Multi-Omic Correlation Graphs

Dvir Netanely, Zohar Yakhini, Ron Shamir
January 2018

Shamir’s Group
Tel Aviv University
Multi-Omics Datasets
Integrating Multi-Omic data

• Datasets differ in various ways
  – Value distribution
  – Number of features
  – Specific Biological meaning (Gene, CpG, miRNA)
  – May include contrasting signals

• Some methods
  – Work with a small subset of the features (most variable), ignore the rest of the information.
  – Assume there is a sample partition induced by the dominant feature signal.
Multi-Omic Correlation Graph

• A method for **early integration** of multi-omic datasets describing the same group of patients.

• Crosses information from the various datasets to pinpoint important key-features.
Multi-Omic Correlation Graph

• Uses inter-omic correlation to identify key-features that are likely to play a role in the biological variance observed among dataset samples.

• We assume that features that exhibit high variability across dataset samples, and strong correlation across several different technologies, represent modules of biological regulation which differ between the various samples.
Approach outline

• Filter features in each dataset based on variance (But can keep a lot more than in normal clustering)
• Unite similar features into unique Profiles (based on Pearson correlation, Threshold R1).
• Calculate the Spearman correlation between profile pairs (On sample intersection)
• Build a correlation graph using extreme correlations as edges.
• Extract interesting Modules from the graph
The Multi-omic breast cancer dataset collection

<table>
<thead>
<tr>
<th>Technology</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq (Gene Expression)</td>
<td>20,530</td>
</tr>
<tr>
<td>Copy Number</td>
<td>24,776</td>
</tr>
<tr>
<td>miRNA</td>
<td>1,046</td>
</tr>
<tr>
<td>DNA Methylation</td>
<td>107,639</td>
</tr>
</tbody>
</table>
The Multi-omic breast cancer dataset collection

<table>
<thead>
<tr>
<th>Technology</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq (Gene Expression)</td>
<td>20,530</td>
</tr>
<tr>
<td>Copy Number</td>
<td>24,776</td>
</tr>
<tr>
<td>miRNA</td>
<td>1,046</td>
</tr>
<tr>
<td>DNA Methylation</td>
<td>107,639</td>
</tr>
</tbody>
</table>

Total #Features: 618
Variance based Feature Filtering

- RNA_Seq: 20530 -> 2000
- Meth450: 107639 -> 200
- miRNA-HiSeq: 1046 -> 500
- CNV_Thresh: 24776 -> 2000
Step 1: Unite similar features into unique Profiles (based on Pearson correlation, Threshold R1)
Step 1: Unite similar features into unique Profiles (based on Pearson correlation, Threshold R1)
Step 1: Unite similar features into unique Profiles (based on Pearson correlation, Threshold R1)
Step 1: Unite similar features into unique Profiles (based on Pearson correlation, Threshold R1)
Step 1: Unite similar features into unique Profiles (based on Pearson correlation, Threshold R1)
Step 1: Unite similar features into unique Profiles (based on Pearson correlation, Threshold R1)

<table>
<thead>
<tr>
<th>Profile Size</th>
<th>mRNA</th>
<th>Meth450</th>
<th>miRNA</th>
<th>CNV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1100</td>
<td>661</td>
<td>202</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>111</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>35</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>8</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>75</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL:</strong></td>
<td><strong>1212</strong></td>
<td><strong>833</strong></td>
<td><strong>260</strong></td>
<td><strong>21</strong></td>
</tr>
</tbody>
</table>
The Multi-omic breast cancer dataset collection

<table>
<thead>
<tr>
<th>Technology</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq (Gene Expression)</td>
<td>20,530</td>
</tr>
<tr>
<td>Copy Number</td>
<td>24,776</td>
</tr>
<tr>
<td>miRNA</td>
<td>1,046</td>
</tr>
<tr>
<td>DNA Methylation</td>
<td>107,639</td>
</tr>
</tbody>
</table>
The Multi-omic breast cancer dataset collection

<table>
<thead>
<tr>
<th>Technology</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq (Gene Expression)</td>
<td>20,530</td>
</tr>
<tr>
<td>Copy Number</td>
<td>24,776</td>
</tr>
<tr>
<td>miRNA</td>
<td>1,046</td>
</tr>
<tr>
<td>DNA Methylation</td>
<td>107,639</td>
</tr>
</tbody>
</table>

Total features: 618
Step 2: calculate cross-omic profile correlations using Spearman correlation.
Step 2: calculate cross-omic profile correlations using Spearman correlation

Expression

miRNA

Methylation

Copy Number Variation

Other

R = 0.8

R = 0.51

R = -0.6

140

90

79
Step 3: Building the multi-omic correlation graph

- Inter-Omic correlations exceeding a certain threshold will be added to the graph as edges.
- Only Negative intra-Omic correlations will be added.
- How should the threshold be determined?
- We experimented with several approaches...
Determining the correlation thresholds

- **Attempt 1**: thresholds are calculated as the top P% correlation R Values for every omics pair.
- Will always yield 2P% edges even on random data.

<table>
<thead>
<tr>
<th>Omics Pair</th>
<th>Number of Edges</th>
<th>Correlation Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq - Meth450</td>
<td>10,096</td>
<td>0.317104</td>
</tr>
<tr>
<td>RNA-Seq - miRNA-HiSeq</td>
<td>3,151</td>
<td>0.319237</td>
</tr>
<tr>
<td>RNA-Seq - CNV_Thresh</td>
<td>255</td>
<td>0.267359</td>
</tr>
<tr>
<td>Meth450 - miRNA-HiSeq</td>
<td>2,166</td>
<td>0.274656</td>
</tr>
<tr>
<td>Meth450 - CNV_Thresh</td>
<td>175</td>
<td>0.245832</td>
</tr>
<tr>
<td>miRNA-HiSeq - CNV_Thresh</td>
<td>55</td>
<td>0.250612</td>
</tr>
</tbody>
</table>
Circular layout
Force Layout (First look)
First sanity check – modules seem to have a biological meaning

Paper from 2012: “Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors”, Niinuma et. al

FLJ12825 (Uncharacterized LOC440101) is an RNA Gene, and is affiliated with the ncRNA class

The Hox genes, which are organized into clusters on different chromosomes, are key regulators of embryonic anterior-posterior (A-P) body pattern formation and are expressed at specific times and in specific positions in developing vertebrate embryos.
Thresholds affect graph density
Testing correlation between node centrality and supervised differential capacity

• We applied a supervised test (Kruskal-wallis) on the mean profile of each node, testing its ability to differentiate the 5 PAM50 subtypes on the 618 samples.

• A pValue was assigned to each node.

• We calculated the correlation between this pValue and several node centrality measures.

• **Conclusion:** Node centrality measures is beneficial in identifying informative modules.
Comparison to a randomly edge-shuffles graph

• Based on Tom’s suggestion, we repeated the calculation on a random graph in which edges were randomly shuffled between nodes of the same color.

• Here, the correlation did not exist.
Determining the correlation thresholds

- **Attempt 2**: FDR on background distribution generated by shuffling the column of the correlation matrix.
RNA_Seq - RNA_Seq

- Building graph... 1 RNA_Seq - RNA_Seq 1
- 412 edges exceed threshold of 0.545900 which represent the 99.900 percentile over absolute values
  - 206 edges are above 0.579129 which is the 99.950 percentile of real data
  - 206 edges are below -0.431877 which is the 0.050 percentile of real data
- 111050 edges are above 0.132 which is the 99.950 percentile on the background
  - 44960 edges are below -0.133 which is the 0.050 percentile on the background
- 58956 edges are above 0.201 which is MEAN+5STD on the background
- 18492 edges are below -0.201 which is MEAN-5STD on the background
RNA Seq - Meth450

- Building graph... 1 RNA Seq - Meth450 2
- 306 edges exceed threshold of 0.438617 which represent the 99.900 percentile over absolute values
  - 153 edges are above 0.436103 which is the 99.950 percentile of real data
  - 153 edges are below -0.440766 which is the 0.050 percentile of real data
- 37544 edges are above 0.134 which is the 99.950 percentile on the background
- 44501 edges are below -0.132 which is the 0.050 percentile on the background
- 14478 edges are above 0.203 which is MEAN+5STD on the background
- 16870 edges are below -0.202 which is MEAN-5STD on the background
Distribution of background permuted inter-omic profile correlations RNA_SEQ vs. Methylation

- **Background**
- **Actual**
- **data1**
- **data2**
- **data3**
- **data4**
- **data5**
- **data6**
RNA Seq - miRNA-HiSeq

- Building graph... 1 RNA Seq - miRNA-HiSeq 3
- 167 edges exceed threshold of 0.449507 which represent the 99.900 percentile over absolute values
- 83 edges are above 0.477759 which is the 99.950 percentile of real data
- 83 edges are below -0.417121 which is the 0.050 percentile of real data
- 17316 edges are above 0.133 which is the 99.950 percentile on the background
- 21829 edges are below -0.133 which is the 0.050 percentile on the background
- 6550 edges are above 0.202 which is MEAN+5STD on the background
- 7597 edges are below -0.202 which is MEAN-5STD on the background
RNA_Seq - CNV_Thresh

- Building graph... 1 RNA_Seq - CNV_Thresh 4
- 8 edges exceed threshold of 0.336666 which represent the 99.900 percentile over absolute values
- 4 edges are above 0.314983 which is the 99.950 percentile of real data
- 4 edges are below -0.347800 which is the 0.050 percentile of real data
- 457 edges are above 0.131 which is the 99.950 percentile on the background
- 750 edges are below -0.133 which is the 0.050 percentile on the background
- 81 edges are above 0.201 which is MEAN+5STD on the background
- 190 edges are below -0.201 which is MEAN-5STD on the background
Before (left) and after (right) changing to background distribution.
Generation based on subtype preserving permutations.
Before (left) and after (right) changing to background distrib. Generation based on subtype preserving permutations.
Before (left) and after (right) changing to background distrib. Generation based on subtype preserving permutations
Before (left) and after (right) changing to background distrib. Generation based on subtype preserving permutations.
Determining the correlation thresholds

• **Attempt 3:** Ratio of profile correlations and all profiles correlations at given FDR.

• New Implementation – matrix based instead of graph based.

• Graphs objects are created for each module separately.
FDR Ratio

• We examine the difference between the profile correlation distribution and the background distribution.
• Blue – general background distribution
• Orange – Distribution of correlations for one profile from Omic 1 vs. all other profiles on Omic2.
FDR Ratio

- These graphs show the cumulative distributions from left to right and from right to left.
- The cyan vertical line represents the 0.05 percentiles on the background distribution on the left, and the 1-0.05 percentile on the right.
- In the legend
  R represents the cumulative value of the profile neighbours, divided by the background distribution.
  Thresh represents the percentile value.
Uninteresting case 1
Interesting on both sides 1
Interesting mostly on the right
Interesting on the left
Module Score Calculation

Modules were extracted based on:
- Filtering modules having at least 3 scores above 1.5 (having color degree>=3)
- Sorting the resulting profiles based on the total score (sum of bests scored)

<table>
<thead>
<tr>
<th>Omic1 Profile</th>
<th>Left (Negative Corr.)</th>
<th>Right (Positive Corr.)</th>
<th>Best=Max(Left,Right)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Omic2</td>
<td>Omic3</td>
<td>Omic4</td>
<td>Omic2</td>
</tr>
<tr>
<td>1</td>
<td>0.09</td>
<td>0.30</td>
<td>0.94</td>
<td>0.10</td>
</tr>
<tr>
<td>2</td>
<td>1.98</td>
<td>3.19</td>
<td>1.88</td>
<td>2.05</td>
</tr>
<tr>
<td>3</td>
<td>1.18</td>
<td>2.97</td>
<td>1.88</td>
<td>1.39</td>
</tr>
<tr>
<td>4</td>
<td>3.34</td>
<td>7.23</td>
<td>4.71</td>
<td>5.09</td>
</tr>
<tr>
<td>5</td>
<td>2.12</td>
<td>3.88</td>
<td>2.83</td>
<td>1.53</td>
</tr>
<tr>
<td>6</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>0.33</td>
<td>1.67</td>
<td>1.88</td>
<td>0.62</td>
</tr>
<tr>
<td>8</td>
<td>1.84</td>
<td>0.46</td>
<td>2.83</td>
<td>2.13</td>
</tr>
<tr>
<td>9</td>
<td>2.26</td>
<td>4.49</td>
<td>2.83</td>
<td>3.65</td>
</tr>
<tr>
<td>10</td>
<td>3.49</td>
<td>0.38</td>
<td>2.83</td>
<td>1.31</td>
</tr>
</tbody>
</table>
Step 4: Extracting interesting graph modules

- What is a module?
  - For each node, extract the subgraph of immediate its neighbors
    - Density based cuts seem less relevant
    - Long modules have little biological interpretation
Step 4: Extracting interesting graph modules

• Based on our initial motivation, interesting modules should be composed of highly correlated nodes originating from multiple omics.
  – Selecting profiles having color degree of at least X
Examples for Low score Modules
GO analysis summary for module 1_RNA_Seq_Profile_125

- regulation of cell differentiation - GO:0045595
- negative regulation of canonical Wnt receptor signaling pathway - GO:
- electron carrier activity - GO:0009055
GO analysis summary for module 4_CNV_Thresh_Profile_15

- Oxygen transporter activity - GO: 0005344
- Serine-type endopeptidase activity - GO: 0004252
- Protein C-terminus binding - GO: 0008022
Examples for High score Modules
GO analysis summary for module 2_Meth450_Profile_31

- regulation of leukocyte activation - GO:002694
- regulation of immune system process - GO:0002682
- leukocyte activation - GO:0045321
- immune response-regulating cell surface receptor signaling pathway - GO:002768
- locomotion - GO:0040011
- calcium ion binding - GO:0005509
- endopeptidase activity - GO:0004175
- cytokine receptor binding - GO:0005126
- serine-type endopeptidase activity - GO:004252
- cell proliferation - GO:0008283
- embryo development - GO:0009790
- brain development - GO:0007420
- organ morphogenesis - GO:009887
- pattern specification process - GO:0007389
- regionalization - GO:0030012
- embryonic organ development - GO:0048568
- regulation of cell differentiation - GO:0045595
- regulatory region DNA binding - GO:000975
- neuron differentiation - GO:0030182
- mesenchymal cell differentiation - GO:0048762
- intermediate filament cytoskeleton - GO:0045111
- structural molecule activity - GO:0005198
- cytokine receptor binding - GO:0005126
- eosinophil chemotaxis - GO:0048245
- intermediate filament organization - GO:0045109
- cellular calcium ion homeostasis - GO:0006874
- oxygen transporter activity - GO:0005344
- cytoskeleton organization - GO:0007010
- epidermis development - GO:0008544
The scoring is not perfect (should be capped)
Application

• So, we have now have a list of X correlation modules, integrating multi-omic datasets.
• We picked the most “interesting” modules, assuming they hold informative biological information.
• How can we use them?
1. Direct biological interpretation

- Test each module for gene set enrichment
- Each module represents a biological signal, backed by at least 3 different measuring technologies.
- Example: The HoxC module
2. Cluster samples based on leveled profiles of the central nodes

• Use the graph as a feature selection method before clustering the samples
• Level the profile of each central node (may come from different technologies)
• Unite the profiles into a matrix
• Cluster the samples using the central nodes as features.
  – Feature selection
  – Dimension reduction
  – Signal boosting
Data Leveling
Converting to levels (RNA-Seq)
Converting to levels (Meth450)
Converting to levels (miRNA)
Converting to levels (CNV thresholded)
Clustering breast tumors based on the leveled profiles of central nodes of strongest modules
Matrix of 293 central node profiles x 618 samples
Comparison to PAM50
### Unified Table of Significant Clinical Labels - Labels (p-Value Threshold: 1.00e-03)

<table>
<thead>
<tr>
<th>Label</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER_Status_nature012</td>
<td>Positive</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>HER2_Final_Status_nature012</td>
<td>LumB</td>
<td>LumA</td>
<td>Basal</td>
</tr>
<tr>
<td>PAM50/CallRNAseq</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAM50/mRNA_nature012</td>
<td>Basal like</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PR_Status_nature012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breast_carcinoma_estrogen_receptor_status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>breast_carcinoma_progestosterone_receptor_status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>histological_type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>icd_0_3_histology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>initial_pathologic_diagnosis_method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lab_proc_henz_neu_immunohistochemistryreceptor_status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UpdatedPAM50_Call</td>
<td>LumB</td>
<td>LumA</td>
<td>Basal</td>
</tr>
<tr>
<td>UpdatedPAM50_B_Call</td>
<td>LumB</td>
<td>LumA</td>
<td>Basal</td>
</tr>
<tr>
<td>UpdatedPAM50_LumAR_Call</td>
<td>LumB</td>
<td>LumA-Rc</td>
<td>Basal</td>
</tr>
<tr>
<td>sampleSets MeTree378 LumAclusteredT 1 based on self.top 2000, Inf 1, GS only NORMALIZE.Dxt</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>sampleSets MeTree679 Her2BasLumAclusteredT 1 based on self.top 2000, Inf 1, GS only NORMALIZE.Dxt</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Sampleset RNA-Seq 1148 Normal-Tx clusteredT 1 NORMALIZE.Dxt</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>er_level_percentatagerecategory</td>
<td></td>
<td></td>
<td>&lt;20%</td>
</tr>
<tr>
<td>progesterone_receptor_level_percentatagerecategory</td>
<td></td>
<td></td>
<td>&lt;10%</td>
</tr>
<tr>
<td>UpdatedPAM50_ROR_S_Group/SubtypeOnly</td>
<td>med</td>
<td>high</td>
<td>med</td>
</tr>
<tr>
<td>UpdatedPAM50_ROR_P_Group/Subtype + Proliferation</td>
<td>low</td>
<td>high</td>
<td>med</td>
</tr>
<tr>
<td>number_of_lymphnodes_positive_by_hc</td>
<td>1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>number_of_lymphnodes_positive_by_hch</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UpdatedPAM50_Conference</td>
<td>1.0</td>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td>UpdatedPAM50_ROR_S_SubtypeOnly</td>
<td>-1.9</td>
<td>12.6</td>
<td>15.1</td>
</tr>
<tr>
<td>UpdatedPAM50_ProliferationScore</td>
<td>-0.8</td>
<td>-0.4</td>
<td>-0.4</td>
</tr>
<tr>
<td>UpdatedPAM50_ROR_P (Subtype + Proliferation)</td>
<td>-2.3</td>
<td>16.6</td>
<td>18.9</td>
</tr>
<tr>
<td>UpdatedPAM50_ER</td>
<td>1.7</td>
<td>1.6</td>
<td>1.5</td>
</tr>
<tr>
<td>UpdatedPAM50_Her2</td>
<td>0.4</td>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td>Infiltration_percent_normalcells_TOPt</td>
<td>3.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infiltration_percent_tumor_nuclei_TOPt</td>
<td>72.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survival - vs All Others</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recurrence - vs All Others</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Clusters
Summary

• Multi Omic correlation graph identifies features that may play a significant role in the underlying biological theme of the analyzed dataset.

• The generated multi-omic correlation graph can be used as an infrastructure for mining several omic datasets together.

• The method is highly extensible and makes no assumptions regarding the specific omic data it is applied to.