Metagenomics

02-715 Advanced Topics in Computational Genomics
Metagenomics

• Population sequencing

• Goal: recover the genomic sequences of the species in genetically diverse environmental samples
  – Human gut, honey bees, corals, ecosystems
  – Cancer tumor cells, pathogen populations such as HIV viral strains
  – Potential discovery of new species
Metagenomics

• Challenges
  – Assembly of a large number of relatively short and noisy reads of the DNA in a sample from next generation sequencers
  – Uncertainty of the population’s size and composition
  – Uneven coverage across species: coverage is affected by the species’ frequency in the sample
  – Reconstructing sequences even for the low-coverage species
Single Genome Sequencing

• The sequence assembly algorithm for single genome sequencing does not work well for metagenome assembly, in the presence of sequencing errors
Genovo

- De novo sequence assembler

- Model-based approach
  - a generative probabilistic model of read generation from environmental samples is specified
  - Captures the uncertainty of the population structure and noise model of the sequencing technology
  - Chinese restaurant process for the unknown number of genomes in the sample
Generative Model for Metagenome Assembly

- Generative models for
  - **Contigs** $\{b_{so}\}$ for letters at positions $o$ of contig $s$
  - **Reads** $x_i$’s within contig number $s_i$ and starting location $o_i$ within the contig
  - **Alignment** $y_i$ (orientation, insertions, deletions) for matching $x_i$ to the contig

- Assume an infinite number of contigs in prior, but in the posterior only a finite number of contigs are supported by reads
Generative Model for Metagenome Assembly I

- Contigs: Infinitely many letters in infinitely many contigs are sampled uniformly
  \[ b_{so} \sim \text{Uniform}(B) \quad \forall s = 1\ldots\infty, \forall o = -\infty\ldots\infty \]
  \[ B = \{A,C,G,T\} \]
  - Assume an infinite number of contigs with infinite number of nucleotides in each contig
Generative Model for Metagenome Assembly II

• Assignment of reads to contigs: $N$ empty reads are partitioned between the contigs

$$s \sim \text{CRP}(\alpha, N)$$

– Assignment of reads to contigs using Chinese restaurant process (preferential attachment)

$$p(s_N = s \mid s_{-N}) = \frac{1}{N - 1 + \alpha} \cdot \begin{cases} N_s & s \text{ is an existing cluster} \\ \alpha & s \text{ represents a new cluster} \end{cases}$$

• $N_{-i,s}$: the number of reads in contig (cluster) $s$, excluding read $i$

• The probability of assigning the given read to a new contig is proportional to $\alpha$
Generative Model for Metagenome Assembly III

• Read position/alignment within the contig $s_i$
  – Reads are assigned a starting point $o_i$ within each contig
    
    \[
    \rho_s \sim \text{Beta}(1, 1 + \beta) \quad \forall s \text{ that is not empty}
    \]
    
    \[
    o_i \sim G(\rho_s) \quad \forall i = 1..N
    \]

• The locations $o_i$ are centered at 0 and can be both negative and positive from the symmetric geometric distribution

\[
G(o; \rho) = \begin{cases} 
0.5(1 - \rho)^{|o|} \rho & o \neq 0 \\
\rho & o = 0
\end{cases}
\]
Generative Model for Metagenome Assembly IV

– Each read is assigned a length $l_i$

$$l_i \sim \mathcal{L} \quad \forall i = 1..N$$

• $\mathcal{L}$ is an arbitrary distribution

– the alignment $y_i$ and read letters $x_i$ for each read is generated from

$$x_i, y_i \sim A(l_i, s_i, o_i, b, p_{ins}, p_{del}, p_{mis}) \quad \forall i = 1..N$$

• $A$ is the noise model known for the sequencing technology (454, Illumina, etc.): noise can be introduced through insertions, deletions, mismatches

• Log likelihood $\log p(x_i, y_i|o_i, s_i, l_i, b)$ is given as

$$\log 0.5 + n_{hit} \log(1 - p_{mis}) + n_{mis} \log \left( \frac{p_{mis}}{|B| - 1} \right) + n_{ins} \log(p_{ins}) + n_{del} \log(p_{del})$$
Learning Algorithm

• Iterated conditional modes (ICM) algorithm
  – Maximize local conditional probabilities sequentially: hill-climbing method
  – Find MAP solution
  – Iterate until convergence (200-300 iterations)
  – Initialization: each read is in its own contig

• Consensus sequence
  – Given aligned reads $b_{so} = \arg\max_{b \in B} a_{so}^b$, where $a_{so}^b$ is the number of reads that align to the location
Learning Algorithm

• Read mapping \((s_i, o_i, y_i)\) for (contig, coordinate, alignment)
  - Sample from the joint posterior
    \[ p(s_i = s, o_i = o, y_i = y|x_i, y_{-i}, s_{-i}, o_{-i}, b, \rho) \]
  - Alignment \(y_i\): for each location and contig, use Smith-Waterman algorithm
    \[ y^*_{so} = \arg\max_y p(x_i, y|s_i = s, o_i = o, b) \]
  - Given the best alignment \(y^*_{so}\) * at each location and contig, determine the read mapping by sampling from
    \[ p(s_i = s, o_i = o, y_{so}^*|s) \propto p(s_i = s|s_{-i})p(o_i = o|s_i = s, \rho_s)p(x_i, y_{so}^*|s_i = s, o_i = o, b) \]
    \[ \propto N_s \cdot G(o; \rho_s) \cdot p(x_i, y_{so}^*|s_i = s, o_i = o, b) \]
    - \(N_s\): the number of reads in each contig
  • Filtering some of the mappings \((s_i, o_i, y_i)\) by 10-mer matching
Learning Algorithm

• Global moves to improve convergence
  – Merge: merge two contigs whose ends overlaps, if it improves the likelihood
  – Center: change the coordinate system of each sequence to maximize the $p(o)$ component of the likelihood
Learning Algorithm
Evaluation Metric

• BLAST profile
  – Estimation of the number of genome bases that the contig spans
  – BLAST the contigs and score each nucleotides in the contigs based on the BLAST scores

• PFAM profile
  – The total number of decoded amino acids matched by PFAM profiles after decoding the contigs into protein sequences and annotating them with PFAM profile detection tools
  – Examine the functional annotation of the contigs
Evaluation Metric

- Likelihood-based scores

\[ \sum_i \text{score}_i^{\text{READ}} - \log(|\mathcal{B}|) L + \log(|\mathcal{B}|) V_0 S \]

- L: the total length of all contigs
- S: the number of contigs
  - First term: penalization for read errors
  - Second term: penalization for contig length for the trade-off between contig length and accuracy for a good assembly
  - Third term: ensuring a minimal overlap of $V_0$ bases between two consecutive reads
Experiments

- Synthetic datasets
  - 454-250bp reads
  - The dataset was composed of the following sequences (in parenthesis, number of reads)
    - Acidianus filamentous virus 1 (14505)
    - Akabane virus segment L (4247)
    - Akabane virus segment M (2636)
    - Black queen cell virus (5309)
    - Cactus virus X (3523)
    - Chinese wheat mosaic virus RNA1 (3300)
    - Chinese wheat mosaic virus RNA2 (1649)
    - Cucurbit aphid-borne yellows virus (2183)
    - Equine arteritis virus (4832)
    - Goose paramyxovirus SF02 (4714)
    - Human papillomavirus - 1 (1846)
    - Okra mosaic virus (1016)
    - Pariacoto virus RNA1 (240)
Experiments

• Datasets

<table>
<thead>
<tr>
<th>name</th>
<th>description (source)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bee1(19k), Bee2(36k) [8]</td>
<td>Samples from two bee colonies. Data obtained by J. DeRisi Lab.</td>
</tr>
<tr>
<td>Coral(40k) [23]</td>
<td>Samples from viral fraction from whole Porites compressa tissue extracts (SRR001078).</td>
</tr>
<tr>
<td>Tilapia1(50k), Tilapia2(64k) [10]</td>
<td>Samples from Kent SeeTech Tilapia farm containing microbial (SRR001069) and viral (SRR001066) communities isolated from the gut contents of hybrid striped bass.</td>
</tr>
<tr>
<td>Peru(84k) [3]</td>
<td>Marine sediment metagnome from the Peru Margin subseafloor (SRR001326).</td>
</tr>
<tr>
<td>Microbes(135k) [4]</td>
<td>Samples from the Rios Mesquites stromatolites in Cuatro Cienagas, Mexico (SRR001043).</td>
</tr>
<tr>
<td>Chicken(311k) [20]</td>
<td>Samples of microbiome from chicken cecum. Dataset at <a href="http://metagenomics.nmpdr.org">http://metagenomics.nmpdr.org</a>, accession 4440283.3</td>
</tr>
<tr>
<td>Synthetic(50k)</td>
<td>Metagenomic samples of 13 virus strains, generated using Metasim [21], a 454 simulator. See Appendix for list.</td>
</tr>
</tbody>
</table>
Results

- Evaluation based on BLAST profiles
• Evaluation of methods based on PFAM
Results

- Evaluation based on reads’ consistency in assembly
Human Gut Microbiome Studies
(Qin et al., Nature, 2010)

• MetaHIT (Metagenomics of the human intestinal tract) project
  – Characterize the content, diversity, and function of the gut microbiome among different individuals
  – The gut microbiomes contribute to energy harvest from food
  – Changes of gut microbiome may be associated with bowel diseases or obesity
Data Generation

• Faecal specimens from 124 healthy, over-weight and obese human adults, and inflammatory bowel disease (IBD) patients

• Average 4.5Gb of sequence for each sample

• 42.7% of the Illumina GA reads was assembled into 6.58 million contigs of length > 500bp

• Common sequence cores for different individuals: 35% of reads from any one sample could be mapped to contigs from other samples
Analysis of Assembled Microbiomes

• Gene catalogue of the human gut microbiome
  – The assembled contigs contain 14 million ORFs occupying 86.7% of the contigs
  – 2.4 million ORFs were present in less than 20% of samples and 0.3 million were found in at least 50% of individuals

• Common set of bacterial species across samples
  – Comparison of assembled contigs with 650 sequenced bacterial and archael genomes by aligning the reads to these genomes
  – 18 species in all individuals, 57 species in >90% and 75 in >50% of individuals
Related Species in Microbiome

• Clusters of species in the network based on correlation coefficients of 155 species
• Similar constellations of bacteria in different individuals
Bacterial Functions

• Minimal gut genome
  – the parts of genomes for functions necessary for a bacterium to thrive in a gut context
  – Present in almost all gut bacterial species
  – Genes specific for the gut vs. housekeeping genes for all bacteria

• Minimal gut metagenome
  – The parts of genomes involved in the homeostasis of the whole ecosystem
  – Present in most individuals’ gut samples