_i, $\beta_{ji} = -1$
- Note that the PF of the upstream gene g_j is normalized by the number of downstream genes it interacts with, $N_{ds}(g_i)$
- The second term is repeated for every gene g_j that is upstream of gene g_i

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

The 1st term captures the significance of the given pathway P_i as provided by ORA, where p_i corresponds to the probability of obtaining a value of the statistic used at least as extreme as the one observed when the null hypothesis is true

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Because IF should be large for severely impacted pathways (i.e., small p-values), the 1st term uses $1/p_i$ rather than p_i

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Log function is necessary to map the exponential scale of the p-values to a linear scale in order to keep the model linear

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

The 2nd term sums up the values of the PFs for all genes *g* on the given pathway P_i, and is normalized by the number of differentially expressed genes on the given pathway P_i

Impact Factor Analysis

- Note that Eq. 1 essentially describes the perturbation factor PF for a gene g_i as a linear function of the perturbation factors of all genes in a given pathway
- Therefore, the set of all equations defining the PFs for all genes in a given pathway P_i form a system of simultaneous equations
- Expanding and re-arranging Equation 1 for all genes g_1, g_2, …., g_n in a pathway Pi can be re-written as follows:

$$\begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} 1 - \frac{\beta_{11}}{N_{ds(g_1)}} & -\frac{\beta_{21}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n1}}{N_{ds(g_n)}} \\ -\frac{\beta_{12}}{N_{ds(g_1)}} & 1 - \frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n2}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\beta_{1n}}{N_{ds(g_1)}} & -\frac{\beta_{2n}}{N_{ds(g_2)}} & \cdots & 1 - \frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha(g_1) \cdot \Delta E(g_1) \\ \alpha(g_2) \cdot \Delta E(g_2) \\ \cdots \\ \alpha(g_n) \cdot \Delta E(g_n) \end{pmatrix}$$

Impact Factor Analysis

$$\begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} 1-\frac{\beta_{11}}{N_{ds(g_1)}} & -\frac{\beta_{21}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n1}}{N_{ds(g_n)}} \\ -\frac{\beta_{12}}{N_{ds(g_1)}} & 1-\frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n2}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\beta_{1n}}{N_{ds(g_1)}} & -\frac{\beta_{2n}}{N_{ds(g_2)}} & \cdots & 1-\frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha(g_1)\cdot\Delta E(g_1) \\ \alpha(g_2)\cdot\Delta E(g_2) \\ \cdots \\ \alpha(g_n)\cdot\Delta E(g_n) \end{pmatrix}$$

- After computing the PFs of all genes in a given pathway as the solution of this linear system, Eq. 2 is used to calculate the impact factor of each pathway
- The impact factor of each pathway is then used as a score to assess the impact of a given gene expression data set on all pathways (the higher the impact factor the more significant the pathway)

Gene Set Enrichment Analysis (GSEA)

- GSEA can be used with any gene set
- It is available as a standalone program, and versions of GSEA available within R/Bioconductor
- GSEA has many options and is a mix of a competitive and self-contained method
  - Default methods is to use a Kolmogorov Smirnov-type statistic to test the distribution of the gene set in the ranked gene list (competitive)
  - Typically that statistic ("enrichment score") is tested by permuting/ reshuffling the group labels (self-contained)
- Two Key Papers
  - Mootha et al., Nature Genetics 34, 267–273 (2003)
  - Subramanian et al., PNAS 102(43), 15545–15550 (2005).
    - Note - the description of GSEA changed between the two papers.

K-S Test

The Kolmogorov–Smirnov test is used to determine whether two underlying one-dimensional probability distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution, in either case based on finite samples.

Kolmogorov-Smirnov Test

- Based on statistics of 'Brownian Bridge'
  - random walk fixed end
- Maximum difference is test statistic
  - Null distribution known
- Reformulated by GSEA as difference of CDF – uniform from axis

K-S Test Finds Irrelevant Sets

- Sometimes ranks concentrated in middle
  - K-S statistic high, but not meaningful for path change
- Fix: ad-hoc weighting by actual t-scores emphasizes departures at extreme ends
- No theory
- Generate null distribution by permutation

GSEA Algorithm: Step 1

- Calculate an Enrichment Score:
  - Rank genes by their expression difference
  - Compute cumulative sum over ranked genes:
    - Increase sum when gene in set, decrease it otherwise
    - Magnitude of increment depends on correlation of gene with phenotype.
- Record the maximum deviation from zero as the enrichment score

The rows represent the samples or chips, and the columns represent the genes

Highly expressed in diseased

Lowly expressed in diseased

- Genes on the left side are highly expressed on the top half (indicated by red color) and lowly expressed on the bottom half (indicated by blue color). The reverse is shown on the right-most genes
- Created a gradient or ranked list corresponding to the degree of correlation with the two phenotypes

- This is depicted nicely by the graph on the bottom of the figure, where the positive ranks on the left represent the correlation to the Disease phenotype and the negative ranks on the right signify the correlation to the Normal phenotype
- The graph also generates a rank gradient that represents the order of the most up-regulated genes for the Disease sample on the left-most, and the most up-regulated genes for the Normal samples on the right-most

- Now, let's hide the heatmap and replace the middle part of the figure with genes from a specific geneset, say genes from the Glycolysis pathway.
- Each vertical blue bars represents a gene from the pathway, being mapped on the same location as the whole dataset
- Again, genes that are located on the left side are highly expressed on the Disease samples, and the opposite is true for the right-most genes

- Now, we are ready to demonstrate the GSEA algorithm.
- The walk down algorithm basically scans the ranked gene list *L*, and when a member of *S* is encountered, an Enrichment Score (*ES*) is registered. This is illustrated on the top part of the figure below; when the ES started to build upon encountering more genes from the GeneSet *S*.

- The more *S* genes is found, the higher the ES

- But, when no *S* genes were encountered for a long walk down, as indicated on the middle section of the middle plot, the ES will decrease accordingly.
- In other words, a high ES relies intimately with the clustering of *S* genes in close proximity. In this example, we would conclude that the *S* genes have high degree of correlation with the Disease phenotype since most of the ES was gained from the left portion of the plot

GSEA Algorithm: Step 1

- Calculate an Enrichment Score:
  - Rank genes by their expression difference
  - Compute cumulative sum over ranked genes:
    - Increase sum when gene in set, decrease it otherwise
    - Magnitude of increment depends on correlation of gene with phenotype
- Record the maximum deviation from zero as the enrichment score

GSEA Algorithm: Step 1

Subramanian et al., PNAS 102(43), 15545–15550 (2005).

GSEA Algorithm: Step 2

- Assess significance:
  - Permute phenotype labels 1000 times
  - Compute ES score as above for each permutation
  - Compare ES score for actual data to distribution of ES scores from permuted data
- Permuting the phenotype labels instead of the genes maintains the complex correlation structure of the gene expression data

GSEA Algorithm: Step 3

- Adjustment for multiple hypothesis testing:
  - Normalize the ES accounting for size of each gene set, yielding normalized enrichment score (NES)
  - Control proportion of false positives by calculating FDR corresponding to each NES, by comparing tails of the observed and null distibutions for the NES

GSEA Algorithm: Step 4

- The original method used equal weights for each gene
  - The revised method weighted genes according to their correlation with phenotype
  - This may cause an asymmetric distribution of ES scores if there is a big difference in the number of genes highly correlated to each phenotype
- Consequently, the above algorithm is performed twice: one for the positively scoring gene sets and once for the negatively scoring gene sets

Overview of GSEA

Subramanian et al., PNAS 102(43), 15545–15550 (2005).

GSEA results for our data set (using pathway gene sets)

Enrichment in phenotype: lean (10 samples)

- 19 / 44 gene sets are upregulated in phenotype **lean**
- 0 gene sets are significant at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 1 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: obese (9 samples)

- 25 / 44 gene sets are upregulated in phenotype **obese**
- 0 gene sets are significantly enriched at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 3 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 12639 native features
- After collapsing features into gene symbols, there are: 6465 genes

Gene set details

- Gene set size filters (min=25, max=500) resulted in filtering out 595 / 639 gene sets
- The remaining 44 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

List of most significant up-regulated gene sets

Table: Gene sets enriched in phenotype lean (10 samples) [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
1	HSA04910_INSULIN_SIGNALING_PATHWAY	Details ...	51	0.37	1.41	0.036	0.960	0.620	1184
2	CALCINEURIN_NF_AT_SIGNALING	Details ...	32	0.39	1.33	0.074	0.833	0.800	2413
3	HSA04514_CELL_ADHESION_MOLECULES	Details ...	41	0.36	1.26	0.188	0.805	0.880	2038
4	HSA04310_WNT_SIGNALING_PATHWAY	Details ...	52	0.29	1.13	0.278	1.000	0.970	1086
5	HSA04350_TGF_BETA_SIGNALING_PATHWAY	Details ...	29	0.33	1.11	0.302	1.000	0.970	647
6	HSA05215_PROSTATE_CANCER	Details ...	28	0.38	1.11	0.291	0.914	0.970	1360
7	HSA04010_MAPK_SIGNALING_PATHWAY	Details ...	73	0.28	1.03	0.477	1.000	0.990	1482

Table: GSEA Results Summary

Dataset	Pimaunlog2_collapsed_to_symbols.Pima
Phenotype	Pima.cls
Upregulated in class	lean
GeneSet	HSA04910_INSULIN_SIGNALING_PATHWAY
Enrichment Score (ES)	0.3685702
Normalized Enrichment Score (NES)	1.4148962
Nominal p-value	0.035714287
FDR q-value	0.96008533
FWER p-Value	0.62

The Enrichment score is based on the difference of the cumulative distribution of the gene-set minus the expected

This plot is basically the Kolmogorov-Smirnov plot rotated by 45 degrees

Zoom In on Enrichment Plot

Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

GSEA Software

http://www.broad.mit.edu/gsea/

Outlook

- Gene Set and Pathway Analysis is a very active field of research: new methods are published all the time!
- One important aspect: taking pathway structure into account
  - All methods we discuss ignored this structure
  - New methods use and "Impact Factor" (IF), which gives more weight to gene that are key regulators in the pathway (Draghici et al (2007))
- Other Aspects:
  - Study the behavior of pathways across experiments in microarray databases like GEO or Array Express
  - Incorporate other data into the analysis (proteomics, metabolomics, sequence data)

Summary

- There are many popular databases/internet resources for pathways and gene sets
- Many important analysis issues
- It is impossible to explain all existing approaches but many of them are some combinations of the methods we discussed
- This is an active field: improvements and further developments are a really active area of research

Questions?

Pathway/ Gene Set Analysis in Genome-Wide Association Studies

Alison Motsinger-Reif, PhD
Associate Professor
Bioinformatics Research Center
Department of Statistics
North Carolina State University

Goals

- Methods for GWAS with SNP chips
  - Integrating expression and SNP information

Many Shared Issues

- Many of the issues/choices/methodological approaches discussed for microarray data are true across all "-omics"
- Many methods have been readily extended for other omic data
- There are several biological and technological issues that may make just "off the shelf" use of pathway analysis tools inappropriate

Genome-Wide Association Studies

Population resources
- trios
- case-control samples

Whole-genome genotyping
- hundreds of thousands or million(s) of markers, typically SNPs

Genome-wide Association
- single SNP alleles
- genotypes
- multimarker haplotypes

Advantages of GWAS

- Compared to candidate gene studies
  - unbiased scan of the genome
  - potential to identify totally novel susceptibility factors
- Compared to linkage-based approaches
  - capitalize on all meiotic recombination events in a population
    - Localize small regions of the chromosome
    - enables rapid detection causal gene
  - Identifies genes with smaller relative risks

Concerns with GWAS

- Assumes CDCV hypothesis
- Expense
- Power dependent on:
  - Allele frequency
  - Relative risk
  - Sample size
  - LD between genotyped marker and the risk allele
  - disease prevalence
  - .ultiple testing
  - …….
- Study Design
  - Replication
  - Choice of SNPs
- Analysis methods
  - IT support, data management
  - Variable selection
  - Multiple testing

Successes in GWAS Studies

- Over 400 GWAS papers published to date
- Big Finds:
  - In 2005, it was learned through GWAS that age-related macular degeneration is associated with variation in the gene for complement factor H, which produces a protein that regulates inflammation (Klein et al. (2005) Science, 308, 385–389)
  - In 2007, the Wellcome Trust Case-Control Consortium (WTCCC) carried out GWAS for the diseases coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. This study was successful in uncovering many new disease genes underlying these diseases.

More Successes

- Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet. 2007
- Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Wellcome Trust Case Control Consortium* Nature. 2007;447;661-78
- Genomewide association analysis of coronary artery disease. *Samani et al.* N Engl J Med. 2007;357;443-53
- Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Parkes et al.* Nat Genet. 2007;39;830-2
- Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Todd et al.* Nat Genet. 2007;39;857-64
- A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Frayling et al.* Science. 2007;316;889-94
- Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Zeggini et al.* Science. 2007;316;1336-41
- Scott et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science, 316, 1341–1345.
- ……

Limitations

- For many diseases, the amount of trait variation explained by even the successes is way below the estimated heritability.
- Recently, GWAS are under a lot of criticism for relatively few translatable findings given the investment and hype.
- Assumptions underlying GWAS are not true for all diseases.

Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).

TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

Reasons GWAS Can Fail

even if well-powered and well-designed….

- Alleles with small effect sizes
- Rare variants
- Population differences
- Epistatic interactions
- Copy number variation
- Epigenetic inheritance
- Disease heterogeneity
- …….

Missing Heritability

Nature Reviews | Genetics

Lusis et al, 2008

Possible Association Models

1. Each of several genes may have a variant that confers increased risk of disease independent of other genes
2. Several genes in contribute additively to the malfunction of the pathway
3. There are several distinct combinations of gene variants that increase relative risk but only modest increases in risk for any single variant

Hypothetical Disease Mechanism

Hypothetical Disease Mechanism

Hypothetical Disease Mechanism

Hypothetical Disease Mechanism

Hypothetical Disease Mechanism

- For each gene probability of knockout = 0.2^2 = 0.04
- Probability of disease:
  - Pathway knocked out = 0.4
  - Pathway in tact = 0.2
- Sample Size = 2000 cases, 2000 controls
- Power:

Best SNP		Pathway	
Significant	Suggestive	0.001	0.005
0.001	0.05	0.42	0.69

Linear Pathway

- For each gene probability of knockout = 0.2^2 = 0.04
- Probability of disease:
  - Pathway knocked out = 0.4
  - Pathway in tact = 0.2
- Sample Size = 2000 cases, 2000 controls
- Power:

Best SNP		Pathway		Pathway (mis-specified)*	
Significant	Suggestive	0.001	0.005	0.001	0.005
0.002	0.02	0.94	0.98	0.51	0.73

*Tested pathway includes 15 genes not in simulated pathway

Enrichment Testing in GWAS

- Testing pathway enrichment is possible in GWAS data
  - Many of the same issues that exist in gene expression enrichment testing occur in GWAS enrichment testing (e.g. choice of statistics, competitive vs self-contained)
- Primary difference:
  - In expression data the unit of testing is a gene
  - In GWAS data the unit of testing is a SNP
- Challenges:
  - Identifying the SNP (set) -> Gene mapping
  - Summarizing across individual SNP statistics to compute a per-gene measure

Mapping SNPs to Genes

- All SNPs in physical proximity of each gene
  - Pros:
    - All/most genes represented
  - Cons:
    - Varying number of SNPs per gene
    - Many of the SNPs may dilute signal
    - Defining gene proximity can affect results
- eSNPs (Expression associated SNPs)
  - Pros:
    - 1 SNP per gene
    - SNPs functionally associated
  - Cons:
    - Assumes variants effect expression
    - Not all genes have eSNPs
    - eSNPs may be study and tissue dependent

Gene summaries

- Initial studies propose different statistics for summarizing the overall gene association prior to enrichment analysis
  - Number/proportion of SNPs with pvalue < 0.05
  - Mean(-log10(pvalue))
  - Min(pvalue)
  - $1-(1-\text{Min(pvalue)})^{N}$
  - $1-(1-\text{Min(pvalue)})^{(N+1)/2}$

First approaches: combining p-values

- Compute gene-wise p-value:
  - Select most likely variant - 'best' p-value
  - Selected minimum p-value is biased downward
  - Assign 'gene-wise' p-value by permutations (Westfall-Young)
    - Permute samples and compute 'best' p-value for each permutation
    - Compare candidate SNP p-values to this null distribution of 'best' p-values
- Combine p-values by Fisher's method, across SNPs (biased in the presence of correlation)

$$V = -\sum_{g_i \in G} \log(p_i)$$

$$p = \mathrm{P}(\chi^2_{(2k)} > 2V)$$

Next approaches

- Additive model: $\log(\frac{p}{1-p}) = \sum_{g_i \in G} \beta_i n_i$
  - Where n_i indexes the number of allele Bs of a SNP in gene *i* in the gene set *G*
  - Select subset of most likely SNP's
  - Fit by logistic regression (glm() in R)
- Significance by permutations
  - Permute sample outcomes
  - Select genes and fit logistic regression again
    - Assess goodness of fit each time
  - Compare observed goodness of fit

Competitive vs. Self-Contained Tests

- Competitive cutoff tests
  - Require only permuting SNP or Gene labels
  - May only allow to assess relative significance
- Self-contained distribution tests
  - Require permuting phenotype-genotype relationships
  - Resource intensive, may be difficult for large meta-analyses
  - Allow to assess overall significance

Competitive vs. Self-Contained Tests

- Self-contained null hypothesis
  - no genes in gene set are differentially expressed
- Competitive null hypothesis
  - genes in gene set are at most as often differentially expressed as genes not in gene set

What does this mean for SNP data?

Choice of Pathways/Gene Sets

- Relatively less "signal" in GWAS than in gene expression (GE)
  - GE enrichment typically test *which* gene sets/pathways show enrichment
  - GWAS enrichment typically test *if* there is enrichment
- Typically want to be conservative about selecting the number of pathways to test, otherwise will be difficult to overcome multiple testing
- Prioritized Approach:
  - Limited number of specific hypotheses (e.g. gene sets from experiment, co-expression modules, disease-specific pathways/ ontologies)
  - Exploratory analyses such as all KEGG/GO sets

Some Specific Methods

- SSEA
  - SNP Set Enrichment Analysis
- i-GSEA4GWAS
- MAGENTA
  - Meta-Analysis Gene-set Enrichment of variant Associations

SSEA

- Zhong et al. AJHG (2010)
- eSNP analysis to map SNPs to genes
  - More on this later…..
- Pathway statistic = one-sided Kolmogorov-Smirnov test statistic
- Pathway p-value assessed by permuting genotype-phenotype relationship
- FDR used to control error due to the number of pathways tested

i-GSEA4GWAS

- Zhang et al. *Nucl Acids Res* (2010)
- http://gsea4gwas.psych.ac.cn/

- Categorizes genes as significant or not significant
  - Significant: At least 1 SNP in the top 5% of SNPs
  - Does not adjust for gene size

- Pathway score: k/K
  - k = Proportion of significant genes in the geneset
  - K = Proportion of significant genes in the GWAS
- FDR assessed by permuting SNP labels

Results

Pathway/Gene set name	Description	Manhattan plot	P-value	FDR	genes/Selected genes/All genes
HSA04950 MATURITY ONSET DIABETES OF THE YOUNG View Detail	Genes involved in ma..... More...		< 0.001	0.0030	11/23/25
PROSTAGLANDIN AND LEUKOTRIENE METABOLISM View Detail	More...		< 0.001	0.0085	13/27/32
HSA00565 ETHER LIPID METABOLISM View Detail	Genes involved in et..... More...		< 0.001	0.0125	15/28/31
DNA REPAIR View Detail	Genes annotated by t..... More...		< 0.001	0.0135	41/113/125
NTHIPATHWAY View Detail	Hemophilus influenza..... More...		< 0.001	0.0142	12/21/24
NEGATIVE REGULATION OF DEVELOPMENTAL PROCESS View Detail	Genes annotated by t..... More...		< 0.001	0.014571428	66/175/197
HSA04330 NOTCH SIGNALING PATHWAY View Detail	Genes involved in No..... More...		< 0.001	0.016	16/35/47
ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY View Detail	Genes annotated by t..... More...		< 0.001	0.020875	60/136/140

MAGENTA

- Segre et al. *PLoS Genetics* (2010)
- Software download:
  - http://www.broadinstitute.org/mpg/magenta/
  - Requires MATLAB!!
  - Less convenient, but more customizable than iGSEA4GWAS
- Customizable proportion of "significant" genes
- Customizable gene window (upstream & downstream)
- Option for Rank-Sum test
- Gene Summary = min(p)
  - Uses stepwise regression to adjust for multiple possible factors: e.g. gene size, SNP density

MAGENTA Results

	95% Cutoff (Top 5%)				75% Cutoff (Top 25%)			
GS	NOMINAL GSEA PVAL	FDR	EXP # GENES	OBS # GENES	NOMINAL GSEA PVAL	FDR	EXP # GENES	OBS # GENES
positive regulation of osteoblast differentiation	3.36E-01	8.02E-01	1	2	3.00E-04	7.91E-02	6	14
one-carbon metabolic process	2.20E-03	3.55E-01	1	6	1.60E-03	1.44E-01	7	15
placenta development	3.36E-01	8.06E-01	1	2	4.00E-04	1.45E-01	6	14
carbohydrate transport	8.19E-01	9.46E-01	2	1	3.20E-03	3.45E-01	8	16

Adaptations of GSEA

- Order log-odds ratios or linkage p-values for all SNPs
- Map SNPs to genes, and genes to groups
- Use linkage p-values in place of t-scores in GSEA
  - Compare distribution of log-odds ratios for SNPs in group to randomly selected SNP's from the chip

Summary Points for GWAS

- In GWAS, few SNPs typically reach genome-wide significance
- Biological function of those that do can take years of work to unravel
- Incorporating biological information (expression, pathways, etc) can help interpret and further explore GWAS results
- Enrichment tests can be used to explore biological pathway enrichment
  - Different tests tell you different things
- Annotation choices very different that in gene expression data, though still rely on the same resources.... not necessarily so for other 'omics"

Adding in Gene Expression Data

- Many motivating reasons to combine/integrate data from multiple "-omes"
- Expression and SNP data is most commonly done
  - Though methods could be applied to combine other "-omics"
- Generally make assumptions about central dogma

Genetics of Gene Expression

- Schadt, Monks, et al. (*Nature* 2003) & Morley, Molony, et al. (*Nature* 2004) showed that gene expression is a heritable trait under genetic control
- Identifying expression-associated SNPs (eSNPs) can identify SNPs which are associated with biological function
- For significant GWAS "hits" eSNPs can suggest candidate genes and possibly information about direction of association

eSNPs can enrich p-values in GWAS studies
Example: T2D Data

Zhong et al. (2010) Elucidating Networks of eSNPs associated with Type 2 Diabetes.

Considerations on Filtering/Mining Data

- Trade-off between un-biased discovery and improving power (improving enrichment)
- Gold standard for publication is p-value < 5e-8 PLUS replication
- For hypothesis generation or biological data mining might be willing to accept more Type I error
- Possible approaches:
  - Gold standard only
  - Gold standard then mining "biological" SNPs (e.g. all SNPs near genes, eSNPs, eSNPs by tissue, etc)
  - Partitioning SNPs into sets by prior information

Considerations: Multiple Test Correction

- Can be valid to test hypotheses in a partitioned fashion if:
  1. The partitions are specified **before** you look at the data
  2. Your multiple testing procedure controls the overall error rate

5% P-value vs 5% FDR

- P-value -> Over a large number of times the experiment is repeated, 5% of the time we'll identify 1 or more false positive SNPs
- FDR -> 5% of identified SNPs are false positives

Partitioned SNP Testing (p-value)

- Can be beneficial if you have a small number of high(er)-confidence SNPs
- Genomewide significance threshold: 5e-8 = 0.05/1,000,000
- Example: 10,000 eSNPs
  - eSNP threshold: 0.025/10,000 = 2.5e-6
  - Remaining SNP threshold: 0.025/990,000 = 2.53e-8

Partitioned Testing (FDR)

- Simple way to control error over multiple partitions
- Controlling FDR at level ξ in each (non-overlapping) set, results in overall FDR ξ

eSNPs: Computing your own

- eSNP analyses are just GWAS's with continuous traits, but 1000's of them
- Approaches:
  - Frequentist:
    - Linear Regression
      - Outlier sensitive, can adjust for covariates
    - Robust Regression
      - Outlier resistant, can adjust for covariates, more computationally demanding
    - Kruskal-Wallis
      - Nonparametric (outlier resistant), difficult to adjust for covariates
  - Bayesian:
    - More resistant to outlier effects than linear regression, but require setting priors on each parameter
    - Some software available:
      - Bimbam
      - SNPTEST

eSNPs: A note on computation

- eSNP analysis is extremely resource intensive in both processor time and storage
- Computation requires a cluster (not possible on a desktop machine)
- Storage: $N_{markers} x N_{expression\ traits}$ is typically large
  - One approach is to store only results with pvalue < some threshold

eSNP Discovery

- eSNPs near gene location are easier to find
  - Real biological effects (*cis* regulation)
  - Fewer hypothesis tests relative to genomewide
- Typical approach is to identify local (proximal) eSNPs and distant (distal) eSNPs in separate steps
- Controlling each at fixed FDR, ξ, controls the overall FDR at ξ
- Choice of proximal window can effect eSNP discovery

Cis vs Trans Regulation

Aside: Cis/Trans vs Proximal/Distal

- *Cis* element -> Regulates transcription only of copy sharing same DNA strand
- *Trans* element -> Regulates transcription of both DNA strands
- *Trans* elements can be near the gene, *cis* elements can be far from gene (on MB scale)
- Proximal (near) and distal (far) more accurate when referring variants associated with expression

eSNPs: Publically Available

- Databases:
  - www.scandb.org
  - http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/
- Available in Synapse (synapse.sagebase.org):
  - Harvard Brain- Brain, multiple disease
  - Kronos Phase I- Brain, alzheimer's
  - Human Liver Cohort- Liver, population sample
- …

Motivation for Integrated Analysis

- Newer approaches will allow you to not do partitioned/filtered analysis, and leverage information across datatypes
- New technologies allow for more ready integration
  - Ex. RNA-Seq
  - Dropping costs allow for more datatypes to be collected simultaneously
  - Biobanking effort are storing more tissues

Motivation for Integrated Analysis

- Naturally allow Bayesian approaches for identifying priors or jointing modeling data
- Several new approaches proposed
  - Methods that were developed for eSNPs are readily extended across data types
  - Other approaches take into account similarities between/ withing phenotypes
    - Several an ontology jointly representing disease risk factors and causal mechanisms based on GWAS results
    - Proposed ontology is disease-specific (nicotine addiction and treatment) and only applicable to very specific research questions
  - More later on "different issues for –omics"

Motivation for Integrated Analysis

- Methods are largely relying on central dogma assumptions that do not always hold

Summary

- Pathway and gene set analysis has been extended to SNP and SNV data
- Some annotation resources are readily adapted, but a new series of choices are available
- Software packages for GWAS pathway analysis are maturing
- Advances in approximation for permutation testing will make these tools more computationally tractable
- Many of the same issues with missing annotation, etc. are still a concern

Summary

- Integration of SNP level and eSNP data has been highly successful, and helps motivate the integration of other "-omes" in analysis
- Such integration will be dependent on the quality of the annotation that it relies on
- Next, we will talk about specific concerns for different datatypes
- Issues will compound in integrated analysis…

Questions?

motsinger@stat.ncsu.edu

Pathway Analysis in other data types

Alison Motsinger-Reif, PhD
Associate Professor
Bioinformatics Research Center
Department of Statistics
North Carolina State University

New "-Omes"

- Genome
- Transcriptome
- Metabolome
- Epigenome
- Proteome
- Phenome, exposome, lipidome, glycome, interactome, spliceome, mechanome, etc...

Goals

- Pathway analysis in metabolomics
- Pathway analysis in proteomics
- Issues, concerns in other data types
  - Methylation data
  - aCGH
  - Next generation sequencing technologies
- Many approaches generalize, but there are always specific challenges in different data types
- Weighted co-expression analysis

Metabolomics

- While many proteins interact with each other and the nucleic acids, the real metabolic function of the cell relies on the enzymatic interconversion of the various small, low molecular weight compounds (metabolites)
- Technology is rapidly advancing
- The frequent final product of the metabolomics pipeline is the generation of a list of metabolites who's concentrations have been (significantly) altered which must be interpreted in order to derive biological meaning

→ Perfect for pathway analysis

2 routes to Metabolomics

Data processing and annotation

- Preprocessing and the level of annotation is VERY different than in genomic and transcriptomic data
- Many steps in overall experimental design that greatly influence interpretation
- Will breifly cover some of the main issues

Analytical Platform

- Likely GC/LC-MS or NMR as they are the most common
- Choice is normally based more on available equipment, etc. more than experimental design
- GC-MS is an extremely common metabolomics platform, resulting in a high frequency of tools which allow for the direct input of GC-MS spectra.
  - Popularity is due to its relatively high sensitivity, broad range of detectable metabolites, existence of well-established identification libraries and ease of automation
  - separation-coupled MS data requires much processing and careful handling to ensure the information it contains is not artifactual

Targeted vs. Untargeted

- Scientists have been quantifying metabolite levels for over 50 years through targeted analysis…
- With new technologies, the focus can be on untargeted metabolomics
  - Really hard to annotate and interpret
  - Integrated –omics analysis being used to help annotate and understand untargeted metabolites
  - Analogous to candidate gene vs. genome wide testing

Key Issues in Metabolomics

- All of the metabolites within a system cannot be identified with any one analytical method due to chemical heterogeneity, which will cause downstream issues as all metabolites in a pathway have not been quantified
- Not all metabolites have been identified and characterized and so do not exist in the standards libraries, leading to large number of unannotated and/or unknown metabolites of interest
- Organism specific metabolic databases/networks only exist for the highest use model organisms making contextual interpretations difficult for many researchers
- Interpreting the huge datasets of metabolite concentrations under various conditions with biological context is an inherently complex problem requiring extremely in depth knowledge of metabolism.
- The issue of determining which metabolites are actually important in the experimental system in question.

Metabolomic Databases

- Two types of data-bases:
  - top-down (gene to protein to metabolite)
  - bottom-up (chemical entity to biological function) approaches
  - www.metabolomicssociety.org/database
- Most commonly used in biomedical applications:
  - MetaCyc
  - KEGG
    - Subdatabases LIGAND, REACTION PAIR and PATHWAY

Metabolomic Databases

- KEGG and MetaCyc are largest (in terms of number of organisms and most in depth comprehensive (i.e. contains linked information from metabolite to gene)
- Others that are rapidly growing:
  - Reactome (human)
  - KNApSAcK (plants)
  - Model SEED (diverse)
  - BiG [40] (6 model organisms)
  - can be more useful than the large databases if a specific organism is desired

Metabolomic Databases

- KEGG and MetaCyc databases each contain a generalized 'conserved' set of pathways based on metabolic pathways that are more or less the same throughout life in general
  - For KEGG, organism specific annotations are available to query
  - For MetaCyc, individual 'Cyc' databases have been generated for a number of organisms,
    - some just computationally
    - others extensively manually curated such as AraCyc for Arabidopsis
- More recent development are the cheminformatic databases like PubChem
  - provide a chemically ontological approach to cataloguing the ill-defined category of 'small molecules' active in biological systems
  - can provide additional non-biology specific information as well alternative formatting options for datasets *(watch for errors!)*

Enrichment analysis

- These databases are used to create "metabolite sets" for enrichment analysis
- Majority of available tools do early generation over-representation analysis
  - With all the advantages and caveats!
  - For more up to date analysis, will need to work to merge databases, etc. to correctly use more up-to-date approaches

Metabolomics Analysis Tools

- Comprehensive platforms
  - Provide a suite of utilities allowing comprehensive analysis from raw spectral data to pathway analysis
    - MetaboAnalyst
    - MeltDB
- Enrichment Analysis
  - Only works with processed data
    - PAPi
    - MBRole
    - MPEA
    - TICL
    - IMPaLA
- Metabolite Mapping
  - Connects metabolites to genetic/proteomic, etc. resources
    - MetaMapp
    - Masstrix
    - Paintomics
    - VANTED
    - Pathos

Metaboanalyst

- A number of utilities:
  - Data quality checking (useful for batch effects)
  - metabolite ID converter among others are also included.
  - If beginning from raw GC or LC-MS data MetaboAnalyst uses XCMS for peak fitting, identification etc.
  - Once at the peak list (NMR or MS) stage, various preprocessing options such as data-filtering and missing value estimation can be used.
  - A number of normalization, transformation and scaling operations can be performed.
  - Suite of statistical analyses including metabolomics standards like PCA, PLS-DA and hierarchically clustered heatmaps, among many other options.
  - *All these things can be done in other programs, but this is a great tool to get started if you're new to metabolomics!*

Metaboanalyst

- Enrichment Analysis tool of MetaboAnalyst was one of the earliest implementations of GSEA for metabolomics datasets (MSEA)
  - quite biased towards human metabolism unless you make custom background pathways/sets
- Three options for input
  - a single column list of compounds (Over Representation Analysis, ORA)
  - a two column list of compounds AND abundances (Single Sample Profiling, SSP)
  - a multi-column table of compound abundances in classed samples (Quantitative Enrichment Analysis, QEA).

Metaboanalyst

- ORA will calculate whether a particular set of metabolites is statistically significantly higher in the input list than a random list, which can be used to examine ranked or threshold cut-off lists
- SSP is aimed at determining whether any metabolites are above the normal range for common human biofluids
- QEA is the most canonical and will determine which metabolite sets are enriched within the provided class labels, while providing a correlation value and p-value

PAPi

- Pathway Activity Profiling is an R-based tool
- As input it takes a list with abundances (normalized and scaled)
- Works on the assumptions that the detection (i.e. presence in the list) of more metabolites in a pathway and that lower abundances of those metabolites indicates higher flux and therefore higher pathway activity
  - Assumption may not always be true
  - Ex. TCA cycle intermediates can have high abundance even when flux through the reactions in this pathway is also high

PAPi

- PAPi calculates an activity score (AS) for each pathway
- The metabolic pathways are taken from the general KEGG database
- The AS indicates the probability of this pathway being active in the cell
- These scores can then be used to compare experimental and control conditions by performing ANOVA or a t-test to compare two sample types.

MetaMapp

- Performs metabolic mapping for unknown and unannotated metabolites
- Since biochemistry is the interconversion of chemically similar entities, compounds can be clustered solely by their chemical similarity
  - Highly beneficial for metabolites without reaction annotation
- Also uses KEGG reactant pair information
  - chemical similarity misclustered some obviously biologically-related metabolites

MetaMapp

- Can also map metabolites based on their mass spectral similarity (for unknowns)
- Can be used to make custom/novel sets for pathway analysis

Summary on Metabolomics Pathway Analysis

- Metabolomics is a maturing area
- "Easy" implementations of tools often behind best practices in pathway approaches
- Issues with time dependencies, tissue dependencies, etc. are more exaggerated in metabolomics
- As the technology is maturing, we are just getting to understand the biases, sources of variation, etc.
  - Data quality control best practices are evolving
  - Will have major impact on the pathway analysis

Specific Issues for other -omics

- Will consider some issues that are both specific to the "-ome" and to particular technologies

- Proteomics
- Epigenomics
- Array CGH data
- RNA seq
- Next generation sequencing
- …….

Proteomics

- After genomics and transcriptomics, proteomics is the next step in central dogma
- Genome is more or less constant, but the proteome differs from cell to cell and from time to time
- Distinct genes are expressed in different cell types, which means that even the basic set of proteins that are produced in a cell needs to be identified

- It was assumed for a long time that microarrays would capture much of this information → NO!

Proteomics vs. Transcriptomics

- mRNA levels do not correlate with protein content
- mRNA is not always translated into protein
- The amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell
- Many proteins are also subjected to a wide variety of chemical modifications after translation
  - Affect function
  - Ex: phosphorylation, ubiquitination
- Many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications

Proteomics

- Technological advances for proteomics has slowed
  - Like metabolomics, the lack of any PCR-like amplification is limited
  - Unlike metabolomics that has a reasonable search space, there estimated to be more than a million transcripts

Proteomics

- Available technologies have different challenges
  - Protein microarrays vs. mass spec based methods
  - General concerns with reproducibility dampened initial excitement

Proteomics

- The high complexity and technical instability mean that the level of annotation is often quite low
- Same challenges as with metabolomics, but more exaggerated given the large annotation space
- Many of the same issues …..

Epigenomics

- "Complete" set of epigenetic modifications on the genetic material of a cell
  - epigenetic modifications are reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence
  - DNA methylation and histone modification most commonly assayed
- Rapidly advancing technologies
  - Histone modification assays
  - CHIP-CHIP and CHIP-Seq
  - Methylation arrays

Epigenomics

- Recent studies have focused on issues related to differential numbers of probes in genes
  - Most microarrays were designed with the same number
  - For methylation data, this is not the case, and extreme bias can be seen
  - Bias results in a large number of false positives
- Can be corrected by applying methods that models the relationship between the number of features associated with a gene and its probability of appearing in the foreground list
  - CpG probes in the case of microarrays
  - CpG sites in the case of high-throughput sequencing
  - Chip annotation
- Can also be corrected with careful application of permutation approaches

Next Generation Sequencing

- Variant calling in NGS can detect single nucleotide variants (SNVs) and SNPs
- For SNPs, the exact same pathway methods can be used as designed for GWAS studies (assuming genotyping in genome wide)
- For rare variants, standard approaches are a challenge
  - highly inflated false-positive rates and low power in pathway-based tests of association of rare variants
  - due to their lack of ability to account for gametic phase disequilibrium
  - New area of methods development

Next Generation Sequencing

- RNA-seq data
  - Not truly quantitative
  - With experience, know that there are very different variance distributions at different levels of expression
  - Will matter for methods that test for differences in variance as well as mean
    - Two sided K-S tests….

Summary on Integrated Analysis

- Technology advances across the "omics" is an exciting opportunity for better understanding complexity
- Technologies have unique properties that need to be understood and accounted for in analysis
- Metabolomics resources are rapidly maturing

Summary on Integrated Analysis

- Database development, curation, editing, etc. always lags behind technology
- Issues with incomplete and inaccurate annotation accumulate as more "omes" are considered
- With more complex data, this complexity is not readily captured in the databases the gene set analysis relies on
  - Differences in cell types, exposure, time, etc.
  - Major needs for methods development…..

Questions?

motsinger@stat.ncsu.edu

Pathway & Network Analysis for Omics Data: Networks in Biology

Ali Shojaie

July 2014
Summer Institute for Statistical Genetics
University of Washington

©Ali Shojaie

Why Study Networks?

- Components of biological systems, e.g. genes, proteins, metabolites, interact with each other to carry out different functions in the cell.
- Examples of such interactions include signaling, regulation and interactions between proteins.
- We cannot understand the function and behavior of biological systems by studying individual components ($2 + 2 \neq 4$!).
- Networks provide an efficient representation of complex reaction in the cells, as well as basis for mathematical/statistical models for the study of these systems.

Central Dogma of Molecular Biology (Extended)

Networks in Biology: Gene Regulatory Interactions

Networks in Biology: Gene Regulatory Networks

Networks in Biology: Protein-Protein Interaction

Networks in Biology: Protein-Protein Interaction (PPI) Networks

Networks in Biology: Metabolic Reactions

Networks in Biology: Metabolic Pathways

But Do Networks Matter?

- They Do!
- Recent studies have linked changes in gene/protein networks with many human diseases.

Systems Biology and Emerging Technologies

Gene Networks and microRNAs Implicated in Aggressive Prostate Cancer

Liang Wang,[1] Hui Tang,[2] Venugopal Thayanithy,[3] Subbaya Subramanian,[3] Ann L. Oberg,[2] Julie M. Cunningham,[1] James R. Cerhan,[2] Clifford J. Steer,[4] and Stephen N. Thibodeau[1]

[1]Departments of Laboratory Medicine and Pathology and [2]Health Sciences Research, Mayo Clinic, Rochester, Minnesota; and Departments of [3]Laboratory Medicine and Pathology, [4]Medicine, and Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, Minnesota

But Do Networks Matter?

0888-8809/07/$15.00/0
Printed in U.S.A.

Molecular Endocrinology 21(9):2112–2123
Copyright © 2007 by The Endocrine Society
doi: 10.1210/me.2006-0474

Estrogen-Regulated Gene Networks in Human Breast Cancer Cells: Involvement of E2F1 in the Regulation of Cell Proliferation

Joshua D. Stender, Jonna Frasor, Barry Komm, Ken C. N. Chang, W. Lee Kraus, and Benita S. Katzenellenbogen

Departments of Biochemistry (J.D.S.) and Molecular and Integrative Physiology (J.F., B.S.K.), University of Illinois at Urbana-Champaign, Urbana, Illinois 61801-3704; Women's Health and Musculoskeletal Biology (B.K., K.C.N.C.), Wyeth Research, Collegeville, Pennsylvania 19426; and Department of Molecular Biology and Genetics (W.L.K.), Cornell University, Ithaca, New York 14853-4203

But Do Networks Matter?

Cell PRESS

Cancer Cell
Article

A Transcriptional Signature and Common Gene Networks Link Cancer with Lipid Metabolism and Diverse Human Diseases

Heather A. Hirsch,[1,7] Dimitrios Iliopoulos,[1,7] Amita Joshi,[1,7] Yong Zhang,[2] Savina A. Jaeger,[3] Martha Bulyk,[3,4,5] Philip N. Tsichlis,[6] X. Shirley Liu,[2] and Kevin Struhl[1,*]
[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA
[2]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA
[3]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[4]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[5]Harvard/MIT Division of Health Sciences and Technology (HST), Harvard Medical School, Boston, MA 02115, USA
[6]Molecular Oncology Research Institute, Tufts Medical Center, Boston, MA 02111, USA
[7]These authors contributed equally to this work
*Correspondence: kevin@hms.harvard.edu
DOI 10.1016/j.ccr.2010.01.022

But Do Networks Matter?

And, incorporating the knowledge of networks improves our ability to find causes of complex diseases.

Molecular Systems Biology 3; Article number 140; doi:10.1038/msb4100180
Citation: *Molecular Systems Biology* 3:140
© 2007 EMBO and Nature Publishing Group All rights reserved 1744-4292/07
www.molecularsystemsbiology.com

molecular systems biology

REPORT

Network-based classification of breast cancer metastasis

Han-Yu Chuang[1,5], Eunjung Lee[2,3,5], Yu-Tsueng Liu[4], Doheon Lee[3] and Trey Ideker[1,2,4,*]

[1] Bioinformatics Program, University of California San Diego, La Jolla, CA, USA, [2] Department of Bioengineering, University of California San Diego, La Jolla, CA, USA, [3] Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea and [4] Cancer Genetics Program, Moores Cancer Center, University of California San Diego, La Jolla, CA, USA
[5] These authors contributed equally to this work
* Corresponding author. Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA. Tel.: +1 858 822 4558; Fax: +1 858 534 5722; E-mail: trey@bioeng.ucsd.edu

Why Do We Need Network Inference?

- Despite progress, our knowledge of interactions in the genome is limited.
- The entire genome is a vast landscape, and experiments for discovering networks are very expensive
- From a statistical point of view, network estimation is related to estimation of covariance matrices, which has many independent applications in statistical inference and prediction (*more about this later*)
- Finally, and perhaps most importantly, gene and protein networks are dynamic and changes in these networks have been attributed to complex diseases.

Networks: A Short Premier

- A network is a collection of nodes V and edges E.
- We assume there are p nodes in the network, and that the nodes correspond to random variables $X_1, \ldots X_p$.
- Edges in the network can be directed $X \to Y$ or undirected $X - Y$.

- In all these example, the nodes are $V = \{1, 2, 3\}$.
- The edges are:

$$\begin{aligned} E_1 &= \{1 - 2, 2 - 3\} \\ E_2 &= \{1 \to 3, 3 \to 2\} \\ E_3 &= \{1 - 2, 1 \to 3\} \end{aligned}$$

Networks: A Short Premier

- A convenient way to represent the edges of the network is to use an adjacency **matrix** A
- A matrix is a rectangular array of data (similar to a table)
- Values in each entry are shown by indeces of row and column

$$A = \begin{bmatrix} . & \mathbf{x} & . \\ . & . & . \\ . & . & . \end{bmatrix}$$ Here, **x** is in row 1 and column 2

- Adjacency matrix is a square matrix, which has a **1 if there is an edge** from a node in one row to a node in another column, and **0** otherwise
- For undirected edges, we add a **1** in both directions

Networks: A Short Premier

$$A = \begin{bmatrix} 0 & \mathbf{1} & 0 \\ \mathbf{1} & 0 & \mathbf{1} \\ 0 & \mathbf{1} & 0 \end{bmatrix} \quad A = \begin{bmatrix} 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 \end{bmatrix} \quad A = \begin{bmatrix} 0 & \mathbf{1} & \boxed{1} \\ \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 \end{bmatrix}$$

What Do Edges in Biological Networks Mean?

- In gene regulatory networks, an edge from gene i to gene j often means that i affects the expression of j; i.e. as i's expression changes, we expect that expression of j to increase/decrease.
- In protein-protein interaction networks, an edge between proteins i and j often means that *the two proteins bind together and form a protein complex*. Therefore, we expect that these proteins are generated at similar rates.
- In metabolic networks, an edge between compound i and j often means that *the two compounds are involved in the same reaction*, meaning that they are generated at relative rates.
- Thus, edges represent some type of association among genes, proteins or metabolites, defined generally to include *linear or nonlinear* associations; more later....

Statistical Models for Biological Networks

- We use the framework of graphical models
- In this setting, nodes correspond to "random variables"
- In other words, each node of the network represents one of the variables in the study
  - In gene regulatory networks, nodes $\equiv$ genes
  - In PPI networks, nodes $\equiv$ proteins
  - In metabolic networks, nodes $\equiv$ metabolites
- In practice, we observe n measurements of each of the variables (genes/proteins/ metabolites) for say different individuals, and want to determine which variables are connected, or use their connection for statistical analysis

Correlation

- Correlation is a simple measure of **linear** association between two random variables.
- Suppose we observe values of X and Y over n samples, correlation determines the extent to which larger/smaller values of Y tend to be associated with larger/smaller values of X:
  - if **larger** values of X are associated with larger values of Y, then X and Y will have a positive correlation
  - if **larger** values of X are associated with smaller values of Y, then X and Y will have a negative correlation
  - if **larger** values of X are not associated with larger/smaller values of Y, then X and Y will have a near 0 correlation
- Correlation between X and Y is often denoted as r_{xy}
- r_{xy} measures the strength of linear relationship between X and Y, and is a number between -1 and 1

Correlation

- $r_{xy} = 1$ means X and Y change linearly in the same direction
- $r_{xy} = -1$ means X and Y change linearly in opposite direction
- $r_{xy} = 0$ means X and Y have no linear relationships
- $r_{xy} = r_{yx}$ i.e. correlation is symmetric

An Overview of Methods for Network Inference

Network Inference Methods Can be categorized into two general classes:

- Methods based on marginal measures of association:
  - Co-expression Networks (uses linear measures of association)
  - Methods based on mutual information (can accommodate non-linear associations)
- Methods based on conditional measures of association:
  - Methods assuming multivariate normality/normality (`glasso`, etc)
  - Generalizations to allow for nonlinear dependencies (`nonparanormal`, etc)

Our Plan

In the remainder of this module, we will discuss the following topics

- Methods for reconstructing undirected networks
  - Co-expression Networks (WGCNA)
  - ARACNE
  - Conditional Independence Graphs (`glasso`, `nonparanormal`, etc)
- Methods for reconstructing directed networks
  - Bayesian Networks (basic concepts, reconstruction algorithm)
  - Reconstructing directed networks from time-course data (dynamic Bayesian networks)
  - Reconstructing directed networks from perturbation screens
- Topology-based pathway enrichment analysis

Undirected Graphical Models – I

(Network Inference using Marginal Association)

Alison Motsinger-Reif, PhD
Bioinformatics Research Center
Department of Statistics
North Carolina State University
motsinger@stat.ncsu.edu

Agenda

1. Weighted Co-Expression network analysis
   - Basic concepts
   - Software details
2. Mutual information for network analysis
   - ARACNE
   - Software details

Philosophy of Weighted Gene Co-Expression Network Analysis

- Understand the "system" instead of reporting a list of individual parts
  - Describe the functioning of the engine instead of enumerating individual nuts and bolts
- Focus on modules as opposed to individual genes
  - this greatly alleviates multiple testing problem
- Network terminology is intuitive to biologists

How to construct a weighted gene co-expression network?

Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17.

Steps for constructing a co-expression network

A) Microarray gene expression data
B) Measure concordance of gene expression with a Pearson correlation
C) The Pearson correlation matrix is either dichotomized to arrive at an adjacency matrix → unweighted network

Or transformed continuously with the power adjacency function → weighted network

Power adjacency function results in a weighted gene network

$$a_{ij} = |cor(x_i, x_j)|^{\beta}$$

Often choosing beta=6 works well but in general use the "scale free topology criterion" described in Zhang and Horvath 2005.

Comparing adjacency functions

Comparing the power adjacency function to the step function

- While the network analysis results are usually highly robust with respect to the network construction method there are several reasons for preferring the power adjacency function.
  - Empirical finding: Network results are highly robust with respect to the choice of the power beta
    - Zhang B and Horvath S (2005)
  - Theoretical finding: Network Concepts make more sense in terms of the module eigengene.
    - Horvath S, Dong J (2008) Geometric Interpretation of Gene Co-Expression Network Analysis. PloS Computational Biology

How to detect network modules?

Module Definition

- Numerous methods have been developed
  - Remember the clustering lectures
- Commonly use average linkage hierarchical clustering coupled with the topological overlap dissimilarity measure
- Once a dendrogram is obtained from a hierarchical clustering method, we choose a height cutoff to arrive at a clustering
- Modules correspond to branches of the dendrogram

The topological overlap dissimilarity is used as input of hierarchical clustering

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

- Generalized in Zhang and Horvath (2005) to the case of weighted networks
- Generalized in Yip and Horvath (2006) to higher order interactions

Using the topological overlap matrix (TOM) to cluster genes

– Here modules correspond to branches of the dendrogram

TOM plot

Vol 452 | 27 March 2008 | doi:10.1038/nature06758 nature

ARTICLES

Genetics of gene expression and its effect on disease

Valur Emilsson[1,2], Gudmar Thorleifsson[1], Bin Zhang[2], Amy S. Leonardson[2], Florian Zink[1], Jun Zhu[2], Sonia Carlson[2], Agnar Helgason[1], G. Bragi Walters[1], Steinunn Gunnarsdottir[1], Magali Mouy[1], Valgerdur Steinthorsdottir[1], Gudrun H. Eiriksdottir[1], Gyda Bjornsdottir[1], Inga Reynisdottir[1], Daniel Gudbjartsson[1], Anna Helgadottir[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Unnur Styrkarsdottir[1], Solveig Gretarsdottir[1], Kristinn P. Magnusson[1], Hreinn Stefansson[1], Ragnheidur Fossdal[1], Kristleifur Kristjansson[1], Hjortur G. Gislason[3], Tryggvi Stefansson[3], Bjorn G. Leifsson[3], Unnur Thorsteinsdottir[1], John R. Lamb[2], Jeffrey R. Gulcher[1], Marc L. Reitman[4], Augustine Kong[1], Eric E. Schadt[2]* & Kari Stefansson[1]*

Different Ways of Depicting Gene Modules

1) Rows and columns correspond to genes

2) Red boxes along diagonal are modules

3) Color bands=modules

Idea:
Use network distance in MDS

Topological Overlap Plot

Gene Functions

Multi Dimensional Scaling

Traditional View

Heatmap view of module

Columns= tissue samples

Rows=Genes
Color band indicates module membership

Message: characteristic vertical bands indicate tight co-expression of module genes

Implementation

- Implemented in the R-package WGCNA:
- Install Package

```
source("http://bioconductor.org/biocLite.R")
biocLite("impute")
install.packages("WGCNA")
```

- Main estimation function

```
adjacency(datExpr, selectCols = NULL, type = "unsigned", power = if
    (type=="distance") 1 else 6, corFnc = "cor", corOptions = "use = 'p'",
    distFnc = "dist", distOptions = "method = 'euclidean'")
```

- Suggested to determine the power so that the network has scale-free distribution, need to search for multiple powers

Relating modules to external data & weighted gene co-expression network analysis

Construct a network
Rationale: make use of interaction patterns between genes

Identify modules
Rationale: module (pathway) based analysis

Relate modules to external information
Array Information: Clinical data, SNPs, proteomics
Gene Information: gene ontology, EASE, IPA
Rationale: find biologically interesting modules

Study Module Preservation across different data
Rationale:
- Same data: to check robustness of module definition
- Different data: to find interesting modules.

Find the key drivers in *interesting* modules
Tools: intramodular connectivity, causality testing
Rationale: experimental validation, therapeutics, biomarkers

What is different from other analyses?

- Emphasis on modules (pathways) instead of individual genes
  - Greatly alleviates the problem of multiple comparisons
    - Less than 20 comparisons versus 20000 comparisons
- Use of intramodular connectivity to find key drivers
  - Quantifies module membership (centrality)
  - Highly connected genes have an increased chance of validation
- Module definition is based on gene expression data
  - No prior pathway information is used for module definition
  - Two module (eigengenes) can be highly correlated

What is different from other analyses?

- Emphasis on a unified approach for relating variables
  - Default: power of a correlation
  - Rationale:
    - puts different data sets on the same mathematical footing
    - Considers effect size estimates (cor) and significance level
    - p-values are highly affected by sample sizes (cor=0.01 is highly significant when dealing with 100000 observations)
- Technical Details: soft thresholding with the power adjacency function, topological overlap matrix to measure interconnectedness

Case Study:
Finding brain cancer genes

Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Shu, Q, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS (2006) "Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target", PNAS | November 14, 2006 | vol. 103 | no. 46

Different Ways of Depicting Gene Modules

1) Rows and columns correspond to genes

2) Red boxes along diagonal are modules

3) Color bands=modules

Topological Overlap Plot

Gene Functions

Multi Dimensional Scaling

Traditional View

Comparing the Module Structure in Cancer and Normal tissues

Messages:
1) Cancer modules can be independently validated

2) Modules in brain cancer tissue can also be found in normal, non-brain tissue
-->
Insights into the biology of cancer

Mean Prognostic Significance of Module Genes

Message: Focus the attention on the brown module genes

Module hub genes predict cancer survival

1. Cox model to regress survival on gene expression levels
2. Defined prognostic significance as –log10(Cox-p-value) the survival association between each gene and glioblastoma patient survival
3. *A module-based measure of gene connectivity significantly and reproducibly identifies the genes that most strongly predict patient survival*

Test set – 55 gbms
r = 0.56; p-2.2 x 10^{-16}

Validation set – 65 gbms
r = 0.55; p-2.2 x 10^{-16}

The fact that genes with high intramodular connectivity are more likely to be prognostically significant facilitates a novel screening strategy for finding prognostic genes

- <u>Focus on those genes with significant Cox regression p-value AND high intramodular connectivity.</u>
  - It is essential to to take a module centric view: focus on intramodular connectivity of disease related module
- Validation success rate= proportion of genes with independent test set Cox regression p-value<0.05.
- Validation success rate of network based screening approach (68%)
- Standard approach involving top 300 most significant genes: 26%

The network-based approach uncovers novel therapeutic targets

Five of the top six hub genes in the mitosis module are already known cancer targets: topoisomerase II, Rac1, TPX2, EZH2 and KIF14.

Hypothesized that the 6-th gene ASPM gene is novel therapeutic target. ASPM encodes the human ortholog of a drosophila mitotic spindle protein.

Biological validation: siRNA mediated inhibition of ASPM

References

- Basso, K., Margolin, A. a, Stolovitzky, G., Klein, U., Dalla-Favera, R., & Califano, A. (2005). **Reverse engineering of regulatory networks in human B cells.** Nature genetics, 37(4), 382-90. doi:10.1038/ng1532
- Margolin, A. a, Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** BMC bioinformatics, 7 Suppl 1, S7. doi:10.1186/1471-2105-7-S1-S7
- Margolin, A. a, Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., & Califano, A. (2006). **Reverse engineering cellular networks.** Nature protocols, 1(2), 662-71. doi:10.1038/nprot.2006.106
- Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., Sulman, E. P., et al. (2010). **The transcriptional network for mesenchymal transformation of brain tumours.** Nature, 463(7279), 318-25. Nature Publishing Group. doi:10.1038/nature08712

ARACNE

- Method for reconstructing biological network using information theoretic method to reduce false positives which are predicted through indirect interactions

  1. Identifies statistically significant gene-gene coregulation by mutual information

  2. It then eliminates indirect relationships in which two genes are coregulated through one or more intermediates

Algorithm for the Reconstruction of Accurate Cellular NEtworks

– "Reverse engineering" or "deconvolution" problem

ARACNE

- ARACNE starts with a network graph where each triplet of genes is connected by an edge.
- The algorithm then examines each gene triplet for which all pairwise MIs are greater than a cut-off and removes the edge with the smallest value.
- Each triplet is analyzed irrespectively of whether its edges have been matched for removal by prior DPI applications to different triplets.
- The DPI states that if genes g1 and g3 interact only through a third gene g2, then I(g1,g3) <= min[I(g1,g2); I(g2,g3)].
- Thus the least of the three MIs can come from indirect interactions only, and checking against the DPI may identify those gene pairs that are not independent but still do not interact.

ARACNE

- An interaction is retained if and only if there exists no alternate paths, via one or more intermediaries or branches on the network graph, which are a better explanation for the information exchange between two genes.
- Since biochemical dynamics is inherently stochastic, statistical interactions over more than a few separating edges are generally weak.
- Will open all three gene loop along the weakest interaction, and therefore introduce false negatives for triplets of interacting genes (although some may be preserved when a non-zero DPI threshold is used).
  - Inability to infer edge directionality.

Removing false positives
Data Processing Inequality (DPI)

$$I(A,C) \leq \min[I(A,B), I(B,C)]$$

- ARACNE: Look at every triplet and remove the weakest link

Model Dependence

No time series → no directionality, steady state statistical dependencies only.

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \qquad + \ldots$$

- Use MaxEnt to define φ
- Can only evaluate two-way marginals
- Truncate at 2nd order potential (cannot reconstruct XOR)
- Mutual information is enough to establish dependencies

$$I(g_i, g_j) = \langle \log P(g_i, g_j)/P(g_i)P(g_j) \rangle$$

Guarantees

- Theorem: If MIs can be estimated with no errors, then ARACNE reconstructs the underlying interaction network exactly, provided this network is a tree and has only pairwise interactions.
- Theorem: The maximum MI spanning tree is a subnetwork of the network built by ARACNE.

Theorem. Let π_{ik} be the set of nodes forming the shortest path in the network between nodes i and k. Then, if MIs can be estimated without errors, ARACNE reconstructs an interaction network without false positives edges, provided: (a) the network consists only of pairwise interactions, (b) for each $j \in \pi_{ik}$, $I_{ij} \geq I_{ik}$. Further, ARACNE does not produce any false negatives, and the network reconstruction is exact iff (c) for each directly connected pair ij and for any other node k, we have $I_{ij} > \min[I_{ik}, I_{jk}]$.

Why should it work?

- Higher order interactions project into lower order ones
- Large loops are locally tress (biological signals decorrelate very fast)
- Small loops (e.g. feed forward) are often transient

Gaussian Kernel for MI Estimation

$$f(x,y) = \frac{1}{2\pi h^2 M}\sum_i \exp\left[-\frac{(x-x_i)^2+(y-y_i)^2}{2h^2}\right]$$

$$f(x) = \frac{1}{\sqrt{2\pi} h M}\sum_i \exp\left[-\frac{(x-x_i)^2}{2h^2}\right]$$

Consistent for $M \to \infty$ for $h(M) \to 0$ and $[h(M)]^2 M \to \infty$.

How to select h?
Maybe $h = h(x,y)$?

Copula Transform

No need to have $h = h(x, y)$.

MI error vs. ranking error

- Can use universal best *h*

Synthetic Networks

$$\frac{dx_i}{dt} = a_i \prod_j \frac{I_{0,j}^{\nu_j}}{I_j^{\nu_j} + I_{j,0}^{\nu_j}} \prod_j \left(1 + \frac{A_j^{\nu_j}}{A_j^{\nu_j} + A_{j,0}^{\nu_j}}\right) - b_i x_i$$

To simulate phenotypes and conditions, randomize a_i and b_i (and, possibly, $I_{j,0}$, $A_{j,0}$).

Benchmarks

$N_{TP} - N_{FP} = \max$ at $p = 10^{-4}$.

ARACNE

- Advantages:
  - Less false positives without losing many true positives.
- Disadvantages:
  - You cannot determine the direction of the interaction.
  - From a loop of three genes one edge will always be removed

Implementation

- Implemented in the R-package minet:
- Install Package

```
source("http://bioconductor.org/biocLite.R")
biocLite(”minet")
```

- Main estimation function

```
aracne(mim, eps=0)
```

  - mim: mutual information matrix

```
mim <- build.mim(syn.data,estimator="spearman")
```

  - eps: threshold for setting an edge to zero

Applied to B lymphocytes expression data

Applied to B lymphocytes expression data

- MYC (proto-oncogene) subnetwork (2063 genes)
- 29 of the 56 (51.8%) predicted first neighbors biochemically validated as targets of the MYC transcription factor.
- New candidate targets were identified, 12 experimentally validated.
  - 11 proved to be true targets.
- The candidate targets that have not been validated are possibly also correct.

Summary

- Mutual information is an intuitive measure of network dependence
- ARACNE is a well established methodology for discovering networks
  - More recent methods expand to model directionality
- Weighted co-expression networks are powerful modeling approaches for building networks and testing for associations with important outcomes
- Next lectures will go into more details of the latest network methodologies

Pathway & Network Analysis for Omics Data: Undirected Graphical Models - II

Ali Shojaie

July 2014
Summer Institute for Statistical Genetics
University of Washington

©Ali Shojaie

Recap: Co-Expression Networks

- This is the simplest (and most-widely used) method for estimating networks; it assumes that edges correspond to large correlation magnitudes
- Let $r(i,j)$ be correlation between X_i and X_j; we claim an edge between i and j if $|r(i,j)| > \tau$.
- Here, τ is a user-specified threshold, and is the tuning parameter for this method.
- Alternatively, we can test $H_0 : r_{xy} = 0$
  - A commonly used test is given by the Fisher transformation

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = \text{artanh}(r) \sim_{H_0} N(0, \frac{1}{\sqrt{n-3}})$$

- By construction, this is an undirected network (correlation is symmetric).
- Related methods, e.g. *weighted* co-expression networks, use similar ideas.

Limitations of Co-Expression Networks

- The estimation is highly dependent on the choice of τ
- However, they may not correctly detect the edges in biological networks: two genes/proteins can have high correlations, even if they don't interact with each other!

Limitations of Co-Expression Networks

- The estimation is highly dependent on the choice of τ
- However, they may not correctly detect the edges in biological networks: two genes/proteins can have high correlations, even if they don't interact with each other!

Limitations of Co-Expression Networks

- The estimation is highly dependent on the choice of τ
- However, they may not correctly detect the edges in biological networks: two genes/proteins can have high correlations, even if they don't interact with each other!

Limitations of Co-Expression Networks

- The estimation is highly dependent on the choice of τ
- However, they may not correctly detect the edges in biological networks: two genes/proteins can have high correlations, even if they don't interact with each other!

Partial Correlations

- Partial correlation measures the correlation between i and j when the effect of the other variables are removed.
- In our example, this means that we would be taking into account that the *"secret" was passed through mutual friends, and not directly*.
- This gives a more direct connection to biological networks; in PPI networks: if protein A binds with B and C, but B and C don't bind, then the correlation between B and C will be removed once conditioned on A.
- Mathematically, the partial correlation between X_i and X_j given X_k is given by:

$$\rho_{ij \cdot k} \equiv \rho(X_i X_j | X_k) = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2}}.$$

Partial Correlation

- Partial correlation is also symmetric
- Partial correlation is also a number between -1 and 1
- In partial correlation networks, we draw an edge between X and Y, if the partial correlation between them is large
- Calculation of partial correlation is more difficult
- Again, we can determine this using testing, however, we need a larger sample size
- New statistical methods have been proposed in the past couple of years to make this possible...(active area of research)

A simple example

$$\textit{Correlation} = \begin{bmatrix} 1 & -.8 & .7 \\ -.8 & 1 & -.8 \\ .7 & -.8 & 1 \end{bmatrix} \quad \textit{PartialCorr} = \begin{bmatrix} 1 & .6 & 0 \\ .6 & 1 & .6 \\ 0 & .6 & 1 \end{bmatrix}$$

A larger example

- A network with 10 nodes and 20 edges
- $n = 100$ observations
- Estimation using correlation & partial correlation (20 edges)

Partial Correlation for Gaussian Random Variables

- It turns out, we can calculate the partial correlation between X_i and X_j given all other variables, by calculating the inverse of the empirical covariance matrix S.
- In other words, the (i, j) entry in Σ^{-1} given the partial correlation between X_i and X_j given all other variables $X_{\setminus i,j}$.
- Now suppose the variables are connected by a graph G, then if $X \sim N(0, \Sigma)$, the nonzero entries in the inverse covariance matrix correspond to the edges of G: $(i, j) \in E$ iff $\Sigma^{-1}_{ij} \neq 0$

Partial Correlation for Gaussian Random Variables

$$\begin{pmatrix} - & x & 0 \\ x & - & x \\ 0 & x & - \end{pmatrix} \quad \begin{pmatrix} - & x & x & 0 \\ x & - & x & 0 \\ x & x & - & 0 \\ 0 & 0 & 0 & - \end{pmatrix}$$

$$\begin{pmatrix} - & x & 0 & x \\ x & - & x & 0 \\ 0 & x & - & x \\ x & 0 & x & - \end{pmatrix} \quad \begin{pmatrix} - & 0 & 0 & x \\ 0 & - & x & 0 \\ 0 & x & - & x \\ x & 0 & x & - \end{pmatrix}$$

Estimation

Therefore, to estimate the edges in the graph G,

- First, calculate the empirical covariance matrix of the observations $S = 1/(n-1)X^{\mathsf{T}}X$ (remember X is $n \times p$).
- Then, find the inverse of S. Non-zero values of this matrix determine where there are edges in the network.
- This seems pretty simple, however, in practice this may not work that well, even if the sample size is very large!!

Difficulties in HD

- A number of problems arise in high dimensional settings, especially when $p \gg n$.
- First, S is not invertible if $p > n$!
- Even if $p < n$, but n is not very large, we may still get poor estimates, and we may get more false positives and false negatives.

Estimation in High Dimensions – Method 1

- A number of methods have been proposed for estimation of conditional independence graphs from Gaussian observations in high dimensions.
- The main idea in most of these methods is to use a regularization penalty, like the lasso.
- The idea in the first method, called neighborhood selection, is to estimate the graph by fitting a penalized regression of each variable on all other variables.
- In other words, we solve, for $j = 1, \ldots, p$

$$\|X_j - \sum_{k \neq j} X_k \beta_k\|^2 + \lambda \sum_{k \neq j} |\beta_k|$$

- The final estimate of the graph is obtained by getting all of the edges fond from these individual regression problems.

Estimation in High Dimensions – Method 2

- In the second approach, called graphical lasso, we directly estimate the inverse covariance matrix by maximizing the l_1 penalized log likelihood
- It turns out that, the log likelihood function for Gaussian random variables can be written as

$$\text{logdet}(\Theta) - \mathrm{tr}(S\Theta),$$

where Θ is the $p \times p$ inverse covariance matrix (also known as precision matrix).
- Therefore, we can estimate Θ by maximizing the penalized log-likelihood objective function

$$\text{logdet}(\Theta) - \mathrm{tr}(S\Theta) - \lambda\|\Theta\|_1,$$

- Here, logdet gives the logarithm of determinant of matrix; tr gives the trace of the matrix, or some of its diagonal values; and λ is the tuning parameter.

Comparing the Two Approaches

- It turns out that the neighborhood selection approach is an approximation to the graphical lasso problem:
  - Consider regression of X_j on $X_k, j \neq k$
  - Then the regression coefficient for neighborhood selection is related to the j, k element of Θ:

$$\beta_k = -\frac{\Theta_{jk}}{\Theta_{jj}}$$

- A main difficulty with the neighborhood selection approach is that the resulting graph is not necessarily symmetric.
- To deal with this, we can take the union or intersection of edges from regressing X_k on X_k and X_j on X_k; however, this is an ad hoc solution.
- On the other hand, neighborhood selection is computationally more efficient, and may gives better estimates.

A Real Example

- Flow cytometry allows us to obtain measurements of proteins in individual cells, and hence facilitates obtaining datasets with large sample sizes.
- Sachs et al (2003) conducted an experiment and gathered data on $p = 11$ proteins measured on $n = 7466$ cells

Choice of tuning parameter

- Unlike supervised learning, choosing the right λ is very difficult in this case.
- As the previous example shows, as λ gets larger, we get sparser graphs.
- However, there is no systematic way of choosing the right λ.
- A number of methods have been proposed, based on the idea of trying to control the false positives, but this is still the topic of ongoing research.
- One option for choosing λ controls the probability of falsely connecting disconnected components at level α (Banerjee et al, 2008). When variables are standardized, this gives:

$$\lambda(\alpha) = \frac{t_{n-2}(\alpha/2p^2)}{\sqrt{n-2+t_{n-2}(\alpha/2p^2)}},$$

where $t_{n-2}(\alpha)$ is the $(100-\alpha)\%$ quantile of t-distribution with $n-2$ d.f.

Some Comments

- The penalized estimation problems discussed above allow estimation of graphical models in the $p \gg n$ settings, e.g. when p is in 1000's and n is in 100's.
- However, both of these methods, and most other methods for estimation of conditional independence networks, work when the network is **sparse**.
- Sparsity means that there are not many edges in the network, and the network is far from fully connected.
- Good new is that biological networks are believed to be "sparse". However, all of these concepts are theoretical and it is difficult to assess how things work on real networks.

Computation

- As we saw previously, the neighborhood selection problem is an approximation to the graphical lasso problem.
- It turns out that this relationship can be used for solving the graphical lasso problem efficiently.
- The idea is to turn the problem into iterating over P regression problems, one for each column of the precision matrix.
- This results in a very efficient algorithm for solving this problem, and in practice, we can solve problems with p in 1000's and n in 100's in a few minutes.
- The algorithm, as well as the approximation for the neighborhood selection problem, is implemented in the R-package `glasso`.
- In practice, it is often better to use the empirical correlation matrix

An Example in R

- Download the empirical covariance matrix from `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`
- Install the R-package `glasso`

```
library(glasso)

##Read the covariance matrix
sachs <- as.matrix(read.table("sachscov.txt"))
dim(sachs)

##glasso
est.1 <- glasso(s=sachs, rho=5, approx=FALSE, penalize.diagonal=FALSE)

##neighborhood selection
est.2 <- glasso(s=sachs, rho=5, approx=TRUE, penalize.diagonal=FALSE)
```

Exercise

- Estimate the graph from the previous example with different values of tuning parameter (Note: this is denoted by `rho` in the code).
- Try the estimation with and without setting `penalize.diagonal=FALSE`. What do you see?
- Try the estimation with the empirical correlation matrix instead (you may find the function `cov2cor()` useful). What do you see?

Non-linear associations

- Recall that correlation is a measure of linear dependence, this is also true about partial correlation.
- Therefore (partial) correlation may miss associations among variables:

- Mutual information-based methods (ARACNE etc) try to address this issue

Conditional Independence Graphs

- As we saw, in the case of Gaussian variables, the relationships are linear, so (partial) correlation works well.
- In case of Gaussian variables, $\Theta_{jk} = 0$ implies that X_j and X_k are conditionally independent.
- Conditional dependence is a general condition that defines the class of conditional independent graphs (CIG). In CIG,
  - $X \perp\!\!\!\perp Y \mid Z$ iff $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$
  - If X and Y are neighbors ($X - Y$), they are conditionally dependent
  - Let $\mathrm{ne(X)}$ be the neighbors of X, then X is conditionally dependents on $\mathrm{ne(X)}$
  - X is conditionally independent of all other nodes, given $\mathrm{ne(X)}$, i.e. if $Z \notin \mathrm{ne(X)}$, then $X \perp\!\!\!\perp Z \mid \mathrm{ne(X)}$

Non-Gaussian Graphical Models

- In case of Gaussian variables, conditional dependence is easily estimated from the inverse covariance matrix; this is not the case for non-linear associations (other distributions).
- However, this requires multivariate normality, which is a strong assumption!
- How to characterize conditional independence for non-Gaussian data?

Nonparanormal (Gaussian Copula Models)

- Suppose $X \nsim N(0, \Sigma)$, but there exists monotone functions $f_j, j = 1, \ldots p$ such that $[f_1(X_1), \ldots f_p(X_p)] \sim N(0, \Sigma)$
- Liu et al (2009) proposed a method for estimation of conditional independence graphs using this model
- A closely related idea (Liu et al, 2012) is to use rank-based correlation (i.e. Spearman's ρ or Kendall's τ) instead of Pearson's correlation
- Both of these ideas have been implemented in the `R`-package `huge`.
- These methods assume that we can get multivariate normality with marginal transformation; this is a bit more flexible, but still somewhat restrictive.
- We can directly model the nonlinear associations (`R`-package `spacejam`) Voorman et al (2013).

Graphical Models for Discrete Random Variables

- CIG's can also be defined for other distributions, e.g. binomial, poisson, etc
- A special case is binary random variables, which can be useful in genetic applications
- A popular model for estimation of CIG's for binary r.v.'s is the Ising model
- Estimation of conditional independence for Ising model is also relatively easy
- In Ising model, $X \perp\!\!\!\perp Y \mid Z$ iff $\mathrm{logOddsRatio(X, Y|Z = z) = 0}$.
- Based on this, HD Ising models can be estimated using neighborhood selection via penalized logistic regression.

Summary

- Estimation of graphical models is an important but challenging problem.
- Partial correlations provide a better representation of edges in biological networks.
- Estimating the conditional independence graph is almost as costly as estimating the co-expression network (we can obtain a good approximation using the neighborhood selection approach at similar computational cost).
- Choosing the tuning parameter is a challenging problem in both cases, after all, we are still doing unsupervised learning!
- It is often difficult to validate the estimates; however, in case of biological networks, we can compare our findings with known interactions from literature.

Pathway & Network Analysis for Omics Data: Bayesian Networks – Basic Concepts

Ali Shojaie

Feb 2014
Summer Institute for Statistical Genetics
University of Washington

©Ali Shojaie

Bayesian Networks

Bayesian Networks

- Bayesian networks are a special class of graphical models defined on directed acyclic graphs.

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Bayesian Networks

- Bayesian networks are a special class of graphical models defined on directed acyclic graphs.
- Directed acyclic graphs (DAGs) are defined as graphs that:
  i) only have directed edges, i.e. if $A_{ij} \neq 0$, $A_{ji} = 0$;
  ii) there are no cycles in the network.

Bayesian Networks

- Bayesian networks are a special class of graphical models defined on directed acyclic graphs.
- Directed acyclic graphs (DAGs) are defined as graphs that:
  i) only have directed edges, i.e. if $A_{ij} \neq 0$, $A_{ji} = 0$;
  ii) there are no cycles in the network.
- Bayesian networks are widely used to model causal relationships between variables.

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Bayesian Networks

- Bayesian networks are a special class of graphical models defined on directed acyclic graphs.
- Directed acyclic graphs (DAGs) are defined as graphs that:
  i) only have directed edges, i.e. if $A_{ij} \neq 0$, $A_{ji} = 0$;
  ii) there are no cycles in the network.
- Bayesian networks are widely used to model causal relationships between variables.
- Note that correlation $\neq$ causation!

Bayesian Networks

- Bayesian networks are a special class of graphical models defined on directed acyclic graphs.
- Directed acyclic graphs (DAGs) are defined as graphs that:
  i) only have directed edges, i.e. if $A_{ij} \neq 0$, $A_{ji} = 0$;
  ii) there are no cycles in the network.
- Bayesian networks are widely used to model causal relationships between variables.
- Note that correlation $\neq$ causation!
- Therefore, we (usually) cannot estimate Bayesian networks from (partial) correlations

Why Bayesian Networks?

Why Bayesian Networks?

Many biological networks include directed edges:

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Why Bayesian Networks?

Many biological networks include directed edges:

- In gene regulatory networks, protein products of transcription factors can alter the expression of target genes, but the target genes (usually) don't have a direct effect on the expression of transcription factors

Why Bayesian Networks?

Many biological networks include directed edges:

- In gene regulatory networks, the signal from the cell's environment is transducted into the cell, and results e.g. in (global) changes in gene expression, but gene expression may not affect the environmental factors

Why Bayesian Networks?

Many biological networks include directed edges:

- Biochemical reactions in metabolic networks, may not reversible, and in that case, one metabolite may affect the other, but the relationship is ont reciprocated

Why Bayesian Networks?

However, biological networks may not be DAGs:

Why Bayesian Networks?

However, biological networks may not be DAGs:

- Gene regulatory networks, signaling networks and metabolic networks, may all contain feedback loops (positive/negative)

which make estimation even more difficult!

What's the Difference?

What's the Difference?

- Bayesian networks are widely used to model causal relationships between variables.

What's the Difference?

- Bayesian networks are widely used to model causal relationships between variables.
- Undirected networks (e.g. GGM) provide information about associations among variables; while this greatly helps in the study of biological systems, in some cases, they are not enough (e.g. drug development).

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

What's the Difference?

- Bayesian networks are widely used to model causal relationships between variables.
- Undirected networks (e.g. GGM) provide information about associations among variables; while this greatly helps in the study of biological systems, in some cases, they are not enough (e.g. drug development).
- The main difference is of course the direction of the edges; however, it turns out that there are also some differences in terms of structure/skeleton of the network (more on this later).

What's the Difference?

- Bayesian networks are widely used to model causal relationships between variables.
- Undirected networks (e.g. GGM) provide information about associations among variables; while this greatly helps in the study of biological systems, in some cases, they are not enough (e.g. drug development).
- The main difference is of course the direction of the edges; however, it turns out that there are also some differences in terms of structure/skeleton of the network (more on this later).
- We can estimate undirected networks from observational data, i.e. steady-state gene expression data, but usually they are not enough for estimation of directed networks
- Finally, estimation of directed networks is often much more difficult

Why is estimation more difficult?

- Estimation of Bayesian networks requires estimating both the skeleton of the network (i.e. whether there is an edge between i and j) and also the direction of the edges.

Why is estimation more difficult?

- Estimation of Bayesian networks requires estimating both the skeleton of the network (i.e. whether there is an edge between i and j) and also the direction of the edges.
- While estimation of skeleton is possible, direction of edges cannot be in general learned from observational data, no matter how many samples we have (this is referred to as *observational equivalence*). Consider this simple graph:

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Why is estimation more difficult?

- Estimation of Bayesian networks requires estimating both the skeleton of the network (i.e. whether there is an edge between i and j) and also the direction of the edges.
- While estimation of skeleton is possible, direction of edges cannot be in general learned from observational data, no matter how many samples we have (this is referred to as *observational equivalence*). Consider this simple graph:

- Then, no matter what n is, we cannot distinguish between $X_1 \to X_2$ and $X_2 \to X_1$, so basically what we see is:

Outline

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Outline

- **Lecture 7**: Basics of Bayesian networks, including

Outline

- **Lecture 7**: Basics of Bayesian networks, including
  - directed acyclic graphs (DAGs)
  - conditional independence in DAGs, d-separation, and moral graphs
  - probability distributions over DAGs
  - structural equation models (SEM)
  - additional topics (faithfulness, Markov equivalence, ...)

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Outline

- **Lecture 7**: Basics of Bayesian networks, including
  - directed acyclic graphs (DAGs)
  - conditional independence in DAGs, d-separation, and moral graphs
  - probability distributions over DAGs
  - structural equation models (SEM)
  - additional topics (faithfulness, Markov equivalence, ...)
- **Lecture 8**: Estimation of Bayesian networks from observational data

Outline

- **Lecture 7**: Basics of Bayesian networks, including
  - directed acyclic graphs (DAGs)
  - conditional independence in DAGs, d-separation, and moral graphs
  - probability distributions over DAGs
  - structural equation models (SEM)
  - additional topics (faithfulness, Markov equivalence, ...)
- **Lecture 8**: Estimation of Bayesian networks from observational data
- **Lecture 9**: Estimation of Bayesian networks from perturbation and time-course data

Directed Graphs: Some Terminology

- nodes in directed networks represent random variables; we denote the set of nodes by V
- edges are directed, and represent causal relationships among variables; we denote the set of edges by E

Directed Graphs: Some Terminology

- nodes in directed networks represent random variables; we denote the set of nodes by V
- edges are directed, and represent causal relationships among variables; we denote the set of edges by E
- The parents of node j are $\{k : k \to j\}$, we denote this by pa_j or $\mathrm{pa}(j)$

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Directed Graphs: Some Terminology

- nodes in directed networks represent random variables; we denote the set of nodes by V
- edges are directed, and represent causal relationships among variables; we denote the set of edges by E
- The parents of node j are $\{k : k \to j\}$, we denote this by pa_j or $\mathrm{pa}(j)$
- The children of node j are $\{k : j \to k\}$

Directed Graphs: Some Terminology

- nodes in directed networks represent random variables; we denote the set of nodes by V
- edges are directed, and represent causal relationships among variables; we denote the set of edges by E
- The parents of node j are $\{k : k \to j\}$, we denote this by pa_j or $\mathrm{pa}(j)$
- The children of node j are $\{k : j \to k\}$
- Two vertices connected by an edge are called adjacent

Directed Graphs: Some Terminology

Directed Graphs: Some Terminology

- $\mathrm{pa}(1) = \emptyset$, $\mathrm{pa}(2) = 1$, $\mathrm{pa}(3) = \mathrm{pa}(4) = \{2\}$, $\mathrm{pa}(5) = \{3, 4\}$

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Directed Graphs: Some Terminology

- $\mathrm{pa}(1) = \emptyset$, $\mathrm{pa}(2) = 1$, $\mathrm{pa}(3) = \mathrm{pa}(4) = \{2\}$, $\mathrm{pa}(5) = \{3, 4\}$
- What are children of $\{1, \ldots 5\}$?

Directed Graphs: Some Terminology

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Directed Graphs: Some Terminology

- A path between two nodes i and j is a sequence of distinct adjacent nodes:
  - e.g. $i \leftarrow k_1 \rightarrow k_2 \rightarrow k_3 \leftarrow j$
  - In a DAG with p nodes, there cannot be a path longer than $p - 1$ (why?)
  - There can be multiple paths between two nodes

Directed Graphs: Some Terminology

- A path between two nodes i and j is a sequence of distinct adjacent nodes:
  - e.g. $i \leftarrow k_1 \rightarrow k_2 \rightarrow k_3 \leftarrow j$
  - In a DAG with p nodes, there cannot be a path longer than $p-1$ (why?)
  - There can be multiple paths between two nodes
- i is an ancestor of j if there is a directed path of length ≥ 1 from i to j: $i \rightarrow \cdots \rightarrow j$ (or if $i = j$)

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Directed Graphs: Some Terminology

- A path between two nodes i and j is a sequence of distinct adjacent nodes:
  - e.g. $i \leftarrow k_1 \rightarrow k_2 \rightarrow k_3 \leftarrow j$
  - In a DAG with p nodes, there cannot be a path longer than $p-1$ (why?)
  - There can be multiple paths between two nodes
- i is an ancestor of j if there is a directed path of length ≥ 1 from i to j: $i \rightarrow \cdots \rightarrow j$ (or if $i = j$)
- If i is an ancestor of j, then j is said to be a descendant of i

Directed Graphs: Some Terminology

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Directed Graphs: Some Terminology

- What are paths between 1&4, 3&4, 2&6?

Directed Graphs: Some Terminology

- What are paths between 1&4, 3&4, 2&6?
- What are ancestors of $\{1, \ldots 5\}$?

Directed Graphs: Some Terminology

An important concept in DAGs is that of colliders (aka "inverted forks"):

Directed Graphs: Some Terminology

An important concept in DAGs is that of colliders (aka "inverted forks"):

- k is a collider on a path between i and j if it is a not an end-point of the path, and the path is of the form

$$i \ldots \rightarrow k \leftarrow \ldots j$$

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Directed Graphs: Some Terminology

An important concept in DAGs is that of colliders (aka "inverted forks"):

- k is a collider on a path between i and j if it is a not an end-point of the path, and the path is of the form

$$i \ldots \rightarrow k \leftarrow \ldots j$$

- k is an non-collider if it is not an end-point, and is not a collider on a path:
  - $i \ldots \leftarrow k \leftarrow \ldots j$
  - $i \ldots \rightarrow k \rightarrow \ldots j$
  - $i \ldots \leftarrow k \rightarrow \ldots j$

Directed Graphs: Some Terminology

An important concept in DAGs is that of colliders (aka "inverted forks"):

- k is a collider on a path between i and j if it is a not an end-point of the path, and the path is of the form

$$i \ldots \rightarrow k \leftarrow \ldots j$$

- k is an non-collider if it is not an end-point, and is not a collider on a path:
  - $i \ldots \leftarrow k \leftarrow \ldots j$
  - $i \ldots \rightarrow k \rightarrow \ldots j$
  - $i \ldots \leftarrow k \rightarrow \ldots j$
- Note: colliders and non-colliders are defined w.r.t. paths; a collider in one path can be a non-collider in another!

Directed Graphs: Some Terminology

Directed Graphs: Some Terminology

- What are the colliders on paths between 1&4, 3&4, 2&6?

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Directed Graphs: Some Terminology

- What are the colliders on paths between 1&4, 3&4, 2&6?
- What are the non-colliders on paths between 1&4, 3&4, 2&6?

Factorization of Probability Distributions over DAGs

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Factorization of Probability Distributions over DAGs

- First, note that for any set of random variables, not necessarily on a DAG, we can write:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_1 \mid X_2, X_3) P(X_2 | X_3) P(X_3) \\ &= P(X_3 \mid X_1, X_2) P(X_2 | X_1) P(X_1) \\ &= \cdots \end{aligned}$$

Factorization of Probability Distributions over DAGs

- First, note that for any set of random variables, not necessarily on a DAG, we can write:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_1 \mid X_2, X_3) P(X_2 | X_3) P(X_3) \\ &= P(X_3 \mid X_1, X_2) P(X_2 | X_1) P(X_1) \\ &= \cdots \end{aligned}$$

- Now, consider this simple DAG

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Factorization of Probability Distributions over DAGs

- First, note that for any set of random variables, not necessarily on a DAG, we can write:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_1 \mid X_2, X_3) P(X_2 | X_3) P(X_3) \\ &= P(X_3 \mid X_1, X_2) P(X_2 | X_1) P(X_1) \\ &= \cdots \end{aligned}$$

- Now, consider this simple DAG

- Then, the probability distribution can be factorized as

$$P(X_1, X_2, X_3) = P(X_3 \mid X_2) P(X_2 | X_1) P(X_1)$$

Factorization of Probability Distributions over DAGs

- In general, for any set of random variables on a DAG $G = (V, E)$, and for any probability distribution P (Markov relative to G) we have

$$P(V) = \prod_{j \in V} P(X_j \mid \mathrm{pa}_j)$$

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Factorization of Probability Distributions over DAGs

- In general, for any set of random variables on a DAG $G = (V, E)$, and for any probability distribution P (Markov relative to G) we have

$$P(V) = \prod_{j \in V} P(X_j \mid \mathrm{pa}_j)$$

- Compare this with the general probability decomposition

$$P(V) = \prod_{j \in V} P(X_j \mid X_1, \ldots, X_{j-1})$$

Factorization of Probability Distributions over DAGs

- In general, for any set of random variables on a DAG $G = (V, E)$, and for any probability distribution P (Markov relative to G) we have

$$P(V) = \prod_{j \in V} P(X_j \mid \mathrm{pa}_j)$$

- Compare this with the general probability decomposition

$$P(V) = \prod_{j \in V} P(X_j \mid X_1, \ldots, X_{j-1})$$

- This means that on DAGs we have

$$P(X_j \mid X_1, \ldots, X_{j-1}) = P(X_j \mid \mathrm{pa}_j)$$

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Factorization of Probability Distributions over DAGs

- In general, for any set of random variables on a DAG $G = (V, E)$, and for any probability distribution P (Markov relative to G) we have

$$P(V) = \prod_{j \in V} P(X_j \mid \mathrm{pa}_j)$$

- Compare this with the general probability decomposition

$$P(V) = \prod_{j \in V} P(X_j \mid X_1, \ldots, X_{j-1})$$

- This means that on DAGs we have

$$P(X_j \mid X_1, \ldots, X_{j-1}) = P(X_j \mid \mathrm{pa}_j)$$

- In other words, the probability distribution for each variable depends only on its parents

Structural Equation Models

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Structural Equation Models

- A popular way to represent causal relationships on DAGs is via structural equation models

$$X_j = f_j(pa_j, \gamma_j), \quad j = 1, \ldots, p$$

Structural Equation Models

- A popular way to represent causal relationships on DAGs is via structural equation models

$$X_j = f_j(pa_j, \gamma_j), \quad j = 1, \ldots, p$$

  - f_j can be in general any function relating j to its parents

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Structural Equation Models

- A popular way to represent causal relationships on DAGs is via structural equation models

$$X_j = f_j(pa_j, \gamma_j), \quad j = 1, \ldots, p$$

  - f_j can be in general any function relating j to its parents
  - γ_j's represent the independent component of jth variable (i.e. the part that doesn't depend on pa_j

Structural Equation Models

- A popular way to represent causal relationships on DAGs is via structural equation models

$$X_j = f_j(pa_j, \gamma_j), \quad j = 1, \ldots, p$$

  - f_j can be in general any function relating j to its parents
  - γ_j's represent the independent component of jth variable (i.e. the part that doesn't depend on pa_j
- For Gaussian random variables, f_i is linear

$$X_j = \sum_{j' \in pa_j} \rho_{jj'} X_{j'} + \gamma_j, \quad j = 1, \ldots, p$$

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs

Structural Equation Models

- A popular way to represent causal relationships on DAGs is via structural equation models

$$X_j = f_j(pa_j, \gamma_j), \quad j = 1, \ldots, p$$

  - f_j can be in general any function relating j to its parents
  - γ_j's represent the independent component of jth variable (i.e. the part that doesn't depend on pa_j
- For Gaussian random variables, f_i is linear

$$X_j = \sum_{j' \in pa_j} \rho_{jj'} X_{j'} + \gamma_j, \quad j = 1, \ldots, p$$

  - here, $\rho_{jj'}$ denotes the magnitude of effect of j' on j, or their partial correlation

A Toy Example

A Toy Example

Assuming normality we can write:

$$
\begin{aligned}
X_1 &= \gamma_1 \\
X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2 \\
X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3
\end{aligned}
$$

A Toy Example

Assuming normality we can write:

$$\begin{aligned} X_1 &= \gamma_1 \\ X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2 \\ X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3 \end{aligned}$$

For non-Gaussian variables, these equations will involve non-linear relationships.

Independence (unconditional)

- Recall the following (equivalent) characterizations of independence, $X \perp\!\!\!\perp Y$:

Independence (unconditional)

- Recall the following (equivalent) characterizations of independence, $X \perp\!\!\!\perp Y$:
  - $P(X = x, Y = y) = P(X = x)P(Y = y)$
  - $P(X = x | Y = y) = P(X = x)$ (is symmetric)

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Independence (unconditional)

- Recall the following (equivalent) characterizations of independence, $X \perp\!\!\!\perp Y$:
  - $P(X = x, Y = y) = P(X = x)P(Y = y)$
  - $P(X = x | Y = y) = P(X = x)$ (is symmetric)
- Intuitively, if $X \perp\!\!\!\perp Y$ then knowledge of X provides no information about Y.

Independence (unconditional)

- Recall the following (equivalent) characterizations of independence, $X \perp\!\!\!\perp Y$:
  - $P(X = x, Y = y) = P(X = x)P(Y = y)$
  - $P(X = x | Y = y) = P(X = x)$ (is symmetric)
- Intuitively, if $X \perp\!\!\!\perp Y$ then knowledge of X provides no information about Y.
- These can be generalized for vectors.

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Independence (unconditional)

- Recall the following (equivalent) characterizations of independence, $X \perp\!\!\!\perp Y$:
  - $P(X = x, Y = y) = P(X = x)P(Y = y)$
  - $P(X = x | Y = y) = P(X = x)$ (is symmetric)
- Intuitively, if $X \perp\!\!\!\perp Y$ then knowledge of X provides no information about Y.
- These can be generalized for vectors.
- If X and Y are jointly Gaussian $X \perp\!\!\!\perp Y$ iff $Corr(X, Y) = 0$.

Independence (unconditional)

- Recall the following (equivalent) characterizations of independence, $X \perp\!\!\!\perp Y$:
  - $P(X = x, Y = y) = P(X = x)P(Y = y)$
  - $P(X = x | Y = y) = P(X = x)$ (is symmetric)
- Intuitively, if $X \perp\!\!\!\perp Y$ then knowledge of X provides no information about Y.
- These can be generalized for vectors.
- If X and Y are jointly Gaussian $X \perp\!\!\!\perp Y$ iff $Corr(X, Y) = 0$.
- If X and Y are binary, $X \perp\!\!\!\perp Y$ iff $logOR(X, Y) = 0$.

Conditional Independence

- Conditional independence $X \perp\!\!\!\perp Y \mid Z$ has similar characterizations:

Conditional Independence

- Conditional independence $X \perp\!\!\!\perp Y \mid Z$ has similar characterizations:
  i) $P(X = x, Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$
  ii) $P(X = x | Y = y, Z = z) = P(X = x | Z = z)$ (is symmetric)

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Conditional Independence

- Conditional independence $X \perp\!\!\!\perp Y \mid Z$ has similar characterizations:
  i) $P(X = x, Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$
  ii) $P(X = x | Y = y, Z = z) = P(X = x | Z = z)$ (is symmetric)
- We also have,

$$P(X = x, Y = y, Z = z) = \frac{P(X = x, Z = z) P(Y = y, Z = z)}{P(Z = z)}.$$

Conditional Independence

- Conditional independence $X \perp\!\!\!\perp Y \mid Z$ has similar characterizations:
  i) $P(X = x, Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$
  ii) $P(X = x | Y = y, Z = z) = P(X = x | Z = z)$ (is symmetric)
- We also have,

$$P(X = x, Y = y, Z = z) = \frac{P(X = x, Z = z) P(Y = y, Z = z)}{P(Z = z)}.$$

- Intuitively, if $X \perp\!\!\!\perp Y$ then if Z is known, knowledge of X provides no information about Y.

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Conditional Independence

- Conditional independence $X \perp\!\!\!\perp Y \mid Z$ has similar characterizations:
  i) $P(X = x, Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$
  ii) $P(X = x | Y = y, Z = z) = P(X = x | Z = z)$ (is symmetric)
- We also have,

$$P(X = x, Y = y, Z = z) = \frac{P(X = x, Z = z) P(Y = y, Z = z)}{P(Z = z)}.$$

- Intuitively, if $X \perp\!\!\!\perp Y$ then if Z is known, knowledge of X provides no information about Y.
- These can be generalized for vectors.

Conditional Independence

- If X & Y are binary, $X \perp\!\!\!\perp Y | Z$ iff $logOR(X, Y|Z) = 0$
  - This is the coefficient in logistic regression of (say) Y on X, Z.

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Conditional Independence

- If X & Y are binary, $X \perp\!\!\!\perp Y | Z$ iff $logOR(X, Y|Z) = 0$
  - This is the coefficient in logistic regression of (say) Y on X, Z.
- If X & Y are jointly Gaussian, $X \perp\!\!\!\perp Y | Z$ iff $Corr(X, Y|Z) = 0$.
  - This is the coefficient in linear regression of (say) Y on X, Z.

The Toy Example, Revisited

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

The Toy Example, Revisited

- Recall that $P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$

The Toy Example, Revisited

- Recall that $P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$
- This implies that $X_3 \perp\!\!\!\perp X_1 | X_2$ (by (i))

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

The Toy Example, Revisited

- Recall that $P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$
- This implies that $X_3 \perp\!\!\!\perp X_1 | X_2$ (by (i))
- However, this is not always the case on DAGs!

The Toy Example, Revisited

- Recall that $P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$
- This implies that $X_3 \perp\!\!\!\perp X_1|X_2$ (by (i))
- However, this is not always the case on DAGs!
- How can we read conditional independence relations from the graph?

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

The Toy Example, Revisited

- Recall that $P(X_1, X_2, X_3) = P(X_3|X_2)P(X_2|X_1)P(X_1)$
- This implies that $X_3 \perp\!\!\!\perp X_1|X_2$ (by (i))
- However, this is not always the case on DAGs!
- How can we read conditional independence relations from the graph?
- We can do this using a concept called d-separation?

An example from genetics

Consider an example from population genetics:

- We have genetic information for *M*other, *F*ather, *D*aughter and *S*on in form of dominant/recessive genotype (A/a) for a single gene
- Then each individual can have one of three states: AA, aa, Aa

An example from genetics

Consider an example from population genetics:

- Now, it is natural to assume that given the parents' genetic information, the genotypes of *S*on and *D*aughter are independent $\Rightarrow S \perp\!\!\!\perp D \mid \{M, F\}$

An example from genetics

Consider an example from population genetics:

- Also, one can assume independence among genotypes of M and $F \Rightarrow M \perp\!\!\!\perp F$
- However, if we know that e.g. Son has Aa, and Mother has aa, then Father should have Aa or AA $\Rightarrow M \not\perp\!\!\!\perp F | S$

d-separation

A path π is said to be d-separated (or blocked) by a set of nodes Z, iff

1. π includes a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle note is in Z, or
2. π contains a collider (or inverted fork) $i \rightarrow m \leftarrow j$ such that neither the middle node m nor its descendants are <u>NOT</u> in Z.

d-separation

A path π is said to be d-separated (or blocked) by a set of nodes Z, iff

1. π includes a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle note is in Z, or
2. π contains a collider (or inverted fork) $i \to m \leftarrow j$ such that neither the middle node m nor its descendants are NOT in Z.

How is this used?

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

d-separation

A path π is said to be d-separated (or blocked) by a set of nodes Z, iff

1. π includes a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle note is in Z, or
2. π contains a collider (or inverted fork) $i \to m \leftarrow j$ such that neither the middle node m nor its descendants are NOT in Z.

How is this used?

- If i and j are d-separated given Z, then $X_i \perp\!\!\!\perp X_j | Z$ for any probability distribution P factorizing according to G

d-separation

A path π is said to be d-separated (or blocked) by a set of nodes Z, iff

1. π includes a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle note is in Z, or
2. π contains a collider (or inverted fork) $i \to m \leftarrow j$ such that neither the middle node m nor its descendants are NOT in Z.

How is this used?

- If i and j are d-separated given Z, then $X_i \perp\!\!\!\perp X_j | Z$ for any probability distribution P factorizing according to G
- If i and j are d-separated given $\emptyset$, then $X_i \perp\!\!\!\perp X_j$ for any probability distribution P factorizing according to G

Genetics example, revisited

Consider an example from population genetics:

Genetics example, revisited

Consider an example from population genetics:

- $\{M, F\}$ block all paths from S to $D \Rightarrow D \perp\!\!\!\perp S \mid \{M, F\}$

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Genetics example, revisited

Consider an example from population genetics:

- $\{M, F\}$ block all paths from S to $D \Rightarrow D \perp\!\!\!\perp S \mid \{M, F\}$
- Is $M \perp\!\!\!\perp F$?

Genetics example, revisited

Consider an example from population genetics:

- $\{M, F\}$ block all paths from S to $D \Rightarrow D \perp\!\!\!\perp S \mid \{M, F\}$
- Is $M \perp\!\!\!\perp F$?
- Is $M \perp\!\!\!\perp F \mid \{S, D\}$, $\mid S$, $\mid D$?

Moral Graphs

- Reading conditional independence relations from DAGs can be difficult
- An alternative approach is to use a modified version of the network, called the moral graph of DAG
- To get the moral graph $\tilde{G}$ of G
  - join ("marry") common parents of each node
  - remove all the directions
- Then, $X_i \perp\!\!\!\perp X_j | Z$ iff Z separates i and j in $\tilde{G}$

Genetics example, revisited (again)

Consider an example from population genetics:

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Genetics example, revisited (again)

Consider an example from population genetics:

- Is $S \perp\!\!\!\perp D \mid \{M, F\}$
- Is $M \perp\!\!\!\perp F$?
- Is $M \perp\!\!\!\perp F \mid \{S, D\}$, $\mid S$, $\mid D$?

Genetics example, revisited (again)

Consider an example from population genetics:

Introduction
Probability Distribution over DAGs
Conditional Independence in DAGs
d-separation
Moral Graphs

Genetics example, revisited (again)

Consider an example from population genetics:

- Is $S \perp\!\!\!\perp D \mid \{M, F\}$
- Is $M \perp\!\!\!\perp F$?
- Is $M \perp\!\!\!\perp F \mid \{S, D\}$, $\mid S$, $\mid D$?

A More Complex Example

What are conditional independence relations in this graph?

A More Complex Example

What are conditional independence relations in this graph?

Pathway & Network Analysis for Omics Data: Bayesian Networks – Estimation from Observational Data

Ali Shojaie

July 24-26, 2013
Summer Institute for Statistical Genetics
University of Washington

©Ali Shojaie

Estimation of DAGs in Biological Settings

- Estimation of DAGs is (in general) computationally very hard (in fact, it's NP-hard): there are $\sim 2^{p^2}$ DAGs with p nodes!
- Three different types of biological data can be used for estimation of directed graphs:
  - i) observational data: steady-state data, or data comparing normal & cancer cells
  - ii) time-course data: time-course gene expression data
  - iii) perturbation data: data from knockouts experiments
- This lecture, we will cover (i), next lecture we will cover (ii) and (iii)

Estimation of DAGs from Observational Data

Algorithms for estimation of DAGs can be broadly categorized into two groups:

- constraint-based methods
  - often based on tests for CI & provide theoretical guarantees
  - PC algorithm, Grow-Shrink
- score & search methods
  - They assign a "score" to each estimated graph (e.g. based on likelihood, Bayes factor, AIC etc)
  - Then do a (greedy) search to find the best scoring graph
  - Hill Climbing algorithm
- "hybrid" methods
  - Usually first find the Markov blanket (e.g. the moral graph)
  - Then perform a search in a restricted space
  - Max-Min Hill Climbing algorithm

Constraint-Based Methods

- Need a conditional independence test (to test if $X \perp\!\!\!\perp Y \mid Z$)
  - For Gaussian data, we can use partial correlation (or the Fisher's Z-transformation of it)
  - For Binary data, we can use logOR
  - In general, we can use conditional mutual information
- The idea is to see if there exists a set S, for each pair of nodes j, j', such that $X_j \perp\!\!\!\perp X_{j'} \mid S$
  - S can have 0 to p-2 members! usually stop at some $k \ll p$
  - I.e., for each pair of variables (all $\binom{p}{2}$ of them), we need to look at all possible subsets of remaining variables!!
- Recall that conditional independence is symmetric $\Rightarrow$ undirected graph!!
- So, these methods find the structure/skeleton of the DAG (will talk about direction later)

PC Algorithm (Spirtes et al, 1993)

- One of the first algorithms for learning structure of DAGs
- Efficient implementations that allow for learning DAG structures with p up to ~ 1000
  - R-package `pcalg` (Kalisch & Buhlmann, 2007)
- The algorithm starts with a complete graph (i.e. a fully connected graph)
- Then for each pair of nodes j, j' it finds a separating set, S such that $X_j \perp\!\!\!\perp X_{j'} \mid S$
- If a set is found, then remove the edge, otherwise, $j - j'$

PC Algorithm (Spirtes et al, 1993)

Start with a complete undirected graph, and set $i = 0$
Repeat

- For each $j \in V$
- For each $j' \in \mathrm{ne}(j)$
- Determine if $\exists S \subset \mathrm{ne}(j) \backslash \{j'\}$ with $|S| = i$
  - Test for CI: is $X_j \perp\!\!\!\perp X_{j'} \mid S$?
  - If such an S exists, then set $S_{jj'} = S$, remove $j - j'$ edge
- $i = i + 1$

Until $|\mathrm{ne}(j)| < i$ for all j

Example

Example

Example

$i = 0$

Example

$i = 0 \quad S_{1,2} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$
$\quad\quad\quad S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$S_{1,4} = \emptyset$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$S_{1,4} = \emptyset$

$i = 1$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$S_{1,4} = \emptyset$

$i = 1$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$S_{1,4} = \emptyset$

$i = 1 \quad S_{3,4} = \{2\}$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$\quad\quad\quad S_{1,4} = \emptyset$

$i = 1 \quad S_{3,4} = \{2\}$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$\quad\quad\quad S_{1,4} = \emptyset$

$i = 1 \quad S_{3,4} = \{2\}$

$i = 2$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$\quad\quad\quad S_{1,4} = \emptyset$

$i = 1 \quad S_{3,4} = \{2\}$

$i = 2 \quad S_{1,5} = \{3, 4\}$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$\quad\quad\quad S_{1,4} = \emptyset$

$i = 1 \quad S_{3,4} = \{2\}$

$i = 2 \quad S_{1,5} = \{3, 4\}$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$\quad\quad\quad S_{1,4} = \emptyset$

$i = 1 \quad S_{3,4} = \{2\}$

$i = 2 \quad S_{1,5} = \{3, 4\}$

$\quad\quad\quad S_{2,5} = \{3, 4\}$

Example

$i = 0 \quad S_{1,2} = \emptyset$

$\quad\quad\quad S_{1,4} = \emptyset$

$i = 1 \quad S_{3,4} = \{2\}$

$i = 2 \quad S_{1,5} = \{3, 4\}$

$\quad\quad\quad S_{2,5} = \{3, 4\}$

Example

$i = 0$ $S_{1,2} = \emptyset$
$S_{1,4} = \emptyset$
$i = 1$ $S_{3,4} = \{2\}$
$i = 2$ $S_{1,5} = \{3, 4\}$
$S_{2,5} = \{3, 4\}$
$i = 3$ STOP ($|\mathrm{ne}_i| < 3 \forall j$)

Analysis of Protein Flow Cytometry using pcalg

```
> dat <- read.table('sachs.data')
> p <- ncol(dat)
> n <- nrow(dat)
## define independence test (partial correlations)
> indepTest <- gaussCItest
## define sufficient statistics
> suffStat <- list(C=cor(dat), n=n)
## estimate CPDAG
> pc.fit <- pc(suffStat, indepTest, p, alpha=0.1, verbose=FALSE)
> plot(pc.fit, main='PC Algorithm')
```

- Need to determine the type of CI test (`indepTest`), and sufficient statistics (`suffStat`)
- Also need to choose α (`alpha`), the probability of false positive for selecting edges.
  - Larger values of α allow more edges (not adjusted for multiple comparisons)
  - The algorithm works faster when α is small

Analysis of Protein Flow Cytometry using `pcalg`

But wait, where did the directions come from? And why are only some of the edges directed?

Markov Equivalence

Consider the following 4 graphs

Markov Equivalence

Consider the following 4 graphs

Which graphs satisfy $X_1 \perp\!\!\!\perp X_3 \mid X_2$?

Markov Equivalence

Consider the following 4 graphs

Markov Equivalence

Consider the following 4 graphs

In the first 3 graphs, $X_1 \perp\!\!\!\perp X_3 \mid X_2$?
Two graphs that imply the same CI relationships via d-separation are called Markov equivalent

Representation of Markov Equivalence

- Markov equivalent graphs correspond to the same probability distribution and cannot be distinguished from each other based on observations!
- Therefore, the direction of edges that correspond to Markov equivalent graphs cannot be determined
- We show these edges using undirected edges in the graph
- The resulting graph is a CPDAG (completed partially directed acyclic graph), and is really the best we can do!

CPDAGs

Introduction
PC Algorithm
Other Estimation Methods
Summary
Structure Learning
Markov Equivalence and PCDAGs

CPDAGs

Finding Partial Directions in DAGs

- Partial directions in DAGs can be determined from unmarried colliders:
  - For each unmarried collider $i - k - j$
  - If $k \notin S_{ij}$, orient $i - k - j$ as $i \to k \leftarrow j$
- In addition to the above rule
  - Orient each remaining unmarried collider $i \to k - j$ as $i \to k \to j$
  - If $i \to k \to j$ and $i - j$ then orient as $i \to j$
  - If $i - m - j$ and $i \to k \leftarrow j$ are unmarried colliders and $m - k$, then orient as $m \to k$

Example

$i = 0$ $\quad S_{1,2} = \emptyset$

$\quad S_{1,4} = \emptyset$

$i = 1$ $\quad S_{3,4} = \{2\}$

$i = 2$ $\quad S_{1,5} = \{3, 4\}$

$\quad S_{2,5} = \{3, 4\}$

The bnlearn package

- There are a number of R-packages for learning the structure of DAGs, including pclag, bnlearn, deal
- bnlearn implements a number of estimation methods, both constraint-based and search-based:
  - constraint-based:
    - Grow-Shrink (GS);
    - Incremental Association Markov Blanket (IAMB);
    - Fast Incremental Association (Fast-IAMB);
    - Interleaved Incremental Association (Inter-IAMB);
  - the following score-based structure learning algorithms:
    - Hill Climbing (HC);
    - Tabu Search (Tabu);
  - the following hybrid structure learning algorithms:
    - Max-Min Hill Climbing (MMHC);
    - General 2-Phase Restricted Maximization (RSMAX2);

Analysis of Protein Flow Cytometry using bnlearn

```
> dag1 <- gs(dat, alpha=0.01)    #GS method
> dag2 <- hc(dat2)               #Hill-Climbing search
>
> par(mfrow= c(1,2))
> plot(dag1)
> plot(dag2)
>
> compare(dag1, dag2)            #compare the two DAGs
```

- For GS need to choose α (alpha), the false positive probability for selecting edges
- gs (and other structure-based methods) find a PCDAG
- hc gives a directed graph (with highest score)
  - A number of criteria for choosing the "best" graph are implemented
  - To "search" the space either a new edge is added, or a current edge is removed, or reversed (if no cycles)

Analysis of Protein Flow Cytometry using bnlearn

```
> dag1
  Bayesian network learned via Constraint-based methods

  model:
    [partially directed graph]
  nodes:                                  11
  arcs:                                   26
    undirected arcs:                      3
    directed arcs:                        23
  average markov blanket size:            6.00
  average neighbourhood size:             4.73
  average branching factor:               2.09

  learning algorithm:                     Grow-Shrink
  conditional independence test:          Pearson's Linear Correlation
  alpha threshold:                        0.01
  tests used in the learning procedure:   2029
  optimized:                              TRUE
```


Analysis of Protein Flow Cytometry using bnlearn

```
> dag2
  Bayesian network learned via Score-based methods

  model:
   [PKC][pjnk|PKC][P44|pjnk][pakts|P44:PKC:pjnk][praf|P44:pakts:PKC][PIP3|pakts
   [plcg|praf:PIP3:P44:pakts:pjnk][pmek|praf:plcg:PIP3:P44:pakts:pjnk]
   [PIP2|plcg:PIP3:PKC][PKA|praf:pmek:plcg:P44:pakts:pjnk]
   [P38|pmek:plcg:pakts:PKA:PKC:pjnk]
  nodes:                                  11
  arcs:                                   35
    undirected arcs:                      0
    directed arcs:                        35
  average markov blanket size:            8.00
  average neighbourhood size:             6.36
  average branching factor:               3.18

  learning algorithm:                     Hill-Climbing
  score:
                                         Bayesian Information Criterion (Gaussia
  penalization coefficient:               4.459057
  tests used in the learning procedure:   505
  optimized:                              TRUE
```

Analysis of Protein Flow Cytometry using bnlearn

The two graphs are quite different

```
> compare(dag1,dag3)
$tp
[1] 9
$fp
[1] 26
$fn
[1] 17
```


Comparison of Results for Protein Flow Cytometry Data

Comparison of Results for Protein Flow Cytometry Data

- The estimated graphs are quite different
- The constrained-based methods seem to have more similarities (at least in terms of structure)
- The estimate from HC has more edges; we can change e.g. the score, but cannot directly control the sparsity

Penalized Likelihood Estimation of DAGs

- Recall that structural equation models can be used to represent causal relationships (and probability distributions) on DAGs

$$X_i = f_i(\mathrm{pa}_i, \gamma_i), \quad i = 1, \cdots, p$$

- And, for Gaussian random variables, we can write

$$X_i = \sum_{j \in \mathrm{pa}_i} \rho_{ji} X_j + \gamma_i, \quad i = 1, \cdots, p$$

Penalized Likelihood Estimation of DAGs

- Recall that structural equation models can be used to represent causal relationships (and probability distributions) on DAGs

$$X_i = f_i(\mathrm{pa}_i, \gamma_i), \quad i = 1, \cdots, p$$

- And, for Gaussian random variables, we can write

$$X_i = \sum_{j \in \mathrm{pa}_i} \rho_{ji} X_j + \gamma_i, \quad i = 1, \cdots, p$$

Penalized Likelihood Estimation of DAGs

Penalized Likelihood Estimation of DAGs

$$
\begin{aligned}
X_1 &= \gamma_1 \\
X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2 \\
X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3
\end{aligned}
$$

Penalized Likelihood Estimation of DAGs

$$
\begin{aligned}
X_1 &= \gamma_1 \\
X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2 \\
X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3
\end{aligned}
$$

Thus $X = \Lambda\gamma$ where

$$
\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & 1 & 0 \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}
$$

Penalized Likelihood Estimation of DAGs

Penalized Likelihood Estimation of DAGs

- It turns out that $\Lambda = (I - A)^{-1}$, where A is the weighted adjacency matrix of the DAG (Shojaie & Michailidis, 2010)
- if $\sigma_i^2 = 1$, we can write

$$\Omega = \Sigma^{-1} = (\Lambda\Lambda^{\mathsf{T}})^{-1} = (I - A)^{\mathsf{T}}(I - A)$$

- Thus, for Gaussian random variables, we can rewrite the problem of estimation of the inverse covariance matrix (i.e. graphical lasso), as a problem in terms of A
- In particular, if we know the ordering of the variables (which is a BIG assumption!), it can be shown that the graphical lasso problem can be solved directly as a function of A:

$$\hat{A} = \underset{A \in \mathcal{A}}{\arg\min} \left\{ \mathrm{tr}\left[(I - A)^{\mathsf{T}}(I - A)S \right] \right\}$$

Penalized Likelihood Estimation of DAGs

- In high dimensions, we can solve a penalized version of this problem, e.g. using the lasso (or adaptive lasso) penalty

$$\hat{A} = \underset{A \in \mathcal{A}}{\arg\min} \left\{ \mathrm{tr}\left[(I - A)^{\mathsf{T}}(I - A)S\right] + \lambda \sum_{i<j} |A_{ij}| \right\}$$

- As in `glasso`, λ is a tuning parameter that controls the amount of sparsity; $\lambda = \frac{2}{\sqrt{n}} Z_{\alpha/(2p^2)}$ controls a false positive probability at level α
- In fact, this can be reformulated as $p - 1$ ℓ_1 lasso regression problems problems, which can be solved very efficiently:

$$\hat{A}_{k,1:k-1} = \underset{\theta \in \mathbb{R}^{k-1}}{\arg\min} \left\{ n^{-1} \|X_{1:k-1}\theta - X_{,k}\|_2^2 + \lambda \sum_{j=1}^{k-1} |\theta_j| w_j \right\}$$

Computational Complexity

- Compared to `pcalg`, this method runs much faster: $\sim np^2$ operations vs $\sim p^q$ (q is the max degree)
- Can be easily implemented in `R` as $p - 1$ regressions using `glmnet`. A more general version is available in the `spacejam` package, which also includes estimation for non-Gaussian data

Computational Complexity

- Compared to `pcalg`, this method runs much faster: $\sim np^2$ operations vs $\sim p^q$ (q is the max degree)
- Can be easily implemented in `R` as $p-1$ regressions using `glmnet`. A more general version is available in the `spacejam` package, which also includes estimation for non-Gaussian data

Simulation Studies

- Settings:
  $p = 50, 100, 200$
  $n = 100$
  Total number of edges in the network $= n$
  100 repetitions

- Performance Criteria
  1. Matthew's Correlation Coefficient (MCC): ranges between -1 (worst fit) and 1 (best fit), similar to F_1
  2. Structural Hamming Distance (SHD): sum of false positive and false negatives
  3. True positive and false positive rates

- Tuning parameter for both PC-Algorithm and penalized likelihood method based on false positive error α

Gaussian Observations

Introduction
PC Algorithm
Other Estimation Methods
Summary
The bnlearn package
Penalized Likelihood Estimation

Gaussian Observations

Random Ordering of Variables

Regulatory Network of E-Coli

- Regulatory network of E-coli with $p = 49$ genes (7 TFs)
- Want to identify regulatory interactions among TFs and regulated genes

Regulatory Network of E-Coli

- Regulatory network of E-coli with $p = 49$ genes (7 TFs)
- Want to identify regulatory interactions among TFs and regulated genes

Summary

- Estimation of DAGs from observational data is both conceptually and computationally difficult
- Constraint-based and search-based algorithms become slow in high dimensions
- Also, may not be able to distinguish DAGs from observational data (Markov equivalence)
- Efficient penalized likelihood methods can estimate DAGs if the ordering is known
- Efficient implementations in R available for most methods
- Different methods need different tuning parameters...

Pathway & Network Analysis for Omics Data: Reconstruction of Regulatory Networks from Time-Course & Perturbation Data

Ali Shojaie

Feb 2014
Summer Institute for Statistical Genetics
University of Washington

©Ali Shojaie

MAPK/ERK Pathway

Estimation of Gene Regulator Networks

- Using steady-state gene expression data:
  - undirected association graphs: Graphical lasso (`glasso`), ARACNE, ...
  - DAGs or CPDAGs: PC-Algorithm, ...
- Using time-course gene expression data
  - Dynamic Bayesian networks
  - Granger causality
- Using perturbation screens, obtained by "perturbing" the biological system, often in the form of knockout or knockdown experiments, where in each experiment one or more genes are perturbed.
  - Model-based approaches: Nested Effect Models (NEM), methods of causal inference
  - Heuristic approaches: e.g. *Pinna et al* (2010),

Gene Regulatory Networks

Consider a simple regulatory network, with two transcription factors and one gene:

Gene Regulatory Networks

Consider a simple regulatory network, with two transcription factors and one gene:

- g_1 : Inducer
- g_2 : Inhibitor
- g_3 : Regulated Gene

Gene Regulatory Networks

The temporal expressions patterns of g_1, g_2 and g_3 may look like:

Gene Regulatory Networks

The temporal expressions patterns of g_1, g_2 and g_3 may look like:

Temporal patterns in Gene Regulatory Networks

- g_1 : Inducer
- g_2 : Inhibitor
- g_3 : Regulated Gene

Temporal patterns in Gene Regulatory Networks

- g_1 : Inducer
- g_2 : Inhibitor
- g_3 : Regulated Gene

Temporal patterns in Gene Regulatory Networks

- g_1 : Inducer
- g_2 : Inhibitor
- g_3 : Regulated Gene

Temporal patterns in Gene Regulatory Networks

- g_1 : Inducer
- g_2 : Inhibitor
- g_3 : Regulated Gene

Estimation of Gene Regulatory Networks from Time-Course Data

- The goal is Discover interactions among genes from time-course data
- This is achieved by observing the patterns of expressions over time
- A suitable framework for inferring such mechanisms is Granger causality:
  - the idea is to see if changes in expression of gene X are predictive of those in Y
  - this model is closely related to the Dynamic Bayesian Networks (DBNs)
  - can handle self-regulatory effects and feedback loops

Granger Causality

Granger Causality

Granger Causality

Granger Causality

Granger Causality

Granger Causality

We say X is Granger-causal for Y

Granger Causality

We say X is Granger-causal for Y

$$Y_t = 0.7Y_{t-1} + 0.4X_{t-1} + 0.2X_{t-2} + \varepsilon_t$$

Granger Causality

- A time series X is said to be Granger-causal for Y if past values of X provide statistically significant information about future values of Y
- This is traditionally checked using a series of F-tests, on lagged values of X
- Granger causality $\neq$ causality : Granger causality is about prediction and does not imply true causal effects
- Recent work extends this framework beyond Gaussian random variables
- We focus on extension of this idea to high dimensional settings, which we refer to as Network Granger Causality

Network Granger Causality: Illustration

p variables observed over T time points

Network Granger Causality: Illustration

p variables observed over T time points

Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top$
- Network Granger Causality Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Truncating Lasso Penalty
Lasso w Adaptive Thresholding

Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top$
- Network Granger Causality Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.
- DAG with $(d+1) \times p$ variables

Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top$
- Network Granger Causality Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.
- DAG with $(d+1) \times p$ variables
- alternatively, a vector autoregressive model of order d (VAR(d)) with p variables.

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Truncating Lasso Penalty
Lasso w Adaptive Thresholding

Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top$
- Network Granger Causality Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.
- DAG with $(d+1) \times p$ variables
- alternatively, a vector autoregressive model of order d (VAR(d)) with p variables.
- Often $d \ll T$, but not known:
  - usually, d is "guessed", and is set to $d = 1$ (especially in applications of DBN), which can result in loss of information
  - the alternative is to include all previous time points (set $d = T - 1$) but that would result in too many variables

Network Granger Causality: Definition

- $X_1, \ldots, X_p$ stochastic processes and $\mathbf{X}^t = (X_1^t, \ldots, X_p^t)^\top$
- Network Granger Causality Model:

$$\mathbf{X}^T = A^1 \mathbf{X}^{T-1} + \cdots + A^d \mathbf{X}^{T-d} + \varepsilon^T$$

- X_j^{T-t} is Granger-causal for X_i^T if $A_{i,j}^t \neq 0$.
- DAG with $(d+1) \times p$ variables
- alternatively, a vector autoregressive model of order d (VAR(d)) with p variables.
- Often $d \ll T$, but not known:
  - usually, d is "guessed", and is set to $d = 1$ (especially in applications of DBN), which can result in loss of information
  - the alternative is to include all previous time points (set $d = T - 1$) but that would result in too many variables
- Recent work has focused on simultaneous estimation of d and network.

Previous work on NGC in high dimensional settings

- The concept of Granger causality has been used in discovering gene regulatory interactions by Fujita et al (2007) and Mukhopadhyay and Chatterjee (2007)
- A number of recent work have considered penalized regression models for estimation of Granger-causal models:
  - lasso regression used in Arnold et al (2007) in a financial application
  - group lasso used in Lozano et al (2009) for grouping effects over time
  - *truncating* lasso Shojaie & Michailidis (2010) to estimate d and network simultaneously
  - lasso w *adaptive thresholding* used in Shojaie, Basu & Michailidis (2012) for improved estimation of d and network

Truncating Lasso Penalty

Shojaie & Michailidis, 2010

$\mathcal{X}^t$: data at time t

$$\arg\min_{\theta^t \in \mathbb{R}^p} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t}\theta^t\|_2^2 + \lambda \sum_{t=1}^{d} \Psi^t \sum_{j=1}^{p} |\theta_j^t| w_j^t$$

$$\Psi^1 = 1, \quad \Psi^t = M^{I\{\|A^{(t-1)}\|_0 < p^2\beta/(T-t)\}}, \ t \geq 2$$

where M is a large constant, and β is the user-specified false negative rate (FNR).

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Truncating Lasso Penalty
Lasso w Adaptive Thresholding

Truncating Lasso Penalty

Shojaie & Michailidis, 2010

$\mathcal{X}^t$: data at time t

$$\arg\min_{\theta^t \in \mathbb{R}^p} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t}\theta^t\|_2^2 + \lambda \sum_{t=1}^{d} \Psi^t \sum_{j=1}^{p} |\theta_j^t| w_j^t$$

$$\Psi^1 = 1, \quad \Psi^t = M^{I\{\|A^{(t-1)}\|_0 < p^2\beta/(T-t)\}}, \ t \geq 2$$

where M is a large constant, and β is the user-specified false negative rate (FNR).

- Can use the following value of λ that controls a version of false positive rate (FPR) at the level α:

$$\lambda(\alpha) = 2n^{-1/2} Z^*_{\frac{\alpha}{2dp^2}}$$

Truncating Lasso Penalty

Shojaie & Michailidis, 2010

$\mathcal{X}^t$: data at time t

$$\arg\min_{\theta^t \in \mathbb{R}^p} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t}\theta^t\|_2^2 + \lambda \sum_{t=1}^{d} \Psi^t \sum_{j=1}^{p} |\theta_j^t| w_j^t$$

$$\Psi^1 = 1, \quad \Psi^t = M^{I\{\|A^{(t-1)}\|_0 < p^2\beta/(T-t)\}}, \; t \geq 2$$

where M is a large constant, and β is the user-specified false negative rate (FNR).

- Can use the following value of λ that controls a version of false positive rate (FPR) at the level α:

$$\lambda(\alpha) = 2n^{-1/2} Z^*_{\frac{\alpha}{2dp^2}}$$

- This method assumes that influences decay over time

An Illustrative Example

Example I: Gene Network of HeLa Cells

9 genes, 47 time points
$d = 3$

Example II: Gene Regulatory Networks of Yeast

5 Transcription Factors, 37 genes ($p = 42$), 8 time points
$d = 2$

Non-decaying Granger-causal effects

Non-decaying Granger-causal effects

Regulatory Network of T-Cell Activation

- Data from Rangel et al (2004) on activation of T-cells
- $p = 58$ genes, $n = 44$ samples, and $T = 10$ time points
- Goal is to estimate the regulatory interactions

Adjacency Matrices of Estimated Networks

Adjacency Matrices of Estimated Networks

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Truncating Lasso Penalty
Lasso w Adaptive Thresholding

Adjacency Matrices of Estimated Networks

Estimated Regulatory Networks

Estimated Regulatory Networks

	Alasso	TAlasso	Thlasso
Alasso	(96)	–	–
TAlasso	99	(101)	–
Thlasso	35	102	(79)

Adaptively Thresholded Lasso Estimate: Main Idea

- Logic: Lasso is in general biased, and cannot achieve structure and norm consistency simultaneously

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Truncating Lasso Penalty
Lasso w Adaptive Thresholding

Adaptively Thresholded Lasso Estimate: Main Idea

- Logic: Lasso is in general biased, and cannot achieve structure and norm consistency simultaneously
- In short, the idea is to start with lasso estimates, and then remove "small" values from the adjacency matrix

Adaptively Thresholded Lasso Estimate: Main Idea

- Logic: Lasso is in general biased, and cannot achieve structure and norm consistency simultaneously
- In short, the idea is to start with lasso estimates, and then remove "small" values from the adjacency matrix
- Consider two levels of thresholding, one for each element of adjacency matrix, and the second for whole adjacency matrices at a given time point

Method Details

(i) Obtain the regular lasso estimate $\tilde{A}^t(\lambda_n)$ by solving

$$\arg\min_{\theta^t \in \mathbb{R}^p} n^{-1} \|\mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t}\theta^t\|_2^2 + \lambda \sum_{t=1}^{T-1} \sum_{j=1}^{p} |\theta_j^t| w_j^t$$

(ii) Let $\Psi^t = \exp\Big(M \mathbf{1}_{\{\|\tilde{A}^t\|_0 < p^2\beta/(T-1)\}}\Big)$, and define the thresholded estimate:

$$\hat{A}_{ij}^t = \tilde{A}_{ij}^t \mathbf{1}_{\{|\tilde{A}_{ij}^t| \geq \tau \Psi^t\}}$$

Here M is a large constant and τ is tuning parameter for thresholding.

(iii) Estimate the order of the time series by setting

$$\hat{d} = \max_t \{t : \|\hat{A}^t\|_0 \geq p^2\beta/(T-1)\}$$

Illustrative Ex I: Under Decay Assumption

Illustrative Ex II: Decay Assumption Violated

Comments

- Benefits:
  - The optimization problem is convex, and can be solved efficiently.
  - Does not require structural assumptions (no decay assumption)
- Drawbacks:
  - Requires more tuning parameters
  - Can be less efficient than truncating lasso if the decay assumption holds
- The tuning parameters can be chosen so that the method has desirable performance
- Penalized methods implemented in the `R` package `ngc`

Data from Perturbation Screens

- Steady-state data are easy to obtain, but only represent *association* among genes and hence have insufficient informational content
- Perturbation data provide direct information on causal directions, but are expensive to obtain. This becomes more complicated if perturbing a particular gene is lethal.
- Data is obtained by knockout or knockdown experiments on one or more genes at a time. The data then measures the effect of the experiments on other genes in the network.

Data from Perturbation Screens

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Nested Effect Models
The RIPE Algorithm

Data from Perturbation Screens

- In practice, due to limited sample size, the perturbation data are often *discretized*: genes are categorized as up/down regulated or active/inactive.

Data from Perturbation Screens

- In practice, due to limited sample size, the perturbation data are often *discretized*: genes are categorized as up/down regulated or active/inactive.

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Nested Effect Models
The RIPE Algorithm

Data from Perturbation Screens

- In practice, due to limited sample size, the perturbation data are often *discretized*: genes are categorized as up/down regulated or active/inactive.

- The *discretized* perturbation data
  (i) do not provide enough information to construct the structure of regulatory networks.

Data from Perturbation Screens

- In practice, due to limited sample size, the perturbation data are often *discretized*: genes are categorized as up/down regulated or active/inactive.

- The *discretized* perturbation data
  (i) do not provide enough information to construct the structure of regulatory networks.
  (ii) provide enough information to determine causal (topological) ordering(s) of nodes.

Methods for Estimation of Regulatory Networks from Perturbation Data

- Nested Effect Model (NEM): defines a probability distribution for perturbed (knockout) genes, and estimates the networks using a Bayesian framework
- Heuristic approaches: start with the network of significant effects of genes on all other genes (based on the perturbation data) and try to trim this network using features of observed networks
- Causal inference methods: in particular, using the intervention calculus (Pearl, 2000) which describes the joint probability distribution of random variables in the setting of experiments

Nested Effect Models

- Motivated by RNAi experiment settings: few knocked-out genes (called *S* genes), and a larger number of affected genes (called *E* genes)

  - Assumes that each *S* gene affect few *E* genes
  - More importantly, assumes that each *E* genes is only affected by one *S* gene
  - The network of *S* gens is arbitrary, but there is no association among *E* genes (condition on *S* genes)
- Considers the setting where *S* genes are (potentially) not observable, but *E* genes are observed
- The goal is to learn the relationship among *S* genes, based on the patterns of *E* genes, which is a difficult problem!

Nested Effect Models

- Works with discretized data: there is either an effect (1) from knocking out of S_i on E_j or not (0)
- Assumes there are positive and negative control samples
- Allows for presence of false positives and false negatives in the discretized data

Nested Effect Models

- In the simplest form (a) a chain with 3 nodes is assumed, and the model tries to learn the relationship between S genes based on the E genes that are affected by each perturbation (b)
- The matrix Φ is the influence matrix discussed before
- To simplify computation, the task of structure learning is broken down into triplets of S genes

Nested Effect Models

- Reconstruction of network of S genes is performed by first clustering the E genes into groups with similar patterns
- It is then decided whether a cluster is up-stream or down-stream the other one based on the patterns of effects (subset relationships)

Nested Effect Models

```
> library(nem)
> data("BoutrosRNAi2002")
> disc <- nem.discretize(D=BoutrosRNAiExpression,neg=1:4,pos=5:8)
> res <- nem(D=disc$dat,para=disc$para,inference="search")

nem(D, ...)
D data matrix with experiments in the columns (binary or continuous)
```

- R package `nem` implements the original NEM model, as well as some of its extensions
- The package works well for up to ~ 100 S genes (though very slow), but may not work for larger experiments

The RIPE* Algorithm

∗ Regulatory Network Inference from joint Perturbation and Expression data (Shojaie et al, 2013)

The RIPE Algorithm

RIPE integrates two sources of data, from perturbation screens and steady-state expression profiles, to give better estimates of regulatory networks

I) Use perturbation data to determine causal ordering(s) among nodes

II) For each ordering from step (I), use steady-state gene expression data to estimate the structure of the graph

III) Use model averaging to construct a consensus graph

Step I) Determining Causal Orderings

- First, obtain the influence graph P from the perturbation data (this can be done many different ways: p-value cutoff/fold-change cutoff etc)

Step I) Determining Causal Orderings

- First, obtain the influence graph P from the perturbation data (this can be done many different ways: p-value cutoff/fold-change cutoff etc)
- In absence of noise, the influence graph is obtained from the original graph by connecting node i to j if there is a directed path from i to j

Step I) Determining Causal Orderings

- First, obtain the influence graph P from the perturbation data (this can be done many different ways: p-value cutoff/fold-change cutoff etc)
- In absence of noise, the influence graph is obtained from the original graph by connecting node i to j if there is a directed path from i to j
- In practice, the influence graph will likely include false positive and false negative edges.

Step I) Determining Causal Orderings

- Create a hyper-graph of strong connected components (SCC), where each node is a collection of ≥ 1 nodes that cannot be further ordered (i.e. there is a cycle).
- Find an ordering (topological sorting) of the SCC graph (note, this is by construction a DAG) using Depth First Search algorithm (DFS).
- Find all possible orderings of each connected component (using backtracking algorithm of Knuth, or Monte Carlo DFS MC-DFS)

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Nested Effect Models
The RIPE Algorithm

Step I) Determining Causal Orderings

- Create a hyper-graph of strong connected components (SCC), where each node is a collection of ≥ 1 nodes that cannot be further ordered (i.e. there is a cycle).
- Find an ordering (topological sorting) of the SCC graph (note, this is by construction a DAG) using Depth First Search algorithm (DFS).
- Find all possible orderings of each connected component (using backtracking algorithm of Knuth, or Monte Carlo DFS MC-DFS)

Step II) Estimation of the Structure

- Given a topological ordering of nodes, the nodes of the graph can be rearranged to form a DAG
- For each ordering, estimate (the structure of) one DAG using the penalized likelihood method of the previous lecture, (by solving $p-1$ lasso regression problems):

$$\hat{A}_{k,1:k-1} = \underset{\theta \in \mathbb{R}^{k-1}}{\arg\min} \left\{ n^{-1} \| X_{1:k-1}\theta - X_{,k} \|_2^2 + \lambda \sum_{j=1}^{k-1} |\theta_j| w_j \right\}$$

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Nested Effect Models
The RIPE Algorithm

Step II) Estimation of the Structure

- Given a topological ordering of nodes, the nodes of the graph can be rearranged to form a DAG
- For each ordering, estimate (the structure of) one DAG using the penalized likelihood method of the previous lecture, (by solving $p-1$ lasso regression problems):

$$\hat{A}_{k,1:k-1} = \underset{\theta \in \mathbb{R}^{k-1}}{\arg\min} \left\{ n^{-1} \| X_{1:k-1}\theta - X_{,k} \|_2^2 + \lambda \sum_{j=1}^{k-1} |\theta_j| w_j \right\}$$

Step III) Building a Consensus Graph

- For each ordering, the estimated graph is a DAG
- However, the true graph may include cycles. Also, results from one ordering may be inaccurate (noise...)
  - L_q: lower qth quantile of the (penalized) negative log-likelihoods
  - $Q = \{o \in \mathcal{O} : \ell(o) \leq L_q\}$ set of orderings for these likelihoods

$$\hat{A}^c_{i,j} = \frac{1}{|Q|} \sum_{k \in Q} \mathbf{1}_{\{|A^k_{i,j}|>0\}}$$

  - $\hat{E} = \{(i,j) : \hat{A}^c_{i,j} \geq \tau\}$

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Nested Effect Models
The RIPE Algorithm

Step III) Building a Consensus Graph

- For each ordering, the estimated graph is a DAG
- However, the true graph may include cycles. Also, results from one ordering may be inaccurate (noise...)
  - L_q: lower qth quantile of the (penalized) negative log-likelihoods
  - $Q = \{o \in \mathcal{O} : \ell(o) \leq L_q\}$ set of orderings for these likelihoods

$$\hat{A}^c_{i,j} = \frac{1}{|Q|} \sum_{k \in Q} \mathbf{1}_{\{|A^k_{i,j}|>0\}}$$

  - $\hat{E} = \{(i,j) : \hat{A}^c_{i,j} \geq \tau\}$

Simulate Network: DAG of size $p = 20$

Data Generation

Perturbation data: Adjacency matrices of true and noisy influence graphs

Steady-state expression data: generated $n = 50$ Gaussian observations according to the true DAG.

Comparison of F_1 measures

How Many Orderings?

For P_3, there are a total of 3962 orderings using the backtracking algorithm.

High Dimensional Cyclic Graphs ($p = 1000$)

Effect of FP and FN errors

A More Complicated Example: DREAM-4 Challenge

- The DREAM project (Dialogue for Reverse Engineering Assessments and Methods) is an attempt to construct realistic regulatory networks
- DREAM-4 challenge had multiple competitions, including reverse engineering 5 networks of size 100 selected from true regulatory components of yeast and E-coli.
- The perturbation data is simulated based on the true network (using coupled ODE)
- Two types of perturbation data are available: knockout and knockdown experiments
- The algorithm of Pinna et al (PINNA) was the winner of the high dimensional reconstruction challenge (on networks of size 100)

DREAM Network 1 (Simplest)

DREAM Network 5 (Most Difficult!)

Comparison of F_1 Measures

Example of estimated modules

Largest cyclic component in DREAM1 network

When the perturbation data includes cycles, the consensus graph will be cyclic.

Network of Yeast Transcription Factors

- 269-node corresponding to known yeast TF's ($p = 269$)
- Perturbation data: knockout experiments from Hu et al (2007, Nat Genetics)
- Steady-state expression data: $n = 200$ day-to-day variation samples of yeast (publicly available), not really iid!
- Used 10,000 orderings
- To evaluate: use available data on yeast regulatory network, which is (most likely) incomplete. Therefore, "false positives" may be true edges

Network of Yeast Transcription Factors

- Significance of true positives (TP), in comparison to the BioGrid network
- Histograms show number of TP's in random networks of equal sizes

Extension: $k \ll p$

- In many biological experiments, perturbation screens are only run on a subset of genes (k out of p)
- If perturbation is available on TFs, the RIPE algorithm can be modified to estimate the network

Estimation from Time Course Data
Network Estimation from Perturbation Data
Summary
Introduction
Nested Effect Models
The RIPE Algorithm

Extension: $k \ll p$

- In many biological experiments, perturbation screens are only run on a subset of genes (k out of p)
- If perturbation is available on TFs, the RIPE algorithm can be modified to estimate the network

Summary

- Estimation of regulatory networks is difficult! In addition to need for causal inference, the presence of feedback loops, and the small sample size of biological experiments hinder estimation of directed regulatory networks
- Available data differ in informational content and available sample size (and hence noise level)
- Time-course and perturbation data offer greater potential for learning the structure of DAGs; however, they also introduce new challenges.
- Computational complexity is a bottleneck of many proposed methods, many existing methods are approximations of the biology, or make strong assumptions
- This is an active area of research, with many methods being developed and implemented...

Pathway & Network Analysis for Omics Data: Network-Based Pathway Enrichment Analysis

Ali Shojaie

Feb 2014
Summer Institute for Statistical Genetics
University of Washington

©Ali Shojaie

Yeast GAL Pathway

Ideker et al, 2001

Issues of Interest

- Incorporate the network information
- Consider changes in the gene (protein, metabolite) expressions
- Consider changes in the network structure
- Test the "effect" of pre-specified subnetwork/pathway, sharing common biological function, chromosomal location etc
- A general framework for inference in complex experiments

Recap: Gene Set Enrichment Analysis

Subramanian et al. (2005) proposed gene set enrichment analysis (GSEA); *Efron & Tibshirani* (2007) formalized the GSEA approach, and proposed a more efficient test statistic

- Test the significance of *a priori* defined gene sets
- Preserve the correlation among genes in the gene set
- Based on a competitive null hypothesis, where activity of each pathway is compared with other pathways, often using a permutation test
- Competitive tests of enrichment assume that a small number of genes have differential activity, and are very sensitive to the choice of gene sets, they also problem with
- Self-contained tests address these issues, but may be less efficient or sensitive to model assumptions (*Goemen & Buhlmann* (2007), *Ackermann & Strimmer* (2009))

Signaling Pathway Impact Analysis (SPIA)

- Combines classical overrepresentation analysis (ORA) with measure of perturbation of a given pathway under a given condition
- A bootstrap procedure is used to assess the significance of the observed pathway perturbation (difficult to extend to comparison of > 2 conditions)
- Currently not applicable to all pathways (more later)
- Models each pathway separately (ignores connections among pathways)
- Implemented in the Bioconductor package `SPIA`

The SPIA Methodology

The SPIA Methodology

SPIA combines two types of evidence

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis

The SPIA Methodology

SPIA combines two types of evidence

(i) the overrepresentation of DE genes in a given pathway

The SPIA Methodology

SPIA combines two types of evidence

(i) the overrepresentation of DE genes in a given pathway

- measured by the p-value for the given number of DE genes $P_{NDE} = P(X \geq N_{DE} \mid H_0)$

The SPIA Methodology

SPIA combines two types of evidence

The SPIA Methodology

SPIA combines two types of evidence

(ii) the abnormal perturbation of the pathway

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis

The SPIA Methodology

SPIA combines two types of evidence

(ii) the abnormal perturbation of the pathway

- the perturbation for each gene in the pathway is defined as $PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{p} \beta_{ij} \frac{PF(g_j)}{N_{DS}(g_j)}$

The SPIA Methodology

SPIA combines two types of evidence

(ii) the abnormal perturbation of the pathway

- the perturbation for each gene in the pathway is defined as $PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{p} \beta_{ij} \frac{PF(g_j)}{N_{DS}(g_j)}$
  - $PF(g_i)$ is the perturbation factor of gene i (not known)
  - β_{ij} is the magnitude of effect of gene j on gene i; currently, $beta_{ij} = 1$ if $j \rightarrow i$
  - $\Delta E(g_i)$ is the fold change in expression of gene i
  - $N_{DS}(g_j)$ is the number of downstream genes from gene j

The SPIA Methodology

The SPIA Methodology

- The accumulated activity of each gene can then be calculated as $ACC(g_i) = B \cdot (I - B)^{-1} \Delta E$

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis

The SPIA Methodology

- The accumulated activity of each gene can then be calculated as $ACC(g_i) = B \cdot (I - B)^{-1} \Delta E$
  - B is the normalized matrix of β's: $B_{ij} = \beta_{ij} / N_{DS}(g_j)$
  - ΔE is the vector of fold changes
  - Requires B to be invertible; would not work otherwise

The SPIA Methodology

- The accumulated activity of each gene can then be calculated as $ACC(g_i) = B \cdot (I - B)^{-1} \Delta E$
  - B is the normalized matrix of β's: $B_{ij} = \beta_{ij} / N_{DS}(g_j)$
  - ΔE is the vector of fold changes
  - Requires B to be invertible; would not work otherwise
- The total accumulated perturbation of the pathway is then given by $t_A = \sum_i ACC(g_i)$

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis

The SPIA Methodology

- The accumulated activity of each gene can then be calculated as $ACC(g_i) = B \cdot (I - B)^{-1} \Delta E$
  - B is the normalized matrix of β's: $B_{ij} = \beta_{ij} / N_{DS}(g_j)$
  - ΔE is the vector of fold changes
  - Requires B to be invertible; would not work otherwise
- The total accumulated perturbation of the pathway is then given by $t_A = \sum_i ACC(g_i)$
- The p-value for pathway perturbation is given by $P_{PERT} = P(T_A \geq t_A \mid H_0)$, which is calculated using a bootstrap approach

The SPIA Methodology

Gene	ΔE	PF	Acc
A	0	0	0
B	2	2	0
C	0	1	1
D	0	1	1
E	0	0	0
F	4	4	0

Gene	ΔE	PF	Acc
A	1.5	1.5	0
B	2	2.5	0.5
C	0	1.25	1.25
D	0	1.25	1.25
E	0	0.5	0.5
F	0	0.5	0.5

The SPIA Methodology

The SPIA Methodology

SPIA combines two types of evidence

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis

The SPIA Methodology

SPIA combines two types of evidence

- The final p-value for each pathway is calculated based on the p-values from parts (i) and (ii):

The SPIA Methodology

SPIA combines two types of evidence

- The final p-value for each pathway is calculated based on the p-values from parts (i) and (ii):
  - $P_G(i) = c_i - c_i \ln(c_i)$
  - $c_i = P_{NDE}(i) P_{PERT}(i)$

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis

The SPIA Methodology

SPIA combines two types of evidence

- The final p-value for each pathway is calculated based on the p-values from parts (i) and (ii):
  - $P_G(i) = c_i - c_i \ln(c_i)$
  - $c_i = P_{NDE}(i) P_{PERT}(i)$

An Example in R: Data on Colorectal Cancer

```
data(colorectalcancer)

#pathway analysis using SPIA
#use nB=2000 or higher for more accurate results
#uses older version of KEGG signaling pathways graphs
res <- spia(de=DE_Colorectal, all=ALL_Colorectal, organism="hsa", beta=NULL,
    nB=2000, plots=FALSE, verbose=TRUE, combine="fisher")

#now combine pNDE and pPERT using the normal inversion method without
#running spia function again
res$pG=combfunc(res$pNDE,res$pPERT,combine="norminv")
res$pGFdr=p.adjust(res$pG,"fdr")
res$pGFWER=p.adjust(res$pG,"bonferroni")
plotP(res,threshold=0.05)

#highlight the colorectal cancer pathway in green
points(I(-log(pPERT))~I(-log(pNDE)),data=res[res$ID=="05210",],col="green",
    pch=19,cex=1.5)
```


The SPIA Methodology

Network-Based Gene Set Analysis (NetGSA)

- Combines the ideas of gene set analysis methods, and network-based single gene analysis
- Generalizes SPIA, to allow for more complex experiments & incorporate interactions among pathways
- Assesses the overall behavior of arbitrary subnetworks (pathways): changes in gene expression & network structure
- Uses latent variables to model the interaction between genes defined by the network
- Uses mixed linear models for inference in complex data
- Computationally challenging for large networks (e.g. not applicable to whole genome sequencing data) unless, pathways separated (similar to SPIA)
- No stable R-package currently available (current version at `http://www.biostat.washington.edu/~ashojaie/research.html`, a more stable version under development)

Problem Setup

- Gene (protein/metabolite) expression data for K experimental conditions and J_k time points
- Network information (partially) available in the form of a directed weighted graph $G = (V, E)$, with vertex set V corresponding to the genes/proteins/metabolites and edge set E capturing their associations
- Edges in the network can be directed $j \to k$ or undirected $j \leftrightarrow k$
- Edges defines the effect of nodes on their immediate neighbors; the weight associated with each edge corresponds to the value of partial correlation
- Represent the network by its adjacency matrix A: $A_{jk} \neq 0$ iff $k \to j$ & for undirected edges, $A_{jk} = A_{kj}$
- Pathways defined *a priori* based on common biological functions, etc

The Latent Variable Model: Main Idea

$$
\begin{aligned}
X_1 &= \gamma_1 \\
X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2 \\
X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3
\end{aligned}
$$

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis
Modeling Network Interactions
The NetGSA model
Uncertainty in Network Information

The Latent Variable Model: Main Idea

$$
\begin{aligned}
X_1 &= \gamma_1 \\
X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2 \\
X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3
\end{aligned}
$$

Thus $X = \Lambda\gamma$ where

$$
\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & 1 & 0 \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}
$$

The Latent Variable Model

- Let Y be the ith sample in the expression data
- Let $Y = X + \varepsilon$, with X the signal and $\varepsilon \sim N_p(0, \sigma_\varepsilon^2 I_p)$ the noise
- The influence matrix Λ measures the propagated effect of genes on each other through the network, and can be calculated based on the adjacency matrix A
- Using $X = \Lambda\gamma$, we get

$$Y = \Lambda\gamma + \varepsilon, \quad \Rightarrow \quad Y \sim N_p(\Lambda\mu, \sigma_\gamma^2 \Lambda\Lambda' + \sigma_\varepsilon^2 I_p)$$

where $\gamma \sim N_p(\mu, \sigma_\gamma^2 I_p)$ are latent variables

Mixed Linear Model Representation

Rearranging the expression matrix into np-vector $\mathbf{Y}$, we can write

$$\mathbf{Y} = \mathbf{\Psi}\beta + \mathbf{\Pi}\gamma + \varepsilon$$

where β and γ are fixed and random effect parameters and

$$\varepsilon \sim N_{np}(\mathbf{0}, R(\theta_\varepsilon)), \qquad \gamma \sim N_{np}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I_{np}})$$

- Temporal Correlation incorporated through R

Mixed Linear Model Representation

Rearranging the expression matrix into np-vector $\mathbf{Y}$, we can write

$$\mathbf{Y} = \mathbf{\Psi}\boldsymbol{\beta} + \mathbf{\Pi}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are fixed and random effect parameters and

$$\boldsymbol{\varepsilon} \sim N_{np}(\mathbf{0}, R(\theta_\varepsilon)), \qquad \boldsymbol{\gamma} \sim N_{np}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I_{np}})$$

- Temporal Correlation incorporated through R

In general, the design matrices, $\mathbf{\Psi}$ and $\mathbf{\Pi}$ depend on the experimental settings (similar to ANOVA), and are functions of Λ

Estimation of MLM Parameters

MLE for β:

$$\hat{\beta} = \left(\Psi' \hat{W}^{-1} \Psi\right)^{-1} \Psi' \hat{W}^{-1} \mathbf{Y}$$

where $W = \sigma_\gamma^2 \Pi\Pi' + R$.

$\hat{\beta}$ depends on estimates of σ_γ^2 and θ_ε^2 (estimated using restricted maximum likelihood (REML)).

Inference using MLM

- Let ℓ be a contrast vector (a linear combination of fixed effects), and consider the test:

$$H_0 : \ell\beta = 0 \quad vs. \quad H_1 : \ell\beta \neq 0$$

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis
Modeling Network Interactions
The NetGSA model
Uncertainty in Network Information

Inference using MLM

- Let ℓ be a contrast vector (a linear combination of fixed effects), and consider the test:

$$H_0 : \ell\beta = 0 \quad vs. \quad H_1 : \ell\beta \neq 0$$

- Use t-test to test the significance of each hypothesis separately

$$T = \frac{\ell\hat{\beta}}{\sqrt{\ell\hat{C}\ell'}}$$

where $C = (\Psi' W^{-1}\Psi)^{-1}$

- Under the null hypothesis, T is approximately t-distributed with degrees of freedom that needs to be estimated

"Optimal" Choice of Contrast Vector

- One intuitive choice is to use the indicator vector for the members of pathway **b**, but this only reflects changes in the mean vector
- Need to *de-couple the effect of each subnetwork* from other nodes

Introduction
Pathway Impact Analysis
Network-Based Gene Set Analysis
Modeling Network Interactions
The NetGSA model
Uncertainty in Network Information

"Optimal" Choice of Contrast Vector

- One intuitive choice is to use the indicator vector for the members of pathway **b**, but this only reflects changes in the mean vector
- Need to *de-couple the effect of each subnetwork* from other nodes

"Optimal" Choice of Contrast Vector

- One intuitive choice is to use the indicator vector for the members of pathway **b**, but this only reflects changes in the mean vector
- Need to *de-couple the effect of each subnetwork* from other nodes

- Can be shown that $(\mathbf{b}\Lambda \cdot \mathbf{b})\gamma$ is not affected by nodes outside **b**, but includes the effects of nodes in **b** on each other
- In the case-control case, the optimal contrast vector is:

$$\ell^* = \left(-\mathbf{b} \cdot \mathbf{b}\Lambda^C, \mathbf{b} \cdot \mathbf{b}\Lambda^T\right)$$

"Optimal" Choice of Contrast Vector

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & 1 & 0 \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}$$

Consider the set, $\mathbf{b} = (0, 1, 1)$; then

$$(\mathbf{b}\Lambda) = (\rho_{12} + \rho_{12}\rho_{23}, 1 + \rho_{23}, 1)$$

On the other hand,

$$(\mathbf{b}\Lambda \cdot \mathbf{b}) = (0, 1 + \rho_{23}, 1)$$

Comparison in Simulated Data

Subnetwork	Mean	Network Influence
1	$\mu_1 = \mu_2 = 1$	$\rho_1 = \rho_2 = 0.2$
2	$\mu_1 = 1, \mu_2 = 2$	$\rho_1 = \rho_2 = 0.2$
3	$\mu_1 = \mu_2 = 1$	$\rho_1 = 0.2, \rho_2 = 0.7$
4	$\mu_1 = 1, \mu_2 = 2$	$\rho_1 = 0.2, \rho_2 = 0.7$

Yeast Galactose Utilization Pathway

Ideker et al (2001) data on yeast Galactose Utilization Pathway

- Gene expression data for 2 experimental conditions: (gal+) and (gal–)
- Gene-gene and protein-gene interactions as well as association weights found from previous studies
- Q: which pathways respond to the change in growth medium?

Analysis of Yeast GAL Data

- Data:
  - gene expression data for 343 genes
  - 419 interactions found from previous studies and integration with protein expression (association among genes also available)
- Results:
  - GSEA finds *Galactose Utilization Pathway* significant
  - NetGSA finds several other pathways with biologically meaningful functions related to survival of yeast cells in gal–

Environmental Stress Response in Yeast

Gene expression data on Yeast Environmental Stress Response (ESR) (*Gasch et al.*, 2000)

- 3 combinations of experimental factor, heat shock and osmotic changes (sorbitol), over 3 time points
- Temporal correlation
- Network correlation
- Q: Which pathways indicate response to environmental stress
  - in different experimental conditions
  - over time

Yeast ESR Data

Gasch et al (2000)

- Gene Expression Data

Experiment	Obs. Time (after 33C)
Mild heat shock (*29C to 33C*), no sorbitol	5, 15, 30 min
Mild Heat Shock, 1M sorbitol at 29C & 33C	5, 15, 30 min
Mild Heat Shock, 1M sorbitol at 29C	5, 15, 30 min

- Network Data
  - Use YeastNet (*Lee et al.*, 2007) for gene-gene interactions (102,000 interactions among 5,900 yeast genes)
  - Use independent experiments of *Gasch et al.* to estimate weights
  - Pathways are defined using GO functions

Model and Results

- Model: Let j and k be indices for time and levels of sorbitol

$$\mathbb{E}Y_{11} = \Lambda\mu, \qquad \mathbb{E}Y_{jk} = \Lambda(\mu + \alpha_j + \delta_k) \quad j, k = 2, 3$$

- Temporal correlation is modeled directly via R (as $AR(1)$ process)
- Results:
  - $\sim$ 3000 genes,
  - 47 pathways showed significant changes of expression
  - 24 pathways showed changes over time
  - 29 pathways showed changes in response to different sorbitol levels
  - 12 pathways showed both types of changes
  - Significant pathways overlap with the gene functions recognized by *Gasch et al.*

Yeast ESR Network

Significant subnetworks

Expression Profiles

Average Standardized Expression Levels of Pathways

- Induced and Suppressed Pathways
- Can observe the transient patterns of expressions as predicted by *Gasch et al.*

Effect of Noise In Network Information

- Let $\tilde{A}$ be observed network information, and A be the truth.
- It can be shown that, if $\|\tilde{A} - A\|$ is small then, NetGSA still works (is *asymptotically most powerful unbiased test*)

Effect of Noise In Network Information

- Let $\tilde{A}$ be observed network information, and A be the truth.
- It can be shown that, if $\|\tilde{A} - A\|$ is small then, NetGSA still works (is *asymptotically most powerful unbiased test*)

Metabolic Profiling in Bladder Cancer

Targeted metabolic profiling of bladder cancer (BCa) (*Putluri et al.*, 2012)

- 58 bladder cancer and adjacent benign samples
- Pathways information obtained from KEGG

- Varying number of identified metabolites per pathway (3-15)
- Q: Which pathways show differential activity in BCa?

Metabolic Profiling in BCa

- 63 metabolites identified, mapped to 70 pathways
- 27 pathways with at least 3 members

Metabolic Profiling in BCa

- Small pathway sizes & significant overlap among pathways

- Existing methods may not work well...

Metabolic Interaction Network

Significant Pathways

- GSEA does not identify any pathway as differential
- GSA identifies Fatty Acid Biosynthesis as differential
- NetGSA identifies another 7 pathways corresponding to role of Amino Acid Metabolism in BCa, also observed by *Putluri et al* (2012)