Integration of high-throughput biological data

Jean Yang and Vivek Jayaswal

School of Mathematics and Statistics
University of Sydney

Meeting the Challenges of High Dimension: Statistical Methodology, Theory and Applications
1 - 12 Oct, 2012
Integrating biological data

Gene Ontology
Protein: iTRAQ
miRNA data
Literature

SNP
Clinical Data
GeneChip Affymetrix Expression
Array CGH DNA copy number
cDNA / long oligo microarray Expression
DNA

mRNA

Protein

Phenotype

- SNP data
- DNA-seq
- Exome capture
- Microarray
- RNA-seq
- microRNA
- Quantitative data e.g. iTRAQ, SILAC.
- Clinical data

(A) Association
- Pathways
- Gene Ontology
- Transcription factor
- other meta data

(B) Prediction

Central dogma of molecular biology

DNA

mRNA

Protein

Phenotype

genome

transcriptome

proteome

phenome
DNA

mRNA

Protein

Phenotype

• SNP data
• DNA-seq
• Exome capture

• Microarray
• RNA-seq
• microRNA

• Quantitative data e.g. iTRAQ, SILAC.

• Clinical data

(A) Association

(B) Prediction

• Pathways
• Gene Ontology
• Transcription factor
• other meta data

omics data

gene

transcriptome

proteome

phenome

DNA

mRNA

Protein

Phenotype

Central dogma of molecular biology
mRNA Expression data

microarray

Variable (5000-30000 genes) or (2000 miRNAs)

N samples

Expression data

Count data

Technologies

MicroRNA

Next-gen Sequencing
Role or miRNA

- Comprises ~1% of genes in animals (~300+ miRNAs in human)
- ~20-30% of human genes are miRNA targets: each miRNA targets ~200 to 400 genes.
- miRNA molecules target mRNA (gene) transcript along short target sites via complementary base-pairing.
Case study: Stage III Melanoma

Melanomas
- A skin cancer more prevalent in Caucasians
- Most common in sunny climates (often correlated with latitude)

Metastasise (Stage III)
- About 40% die within a year
- 20% die between 1 and 4 years
- 40% are cleared of the cancer (after 4 years)

Samples (n=79) were obtained from A/Professor Graham Mann's group from the Westmead Institute for Cancer Research and Melanoma Institute, Australia.
Survival Time in Years

Number of individuals

Years of Survival

Two survival groups

Bad prognosis: Survival < 1 year and died due to melanoma

Good prognosis: Survival ≥ 4 years with no sign of relapse
Stage III Melanoma data set

- Gene and microRNA expression data obtained for 45 matched samples
  - 23 good prognosis (Alive NSR > 4 yrs)
  - 22 bad prognosis (Dead melanoma < 1 yr)
Data

- Stage III Melanoma patients

Clinical  
48 x 21

Gene expression  
47 x 26085

miRNA  
45 x 1347
Several computational miRNA-target prediction algorithms have been developed e.g. TargetScan, PicTar, microCosm (based on miRanda), and TargetMiner.

- Large variations in results obtained using different algorithms.
- Low sensitivity values.
- Most widely used approach combines the results from multiple target prediction algorithms.
Data

- Stage III Melanoma patients

Gene expression

47 x 26085

Focusing on gene expression data only
Simple approach: gene set test

- Test the hypothesis that a set of genes (rather than a single gene) is differentially expressed

- Current implementations in R e.g. GSEA, PGSEA, and GSA packages do not allow for testing of multiple conditions simultaneously. For example, cannot test a timecourse gene expression data for identifying regulatory miRs

Define the gene set as microRNA targetss
Functional enrichment

- Is a functional gene set $S$ (e.g., target genes of miR-196) over-represented among DE genes?
Functional enrichment

• Is a functional gene set $S$ (e.g., target genes of miR-196) over-represented among DE genes?

<table>
<thead>
<tr>
<th></th>
<th>DE</th>
<th>Not DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>15</td>
<td>200</td>
</tr>
<tr>
<td>Not Target</td>
<td>100</td>
<td>10000</td>
</tr>
</tbody>
</table>

Fisher’s exact test
Functional enrichment

- Is a functional gene set $S$ (e.g., target genes of miR-196) overrepresented among DE genes?

<table>
<thead>
<tr>
<th></th>
<th>DE</th>
<th>Not DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>15</td>
<td>200</td>
</tr>
<tr>
<td>Not Target</td>
<td>100</td>
<td>10000</td>
</tr>
</tbody>
</table>

- Is a functional gene set $S$ “differentially expressed”?

Fisher’s exact test

Wilcoxon Rank Sum test
Single platform gene set approaches

Cut-off specified

Over-representation Test
- Hypergeometric Test
- Binomial Test of Proportions

No cut-off specified

Self-contained

Univariate
- ROAST
- Hotelling’s $T^2$
- N-statistic

Multivariate
- Wilcoxon Rank Sum Test
- Gene Set Enrichment Analysis
- PAGE
- Logistic Regression
- CAMERA

Competitive

Univariate

Multivariate

None found
Data

- Stage III Melanoma patients

Gene expression
47 x 26085

Limitations:
- Only able to Identify “interesting” miRNAs one at a time.
Data

- Stage III Melanoma patients

Gene expression
- 47 x 26085

miRNA
- 45 x 1347

Questions
- Association?
- Network?
- Regulation?

Can we look for relationship between groups / clusters / modules?
Interested in the statistical significance of association between a miRNA cluster and gene cluster even if every miRNA-gene pair does not have a significant association.
Step I: Selection of clusters

Expression Data: $X$  
$[M \times N]$  

Mapping matrix, $M$  
miRNA - mRNA  
$[M \times G]$  

Aimed to group genes based on expression profiles and mapping information.

Unguided clustering  e.g. PAM

Guided clustering: Multivariate Random Forest (MRF)

Unguided clustering  e.g. PAM

Evaluation of two clustering methods

“Significant” clusters

MRF: Xiao and Segal, PLoS Comp Bio, 09
Unguided clustering

• Based on miRNA-mRNA mapping data
  – Generate a matrix of N genes and M miRNAs such that an element \((i, j)\) of the matrix takes the value 1 if miRNA \(I\) is predicted to target mRNA \(j\) and 0, otherwise

• Partition around medoids algorithm
  – Obtain the pairwise distances between genes
  – Assign each gene to a cluster such that the distance between the genes within a cluster is minimized

• Identifies miRNAs that share similar targets or genes that are co-regulated
Guided clustering

• Uses both expression data and miRNA-mRNA mapping information

• More likely to find genes that are grouped together because –
  – share similar targets (or regulators)
  – similar expression profiles

• Based on multivariate random forest (MRF)
Identifying gene clusters using MRF

• Obtain the proximity matrix such that $0 \leq \rho(i, j) \leq 1$

• Obtain dissimilarity matrix = $1 –$ proximity matrix

• Employ PAM to obtain gene clusters

• But there are questions
  – How many clusters are there?
  – How to evaluate whether MRF-based guided clustering is better than unguided clustering?
Guided vs. unguided clustering

- Permutation test
  - For a given cluster, obtain a measure of cluster tightness ($\theta$)
  - Generate a random cluster of the same size as before and estimate cluster tightness, $\theta^*$
  - Obtain $P(\theta^* \leq \theta)$
Step 2: Identify significant association

A miRmR pair is considered to have an association
- if mapping matrix has the value 1.
- evidence of a relationship between miRNA expression and mRNA expression based on a linear model

\[ u = \alpha + \beta v, \]

Testing H0: \( \beta = 0 \) vs. H1: \( \beta \neq 0 \) using a t-test.
Methods for measuring association

**Conditions**

- **Two conditions unbalanced samples – say case (disease) vs. control (healthy)**
- **Multiple conditions (N is large)**
- **5 - 10 conditions – e.g. a short timeseries dataset or different subtypes of MM**

**Methods**

- **Correlation**
  - Pearson’s correlation
  - Canonical correlation
  - Bayesian model
    - GenMir++
- **Linear regression models**
- **Odds-ratio method**
To determine the cluster pairs or modules of interest, we perform a permutation test

- Identify the number of miRNA-mRNA pairs with statistically significant association, $\eta$
- Obtain this number for a randomly selected mRNA cluster, $\eta^*$
- Obtain $P(\eta^* \geq \eta)$
Back to our data

microRNA-gene modules

• Input for MRF based analysis

  • 935 unique genes from a total of 15,568 unique genes
  • 158 microRNAs from a total of 1,347
  • Genes and microRNAs were selected based on
    • Unadjusted p-value for a comparison of GP vs BP < 0.05 and
    • microRNA-gene pair predicted in the union of PicTar, TargetScan and MicroCosm
Some examples of clusters include:

- miR-15a, -16, -195
- let-7a, -7c, -7f, miR-98
- miR-26a, -26b, -410
- miR-107, -25, -32

Enriched for a few KEGG pathways, e.g.:

- RNA degradation
- mRNA surveillance pathway
Some results for the GBM dataset

- **hsa-miR-100, hsa-miR-99a**
  - Gene cluster size = 76

- **hsa-miR-29b, hsa-miR-103, hsa-miR-107**
  - Gene cluster size = 72

- **hsa-miR-192, hsa-miR-635, hsa-miR-613, hsa-miR-136, hsa-miR-203, hsa-miR-23a, hsa-miR-766, hsa-miR-154, hsa-miR-25, hsa-miR-27a, hsa-miR-377, hsa-miR-519e, hsa-miR-656, hsa-miR-380**
  - Gene cluster size = 71
  - Gene cluster size = 95
  - Gene cluster size = 125
  - Gene cluster size = 46
  - Gene cluster size = 40

- **hsa-miR-659, hsa-miR-137, hsa-miR-432, hsa-miR-496, hsa-miR-550, hsa-miR-191, hsa-miR-218, hsa-miR-381, hsa-miR-498, hsa-miR-210, hsa-miR-410, hsa-miR-488, hsa-miR-539, hsa-miR-584, hsa-miR-95, hsa-miR-433, hsa-miR-563, hsa-miR-586**
  - Gene cluster size = 53
  - Gene cluster size = 171
  - Gene cluster size = 93
  - Gene cluster size = 46
  - Gene cluster size = 67

Additional pathways:

1. Pathways in cancer
2. Small cell lung cancer
3. Focal adhesion
4. Integrin cell surface interactions
5. Steroid biosynthesis
6. Alcohol metabolic process
7. Cell adhesion
8. Metabotropic glutamate receptor group III pathway
9. Cellular amino acid and derivative metabolic process
10. Amine metabolic process
11. Pathways in cancer
12. Negative regulation of cell cycle process
13. Regulation of cell cycle
14. Cell-cell signaling
15. Signaling by Rho GTPases
16. 2-arachidonoylglycerol biosynthesis
17. Calcium signaling pathway
18. Long-term potentiation
DNA
mRNA
Protein
Phenotype

• SNP data
• DNA-seq
• Exome capture

• Microarray
• RNA-seq
• microRNA

• Quantitative data e.g. iTRAQ, SILAC.

• Clinical data

(A) Association

• Pathways
• Gene Ontology
• Transcription factor
• other meta data

(B) Prediction

omics data

DNA
mRNA
Protein
Phenotype

genome

transcriptome

proteome

phenome

Central dogma of molecular biology
Data

• Stage III Melanoma patients

Gene expression
47 x 26085

miRNA
45 x 1347

Questions
Association ?
Network ?
Regulation ?

Can we look for relationship between groups / clusters / modules ?

External database
miRNA - mRNA
Data

• Stage III Melanoma patients

Clinical
48 x 21

Gene expression
47 x 26085

miRNA
45 x 1347

Questions
Prediction?
Finding biomarkers?
Data

- Stage III Melanoma patients

Clinical
48 x 21

Questions
Prediction?
Finding biomarkers?
Missing values

- 48 samples and 21 variables comprising of continuous, binary, and ordinal variables
Bootstrap Multiple Imputation (B-MI) procedure

Stage 1:
- m=25 multiple imputations
- $Z^1$, $Z^2$, ..., $Z^{m-1}$, $Z^m$

Stage 2:
- Draw one bootstrap sample
- $Z^1_{b^*_1}$, $Z^2_{b^*_2}$, ..., $Z^{m-1}_{b^*_m}$, $Z^m_{b^*_m}$

Stage 3:
- Feature selection method

Stage 4/5:
- Combine regression estimates using $\tau_{MI}$ and produce $\beta_{b^*_j}$
- Repeat B times and obtain $\beta^*$

Stage 6:
- Select variables for final model depending on $\tau_B$
Results from clinical data

<table>
<thead>
<tr>
<th>BMI procedure</th>
<th>LOOCV error rate*</th>
</tr>
</thead>
<tbody>
<tr>
<td>glm with stepwise selection</td>
<td>39% [9]</td>
</tr>
<tr>
<td>lasso</td>
<td>43% [4]</td>
</tr>
<tr>
<td>Elastic net</td>
<td>35% [13]</td>
</tr>
<tr>
<td>Ridge</td>
<td>41% [21-all]</td>
</tr>
</tbody>
</table>

*In [x] are the number of variables selected

The set of 13 clinical variables which gives lowest CV error rate is used in subsequent integration models
Data

- Stage III Melanoma patients

Clinical
48 x 21

Gene expression
47 x 26085

miRNA
45 x 1347

Questions
Prediction? Finding biomarkers?
Sparse canonical correlation analysis

Data X
\( n \times p \)

Composite latent variable for X

Data Z
\( n \times q \)

Composite latent variable for Z

Outcome Y
\( n \times 1 \)

Maximize
Correlation between
Xu and Zv

Sparse

linear combinations
Xu and Zv

SCCA

Clinical + Var1 = Xu OR Var2 = Zv

Predict outcome
Weighted lasso

- Correlation between Gene and miRNA;
  Weight 1 = 1/correlation

- Non-zero coefficients in a SCCA between Gene and miRNA;
  weights 2 = 1/coef
Weighted lasso (cont)

Gene data [Selected genes] + Clinical

Logistic regression

Final predictive variables
Here the preselected clinical variables + Genes + miRNAs selected through integration methods are used as variables.

<table>
<thead>
<tr>
<th></th>
<th>SCCA</th>
<th>Weighted lasso</th>
<th>Direct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>weight 1</td>
<td>weight 2</td>
</tr>
<tr>
<td>Clinical</td>
<td></td>
<td></td>
<td>35%</td>
</tr>
<tr>
<td>Clinical + G + M</td>
<td>33%</td>
<td>25%</td>
<td>23%</td>
</tr>
</tbody>
</table>
Summary and thoughts

• Based on MRF, we have a framework for finding interesting miRNA-mRNA modules.
• Further work is needed for developing better association measures under different conditions.
• Still very challenging to find approaches for methods evaluation ??
• Methods for integrating high-throughput data largely depends on the question of interest.
• Further research in finding robust and stable models utilizing both external and internal information.
• Biological interpretation.
Acknowledgements

• University of Sydney, School of mathematics and statistics
  – Samuel Müller
  – John Robinson
  – John Ormerod

  Bioinformatics research group
  – Anna Campain
  – Vivek Jayaswal

• St Vincent Centre for Applied Biomedical Research
  – David Ma
  – Mark Lutherborrow
  – Adam Bryant.

• Melanoma program at MIA/WMI Genomics Team
  – Giulietta Pupo
  – Svetlana Pianova
  – Varsha Tembe
  – Sarah-Jane Schramm
  – Graham Mann

• Melanoma Program at MIA/RPA Tumour bank and pathology
  – Cadance Carter
  – Chitra De Silva
  – Ken Lai
  – James Wilmott
  – Richard Scolyer

Supported in part by ARC grant DP0770395 (YHY, AC, VJ)