Parallel taxonomic classification algorithm for metagenomic sequences

Le Van Vinh, Tran Van Hoai, Duong Ngoc Hieu, Bui Xuan Giang, Tran Van Lang, Thoai Nam

Bach Khoa University

March 1, 2017
Metagenomics

- Known as *environmental genomics* or *community genomics*
- Study of *uncultured* microorganisms
- Contribute to advances in many fields, e.g., Earth sciences, life sciences, bioenergy, biotechnology, agriculture.

*Figure:* Acid drainage (wikipedia.org)
Genomic vs. Metagenomics

**Problem:** more than 99% microbes cannot be cultured in the laboratory
Genomics vs. Metagenomics

**Problem**: more than 99% microbes cannot be cultured in the laboratory

**Problem**: lose information about which genes/sequences belong to which genomes/microbes

Source: Gianoulis, Harvard University, US
Problem’s definition

The taxonomic classification aims to group reads into bins and determines phylogenetic relationships between them and known taxa.
Why classification problem?

(1) The primary goals of metagenomic studies [2]
   - Who is out there? (taxonomic content, abundance level?)
   - What are they doing?
   - How do they compare?

(2) An important step in a metagenomic project [1]
Research challenges

Short read length

- Genomic signatures are less preserved in DNA read with length $< 1000bp$ [3]
- RAlphy (2011)[4]: 32% - 36% (Accuracy) for reads with the length of 100bp
Research challenges

Short read length

- Genomic signatures are less preserved in DNA read with length < 1000bp [3]
- RAlphy (2011)[4]: 32% - 36% (Accuracy) for reads with the length of 100bp

Large amount of data

- Require large computational costs
- Work with a huge amount of reference database (GenBank (12/2016): \( \approx 200.000.000 \) sequences with \( \approx 2 \times 10^{13} \) bases)
Related works

- **Composition-based methods**
  - Using genomic signatures (e.g., oligonucleotide frequencies, GC-content)
  - TACOA [6], AKE [7])
  - are fast, but difficult to analyze short reads

- **Homology-based methods**
  - Basing on the similarity between sequences
  - MEGAN [2], CARMA3 [9], MetaCluster-TA [10]
  - work well with both short and long reads, but much computational expense
Related works

Metagenomic applications are based on high-performance computing techniques

- **Map-reduce framework:** MrMC-MinH [11],
- **GPU, multi-core-CPU:** Parallel-META [12]
- **MPI:** mpiBlast [13]
Our previous works

Figure: Classification process of SeMeta (Vinh et al. (2016) [8])
ParSeMeta

Step 1: Clustering

- Clusters
  - core
  - unclustered reads

Computer node

similarity search

Step 2: Taxonomic Assignment

- A part of reference database
- Labling
  - cluster 1
  - cluster 2
  - cluster 3
- Labling other reads

- taxon A
- taxon B
- unassigned reads

Figure: Classification process of ParSeMeta
Experiment setup

- Test on two physical machines (12 CPUs, 120G RAM, and 100GB disk storage)
- Three cases (number of cores, number of virtual machines, memory sizes)
**Figure:** The performance of ParSeMeta and SeMeta with different numbers of core

**Figure:** The performance of ParSeMeta and SeMeta with different numbers of core
Figure: The performance of ParSeMeta with different numbers of virtual machines, with cases of using 3GB RAM and 6GB RAM
### The classification quality

<table>
<thead>
<tr>
<th>Dataset ds1</th>
<th>Species level</th>
<th>Genus level</th>
<th>Family level</th>
<th>Order level</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeMeta</td>
<td>Sen. 42.76%</td>
<td>42.76%</td>
<td>99.71%</td>
<td>99.71%</td>
</tr>
<tr>
<td></td>
<td>Pre. 42.88%</td>
<td>42.88%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>ParSeMeta</td>
<td>Sen. N/A</td>
<td>99.71%</td>
<td>99.71%</td>
<td>99.71%</td>
</tr>
<tr>
<td></td>
<td>Pre. N/A</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset ds2</th>
<th>Species level</th>
<th>Genus level</th>
<th>Family level</th>
<th>Order level</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeMeta</td>
<td>Sen. 24.72%</td>
<td>30.24%</td>
<td>61.94%</td>
<td>61.94%</td>
</tr>
<tr>
<td></td>
<td>Pre. 39.91%</td>
<td>30.34%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>ParSeMeta</td>
<td>Sen. 24.72%</td>
<td>30.24%</td>
<td>61.94%</td>
<td>61.94%</td>
</tr>
<tr>
<td></td>
<td>Pre. 39.91%</td>
<td>30.34%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset ds3</th>
<th>Species level</th>
<th>Genus level</th>
<th>Family level</th>
<th>Order level</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeMeta</td>
<td>Sen. 46.69%</td>
<td>64.84%</td>
<td>64.84%</td>
<td>64.84%</td>
</tr>
<tr>
<td></td>
<td>Pre. 67.09%</td>
<td>93.16%</td>
<td>93.16%</td>
<td>93.16%</td>
</tr>
<tr>
<td>ParSeMeta</td>
<td>Sen. 23.64%</td>
<td>64.84%</td>
<td>64.84%</td>
<td>64.84%</td>
</tr>
<tr>
<td></td>
<td>Pre. 16.45%</td>
<td>93.16%</td>
<td>93.16%</td>
<td>93.16%</td>
</tr>
</tbody>
</table>
Propose a parallel algorithm utilizing the advantages of high performance computing system
ParSeMeta is able to reduce much computational time, still keeps classification quality
Future works: apply on large-scale metagenomic datasets, predicting execution time of the algorithm with different settings (parameters of algorithm, allocated resources)
References

References


Applying Reconfigurable Computing for Internet Security

- HR-IPS: HCMUT Reconfigurable Intrusion Prevention System

- HR-IPS meets line rate full duplex Gbps network link without any drops.
- 5x better throughput compare to bare software solution (Snort IPS).

![Throughput comparison between Snort and HR-IPS](chart)
Low-power wireless water quality monitoring system

- Water quality monitoring system for shrimp farming
- Energy harvesting
- Low-power network protocol
- Data management