Integrative Analysis of Heterogeneous Genomics Data

Xianghong Jasmine Zhou
Molecular and Computational Biology
University of Southern California
The rapid accumulation of genomics data

- **Gene expression data**
  - NCBI GEO and EBI ArrayExpress
- **High-throughput sequencing data**
  - NCBI Sequence Reader Achieve
- **Molecular Interaction Data**
  - BioGrid Database
- **Genome-wide Association Data**
  - NCBI dbGap database

Knowledge discovery by integrating the heterogeneous data (the same or different data types) is a major challenge
In this talk...

• How to integratively analyze multiple genomic datasets of the same type (horizontal data integration)?

• How to analyze multiple genomics datasets of different types but measured on the same set of samples (vertical data integration)?
Horizontal Data Integration

A graph-based approach for the integrative analysis of gene expression data
Our testing system: the large accumulation of expression data in public repositories

• NCBI Gene Expression Omnibus
  >700,000 experiments

• EBI Array Express
  >300,000 experiments
Challenges for Cross-dataset Comparison

• Different technologies
  • Microarray
  • RNA-seq

• Different platforms
  • Microarray: Affy, Illumina, etc
  • RNA-seq: Solexa, 454, SOLiD, etc

• Different laboratory protocols: for RNA-seq data, biases such as the preferential selection of sequences could be introduced during the adapter ligation step, PCR amplification, and sequencing.
Graph-based Approach for the Integrative Gene Expression Analysis

Datasets

Coexpression Networks

Recurrent Patterns

Annotation

Gene Ontology

Functional Annotation
Frequent Subgraph Mining Problem is hard!

Problem formulation: Given $n$ graphs, identify subgraphs which occur in at least $m$ graphs ($m \leq n$)

Our graphs are massive!
The traditional pattern growth approach (expand frequent subgraph of $k$ edges to $k+1$ edges) would not work, since the time and memory requirements increase exponentially with increasing size of patterns and increasing number of networks.
Novel Algorithms to identify diverse frequent network patterns

• CoDense (ISMB 2005)
  – identify frequent dense subgraphs across many massive graphs

• Network Biclustering (ISMB 2007)
  – identify frequent subgraphs across many massive graphs

• Network Modules (ISMB 2007)
  – identify frequent dense vertex sets across many massive graphs

• MOSA (RECOMB 2008)
  – identify phenotype-specific frequent dense vertex sets
However, all these algorithms focused primarily on the unweighted graphs.

- Weighted networks are often perceived as harder to analyze.

- However, weighted networks are obviously more informative than their unweighted counterparts.
Tensor Computation for Integrative Analysis of Multiple Weighted Coexpression Networks
Mathematical Representation of Single Network

Network  →  Matrix
Mathematical Representation of Single Network

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.5</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.8</td>
<td>0.3</td>
<td>0</td>
</tr>
</tbody>
</table>
Mathematical Representation of Multiple Network

Multiple Networks

Tensor
(Multi-dimensional Array)
Tensor Model

**Microarray datasets**  **Co-expression networks**  **Collection of networks**

A tensor model representing co-expression networks from microarray datasets. Each dataset corresponds to a set of networks, and the collection of networks is represented as a 3rd-order tensor $A = (a_{ijk})^{8 \times 8 \times 5}$, with dimensions gene $\times$ gene $\times$ network.
Representation of a Frequent Heavy Subgraph (FHS)

- **$x$: gene membership vector** of FHS
  
  $$x_i = \begin{cases} 
  1, & \text{if the } i\text{-th gene is in FHS} \\
  0, & \text{otherwise} 
  \end{cases}$$

- **$y$: network membership vector** of FHS
  
  $$y_j = \begin{cases} 
  1, & \text{if the } j\text{-th network is in FHS} \\
  0, & \text{otherwise} 
  \end{cases}$$

- The **sum of edges’ weights** of FHS (Tensor $A$)
  
  $$E_A(x,y) = \sum_i \sum_j \sum_k a_{ijk} \; x_i \; x_j \; y_k$$
Identify Frequent Heavy Subgraphs (FHSs)

• Problem Formulation
  – **Intuition:** A FHS with $k_1$ genes and $k_2$ networks must have the largest sum of edges’ weights, among all same-size FHS.
  – **Maximize**
    
    $$E_A(x,y) = \sum_i \sum_j \sum_k a_{ijk} x_i x_j y_k$$
  – **Subject to**
    
    $$x_1 + x_2 + \cdots + x_n = k_1$$
    $$y_1 + y_2 + \cdots + y_m = k_2$$
    $$x_i \text{ and } y_j \text{ are 0 or 1}$$

Non-Linear Integer Programming !!!
Identify Frequent Heavy Subgraphs (FHSs)

- Continuous Relaxation of Integer Programming
  - Relax binary vectors $x, y$ to continuous values

\[ \begin{align*}
  \| x \|_p &= (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p} = 1 \\
  \| y \|_q &= (|y_1|^q + |y_2|^q + \cdots + |y_m|^q)^{1/q} = 1
\end{align*} \]

**Constraint 1:**
\[ x_1 + x_2 + \cdots + x_n = k_1 \]

**Constraint 2:**
\[ y_1 + y_2 + \cdots + y_m = k_2 \]

**Constraint 3:**
$x_i$ and $y_j$ are 0 or 1

**Solution:**
Those $x_i$ and $y_j$ with 1 form FHS

Those large $x_i$ and $y_j$ form FHS
Choice of Vector Norms

The L0 norm leads to sparse solutions where only a few components of the membership vectors are significantly different from zero. The norm generally gives a "smooth" solution where the elements of the optimized vector are approximately equal.

Norm for the gene membership vector: a subset of values should be significantly non-zero and close to each other, while the rest are close to zero.

$\ell_{0,\infty}(x) = \alpha \|x\|_0 + (1 - \alpha) \|x\|_{\infty}$

Norm for network membership vector: as many network membership values as possible are non-zero and close to each other.

$L_{\infty}$
Parameters Selection

- **Parameters of Sparsity Norm: \( p \) and \( \alpha \)**
  - Perform simulation study: generated a number of uniformly random weighted networks collections and randomly place a RHS with different sizes, occurrences and heaviness to a subset of networks. Then we applied our tensor algorithm with different combinations of \( p \) and \( \alpha \) to identify this FHS.
  - \( p = 0.8 \) and \( \alpha = 0.2 \)

- **Parameter of Smooth Norm: \( q \)**
  - As long as \( q \geq 2 \), it can arrive at a very good approximation to
  - We simply choose \( q = 10 \)
Optimization Method

- Non-convex constraint: 
  \[ L_{0,\infty}(x) = \alpha \|x\|_0 + (1 - \alpha) \|x\|_\infty \]
- Convex constraint: 
  \[ L_\infty \]
- But non-linear non-convex optimization is hard!

- Our method: Multi-Stage Convex Relaxation
  - **Concave Duality**: relax non-convex function to convex
  - **Intuition**: construct a sequence of convex relaxation problems that give better and better approximations to the original non-convex problem.
Non-Uniform Sampling for Fast Computation

• Analysis of large-scale networks with many edges need to spend long computation time
  – Edge sampling can provide an efficient approximation to many graph problems
• Non-uniform sampling method
  – Two rules:
    • Rule 1: *Always* samples edges with weights \( \geq \tilde{a} \)
    • Rule 2: Randomly sample edges of weights \(< \tilde{a}\) with probability \( p_{ijk} \)

\[
p_{ijk} = \begin{cases} 
1 & \text{if } a_{ijk} \geq \tilde{a}, \quad \text{(rule 1)} \\
p \left( \frac{a_{ijk}}{\tilde{a}} \right)^b & \text{if } a_{ijk} < \tilde{a}, \quad \text{(rule 2)} 
\end{cases}
\]
Identify Many Frequent Heavy Subgraphs

- Identify a FHS (a subtensor) in the tensor $A$
- Mask this subtensor/FHS with zeros
- Identify the next FHS in the modified tensor $A$
Experimental Results

• Data
  – 129 gene expression datasets
  – Covering 8,504 genes
  – Sampling edges by
    \[ p_{ijk} = \begin{cases} 
    1 & \text{if } a_{ijk} \geq 0.6 \\
    0.1 \left( \frac{a_{ijk}}{0.6} \right)^4 & \text{if } a_{ijk} < 0.6 
    \end{cases} \]
  – Discovered 4,327 FHS with ≥ 5 genes heavily (heaviness≥0.4) occur in ≥ 5 networks

• Results Validation
  – Identify functional, regulatory, and phenotype-specific modules
  – Identify protein complexes
  – Reconstruct transcriptional regulatory networks
**Functional Modules Identified by FHS (1)**

- Gene Ontology (GO) database: Biological process, Cellular component, Molecular function

  - The member genes of a FHS are enriched in a GO term (associated genes<500) with a hypergeometric p-value < 0.001

  - 60% of FHSs are functionally homogenous in terms of GO biological process terms, while 9.3% in random

  - Cover a wide range of biological processes such as translational elongation, mitosis, cell cycle, RNA splicing, ribosome biogenesis, histone modification, chromosome localization, spindle checkpoint, posttranscriptional regulation, protein folding, etc.
Functional Modules Identified by FHS (2)

- **Observation:** higher the heaviness and the recurrence, the more functionally homogenous are the FHS
Protein Complexes Identified by FHS

  - Assess to what degree the RHS modules represent known protein complexes
    - 43.8% of FHSs are enriched of genes from the same protein complexes ($p < 0.001$), while 1% in random
Regulatory Modules Identified by FHS (1)

• Genome-wide ENCODE data: 191 ChIP-seq profiles generated by the Encyclopedia of DNA Elements (ENCODE) consortium.
  – Genome-wide binding of 40 transcription factors (TF), 9 histone modification marks, and 3 other markers (DNase, FAIRE, and DNA methylation) on 25 different cell lines

• ChIP-chip data: We collected 109 ChIP-chip experiments from published papers. There are total 60 TFs.
Regulatory Modules Identified by FHS (2)

- **Genome-wide ENCODE data**
  - 77.1% of FHSs are transcriptional homogenous, while 5.3% in random

- **ChIP-chip data**
  - 32.3% of FHSs are enriched of at least one of the TFs, while 3.7% in random
Regulatory Modules Identified by FHS: an example

- An 8-gene module that is enriched in the binding of multiple regulatory factors: Pol2 (p-value = 3.4E-6), H3K36me3 (p-value = 5.6E-5), E2F4 (p-value = 6E-4), and cFos (p-value = 3.4E-4). The module is active in 8 datasets, and member genes are involved in DNA replication initiation (p-value = 7.0E-5).
Comparisons with unweighted graph based approach

Figure 5. Comparison between weighted and unweighted network analysis: The weighted networks were transformed to unweighted networks by dichotomizing edges with an expression correlation cutoff of 0.6. The proposed tensor method was then applied to both weighted and unweighted networks. We compared rates of functional homogeneity detected in the top $K = 200, 400, \cdots, 2000$ modules, ranked by (a) recurrences or (b) average heaviness in their datasets of occurrence. Weighted graph analysis consistently outperforms unweighted graph analysis.
Discovering of Phenotype-specific modules

For each FHS, we evaluated phenotype specificity by computing the hypergeometric enrichment of specific phenotype concepts present in those datasets where the FHS occurs. If p-value < 0.001, we consider the FHS module to be phenotype-specific.

15.0% of FHSs show phenotype-specific activation patterns, compared to 0.6% of randomly generated FHSs.
High-order cooperativity in protein complex networks

Cell Cycle
Respiratory Chain
Translation
Post-transcription process
Reconstruct Transcriptional Regulatory Network

Frequently occurring tight clusters
Frequently occurring tight clusters

Transcription Factors
Co-occurrence of tight clusters

Co-expression network constructed with the dataset 1
Co-occurrence of tight clusters

Co-expression network constructed with the dataset 2
Co-occurrence of tight clusters

Relevance network constructed with the dataset 3
Co-occurrence of tight clusters

Relevance network constructed with the dataset 4
Co-occurrence of tight clusters

Relevance network constructed with the dataset 5
Three types of transcription cascades

Type I

TF1 → TF2
TF1 → TF3

TF2 → Module 1
TF3 → Module 2

Type II

TF1 → TF2

TF2 → Module 1

Type III

TF1 → TF2

TF2 → Module 1

Transcription regulation

Protein interaction
Reconstruct Transcriptional Regulatory Network

- **DNA metabolic process**
  - MCM complex
  - NF-kappaB
  - STAT1
  - TAF1
  - ELF1

- **Cell cycle**
  - M1304
  - M3092
  - M2099
  - M1617
  - M1585

- **Translational elongation**
  - Ribosome cytoplasmic complex
  - M564
  - M254
  - M1838
  - M121
  - M823
  - M2091
  - M10
  - M1464

- **Electron transport chain**
  - M1617
  - M1585
  - Chromosomal passenger complex
  - PA700-20S-PA28 complex
  - SMN complex
  - Respiratory chain complex

- eIF3 complex
  - M3098
  - M20

- F1F0-ATP synthase mitochondrial
  - M3715
  - M3295
  - M898
  - M3676
  - M768
Systematic reconstruction of splicing regulatory modules by integrating many RNA-Seq datasets
RNA-seq data processing

1. 19 human RNA-seq datasets with at least six experiments from the NCBI SRA database
2. Estimate the expression of exons and genes.
   
   Gene expression = \text{sum(expression of transcripts)}
   Exon expression = \text{sum(expression of transcripts containing the exon)}
3. Calculate the exon inclusion rate = exon expression/gene expression
5. Run our tensor program to identify frequent splicing modules from many co-splicing networks.
Tensor Model

Dataset 1

Dataset 2

Dataset 3

Dataset 4

RNA-seq datasets
Co-splicing networks

Collection of networks

3rd-order tensor with dimensions exon × exon × network
Frequent co-splicing clusters are more likely to represent functional modules.
Splicing regulatory analysis

1. Collect 33 binding motifs of RNA-binding proteins from RBPDB.
2. Retrieve the internal exon and 100bp flanking intron region of co-splicing clusters, perform motif enrichment analysis to identify potential splicing factors.
3. Top 4 enriched splicing factors are RBM4, YTHDC1, YBX1, and SFRS1.
Co-binding on the same splicing modules suggests combinatorial regulation of splicing factors
Co-splicing clusters reveal novel functions

Strikingly, 97.7% recurrent co-splicing clusters have low expression correlations (i.e., average correlations < 0.4), therefore the co-splicing clusters revealed many novel functional groups that can not be identified by co-expression clusters.
Co-splicing clusters reveal novel functions

Spliceosomal complex, RNA splicing

Exon network
Average weight = 0.52

Gene network
Average weight = 0.29
Identifying Coupled Transcription-Splicing Modules from Gene-Exon Two-layer Networks
Transcription and splicing can be functionally coupled to ensure efficient and precise gene expression.

- Transcription of the same gene by different promoters can generate different splicing products, and certain splicing factors are known to stimulate transcription. Thus far, the scope and mechanism of the transcription-splicing coupling is still largely unknown.

- Thus far, it is not clear by what other mechanisms, to what extent, for which cellular functions, and under what conditions transcription-splicing coupling happens.
An RNA-seq Dataset

Gene Expression Profiles

Exon Inclusion Rate Profiles (Normalized Index)

Co-Expression and Co-Splicing Network

Inclusion rate is calculated based on >70% quantile gene expression
Coupling between transcription and splicing

We identify a subset of co-expressed genes, all or some of whose exons are co-splied.
Frequently Occurring in Many Datasets

The reliability of coupled modules are increased
Results

• We applied our method to 19 paired weighted co-expression and co-splicing networks derived from human RNA-seq data, and identified 8,667 Frequent Coupled Modules.

• Each coupled module contains > 5 member genes, > 5 member exons, appears in > 2 RNA-seq datasets, has a “heaviness” > 0.4. The average gene/exon size of these patterns are 12.61/12.64 and the average recurrence is 2.04.

• A large proportion of the identified coupled modules are involved in post-transcriptional processing.
We found that functionally coupled recruitment of transcription and splicing factors is often mediated by PPIs

- We considered not only the direct and one-hop indirect PPIs between two factors.

- We used 109 human transcription factors in ENCODE77 and the TFBS motifs in the JASPAR database. The putative splicing factors for each coupled module were identified based on the data in the SpliceAid database.

- The total number of PPIs between transcription and splicing factors within the same coupled modules is 105, compared to only 14.8 in random modules with the same size of gene and exon sets, with the fold ratio 7.1.

→ This evidence supports that PPI mediated association can be an important mechanism of transcription-splicing coupling.
Vertical Data Integration
Integrative Analysis of Multi-Dimensional Genomics Data
Multi-dimensional genome-wide profiling of same samples

Thanks to the next-generation sequencing, more and more such multi-dimensional genomics datasets will emerge...
Challenges

- Different types of genomics data have different scales and units, and thus can not be simply aggregated for analysis.

- Current methods dealt with 2 different data types measured on the same samples (e.g. eQTL), but to our knowledge, no methods are available to simultaneously deal with >2 data types?

- The sample size is generally not sufficiently large to apply traditional population genetics approaches.
The multi-dimensional genomic datasets can be represented as a set of matrices sharing the same sample dimension.
Our goal

- Identify **multi-dimensional module** across multiple types of genomics data

![Diagram showing DNA methylation, MicroRNA expression, and Gene expression across 384 samples]
Our Approach

- Develop a joint Non-negative Matrix Factorization (NMF) approach

\[
\min_{W, H_1, H_2, H_3 \geq 0} \sum_{i=1,2,3} \|X_i - WH_i\|_F^2
\]

- $X_i$: the data matrix of the $i$-th type of genomics data
- $W$: the component matrix
- $H_i$: the association matrix of the $i$-th type of genomics data
NMF

• Given a non-negative matrix $X$, find non-negative matrix factors $W$ and $H$ such that:

\[ X \approx WH \]

• $X = n \times m$
• $W = n \times k$
• $H = k \times m$ \hspace{1cm} (n+m)k < nm so that data is compressed
NMF

• NMF distinguished from other methods by its non-negativity constraint

• This allows a parts-based representation because only addition is additive combinations are allowed
Coherent patterns in matrices (Multiple NMF)

\[
\min_{W, H_1, H_2, H_3 \geq 0} \sum_{i=1,2,3} \| X_i - WH_i \|_F^2
\]
Interpretation and illustration of the framework
Applying MultiNMF to the TCGA data

- The Cancer Genome Atlas (TCGA) is a comprehensive and coordinated effort to accelerate our understanding of the genetics of cancer using multi-dimensional genome analysis technologies.

- We compiled DNA methylation (15418 markers), MicroRNA (799 miRNAs) and gene expression data (17811 genes) for 385 ovarian tumor samples.

- Applying our method to the TCGA Ovarian cancer data, we identified 200 multidimensional modules, covering 2985 genes, 2008 DNA methylation loci, and 270 miRNAs.
Multi-dimensional modules real multi-level functional synergy

Although the individual dimensions of these modules already exhibit a significant level of functional homogeneity, combining all dimensions reveals an even stronger functional synergy.

GE: Gene Expression
DM: DNA Methylation
ME: miRNA Expression
Combination: GE-DM-ME
Statistical significance of cross-dimensional correlations in the multi-dimensional modules
Multi-dimensional modules can elucidate vertical association mechanisms between different layers of gene regulation.
Multi-level factors cooperatively perturb pathways

(A) Bladder cancer pathway and (B) TGF-β signaling pathway are enriched in the combination of molecules in all 3 dimensions, but not in each dimension.

Molecules in the module participating in the corresponding pathways include those from the gene expression dimension (green), DNA methylation dimension (red), miRNA expression dimension (blue), miRNA targets (gray).
Clinical characterization: survival analysis

Median survival (26.4 vs. 34.1 months)

Median survival (38.2 vs. 33.8 months)
Acknowledgement

• Horizontal Data Integration
  – Wenyuan Li
  – Chao Dai
  – Jim Liu
  – Tong Zhang (Rutgers)
  – Haifeng Li

• Vertical Data Integration
  – Shihua Zhang
  – Jim Liu
  – Wenyuan Li
  – Shen Hui and Prof. Peter Laird

Funding agencies: NIGMS, NIA, NSF Career, Sloan Foundation, Zumberg Foundation