Canadian Bioinformatics Workshops

bioinformatics.ca
Module 9
Data Integration

Anna Goldenberg and Lauren Erdman
Bioinformatics for Cancer Genomics
May 30 – June 3, 2016

gene

epigenetic

mRNA expression

clinical

diet

imaging

protein

questionnaire

Strongly agree

Agree

Disagree

miRNA
Learning Objectives of Module

• Clinical data overview
• Single data type analysis (quick review)
• Data integration methods
  – Concatenate and Cluster
  – iCluster
  – SNF
• Advantages and drawbacks of different integration methods
• Survival analysis
Clinical data

- ID
- Race
- Family history (yes/no)
- Radiation
- Chemo
- Hormone therapy
- Protein IHC
- Stage
- Size
- Age at diagnosis
- Estrogen receptor level
- Progesterone level
- Overall outcome (dead/alive)
- Overall survival time
- Disease specific outcome (dead/alive)
- Disease specific survival time
- Recurrence status (yes/no)
- Time to recurrence
- Time to distant recurrence
- Distant recurrence status (yes/no)
Usage of clinical data alone

• Predict tool: http://www.predict.nhs.uk/predict.html
Usage of clinical data alone

- Predict tool: http://www.predict.nhs.uk/predict.html
Available patient data

- Genetic
- Epigenetic
- mRNA expression
- Clinical
- Diet
- Protein
- Questionnaire

Module 9: Data Integration
Data is available
The Cancer Genome Atlas (TCGA)

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Total</th>
<th>Exome</th>
<th>SNP</th>
<th>Methylation</th>
<th>mRNA</th>
<th>miRNA</th>
<th>Clinical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast invasive carcinoma [BRCA]</td>
<td>1098</td>
<td>1077</td>
<td>1095</td>
<td>1080</td>
<td>1094</td>
<td>1077</td>
<td>1078</td>
</tr>
<tr>
<td>Ovarian serous cystadenocarcinoma [OV]</td>
<td>586</td>
<td>536</td>
<td>579</td>
<td>584</td>
<td>583</td>
<td>582</td>
<td>585</td>
</tr>
<tr>
<td>Glioblastoma multiforme [GBM]</td>
<td>528</td>
<td>512</td>
<td>523</td>
<td>524</td>
<td>508</td>
<td>496</td>
<td>520</td>
</tr>
</tbody>
</table>

Total of 33 cancers.
9 cancers have over 500+ samples
All publicly available!
Why integrate patient data

- Use omic patient data to aid clinical decisions
  - To identify more homogeneous subsets of patients (that might respond similarly to a given drug)
  - To help better predict response to drugs
Single data type analysis

20 GBM, gene expression for ~18,000 genes

- Collect gene expression
- Select most varied genes
- Perform hierarchical clustering using selected genes
- Identify clusters (here, 2 clusters)
- Identify genes associated that differ between clusters using t-test (if 2 clusters) or ANOVA (if more)
- Correct for multiple hypothesis testing

(Liang et al, PNAS, 2005)
Single data type analysis

25 GBM, gene expression for ~18,000 genes

(Liang et al, PNAS, 2005)
Kaplan-Meier Curve

![Kaplan-Meier Curve Diagram]

- **Probability** vs. **Survival (Years)**
- Two groups: GBM Group 1 (orange) and GBM Group 2 (blue)
- The curves show the probability of survival over time for the two groups.

**Module 9: Data Integration**
The probability of surviving for 1 year is 80% for Group 1.
The probability of surviving for 1 year is 20% for Group 2.
The probability of surviving for 1 year is 80% for Group 1
The probability of surviving for 1 year is 20% for Group 2
Single data type driven integration

200 GBMs

mRNA  mutations  CNV  clinical

more genes  more genes

p-value = {0.2, 0.6, 0.5}

(Verhaak et al, Cancer Cell, 2010)
Single data type driven integration

200 GBMs

mRNA → mutations → CNV → clinical

more genes + more genes

Lead to identification of proneural, neural, classical and mesenchymal groups

(Verhaak et al, Cancer Cell, 2010)
Single data type driven integration

200 GBMs

mRNA + mutations + CNV + clinical

What about methylation data?

(Verhaak et al, Cancer Cell, 2010)
More recent GBM study (Sturm et al, 2012)
Integration approaches

1. Concatenate and cluster (commonly used in TCGA analysis)
2. iCluster (Shen et al, 2009)
3. SNF (Wang et al, 2014)
**Concatenation**

<table>
<thead>
<tr>
<th>Patients</th>
<th>Gene expression</th>
<th>Methylation (gene level or probe level)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Module 9: Data Integration
Hierarchical Clustering

![Hierarchical Clustering Table](http://people.revoledu.com/kardi/tutorial/Clustering/Numerical%20Example.htm)

**min distance**
(single linkage)
Hierarchical Clustering

<table>
<thead>
<tr>
<th>Dist</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00</td>
<td>0.71</td>
<td>5.66</td>
<td>3.61</td>
<td>4.24</td>
<td>3.20</td>
</tr>
<tr>
<td>B</td>
<td>0.71</td>
<td>0.00</td>
<td>4.95</td>
<td>2.92</td>
<td>3.54</td>
<td>2.50</td>
</tr>
<tr>
<td>C</td>
<td>5.66</td>
<td>4.95</td>
<td>0.00</td>
<td>2.24</td>
<td>1.41</td>
<td>2.50</td>
</tr>
<tr>
<td>D</td>
<td>3.61</td>
<td>2.92</td>
<td>2.24</td>
<td>0.00</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>E</td>
<td>4.24</td>
<td>3.54</td>
<td>1.41</td>
<td>1.00</td>
<td>0.00</td>
<td>1.12</td>
</tr>
<tr>
<td>F</td>
<td>3.20</td>
<td>2.50</td>
<td>2.50</td>
<td>0.50</td>
<td>1.12</td>
<td>0.00</td>
</tr>
</tbody>
</table>

http://people.revoledu.com/kardi/tutorial/Clustering/Numerical%20Example.htm
Deciding on the number of clusters

- Arbitrarily cutting the dendrogram (by eye)
- Silhouette statistic
- Eigen gap (for spectral clustering)
- Many more (review in Yan et al, PhD thesis, 2005)
Silhouette statistic

- First presented by Rousseeuw (1987) to show graphically how well each pattern is classified to a cluster.
- For each pattern $i$ in class $C_r$

$$Sil_i = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

- $a(i)$ = average distance to all other patterns in $C_r$.
- $b(i)$ = average distance to all other patterns in other clusters.

- $-1 \leq Sil \leq 1$
- $Sil = 1$: good assignment
- $Sil = -1$: bad assignment
- $Sil = 0$: borderline
Silhouette statistic

a. Three clusters in 2 dimensions
b. Three clusters in 10 dimensions, each cluster has 50 observations
c. 4 clusters in 10 dimensions with randomly chosen centers
d. Six clusters in 2 dimensions

(a)  (d)
Silhouette statistic

a. Three clusters in 2 dimensions
b. Three clusters in 10 dimensions, each cluster has 50 observations
c. 4 clusters in 10 dimensions with randomly chosen centers
d. Six clusters in 2 dimensions
Consensus Clustering

- Resampling based method for class discovery and visualization of gene expression microarray data
- Goal: assessing stability
- Method:
  - For a 1000 iterations
    1. Resample data
    2. Cluster with fav. clust. method (hier, k-means)
- Compute consensus matrix
- Partition D based on Consensus Matrix

iCluster (Shen et al, 2009)

- Gaussian latent variable model
- Sparsity regularization (Lasso-type)
- Latent variables Z (embedding is shared)

Package on CRAN: iCluster
Drawbacks of existing methods

• A lot of manual processing (e.g. need to pre-filter genes for iCluster – takes ~1500 genes max)

• Many steps in the pipeline

• Integration mostly done in the feature space – if there is signal in a combination of features, it’ll be lost

• Focusing on similarity across data types – what if there is complementary information?
Similarity Network Fusion (Wang et al, 2014)

- Integrate data in the patient space
  1. Construct patient similarity matrix
  2. Fuse multiple matrices
1. Construct similarity networks

Patients mRNA expression genes

Patients
1. Construct similarity networks

Module 9: Data Integration
2. Combine networks

Similarity Networks

Patient similarity: mRNA-based DNA Methylation-based Supported by all data

Module 9: Data Integration
2. Combine networks

Similarity Networks

Fusion Iterations

Patient similarity:
- mRNA-based
- DNA Methylation-based
- Supported by all data

Module 9: Data Integration

bioinformatics.ca
2. Combine networks

- **Similarity Networks**
- **Fusion Iterations**
- **Fused Similarity Network**

Patient similarity:
- mRNA-based
- DNA Methylation-based
- Supported by all data
Experiments

Data:

- 2 simulations
- 5 TCGA cancers

Compared Methods:

- Concatenation
- iCluster
- PDSB
- Multiple kernel learning
Simulation 1 – complementarity
Simulation 2 - removing noise

Gaussian Noise

Data type 1

Ground truth

Data type 2

Gamma Noise

Dataset

Class 1

Class 2

Module 9: Data Integration
Simulation 2 - removing noise

![Graph showing NMI vs. Scale of Noise σ](image1)

![Graph showing NMI vs. Scale of Noise λ](image2)

- **Fusion + Spectral**
- **iCluster**
- **Concatenation**
- **MKL**
- **PSDF**
## TCGA Data

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Patients</th>
<th>mRNA</th>
<th>Methylation</th>
<th>miRNA</th>
<th>Controls mRNA</th>
<th>Controls Methylation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM</td>
<td>215</td>
<td>12,042</td>
<td>1,491</td>
<td>534</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>BIC</td>
<td>105</td>
<td>17,814</td>
<td>23,094</td>
<td>1,046</td>
<td>63</td>
<td>27</td>
</tr>
<tr>
<td>KRCCC</td>
<td>124</td>
<td>20,532</td>
<td>24,976</td>
<td>1,046</td>
<td>68</td>
<td>199</td>
</tr>
<tr>
<td>LSCC</td>
<td>105</td>
<td>12,042</td>
<td>27,578</td>
<td>1,046</td>
<td>-</td>
<td>27</td>
</tr>
<tr>
<td>COAD</td>
<td>92</td>
<td>17814</td>
<td>27578</td>
<td>705</td>
<td>19</td>
<td>37</td>
</tr>
</tbody>
</table>
Case study: Glioblastoma

DNA methylation data

mRNA expression

miRNA expression
Case study: Glioblastoma

DNA methylation data

mRNA expression

miRNA expression

Cytoscape

Bo Wang

bioinformatics.ca
Clinical properties of the subtypes

- Subtype 1
  - Survival probability: 1.0
  - Survival time (months): 0, 20, 40, 60, 80, 100
  - Survival probability: 1.0, 0.8, 0.6, 0.4, 0.2, 0.0
  - Age (years): 20, 40, 60, 80, 100
- Subtype 2
  - Survival probability: 0.8
  - Survival time (months): 0, 20, 40, 60, 80, 100
  - Survival probability: 0.8, 0.6, 0.4, 0.2, 0.0
- Subtype 3
  - Survival probability: 0.6
  - Survival time (months): 0, 20, 40, 60, 80, 100
  - Survival probability: 0.6, 0.4, 0.2, 0.0

- p-value = 3x10^{-5}
Clustering of the network

Wang et al, 2014
Patient networks: advantages and disadvantages

- Integrative feature selection
- Growing the network requires extra work
- Unsupervised – hard to turn into a supervised problem

✓ Creates a unified view of patients based on multiple heterogeneous sources
✓ Integrates gene and non-gene based data
✓ No need to do gene pre-selection
✓ Robust to different types of noise
✓ Scalable

Package on CRAN: SNFtool
Survival Analysis

• Survival data
  – Hazard Rates
  – Survival Functions
• Kaplan-Meier Estimator
• Log-rank test
• Cox proportional hazards ratio model
Survival data

- Time to a single event (e.g. time to death or time to treatment failure)
- Some data on patients maybe missing, e.g. lost due to the end of the study – such data is called right censored (i.e. death occurred to the right of the last observation)
- For uncensored data we observe death time
- For censored we know that it’s beyond that time
- Assume non-informative (independent) censoring (e.g. moved out of town and lost track of the patient not related to the disease)
Survival data example

days to last follow-up
Two important statistics

• Event time: \( X \)

• Survival function
  \[ S(x) = P(X > x) - \text{probability of a person being alive at } x \]

• Hazard rate
  \[ H(x) = \lim_{\delta x \to 0} \frac{P(x \leq X \leq x + \delta x \mid X > x)}{\delta x} \]

  \text{Probability of a person dying (failing) in the next instant}
Some examples of the hazard rate

• Constant hazard rate – no aging
• Positive hazard rate – the older you are the more likely you are to die
• Negative hazard rate – e.g. dying risk is the highest at birth (infant mortality)
Kaplan Meier Estimator

- Basic estimator Kaplan-Meier (1958)
- \( S(t) \) – probability that a member from a given population will have a lifetime exceeding \( t \)
- \( n_i \) – number of people at risk of dying at time \( t \)
- \( d_i \) – number of actual deaths at time \( t \)

\[
S(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}
\]
Kaplan-Meier Curve

Probability

Survival (Years)

GBM Group 1
GBM Group 2

Module 9: Data Integration
Are Kaplan Meier curves significantly different? Log-rank test

- Comparing 2 groups
- $j = 1..J$ be the distinct events in either group
- For each $j$, $N_{1j}$ and $N_{2j}$ – numbers of people at risk in groups 1 and 2 respectively at the start of period $j$ ($N_j = N_{1j} + N_{2j}$)
- $O_{1j}$ and $O_{2j}$ - observed number of events in each group (corresponds to variable $d_i$ above). $O_j = O_{1j} + O_{2j}$
- Null hypothesis: the distributions (2 curves) are the same. Given that $O_j$ happened across both groups at time $j$, $O_{1j}$ is hypergeometric with expected value and variance:
  $$E_{1j} = \frac{O_j}{N_j} N_{1j} \quad V_j = \frac{O_j (N_{1j} / N_j) (1 - N_{1j} / N_j) (N_j - O_j)}{N_j - 1}$$

  $$Z = \frac{\sum_{j=1}^{J} (O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}} \sim N(0, 1)$$
  - get p-value from look up table
Hazard ratio

Hazard ratio compares two groups differing in treatments or prognostic variables etc. Measures relative survival in two groups based on the complete period studied.

\[ R = \frac{O_1/E_1}{O_2/E_2} \]

R=0.43 – relative risk (hazard) of poor outcome under the condition of group 1 is 43% of that of group 2.

R= 2.0 then the rate of failure in group 1 is twice the rate in the group 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>P Value</th>
<th>Hazard Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at surgery</td>
<td>0.08</td>
<td>0.71 (0.48–1.05)</td>
</tr>
<tr>
<td>Clinical tumor size</td>
<td>0.23</td>
<td>1.26 (0.86–1.86)</td>
</tr>
<tr>
<td>Recurrence score</td>
<td>&lt;0.001</td>
<td>3.21 (2.23–4.61)</td>
</tr>
</tbody>
</table>
Cox proportional hazard (Cox PH) model

- Captures how well multiple variables $x_1...x_p$ (e.g. genes, clinical variables, etc) affect survival
- Estimates the ratio of the risks

Cox regression:  
\[ h_i(t) = h_0(t)e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}} \]

$h_0(t)$ – baseline hazard
other predictors are exponentially increasing the hazard
Some intuition behind the model

\[ h_i(t) = h_0(t)e^{\beta_1 x_1 + \ldots + \beta_p x_p} \]

- \( \beta_j \) represents log hazard ratio increase for one unit increase of the predictor \( x_j \) holding all other predictors constant.

- \( e^{\beta_j} \) – Hazard ratio increase for one unit increase of \( x_j \).

- \( \beta_j < 0 \) – Means increasing \( x_j \) is associated with decreased hazard and longer survival times.
Using and interpreting Cox PH

1. Hazard ratio for a subject i with a set of predictors X compared to a subject j with a set of predictors X*:

\[ HR = \frac{h_i(t)}{h_j(t)} = \frac{e^{\beta X}}{e^{\beta X^*}} = e^{\beta(X - X^*)} \]

2. HR can be interpreted as a percentage change in risk

Example: Let \( x = 1 \) when treatment is active and \( x = 0 \) when treatment is placebo. If \( HR = e^{\beta} = 0.8 \), it means 20% decrease in mortality risk if using treatment compared to placebo.
Concordance index (C-index)

- Captures the ability of the model to order individuals correctly with respect to their survival time

C-index = (#concordant pairs of indiv. + 0.5 * ties)/ all pairs

- Important – no other metric captures the ordering of individuals
Breast Cancer (METABRIC example)

CNV and expression data
Discovery: 997 patients
Validation: 995 patients

Nature, 2012

<table>
<thead>
<tr>
<th></th>
<th>PAM50 (5 clusters)</th>
<th>iCluster (10 clusters)</th>
<th>SNF (5 clusters)</th>
<th>SNF (10 clusters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P value discovery cohort</td>
<td>$3.0 \times 10^{-9}$</td>
<td>$1.2 \times 10^{-14}$</td>
<td>$6.10 \times 10^{-11}$</td>
<td>$3.31 \times 10^{-12}$</td>
</tr>
<tr>
<td>P value validation cohort</td>
<td>$1.7 \times 10^{-9}$</td>
<td>$2.9 \times 10^{-11}$</td>
<td>$5.12 \times 10^{-13}$</td>
<td>$7.86 \times 10^{-12}$</td>
</tr>
<tr>
<td>CI discovery cohort</td>
<td>0.560</td>
<td>0.621</td>
<td>0.638</td>
<td>0.638</td>
</tr>
<tr>
<td>CI validation cohort</td>
<td>0.551</td>
<td>0.605</td>
<td>0.633</td>
<td>0.633</td>
</tr>
</tbody>
</table>

established
Breast Cancer (METABRIC example)

CNV and expression data
Discovery: 997 patients
Validation: 995 patients

So how many subtypes are there really in breast cancer?

Nature, 2012

<table>
<thead>
<tr>
<th></th>
<th>PAM50 (5 clusters)</th>
<th>iCluster (10 clusters)</th>
<th>SNF (5 clusters)</th>
<th>SNF (10 clusters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$ value discovery</td>
<td>$3.0 \times 10^{-9}$</td>
<td>$1.2 \times 10^{-14}$</td>
<td>$6.10 \times 10^{-11}$</td>
<td>$3.31 \times 10^{-12}$</td>
</tr>
<tr>
<td>cohort</td>
<td>$1.7 \times 10^{-9}$</td>
<td>$2.9 \times 10^{-11}$</td>
<td>$5.12 \times 10^{-13}$</td>
<td>$7.86 \times 10^{-12}$</td>
</tr>
<tr>
<td>CI discovery cohort</td>
<td>0.560</td>
<td>0.621</td>
<td>0.638</td>
<td>0.638</td>
</tr>
<tr>
<td>CI validation cohort</td>
<td>0.551</td>
<td>0.605</td>
<td>0.633</td>
<td>0.633</td>
</tr>
</tbody>
</table>
Predicting using the network

Module 9: Data Integration
Predicting using the network Breast Cancer (METABRIC example)

CNV and expression data
Discovery: 997 patients
Validation: 995 patients

Nature, 2012

established
Data integration - future

• Simultaneous feature selection and data integration
• Supervised vs unsupervised approaches
• Weights for contributions of different types of data