Computational Methods in Metagenomics and Microbiome Research

0368-3116-01

Prof. Elhanan Borenstein
School of Computer Science
Semester B, 2019
Outline

A Few Updates about Seminar Logistics

“Metagenomics”

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics
A Few Updates about Seminar Logistics

“Metagenomics”

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics
Seminar Format

- ~15 students registered
- 1 paper per student, 1-2 students/papers per class

Class structure:
- 1:10 Start talk 1 (35 minutes!!)
- 1:45-2:00 Feedback/Discussion
- 2:00-2:10 Break
- 2:10 Start talk 2 (35 minutes!!)
- 2:45-3:00 Feedback/Discussion

Paper selection:
- List posted: Sunday, March 10, 9:00
- Please select your paper by: Monday, 17:00
- First talk: March 20
Presentations

- In your presentation:
  - Emphasize the main task of the paper
  - Cover required background
  - Focus more (but not only) on methods (may need to dig)
  - **Choose wisely what to cover and how deeply**
  - Summarize briefly
  - Add something original

- Remember the dos and don’ts of a good talk and of slide design

- Keep us interested
Other Tasks

- **As a speaker**
  - Come early (or on a different day) to make sure your presentation works
  - Print slides for everybody
  - Send me slides to post online
  - Prepare discussion points (2-3 points, NOT in paper)

- **As a listener**
  - Read/skim the papers before class
  - Listen, take notes, engage
  - Participate in feedback/discussion
  - Learn and have fun
Outline

A Few Updates about Seminar Logistics

“Metagenomics”

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics
Microbial communities
- Hundreds of species!
- 100 trillion microbes! (weighing ~3-4lb)
- 150x more genes (~3,300,000)
metagenomics

“The study of genetic material recovered directly from environmental samples”
(the term was first used by Jo Handelsman in 1998)
Outline

A Few Updates about Seminar Logistics

“Metagenomics”

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics
Genomes and Genes: A two minute conceptual view

Gene 1: cacatgtaccaggagaaatgaaga
Gene 2: caatttgtaacacctttagacttttcatcaacctttctagtgtcactttttgcc
Gene 3: agugucacuuuuuuugcc

Genome: ...cacatgtaccaggagaaatgaagacaatttgtagaacctttcatcaacctttctagtgtcacttttttgcc...

Transcripts: cacauguaccaggagaaauugaaga
Gene 2: uuuuagacuu
Gene 3: agugucacuuuuuuugcc

Molecular Tools:
Genomes and Genes: In Real Life

DNA

Proteins

RNA

Ribosome

Molecular Tool:

Transcripts:

cacuguaccaggagaauga

uccnacuu

agugcacu우우우 우 gotta

Gene 1

Gene 2

Gene 3

Genome: cacatgtaccaggagaaatgaagaacattttaagacttctcatcnaaatttctgctactttgttttggot...

0.34 nm

3.4 nm

0.34 nm

2 nm

Phosphorus
Carbon in sugar-phosphate "backbone"
Hydrogen
Oxygen
Bases

Messenger RNA

4at1 Aspartate Carbamoyltransferase
1rcx Ribulose Bisphosphate Carboxylase/Oxygenase
1n2c Nitrogenase
1gax Valyl-tRNA Synthetase
1g6f Threonyl-tRNA Synthetase
1euq Glutaminyl-tRNA Synthetase
1ttt Elongation Factor

DNA
Variation in Genome Length and # Genes
## Variation in Genome Length and # Genes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size (base pairs)</th>
<th>Protein Coding Genes</th>
<th>Number of Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model Organisms</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>model bacteria <em>E. coli</em></td>
<td>4.6 Mbp</td>
<td>4,300</td>
<td>1</td>
</tr>
<tr>
<td>budding yeast <em>S. cerevisiae</em></td>
<td>12 Mbp</td>
<td>6,600</td>
<td>16</td>
</tr>
<tr>
<td>fission yeast <em>S. pombe</em></td>
<td>13 Mbp</td>
<td>4,800</td>
<td>3</td>
</tr>
<tr>
<td>amoeba <em>D. discoideum</em></td>
<td>34 Mbp</td>
<td>13,000</td>
<td>6</td>
</tr>
<tr>
<td>nematode <em>C. elegans</em></td>
<td>100 Mbp</td>
<td>20,000</td>
<td>12 (2n)</td>
</tr>
<tr>
<td>fruit fly <em>D. melanogaster</em></td>
<td>140 Mbp</td>
<td>14,000</td>
<td>8 (2n)</td>
</tr>
<tr>
<td>model plant <em>A. thaliana</em></td>
<td>140 Mbp</td>
<td>27,000</td>
<td>10 (2n)</td>
</tr>
<tr>
<td>moss <em>P. patens</em></td>
<td>510 Mbp</td>
<td>28,000</td>
<td>27</td>
</tr>
<tr>
<td>mouse <em>M. musculus</em></td>
<td>2.8 Gbp</td>
<td>20,000</td>
<td>40 (2n)</td>
</tr>
<tr>
<td>human <em>H. sapiens</em></td>
<td>3.2 Gbp</td>
<td>21,000</td>
<td>46 (2n)</td>
</tr>
<tr>
<td><strong>Bacteria</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. ruddii</em> (smallest genome of an endosymbiont bacteria)</td>
<td>160 kbp</td>
<td>182</td>
<td>1</td>
</tr>
<tr>
<td><em>M. genitalium</em> (smallest genome of a free living bacteria)</td>
<td>580 kbp</td>
<td>470</td>
<td>1</td>
</tr>
<tr>
<td><em>H. pylori</em></td>
<td>1.7 Mbp</td>
<td>1,600</td>
<td>1</td>
</tr>
<tr>
<td><em>Cyanobacteria S. elongatus</em></td>
<td>2.7 Mbp</td>
<td>3,000</td>
<td>1</td>
</tr>
<tr>
<td>methicillin-resistant <em>S. aureus</em> (MRSA)</td>
<td>2.9 Mbp</td>
<td>2,700</td>
<td>1</td>
</tr>
<tr>
<td><em>B. subtilis</em></td>
<td>4.3 Mbp</td>
<td>4,100</td>
<td>1</td>
</tr>
<tr>
<td><em>S. cellulosum</em> (largest known bacterial genome)</td>
<td>13 Mbp</td>
<td>9,400</td>
<td>1</td>
</tr>
<tr>
<td><strong>Eukaryotes - Multicellular</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pufferfish <em>Fugu rubripes</em> (smallest known vertebrate genome)</td>
<td>400 Mbp</td>
<td>19,000</td>
<td>22</td>
</tr>
<tr>
<td>poplar <em>P. trichocarpa</em> (first tree genome sequenced)</td>
<td>500 Mbp</td>
<td>46,000</td>
<td>19</td>
</tr>
<tr>
<td>corn <em>Z. mays</em></td>
<td>2.3 Gbp</td>
<td>33,000</td>
<td>20 (2n)</td>
</tr>
<tr>
<td>dog <em>C. familiaris</em></td>
<td>2.4 Gbp</td>
<td>19,000</td>
<td>40</td>
</tr>
<tr>
<td>chimpanzee <em>P. troglodytes</em></td>
<td>3.3 Gbp</td>
<td>19,000</td>
<td>48 (2n)</td>
</tr>
<tr>
<td>wheat <em>T. aestivum</em> (hexaploid)</td>
<td>16.8 Gbp</td>
<td>95,000</td>
<td>42 (2n=6x)</td>
</tr>
<tr>
<td>marbled lungfish <em>P. aethiopicus</em> (largest known animal genome)</td>
<td>130 Gbp</td>
<td>unknown</td>
<td>34 (2n)</td>
</tr>
<tr>
<td>herb plant <em>Paris japonica</em> (largest known genome)</td>
<td>150 Gbp</td>
<td>unknown</td>
<td>40 (2n)</td>
</tr>
</tbody>
</table>
A Bit More About Genes (≈1 course)

- Different species differ in the set of genes they encode and in the exact sequence of each gene.

**Homology:**

- Homologues genes are genes that derive from a common ancestor.
- Orthologues genes are homologous in different species (arise via speciation).
- Paralogues are homologous genes in the same species (arise via gene duplication).

- As species evolve, their genomes (and genes) gradually diverge in sequence.
- The closer the sequences of two genes are, the more likely it is the proteins they encode function similarly.
Sequencing, Alignment, Assembly (≈1 course)

...cagccttgcagetaccaggagaaaatgaactttttcatcaactttctagtgtcactttttttgcc...

Replicate

Shred

Sequence

Align

...cagccttgcagetaccaggagaaaacttttttgcc...

Assemble

cagccttgcagetaccaggagaaaacttttttgcc...
Outline

A Few Updates about Seminar Logistics

“Metagenomics”

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics
“Genomics”

Real Genomics

Replicate

Shred

Sequence

Align

Assemble

...cagctttgcagetaccaggagaaatatgaacttttcataacttctcaggtcacatccccccc...

tttcatcaacctt
acttttcact
cttacggg
tgassocggagaa
tcaaacctttttg
tccatcaacctt
tcaaacctttttg
tcccatac
tgcttcagctttttg
...cagctttgcagetaccaggagaaatatgaacttttcataacttctcaggtcacatccccccccccc...

cgctttgcag
tgcageta

aggagaaaaac
cgctttgcag
tgcageta

aggagaaaaa

aggagaaaaac
“Metagenomics”

Who’s there?

What are they doing?

Horns

Wings

“Metagenomics”
‘Real’ Metagenomics

Who’s there?

Microbiome sample

Extract DNA

Targeted sequencing (16S)

Cluster into species or OTUs

Species/Taxonomic composition

Shotgun sequencing

Map to a gene database (annotate)

BLASTx

Functional composition & comparative analysis

What are they doing?

Species/Taxonomic composition

Functional composition & comparative analysis

Whole rail

Sesarose Sx

Agricultural soil

Mouse gut

Human gut 2

Human gut 1

KEGG pathways

Carbohydrate metabolism

Energy metabolism

Nucleotide metabolism

Amino acid metabolism

Metabolism of other amino acids

Glycolysis/Gluconeogenesis

Biosynthesis of polyketides and nonribosomal peptides

Metabolism of cofactors and vitamins

Biosynthesis of secondary metabolites

Xenobiotics biodegradation and metabolism
Outline

A Few Updates about Seminar Logistics

“Metagenomics”

Genomes, Genes, and the Central Dogma

Real Metagenomics

Computational Challenges in Metagenomics
Key Challenge 1

Targeted sequencing (16S)

Extract DNA

Shotgun sequencing

Microbiome sample

Who's there?

Functional composition & comparative analysis

Species/Taxonomic composition

What are they doing?

From 16S Sequences to Taxa Composition

- Clear clusters are not always feasible
- True variation vs. sequencing errors?
- Low resolution
- A fixed threshold doesn’t always reflect the same phylogenetic closeness
  - Different species clustered together
  - Strains clustered separately
Key Challenge 2

Microbiome sample

Extract DNA

Shotgun sequencing

Targeted sequencing (16S)

Cluster into species or OTUs

Species/Taxonomic composition

Who’s there?

From Shotgun Sequencing to Species- and Strain-Level Profile

- Different species encode different genes
- Low abundance strains
- Shared variation

What are they doing?
Key Challenge 3

Microbiome sample

Extract DNA

Targeted sequencing (16S)

- AGAGCTGCTCA
- AGAGCTGCTCA
- AGAGCTGCTCA
- AGAGCTGCTCA
- AGAGCTGCTCA

Cluster into species or OTUs

Species/Taxonomic composition

Who's there?

Shotgun sequencing

- AGGAGCTGCTCA
- AGGAGCTGCTCA
- AGGAGCTGCTCA
- AGGAGCTGCTCA
- AGGAGCTGCTCA

Map to a gene database (annotate)

BLASTx

Comparative analysis

- Compositionality
- Phylogeny
- Distance metrics

What are they doing?

Competing Taxonomic Profiles

- Compositionality
- Phylogeny
- Distance metrics
Key Challenge 4

Who's there?

Microbiome sample

Extract DNA

Targeted sequencing (16S)

Shotgun sequencing

Cluster into species or OTUs

Map to a gene database (annotate)

Inferring Species Interactions

- Not phenotypic data (uncultured species)
- Low abundance species
- Sparsity
- Complex dynamics
- Compositionality