Gene networks
Real-life examples
Biological networks

• Genes/proteins do not exist in isolation
• Interactions between genes or proteins can be represented as graphs
• Examples:
  – metabolic pathways
  – regulatory networks
  – protein-protein interactions (e.g. yeast 2-hybrid)
  – genetic interactions (synthetic lethality)
Gene networks research at UMD

• Active area of research in Carl Kingsford's lab
• Data will be generated in Najib El Sayed's lab
• My own research on microbial communities will translate into such data.
Metagenomics
Human microbiome

  

- Examine all bacteria in an environment (human gut) at the same time using high-throughput techniques
Why the gut biome? We are what we eat

• Majority of human commensal bacteria live in the gut (more bacterial cells than human cells by an order of magnitude – 100 trillion bacterial cells)
• We rely on gut bacteria for nutrition

• Gut bacteria important for our development

• Imbalances in bacterial populations correlate with disease
Environment “exploration”

- **Culture-based**
  - heavily biased (1-5% bacteria easily cultured)
  - amenable to many types of analyses

- **Directed rRNA sequencing**
  - less biased
  - limited analyses possible

- **Random shotgun sequencing**
  - “differently” biased
  - amenable to many types of analyses
  - $$$
Project overview

• Collaboration between TIGR, Stanford, and Washington University (St. Louis)
• Sequenced fecal samples from two healthy individuals (XX, XY) (veg+, veg-) correlation lost due to IRB
• Also performed “traditional” amplified 16S rDNA sequencing

<table>
<thead>
<tr>
<th></th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shotgun reads</td>
<td>65,059</td>
<td>74,462</td>
<td>139,521</td>
</tr>
<tr>
<td>amplified 16S rDNA clones</td>
<td>3,514</td>
<td>3,601</td>
<td>7,115</td>
</tr>
</tbody>
</table>

All shotgun reads from ~ 2 kbp library
Metagenomic pipeline

- Assembly (graph theory, string matching)
  - puzzle-together shotgun reads into contigs and scaffolds
- Gene finding (machine learning)
- Binning (clustering, statistics)
  - assign each contig to a taxonomic unit
- Annotation (natural language processing)
  - gene roles, pathways, orthologous groups, etc
- Analysis (statistics, graph theory, data visualization)
  - diversity
  - comparison between environments
  - metabolic potential
  - etc.
Comparative Assembly (AMOScmp)

<table>
<thead>
<tr>
<th></th>
<th>Genome size</th>
<th>Coverage</th>
<th># contigs</th>
<th># bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bifidobacterium longum</td>
<td>2.26 MB</td>
<td>0.7</td>
<td>789</td>
<td>988,707</td>
</tr>
<tr>
<td>Methanobrevibacter smithii</td>
<td>~1.9 MB</td>
<td>3.5</td>
<td>222</td>
<td>1,538,516</td>
</tr>
</tbody>
</table>

> 50% of archaeal contigs are likely *M. smithii*
## Binning results

<table>
<thead>
<tr>
<th>Order</th>
<th>amplified rRNA clones</th>
<th>shotgun rRNA (bases)</th>
<th>shotgun blastx(bases)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subject 1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Clostridiales</td>
<td>2,777</td>
<td>3,386</td>
<td>70,055</td>
</tr>
<tr>
<td>Bifidobacteriales</td>
<td>30</td>
<td>0</td>
<td>31,443</td>
</tr>
<tr>
<td>Coriobacteriales</td>
<td>4</td>
<td>6</td>
<td>25,781</td>
</tr>
<tr>
<td>Methanobacteriales</td>
<td>0</td>
<td>0</td>
<td>18,188</td>
</tr>
</tbody>
</table>
Metagenomics...

• This work is ongoing at UMD with support from NSF and NIH

• Paid summer internships available – contact me if you are interested.
Assembly with optical maps
Optical mapping data

- Restriction mapping
  (set/bag of fragment sizes)
  - restriction digest
  - spectrum of sizes defines “fingerprint”

<table>
<thead>
<tr>
<th>#</th>
<th>size (stdev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.2 (0.3)</td>
</tr>
<tr>
<td>2</td>
<td>4.1 (0.8)</td>
</tr>
<tr>
<td>3</td>
<td>2.2 (0.5)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Optical mapping
  (list/array of fragment sizes)
  - ordered restriction digest
  - order of fragment sizes defines fingerprint
Contig matching problem

- Find “best” placement of a contig on the map

\[ \chi^2 \text{ score} = \sum_{k=1}^{j} \left( \frac{c_k - o_k}{\sigma_k} \right) \]

- by best we mean:
  - most matched sites
  - best correspondence between fragment sizes
Solution to the matching problem

- Simple dynamic programming ($O(m^2n^2)$)

$$S[i, j] = \max_{0 \leq k \leq i, 0 \leq l \leq j} -C_r \times (i - k + l - j) - \frac{(\sum_{s=k}^{i} c_s - \sum_{t=l}^{j} o_t)^2}{\sum_{t=l}^{j} \sigma_t^2} + S[k-1, l-1]$$

- Main challenge: this procedure always returns a "best" match

- Solution:
  - Compute P-value: likelihood a random match would be as good as the best
  - Randomized bootstrapping: randomly permute contig and find best match...
Results – real data

Yersinia kristensenii
Optical map: 350 sites (AFLII)
Assembly: 86 contigs, 404 sites
48 contigs have > 1 site
45 contigs can be placed
  30 unique matches
  15 placed by greedy
4.4Mb (93%) in scaffold

Yersinia aldovae
Optical map: 360 sites (AFLII)
Assembly: 104 contigs, 411 sites
58 contigs have > 1 site
52 contigs can be placed
  31 have unique matches
  21 placed by greedy
3.7Mb (88%) in scaffold

Un-placed contigs appear to be mis-assemblies

With Niranjan Nagarajan
Voxelation
Voxelation


- Gene expression information in a spatial context
- Combines microarray analysis with computer graphics
Figure 2  Voxelation scheme

• Mouse brain cut up into voxels
• Run a separate microarray experiment on each voxel
Figure 4  Spatial gene expression patterns for the subset of correlated genes

Figure 7 SVD delineates anatomical regions of the brain

Figure 5  Putative regulatory elements shared between groups of correlated and anticorrelated genes
Figure 6 Differentially expressed genes

[Graphs and data plots showing gene expression levels and statistical analyses.]

Research at UMD

• Possible future work with Amitabh Varshney (CS) and Cristian Castillo-Davis (Biology)