Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets
Gaëtan Benoit

To cite this version:
Gaëtan Benoit. Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets. RCAM, Oct 2015, Paris, France. hal-01231795

HAL Id: hal-01231795
https://inria.hal.science/hal-01231795
Submitted on 18 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Simka
Fast kmer-based method for estimating the similarity between numerous metagenomic datasets

Gaëtan BENOIT
PHD student - ANR Hydrogen

GenScale bioinformatics research group
IRISA/INRIA – Rennes - FRANCE
24/09/2015
Tara Oceans
• 200 sampling sites

N metagenomic samples

Whole genome sequencing

De novo comparative metagenomics

Similarity heatmap $N^2$

N MetaG datasets

Tara Oceans
• > 1000 datasets
• > 100 millions reads each
De novo comparative metagenomics

- Sea water
  - < 5% assembled reads
  - < 10% known species

- Compare environmental data and genomic contents
  - Environmental characteristics of Agulhas rings affect interocean plankton transport. Villar et al. Science 2015

Similarities on
  - Barcodes
  - WGS
  → same conclusions
Similarity between 2 samples

Idea: Similarity is given by the size of the intersection

Intersection = number of similar reads

Similar: alignment score > 90%
(Blast-like approaches, all-vs-all sequence alignment)

100M reads
Would require months of computation
State of the art

Commet (Maillet et al. IEEE BIBM 2014)

- "similar": share at least $t$ non-overlapping kmers (words of size $k$
- Example: $t=2$:

```
ACTGACT
CGTTAGT
ACTGACT
CGTTAGT
```

- Computes one intersection in **few hours**
  - Alignment free method
  - Only **indexing and querying kmers**

Does not scale on large metagenomic projects: $N(N-1)$ comparisons
Simka: comparing numerous datasets

New representation of datasets

• **Not** viewed as a **read-set** anymore
• But as a **bag** of its **overlapping kmers**

**Output of kmer counting algorithm**

Kmers (k=6)

In practice k > 30
Simka: comparing numerous datasets

New similarity function

- Pairwise similarity measure based on shared kmers

Abundance-based Jaccard similarity

\[
AJ(A, B) = \frac{\sum_{w \in A \cap B} N_A(w) + N_B(w)}{N_A(w) + N_B(w) - \sum_{w \in A \cup B} N_A(w) + N_B(w)}
\]

We need

- Fast kmer counting algorithm
- Compute faster the N(N-1) comparisons

\[
\begin{align*}
A &= AAAAAA  4 \\
T &= TTTAAA  1 \\
T &= TTTTTT  1 \\
\end{align*}
\]

\[
\begin{align*}
A &= AAAAAA  1 \\
C &= CCCCCA  1 \\
C &= CCCCCCT  2 \\
\end{align*}
\]

\[
AJ = \frac{4 + 1 + 1 + 1 + 1 + 2}{4 + 1 + 1 + 1 + 1 + 2} = \frac{5}{10} = 0.5
\]
Multiset kmer counting

- Count the kmers of N datasets **simultaneously**
  - Based on KMC2 algorithm (Deorowicz et al. Bioinformatics 2015)
  - Available in GATB library (Drezen et al. Bioinformatics 2014)

For each distinct kmer

<table>
<thead>
<tr>
<th>ACGATC</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>4</td>
<td>5</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

- We can now count the kmers of large datasets quickly
- Compute all the intersections simultaneously
Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

\[
AJ(i, j) = \frac{\sum_{w} N_i(w) + N_j(w)}{\sum_{w} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}
\]
Simka algorithm

- Example with 3 datasets on Abundance-based Jaccard similarity

\[
AJ(i, j) = \frac{\sum_{w \in i} N_i(w) + \sum_{w \in j} N_j(w)}{\sum_{w \in i} N_i(w) + \sum_{w \in j} N_j(w)} = \frac{n(i, j)}{u(i, j)}
\]

Kmer shared by A, B and C

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

For each pair (i, j)
Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

\[
AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}
\]

Kmer shared by A, B and C

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

For each pair (i, j)

If kmer is shared \( N_i > 0 \) and \( N_j > 0 \)
Simka algorithm

- Example with 3 datasets on Abundance-based Jaccard similarity

\[ AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)} \]

For each pair \((i, j)\)
- If kmer is shared \(N_i > 0\) and \(N_j > 0\)
- Update intersection \(n[i,j] += N_i + N_j\)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Kmer shared by A, B and C

For each pair \((i, j)\)
- If kmer is shared \(N_i > 0\) and \(N_j > 0\)
  - Update intersection \(n[i,j] += N_i + N_j\)
Simka algorithm

- Example with 3 datasets on Abundance-based Jaccard similarity

\[
AJ(i, j) = \frac{\sum_{w} N_i(w) + N_j(w)}{\sum_{w} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}
\]

Kmer shared by A, B and C

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

For each pair \((i, j)\)

- If kmer is shared \(N_i > 0\) and \(N_j > 0\)
  - Update intersection \(n[i,j] += N_i + N_j\)
  - Update union \(u[i,j] += N_i + N_j\)

\[AJ(A,B) = \frac{6}{6} = 1\]
Simka algorithm

- Example with 3 datasets on Abundance-based Jaccard similarity

\[
AJ(i, j) = \frac{\sum_{w} N_i(w) + N_j(w)}{\sum_{w} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}
\]

For each pair \((i, j)\)

- If kmer is shared \(N_i > 0\) and \(N_j > 0\)
  - Update intersection \(n[i,j] += N_i + N_j\)
  - Update union \(u[i,j] += N_i + N_j\)

Kmer shared by A, B and C

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

AJ(A,B) = 6/6 = 1
AJ(A,C) = 12/12 = 1

AJ(i, j) = \(\frac{N_i(w) + N_j(w)}{N_i(w) + N_j(w)}\)
Simka algorithm

- Example with 3 datasets on Abundance-based Jaccard similarity

\[
AJ(i, j) = \frac{\sum_{w \in i \cap j} \left( N_i(w) + N_j(w) \right)}{\sum_{w \in i \cup j} \left( N_i(w) + N_j(w) \right)} = \frac{n(i, j)}{u(i, j)}
\]

Kmer shared by A, B and C

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

For each pair (i, j)

- If kmer is shared \(Ni > 0\) and \(Nj > 0\)
  - Update intersection \(n[i,j] += Ni + Nj\)
  - Update union \(u[i,j] += Ni + Nj\)

\[
AJ(A, B) = \frac{6}{6} = 1
\]

\[
AJ(A, C) = \frac{12}{12} = 1
\]

\[
AJ(B, C) = \frac{10}{10} = 1
\]
Simka algorithm

• Example with 3 datasets on Abundance-based Jaccard similarity

\[
AJ(i, j) = \frac{\sum_{w} N_i(w) + N_j(w)}{\sum_{w} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}
\]

For each pair \((i, j)\)
- If kmer is shared \(N_i > 0\) and \(N_j > 0\)
  - Update intersection \(n[i,j] += N_i + N_j\)
  - Update union \(u[i,j] += N_i + N_j\)

Kmer specific to A

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

AJ(A,B) = 6/12 = 0.5
AJ(A,C) = 12/18 = 0.66
AJ(B,C) = 10/10 = 1
Simka algorithm

- Example with 3 datasets on **Abundance-based Jaccard similarity**

\[
AJ(i, j) = \frac{\sum_{w \in i \cap j} N_i(w) + N_j(w)}{\sum_{w \in i \cup j} N_i(w) + N_j(w)} = \frac{n(i, j)}{u(i, j)}
\]

<table>
<thead>
<tr>
<th>ACGATC</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

For each pair (i, j)

- If kmer is shared, **Ni > 0 and Nj > 0**
  - Update intersection: \( n[i, j] += N_i + N_j \)
  - Update union: \( u[i, j] += N_i + N_j \)

**AJ(A, B) = 6/12 = 0.5**

**AJ(A, C) = 12/18 = 0.66**

**AJ(B, C) = 10/10 = 1**
Execution time

On 21 Tara samples, 2.1G reads, 210GB data

- Commet (state of the art)
  - On cluster (one node)
    - A few weeks

- Simka
  - On cluster (one node): 3h
    - Counting kmers: 75%
    - Update unions/intersections $\binom{N}{N-1}$: 25%
  - On standard computer: 10h
    - 4 GB memory, 4 cores
Representation of similarity

\[
AJ_{sym}(i, j) = \frac{\sum_{w} \left( \frac{N_i(w) + N_j(w)}{N_i(w) + N_j(w)} \right)}{N_i(w) + N_j(w)}
\]

\[
AJ_{asym}(i, j) = \frac{\sum_{w} \left( \frac{N_i(w)}{N_i(w)} \right)}{N_i(w)}
\]

Hierarchical clustering

Heatmap

100% similar

40% similar

14% similar

Asymmetry can show information about diversity
Results

Two things to compare with Commet

Heatmap absolute similarity

Hierarchical clustering relative similarity
On 21 datasets from Tara Oceans

Commet (k=33)
On 21 datasets from Tara Oceans

Commet (k=33)  Simka (k=31)
Other measures provided

- Transform the **abundance vectors** to get new measures
  - From **abundance** to **presence/absence**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

To boolean vector

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Jaccard similarity:

\[
J_{sym}(A, B) = \frac{A \cap B}{|A|} \quad \text{and} \quad J_{asym}(A, B) = \frac{A \cap B}{|A|}
\]

- Filtering potentially erroneous kmers
  - Discard kmer if **specific** (D=1) and **unique** (Ni=1)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACGATC</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Conclusion

• Simka
  • New pairwise similarity functions based on **shared kmers**
  • Provides measures based on **abundance** and **presence/absence**
  • Results close to read-based methods
  
• Fast and low memory thanks to the GATB library
  • Execution time (21 Tara samples, 2.1G reads, 210GB data)
    • Commet (state of the art): **few weeks**
    • Simka
      • On cluster: **3h**
      • On standard computer: 10h

**GATB**
www.gatb.inria.fr

**Simka**
www.gatb.fr/software/simka
Perspectives

- Add similarity measures well used in ecology (Ex: Bray Curtis)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

- Bootstrapping
  - Test the robustness of Simka
    - Are there differences between using 100M or 10M reads?
    - Provides similarity matrix with confidence intervals
      - $\text{Sim}(A, B) = 23\% \pm 0.6\%$
    - Add confidence levels to dendrogram

- Simka Potara
  - For cluster and cloud
  - MapReduce model
  - Process huge amount of data
    - **130 Tara samples** (13G reads, 1.3TB data) compared in < 1 day
Acknowledgements

PHD Supervisor
---
Claire Lemaitre
Pierre Peterlongo
Dominique Lavenier

GATB devs
---
Erwan Drezen
Guillaume Rizk

Commet
---
Nicolas Maillet

Partners

INRA
---
Stéphane Robin
Sophie Schbath
Mahendra Mariadassou
Julien Chiquet
Julie Aubert
Sarah Ouadah
Emilie Lebarbier

CEA
---
Olivier Jaillon
Jean-Marc Aury
Eric Pelletier
Thomas Vannier

Founded
---
ANR Hydrogen