Kmer2Reads, an associative index for Third Generation Sequencing data

Léa Vandamme, Bastien Cazaux and Antoine Limasset

Univ. Lille, CNRS, UMR 9189 - CRIStAL, F-59000 Lille

SeqBIM

November 21, 2023
Genome assembly is complex

Long reads properties:
- Error rate: from ~0.1% to ~10%
- Length: 10 - 100 kilo bases

Hard task
K-mer to reads index

Select reads similar to our query sequence

- Select reads sharing k-mers with our query sequence
- Use k-mer to reads index
State of the art

➢ Hashing methods (based on hash tables / MPHF)

Use example: Pufferfish (Almodaresi, 2018), SRC (Marchet, 2020), BLight (Marchet, 2021), Variant of FM-index: r-index (Mun, 2020)

➢ Full text indexing (locate occurrences of pattern in a text)
Challenge:

Scale to large genomes (Human genome = 100 million bp, 10 million reads)

- Limit memory and time cost
A naive version?

Similar k-mers

- ATCGATCG
- TCGATCGA
- CGATCGAT
- GATCGATC
- ATCGATCT
- TCGATCTA
- CGATCTAT

Duplicated read IDs

- [1, 56, 89, 115, 745, 987, 1025, 1026]
- [1, 56, 89, 115, 745, 987, 1025, 1026]
- [1, 56, 89, 115, 745, 987, 1025, 1026]
- [1, 56, 89, 115, 745, 987, 1025, 1026]
- [1, 56, 89, 115, 745, 987, 1025, 1026]
- [1, 56, 90, 115, 745, 945, 1025, 1156]
- [1, 56, 90, 115, 745, 945, 1025, 1156]

Large lists of integers
Our contributions

- Lots of similar k-mers \(\rightarrow\) Minimizers
- Duplicated lists of read IDs \(\rightarrow\) Distinct colors
- Large lists of read IDs \(\rightarrow\) Compression
### Similar k-mers?

**Ecoli** (error rate 1%, coverage = 100X, read length = 10000):
- **K-mers**: 4,553,982
- **M-mers**: 443,123

1 order of magnitude

<table>
<thead>
<tr>
<th>Minimizer</th>
<th>K-mers</th>
<th>M-mers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATCGATCG</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
</tr>
<tr>
<td>TCGATCGA</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
</tr>
<tr>
<td>CGATCGAT</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
</tr>
<tr>
<td>GATCGATC</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
</tr>
<tr>
<td>ATCGATCT</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
<td>[1, 56, 89, 115, 745, 987, 1025, 1026]</td>
</tr>
<tr>
<td>TCGATCTA</td>
<td>[1, 56, 90, 115, 745, 945, 1023, 1156]</td>
<td>[1, 56, 90, 115, 745, 945, 1023, 1156]</td>
</tr>
<tr>
<td>CGATCTAT</td>
<td>[1, 56, 90, 115, 745, 945, 1023, 1156]</td>
<td>[1, 56, 90, 115, 745, 945, 1023, 1156]</td>
</tr>
</tbody>
</table>

Diagram:

```
ATCGATCG         [1, 56, 89, 115, 745, 987, 1025, 1026]
TCGATCGA         [1, 56, 89, 115, 745, 987, 1025, 1026]
CGATCGAT         [1, 56, 89, 115, 745, 987, 1025, 1026]
GATCGATC         [1, 56, 89, 115, 745, 987, 1025, 1026]
ATCGATCT         [1, 56, 89, 115, 745, 987, 1025, 1026]
TCGATCTA         [1, 56, 90, 115, 745, 945, 1023, 1156]
CGATCTAT         [1, 56, 90, 115, 745, 945, 1023, 1156]

AT

TCGATCTA         [1, 56, 90, 115, 745, 945, 1025, 1156]
CGATCTAT         [1, 56, 90, 115, 745, 945, 1025, 1156]
```
**False positives?**

**Solution:** count shared k-mers between potential output reads and query sequence.

Dataset: Ecoli (error rate 1%, coverage = 100X, read length = 10000), seuil : 0.6
Query: sequence of length 50bp, extracted from a read
- Nombre reads total : 46.396
- Potential hits : 91
- Reads deleted : 1
Duplicated colors?

Color = List of read IDs

Index distinct colors

Dataset: Ecoli (error rate 1%, coverage = 100X, read length = 10000), seuil : 0.6
Query: sequence of length 50bp, extracted from a read

- M-mer number: 443,123
- Color number: 88,345

1 order of magnitude
Large ID lists?

**TurboPFor**

Delta encoding:

\[ [1, 1000, 1005, 1024, 1025, 1034] \]

\[ \rightarrow [1, 999, 5, 19, 1, 9] \]

1 order of magnitude
Our tool: K2R

Error rate 1%, coverage = 100X, read length = 10.000

<table>
<thead>
<tr>
<th>Dataset</th>
<th>K-mer number</th>
<th>M-mer number</th>
<th>Color number</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.Coli</td>
<td>4,553,982</td>
<td>443,123</td>
<td>88,345</td>
</tr>
<tr>
<td>C.Elegans</td>
<td>94,006,409</td>
<td>6,338,436</td>
<td>4,475,066</td>
</tr>
</tbody>
</table>
Results: memory & time

**Index creation**
- Parameters: error rate 1%, read length = 10,000bp
- **Dataset**
  - C.Elegans: 100X

![Graph showing memory comparison](image1)

![Graph showing time comparison](image2)
Results: comparison

Index creation: error rate: 0.1%

Index creation: error rate: 10%

Memory usage (MiB)

Datasets:
- E. Coli
- S. Cerevisiae
Results: comparison

Index creation: error rate: 0.1%

Index creation: error rate: 10%
Index files on disk

Index creation:
- **C. Elegans (100.3MB)**: error rate 1%, read length = 10,000bp.

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Read FASTA file size (Go)</th>
<th>M-mer binary file size</th>
<th>Color binary file size</th>
<th>Total</th>
<th>Total / input</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.958</td>
<td>0.057</td>
<td>0.168</td>
<td>0.225</td>
<td>0.23</td>
</tr>
<tr>
<td>50</td>
<td>4.7</td>
<td>0.137</td>
<td>0.901</td>
<td>1.038</td>
<td>0.22</td>
</tr>
<tr>
<td>100</td>
<td>9.4</td>
<td>0.218</td>
<td>1.8</td>
<td>2.018</td>
<td>0.21</td>
</tr>
<tr>
<td>150</td>
<td>15</td>
<td>0.278</td>
<td>2.6</td>
<td>2.878</td>
<td>0.19</td>
</tr>
<tr>
<td>200</td>
<td>19</td>
<td>0.324</td>
<td>3.4</td>
<td>3.724</td>
<td>0.19</td>
</tr>
</tbody>
</table>
## Conclusion

<table>
<thead>
<tr>
<th></th>
<th>Construction</th>
<th>Memory</th>
<th>Debit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Connector</td>
<td>✔</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>r-index</td>
<td>✗</td>
<td>✔</td>
<td>~</td>
</tr>
<tr>
<td>K2r</td>
<td>✔</td>
<td>~</td>
<td>✔</td>
</tr>
</tbody>
</table>
Perspectives

Sorting reads

→ Less redundant vectors
→ Improve delta encoding
 Perspectives

➢ Bottleneck : memory (challenge r-index ?)
➢ Soon : article
➢ Applications : open to collaborations :}


Take home messages

Goals
- Associate k-mers to reads, in order to study genome sequences from raw reads

Etat de l’art
- Hashing methods (memory expensive)
- Full-text indexing (hard to construct)
Take home messages

Solutions
- Minimizers (False positives are managed)
- Colors
- Compression

K2R
- Fast
- Dynamic
- Scalable