Comparative Analysis of Metabolic Pathways in Metagenomics

Bo Liu, Adam Bazinet, Marco Adelfio
CMSC858L Spring 2009
Outline

Background
  – Metagenomics, Metabolic Pathway

Data Set and Processing
  – Problem Statement, 6 Lean/Obese Twins,, Functional Annotation

Previous Approach

Our Methods
  1. Distance Calculation
  2. Discriminative Power Analysis
  3. Greedy Search Pathway Signature
  4. Pathway Motifs

Future Work
Metagenomics

• Genomics: DNA sequences of an individual organism

• Metagenomics: DNA sequences obtained directly from environmental samples (soil, ocean, human gut)

• Identify genes and metabolic pathways

• Compare to other microbial communities

http://www.mpi-bremen.de/en/Genomic_Standards_for_the_Future.html
Metabolic Pathway

- **Metabolic Pathway**: a series of biochemical reactions that occur inside a cell
- **Edges**: catalytic enzymes, proteins encoded by genes
- **Nodes**: molecular substrates, such as dietary minerals and vitamins
- **KEGG**: most widely used database
Problem Statement

• Find functional signatures of lean and obese people in the human gut microbiome

• Signatures could be differentially abundant organisms, metabolic pathways, or sets of genes

Turnbaugh et al 2009
## Data Set

<table>
<thead>
<tr>
<th>BMI</th>
<th>Subject ID</th>
<th>Reads Used</th>
<th>KEGG Annotated</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lean</td>
<td>TS1</td>
<td>217,386</td>
<td>56,458</td>
<td>25.97</td>
</tr>
<tr>
<td></td>
<td>TS2</td>
<td>443,640</td>
<td>78,083</td>
<td>17.60</td>
</tr>
<tr>
<td></td>
<td>TS4</td>
<td>414,754</td>
<td>103,979</td>
<td>25.07</td>
</tr>
<tr>
<td></td>
<td>TS5</td>
<td>490,776</td>
<td>94,913</td>
<td>19.34</td>
</tr>
<tr>
<td></td>
<td>TS7</td>
<td>555,853</td>
<td>137,462</td>
<td>24.73</td>
</tr>
<tr>
<td></td>
<td>TS8</td>
<td>414,497</td>
<td>108,086</td>
<td>26.08</td>
</tr>
<tr>
<td>Obese</td>
<td>TS19</td>
<td>498,880</td>
<td>94,089</td>
<td>18.86</td>
</tr>
<tr>
<td></td>
<td>TS20</td>
<td>495,040</td>
<td>105,072</td>
<td>21.22</td>
</tr>
<tr>
<td></td>
<td>TS28</td>
<td>302,780</td>
<td>70,494</td>
<td>23.28</td>
</tr>
<tr>
<td></td>
<td>TS29</td>
<td>502,399</td>
<td>133,639</td>
<td>26.60</td>
</tr>
<tr>
<td></td>
<td>TS49</td>
<td>519,072</td>
<td>93,868</td>
<td>18.08</td>
</tr>
<tr>
<td></td>
<td>TS50</td>
<td>549,700</td>
<td>131,356</td>
<td>23.90</td>
</tr>
</tbody>
</table>
Pathway Annotation

Sequencing Reads NCBI Short Read Archive

Pathway Annotation, KEGG

KO Assignment

EC Assignment

read1

read2

read3

BLASTX on UMIACS Condor, 5-6 Days

KO1

KO2

KO3

KO4

EC1

EC2

EC3

EC4

Path1

Path2

Reaction1

Reaction2

Metabolic Pathway

Reactions
An example KEGG pathway

GLYCOLYSIS / GLUCONEOGENESIS

Starch and sucrose metabolism

α-D-Glucose-1P

D-Glucose (extracellular)

Pentose phosphate pathway

Glycerate-3P

Carbon fixation in photosynthetic organisms
# Final Data Representation

<table>
<thead>
<tr>
<th>Pathway</th>
<th>KO Number</th>
<th>Start Molecule</th>
<th>End Molecule</th>
<th>TS1</th>
<th>TS2</th>
<th>TS4</th>
<th>TS5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>00010</td>
<td>K00001</td>
<td>C00084</td>
<td>C00469</td>
<td>63</td>
<td>84</td>
<td>80</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>00010</td>
<td>K00016</td>
<td>C00022</td>
<td>C00186</td>
<td>13</td>
<td>12</td>
<td>23</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>00010</td>
<td>K00134</td>
<td>C00118</td>
<td>C00236</td>
<td>93</td>
<td>91</td>
<td>77</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>00010</td>
<td>K00128</td>
<td>C00033</td>
<td>C00084</td>
<td>17</td>
<td>19</td>
<td>6</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>00010</td>
<td>K00150</td>
<td>C00118</td>
<td>C00236</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Weighted Pathway
Previous Approach

- Comparisons based on sum of counts in each pathway
Drawbacks of Previous Approach

• Fails to reveal finer-level pathway perturbations
• Disregards topology and molecular interactions
• Ignores between-pathway interactions
Z-score

- \( z = \frac{x - \mu}{\sigma} \)
  - \( z \) is the normalized score
  - \( x \) is the original score
  - \( \mu \) is the population mean
  - \( \sigma \) is the population standard deviation

- aka standard score or normal score
Z-score

• lean samples / obese samples

• raw counts of gene X:
  3  0  6  0  2  10  6  6  22  103  25  78

• Z-normalized counts of gene X:
  -0.56  -0.65  -0.47  -0.65  0.59  -0.35
  -0.47  -0.47   0.01   2.43  0.10   1.68
\( t\)-test

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}
\]

where

\[
S_{X_1X_2} = \sqrt{\frac{S_{X_1}^2 + S_{X_2}^2}{2}}
\]

- Independent two-sample \( t\)-test
- Equal sample sizes, equal variance
- Null hypothesis is that the two samples have the same mean
t-test example

- lean / obese
- raw counts of gene X:
  3 0 6 0 2 10 6 6 22 103 25 78
- t-score: 2.19
- probability: 0.08
- null hypothesis: not rejected (with 95% confidence)
Distance Calculation

• Would like a metric that gives the distance between two individual samples

• Existing network alignment techniques are inapplicable because metabolic network topologies are fixed

• Simple metric: sum up the differences in edge weights between two individual samples

• Normalize by total number of ‘reads mapped’ for a given sample
Normalization Example

- lean samples / obese samples

- raw counts of gene X:
  3 0 6 0 2 10 6 6 22 103 25 78

- counts normalized by ‘reads mapped’:
  0.00011 0.00000 0.00012 0.00000 0.00003 0.00020
  0.00013 0.00012 0.00067 0.00167 0.00056 0.00127

- Z-normalized counts:
  -0.54 -0.75 -0.53 -0.75 -0.69 -0.38
  -0.51 -0.53 0.49 2.31 0.29 1.58
Distance Calculation (naive)

- For each pair of samples, sum the difference in normalized scores across all genes
- Produces a distance matrix

<table>
<thead>
<tr>
<th></th>
<th>LEAN1</th>
<th>LEAN2</th>
<th>LEAN3</th>
<th>LEAN4</th>
<th>LEAN5</th>
<th>LEAN6</th>
<th>OBESE1</th>
<th>OBESE2</th>
<th>OBESE3</th>
<th>OBESE4</th>
<th>OBESE5</th>
<th>OBESE6</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEAN1</td>
<td>000</td>
<td>4285</td>
<td>5792</td>
<td>5023</td>
<td>3970</td>
<td>5818</td>
<td>5215</td>
<td>4461</td>
<td>4478</td>
<td>5104</td>
<td>3987</td>
<td>5385</td>
</tr>
<tr>
<td>LEAN2</td>
<td>4285</td>
<td>000</td>
<td>4278</td>
<td>5391</td>
<td>3812</td>
<td>3813</td>
<td>4134</td>
<td>3895</td>
<td>4767</td>
<td>5332</td>
<td>4836</td>
<td>5430</td>
</tr>
<tr>
<td>LEAN3</td>
<td>5792</td>
<td>4278</td>
<td>000</td>
<td>5692</td>
<td>4762</td>
<td>4981</td>
<td>2986</td>
<td>4483</td>
<td>5874</td>
<td>6262</td>
<td>5715</td>
<td>6157</td>
</tr>
<tr>
<td>LEAN4</td>
<td>5023</td>
<td>5391</td>
<td>5692</td>
<td>000</td>
<td>3968</td>
<td>3982</td>
<td>5104</td>
<td>4466</td>
<td>3874</td>
<td>4728</td>
<td>3338</td>
<td>4194</td>
</tr>
<tr>
<td>LEAN5</td>
<td>3970</td>
<td>3812</td>
<td>4762</td>
<td>3968</td>
<td>000</td>
<td>2726</td>
<td>4460</td>
<td>2915</td>
<td>3764</td>
<td>4623</td>
<td>3606</td>
<td>4359</td>
</tr>
<tr>
<td>LEAN6</td>
<td>3885</td>
<td>3813</td>
<td>4981</td>
<td>3982</td>
<td>2726</td>
<td>000</td>
<td>3987</td>
<td>5339</td>
<td>5936</td>
<td>5336</td>
<td>5869</td>
<td></td>
</tr>
<tr>
<td>OBESE1</td>
<td>5215</td>
<td>4134</td>
<td>2986</td>
<td>5104</td>
<td>4460</td>
<td>4529</td>
<td>000</td>
<td>3987</td>
<td>5339</td>
<td>5936</td>
<td>5336</td>
<td>5869</td>
</tr>
<tr>
<td>OBESE2</td>
<td>4461</td>
<td>3895</td>
<td>4483</td>
<td>4466</td>
<td>2915</td>
<td>3231</td>
<td>3987</td>
<td>000</td>
<td>4258</td>
<td>4951</td>
<td>4215</td>
<td>4907</td>
</tr>
<tr>
<td>OBESE3</td>
<td>4478</td>
<td>4767</td>
<td>5874</td>
<td>3874</td>
<td>3764</td>
<td>3510</td>
<td>5339</td>
<td>4258</td>
<td>000</td>
<td>4192</td>
<td>3413</td>
<td>4134</td>
</tr>
<tr>
<td>OBESE4</td>
<td>5818</td>
<td>5332</td>
<td>6262</td>
<td>4728</td>
<td>4623</td>
<td>4281</td>
<td>5936</td>
<td>4951</td>
<td>4192</td>
<td>000</td>
<td>4009</td>
<td>3497</td>
</tr>
<tr>
<td>OBESE5</td>
<td>4586</td>
<td>4836</td>
<td>5715</td>
<td>3338</td>
<td>3606</td>
<td>3579</td>
<td>5336</td>
<td>4215</td>
<td>3413</td>
<td>4009</td>
<td>000</td>
<td>3246</td>
</tr>
<tr>
<td>OBESE6</td>
<td>5459</td>
<td>5430</td>
<td>6157</td>
<td>4194</td>
<td>4359</td>
<td>3989</td>
<td>5869</td>
<td>4907</td>
<td>4134</td>
<td>3497</td>
<td>3246</td>
<td>000</td>
</tr>
</tbody>
</table>
Distance Calculation (naive)

gsi for **lean** (6 individuals) is: 0.19
P-value based on permutation test (10000 reps) is: 0.197

gsi for **obese** (6 individuals) is: 0
P-value based on permutation test (10000 reps) is: 1
Distance Calculation (improved)

- *t*-test each gene and *only include in the distance calculation if deemed significant* at e.g., 95%

- Z-normalized counts:
  -0.54  -0.75  -0.53  -0.75  -0.69  -0.38
  -0.51  -0.53  0.49  2.31  0.29  1.58

- *t*-score: 2.58  probability: 0.0496

- null hypothesis: *rejected* (with 95% confidence)
Distance Calculation (improved)

\[ \frac{192}{4231} \text{ genes selected} \] at 95% confidence

gsi for \textit{lean} (6 individuals) is: 0.69
P-value based on permutation test (10000 reps) is: 0.004

gsi for \textit{obese} (6 individuals) is: 1
P-value based on permutation test (10000 reps) is: 0.001

Monday, May 4, 2009
Discriminative Power Analysis

• Following Lee et al 2008:


• “For each pathway, an activity level is summarized from the gene expression levels of its condition-responsive genes (CORGs), defined as the subset of genes in the pathway whose combined expression delivers optimal discriminative power for the phenotype.”
Figure 1. A schematic diagram of key gene identification and activity inference. Selected significant pathways are further subject to CORG identification corresponding to the phenotype of interest. Gene expression profiles of patient samples drawn from each subtype of diseases (e.g., good or poor prognosis) are transformed into a “pathway activity matrix”. For a given pathway, the activity is a combined z-score derived from the expression of its individual key genes. After overlaying the expression vector of each gene on its corresponding protein in the pathway, key genes which yield most discriminative activities are found via a greedy search based on their individual power (see Methods). The pathway activity matrix is then used to train a classifier.

Lee et al 2008
Types of Markers (in Lee et al)

- **PAC** - ‘pathway activity inference using condition-responsive genes’ (CORG set)
- **PAC_all** - like PAC, but use *all* genes in each pathway
- **CORGs** - consider CORGs separately, averaging their individual *t*-scores
- **Genes** - like CORGs, but disregard pathway constraints and take top-ranked genes across *all pathways* equal to the total number of CORGs
Types of Markers (our data*)

- **PAC, PAC_all, CORGs, Genes** - as in Lee et al, except set-growing termination condition relaxed

- **PAC_genes** - like PAC, but disregards pathway constraints and uses top-ranked genes across all pathways

- **PAC_modules** - like PAC, but with the constraint that genes must be topologically connected

  * Edges connecting to ‘ubiquitous molecules’ are ignored
Lean/Obese Discriminative Power Values (EC Data)

Legend
- PAC
- PAC_module
- Genes
- CORGs
- PAC_all
- PAC_genes

average t-score ↑

% of pathways used →

Monday, May 4, 2009
Lean/Obes Discriminative Power Values (KO Data)

Legend
- PAC
- PAC_module
- Genes
- CORGs
- PAC_all
- PAC_genes

average t-score ↑

% of pathways used →

Monday, May 4, 2009
Finding Subpathways

• A module, better termed *subpathway*, is simply a CORG set with the additional constraint that the genes (which map to edges) must be connected

• Allow subpathways to grow through edges that don’t change the t-score (likely duplicate ECs or KOs)

• Cycle through genes as many times as necessary to ensure subpathway is of maximal size (illustrate why)
Summary Statistics

- EC data, CORG sets
  average size: 3.32
  average t-score: 3.12

- EC data, subpathways
  average size: 2.03
  average t-score: 2.31

- KO data, CORG sets
  average size: 5.22
  average t-score: 5.31

- KO data, subpathways
  average size: 2.95
  average t-score: 3.49

- Size and t-score are positively correlated
Finding Pathway Signature

• Genes function in concert rather than alone
• Interaction network is more informative than lists of genes
• Potential to detect more subtle signals
• A subnetwork: differentially abundant
Scoring Function

• $t$-test for each gene

<table>
<thead>
<tr>
<th></th>
<th>Lean</th>
<th>Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td>K00001</td>
<td>63</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>84</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>101</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>112</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>188</td>
</tr>
</tbody>
</table>

• Each gene: $t$-value and $p$-value

Ideker et al, 2002
Scoring Function

• Assumption: student distribution is close to normal.

• $z = \text{CDF}^{-1}(1-p)$, larger $z$-scores $\rightarrow$ smaller $p$-values $\rightarrow$ more significant.

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i$$

$$\text{Mean}(Z_A) = \frac{1}{\sqrt{k}} \sum_{i \in A} \text{Mean}(z_i) = 0$$

$$\text{Var}(Z_A) = \frac{1}{\sqrt{k}^2} \sum_{i \in A} \text{Var}(z_i) = 1$$

• $Z_A \sim \text{N}(0,1)$, $Z_A$ is the score of a subnetwork

Ideker et al, 2002
Greedy Search

- Start from a significant gene
- Find the best neighbor
- Calculate $Z_A$
- If significant (p-value: 0.05 = Z-value: 1.645)
- Stop when a subnetwork is not significant
Monday, May 4, 2009
Future Work

• $t$-test alternatives, because many $t$-tests $\rightarrow$ high false positives

• Edge direction

• Advanced search algorithm
Network Motifs

- *Subgraphs* that occur more or less frequently than expected
- Typically identified only using topology
- Topology in metabolic networks is fixed, only edge weights change
- We need a weighted motif measure
Motif Counting

• Standard method:

• Find all small subgraphs that occur much more (or less) often than in a random graph
Motif Intensity

• Instead of counting occurrences, we measure motif intensity:

\[ I(g) = \left( \prod_{(ij) \in \ell_g} w_{ij} \right)^{1/|\ell_g|} \]

• Introduced by Onnela et al. in 2005. Two applications in original paper: stock correlations and throughput of metabolic reactions.
Motif Intensity

• Intuition: motif intensity should gradually approach zero as any single edge weight decreases to zero
Motif Intensity

- Enumerate all 3- and 4-node motifs present in any sample (using EC data)

- Calculate the intensity of every motif for individual samples

- If a particular motif occurs more than once, sum the intensities

- For each motif, perform $t$-test to compare intensities between lean and obese samples
# Results (Absolute Intensity)

<table>
<thead>
<tr>
<th>Motif</th>
<th>Count</th>
<th>P-value</th>
<th>Higher Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Motif 1" /></td>
<td>17</td>
<td>.0137</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image2.png" alt="Motif 2" /></td>
<td>22</td>
<td>.0153</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image3.png" alt="Motif 3" /></td>
<td>24</td>
<td>.0266</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image4.png" alt="Motif 4" /></td>
<td>29</td>
<td>.0297</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image5.png" alt="Motif 5" /></td>
<td>17</td>
<td>.0383</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image6.png" alt="Motif 6" /></td>
<td>13</td>
<td>.0481</td>
<td>Obese</td>
</tr>
</tbody>
</table>

* No significant 3-node motifs
## Results (Relative Intensity)

<table>
<thead>
<tr>
<th>Motif</th>
<th>Count</th>
<th>P-value</th>
<th>Higher Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Motif 1" /></td>
<td>17</td>
<td>.0061</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image2" alt="Motif 2" /></td>
<td>22</td>
<td>.0089</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image3" alt="Motif 3" /></td>
<td>164</td>
<td>.0104</td>
<td>Lean</td>
</tr>
<tr>
<td><img src="image4" alt="Motif 4" /></td>
<td>24</td>
<td>.0135</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image5" alt="Motif 5" /></td>
<td>17</td>
<td>.0180</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image6" alt="Motif 6" /></td>
<td>29</td>
<td>.0298</td>
<td>Obese</td>
</tr>
<tr>
<td><img src="image7" alt="Motif 7" /></td>
<td>52</td>
<td>.0312</td>
<td>Obese</td>
</tr>
</tbody>
</table>
Motifs - Future Work

• Other Topological Measures
  • Measure weighted clustering coefficient
  • Measure coherence instead of intensity

• Algorithm Tweaks
  • Incorporate hyperedge nature of edges
  • Detect multi-graph motifs
Conclusion

• Given samples from the gut microbiome of lean and obese human individuals

• We identified quantitative differences between phenotypes using several methods:
  1. Distance Calculation
  2. Discriminative Power Analysis
  3. Greedy Search Pathway Signature
  4. Pathway Motifs

• Unlike previous approaches, we incorporated pathway topology into our methods