**METAGENOMICS**

**My Lab**

**Traditional Genomics**

- Environmental Sample
- Isolate Organism and Culture

---

**Once cultured -- Sequence**

- Whole Genome Shotgun Sequencing Method
- Only one possible genome in
- So, only one possible out

---

**Metagenomics - No longer “need” to culture**

- Each is different sequence

---

**What’s this good for?**

**Types of Environmental Samples**

- From the ocean: (e.g. Sargasso Sea very salty and thus nutrient deprived)
- Venter et al. (2004) sampled it and found diverse prokaryotic life despite doubts.
- From the human body: (e.g. our gastrointestinal tract)
- From extinct, ancient animals (e.g. separate a Mammoth’s DNA from the other sample’s 40% of bugs and plants)

---

**MICROBES**

**INVISIBLE INVADERS**

**AMAZING ALLIES**

Why do we want to know the sample composition?
Trillions are depending on you (mostly - you’re depending on them):
The Human Microbiome

- You are 100 trillion cells (yours have a nucleus and therefore, you are eukaryotic)
- You contain 1000-2000 trillion Microbes (10/20-to-1) (~ they have no nucleus ~ they are prokaryotic)
- Microbes make up 1-2% of your body mass
- Every gram of your intestines have ~100 billion bacteria
- Microbes in our intestines help produce certain essential vitamins such as vitamin B12.
- Most microbes deny disease-causing microbes the foothold they need to colonise our bodies and do damage

Implications that can revolutionize technology

- Can’t use plants as fuel because don’t know how to break it down
- Termites’ guts can digest cellulose efficiently and “fighteningly quickly”

The new anti-cancer drug -- Tunicates?

Scientists have shown how a microbe that lives inside sea squirts could be used to biosynthesize a chemical compound that may help fight cancer.

(Submitted by the Duquesne Allegheny Institute)

C. Pelagibacter -- most compact genome on earth

- Can’t culture – only metagenomic reads
- Estimated 20 billion Pelagibacter microbes
- No junk
- The fewest genes “life” can get away-with and still be efficient

Microbes in Ants and Fruit Flies

- Dr. Russell’s research
- Co-evolution of arthropod gut microbes with different arthropod lineages
- 3:30 Talk by Noah Whiteman in Stratton 113 on ‘ecology and evolution of host-symbiont interactions’
Precursory Steps to Analysis

- Obtain Environmental Sample (Do not taint!)
- Sequence the Sample / DNA Microarrays (detect gene or gene expression) / Metaproteomics (mass spectroscopy to identify proteins in sample)
- Sequencing: Traditionally (standard for 30 years), Sanger (Chain-Termination Method used)
  - ~850 bp read (usually good because unique and have less “chunks” to sequence)
  - Needs an amplification step (cloning) -- SLOW

New Sequencing Technologies: Short and Fast

- 454 Pyrosequencing
  - 100-250 bp reads
- Solexa Sequencing-by-Synthesis
  - 25-40 bp reads

High Throughput Sequencing

- Capillary electrophoresis (Sanger)
  - Between 96 and 384 samples (76-308 Kohr/run)
- Solexa
  - 30M samples (2Gb / run)
- Solid
  - 40M samples (2-6 Gb / run)

Sequencers generate fragments (sampling underlying distribution)

Problem: If billions of organisms with 100’s of trillions of nucleotides (and we can sample a billion bp on a good day) -- what is Nyquist for this type of problem?

First step: Reads need to be classified

Parallel of Speech Processing to Metagenomic Processing
**Signal Processing for Recognition**

* Speech Recognition (Gene and Function)
* Speaker Identification (Taxa ID)

\[ y = s + n \]  
(signal + noise)

noise is assumed to be Gaussian

Who’s voice? (Which taxa?)

**Noise in Taxa Recognition**

\[ y, \text{DNA reads, are assumed noiseless (not true assumption)} \]

Noise is in the mapping from sequence to Taxa

<table>
<thead>
<tr>
<th>Reads</th>
<th>Burkholderia</th>
<th>Clostridium</th>
<th>E. Coli</th>
<th>Staphylococcus</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAAAAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTAGTTAGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTAGTTAGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTAAGTTAGAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAGTTAGATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Speaker Identification Problems and their parallel - I**

- How do you account for noise in the signal?
  - Microphone noise, environmental “noises”
  - Our equivalent -- Sequencing error, “background DNA” (currently everything is of interest)
  - Sequence-to-Taxa mapping

**Speaker Identification Problems and their parallel - II**

- How do you account for the different volumes of speakers
  - Imbalanced amounts of Taxa
  - Not the nebulous “Normalization of Metagenomic data”
- How do you characterize a voice?
  - Spectral Formant Features are “Fingerprint” of Voice
  - Genomic features, Genes that characterize a Taxa

**Speaker Identification Problems and their parallel - II**

- How does the system handle varying speeds of inputs
  - Time-Frequency trade-off in Speech
  - Same with metagenomic reads -- shorter the read, the “less resolution” we have
- How can you account for imitating speech patterns?
  - Mimicking and Interference
  - Close strains and species

**Classified as what?**

- Taxa
- Genes
- Function
2007: 1.8 million species known to Science

- A species is often defined as a group of organisms capable of interbreeding and producing fertile offspring.
- Does not apply to asexual single-cells
- Horizontal gene transfer violates the getting genes only from parents assumption (Vertical)

Traditional Taxa Classification - 16S

- 1200 bp average length
- 3' end of 16S RNA bind to mRNA to start translation

1. Universally distributed, allowing the comparison of phylogenetic (tree-of-life) relationships (present in all 3 kingdoms)
2. Core of information genes which are only weakly affected by horizontal gene transfer
3. Functionally highly constrained mosaics of sequence stretches ranging from conserved to more variable (most organisms will die with too high of mutation)

Taxonomy Standards Based on 16S

- Intra-species 16S variation -- 5%
- New: Can sometimes be as low as 1% and as high as 5%
- Intra-genus 16S variation -- not well studied but can be 6-9%

16S tells part of the story

- 1200bp out of millions of bases (other parts of genome may contain clues)
- Other parts may be laterally transferred

Extremely Mobile Elements: Horizontal Gene Transfer

- Like your neighbors drug-resilience?
- Get a few plasmids from them

Anomalies
Current 16S Techniques

Multimer vs. BLAST (Composition vs. Homology)

- A Multimer is a string of length N (or what we refer to as Nmer)
  - A= 1mer, AT = 2mer, ATG = 3mer, ATCG = 4mer, etc.
- Multimer approaches “count Nmer frequencies”. These features characterize the full gene.
- Classifies according to Probability of finding Multimers in sequence. The more a sequence diverges, the less likely this is to happen.
- BLAST uses dynamic programming – maximize score between two sequences. (and find sequence similarity)
- Arguably more robust to a high-level of mutations

Least Common Ancestor

Assigns each read to the lowest common ancestor (LCA) of the set of taxa that it hit in the comparison

Techniques for full-genomes

<table>
<thead>
<tr>
<th>Features</th>
<th>Classifier</th>
<th>Published Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homology-based</td>
<td>Nearest Neighbor</td>
<td>BLAST [34]</td>
</tr>
<tr>
<td></td>
<td>Nearest Neighbor &amp; Least Common Ancestor</td>
<td>MEGAN [35]</td>
</tr>
<tr>
<td>Composition-based</td>
<td>Naive Bayesian</td>
<td>RUP classifier (16S sequences only) [56]</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines</td>
<td>Resen et al. [57]</td>
</tr>
</tbody>
</table>

Processing: Identify the Content from Reads

- Identify the Kingdom, Phyla, Class, Genera, and Species distribution of a sample
- Example in Matlab (2008a) Metagenomics Demo:
  - Analyzes first 100 reads of the Sargasso Sea (each around 850 bp)

Tree: BLAST + Least Common Ancestor

MEGAN (MetEraGenome ANalyzer): a graphical tool that uses BLAST reports to assess content of a sample

- Ambiguous hits for SHORT fragments
- Assigns each read to the lowest common ancestor (LCA) of the set of taxa that it hit in the comparison
- Top-percent filter is used to retain only those hits for a given read r whose scores lie within a given percentage of the highest score involving r.