Mapping Algorithms
Matthias Zytnicki
MIAT INRA

Outline
1 Introduction
2 Genome mapping
   Hashing-based tools
   Suffix array-based tools
3 Transcriptome mapping
4 Conclusion

What is mapping

Desired definition
Map a read: predict the locus from which the read originates.

Data
genome read mapping genomic coordinate(s)

Example
@SEQ_ID
GATTT +
ccedd
BWA @SEQ_ID 113 1...

Assumption
A read is likely to map at a locus iff similarity is high.

Implemented definition
Map a read: list the loci with less than k errors.
First implications

Ambiguity in the mapping

Outline

1. Introduction
2. Genome mapping
   - Hashing-based tools
   - Suffix array-based tools
3. Transcriptome mapping
4. Conclusion

Tools...

Outline

1. Introduction
2. Genome mapping
   - Hashing-based tools
   - Suffix array-based tools
3. Transcriptome mapping
4. Conclusion
Seed-and-extend algorithm

Idea
1. Get the $k$-mers of the genome,
2. "Sort" them,
3. Get the $k$-mers of a read,
4. Compare the two,
5. Finish the alignment.

Example
Genome is AACGTAC, read is CCGT

AACGTA = AAC ACG CGT GTA
AAC
ACG
CGT
GTA

Choice of $k$ — Example 1

Example
Consider sequence ACGCGTGTA and read ACACGAGTA.

- With $k = 3$, seed GTA matches.
- With $k = 4$, no seed match.

Pigeon hole principle
With $r$ errors, $r + 1$ seeds.
⇒ small $k$ = more sensitive.

Step 2: Extension with errors

Needleman–Wunsch

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Optimizations
Banded Needleman–Wunsch, FSA, parallel Shift-OR, vectorization, SIMD, . . .

Choice of $k$ — Example 2

Example
Consider sequence ACGACGACGACGACGT and read ACGT.

$k = 3$:
- ACG → 0, 3, 6, 9, 12
- CGA → 1, 4, 7, 10
- CGT → 13
- GAC → 2, 5, 8, 11
⇒ 4 matches, 1 extension succeeds.

$k = 4$:
- ACGA → 0, 3, 6, 9
- ACGT → 12
- CGAC → 1, 4, 7, 10
- GACG → 2, 5, 8, 11
⇒ 1 matches, 1 extension succeeds.

Remark
$k$-mers with many occurrences usually are discarded.
Conclusion so far

Trade-off

- Small \( k \): more sensitive, slower.
- High \( k \): more specific, larger database (size: up to \( 4^k \)).

Tentative complexity

- Pre-process genome (done once).
- Cut a read into \( k \)-mers (fast).
- Map each \( k \)-mer to the table (fast for each \( k \)-mer).
- Consider every position in the genome (variable).
- Extend with errors (slow).

Drawbacks

Slow or not sensitive when:

- accepting many errors,
- accepting highly repeated seeds.

Outline

1. Introduction
2. Genome mapping
   - Hashing-based tools
   - Suffix array-based tools
3. Transcriptome mapping
4. Conclusion

Suffix tree

GATTACA is:

Looking for a read

Idea
Simply follow the right arrow.

Example
Look for TAC. Look for TAT.

Problem
Suffix trees do not fit in memory. Use suffix arrays instead:

- same algorithms,
- condensed data structure.
Handling errors

Example
Look for GAC with 1 error. GAC

Problem
The search space is very large!

Conclusion so far

Tentative complexity
- Pre-process the genome (done once).
- Look for a read (quite fast if no error).

Slow or not sensitive when:
- accepting many errors (small seed).

Outline

1 Introduction
2 Genome mapping
   Hashing-based tools
   Suffix array-based tools
3 Transcriptome mapping
4 Conclusion

Introduction

RNA mapping ≠ DNA mapping

spliced read

genome
**Difficult cases**

- many differences (mutations and/or sequencing errors),
- repeated sequence,
- read on 3+ exons,
- gene or pseudogene?

*Kim et al., Gen. Biol., 2013*

- end of read on another exon,
- read on unknown and poorly expressed junction.

**Algorithms**

---

**a** Exon-first approach

- Exon read mapping
- Spliced read mapping

*Garber et al., Nat. meth., 2011*

**b** Seed-extend approach

- Seed matching
- Seed extend

*Garber et al., Nat. meth., 2011*

---

**Tophat1**

- Map reads to whole genome with Bowtie
- Collect initially unmappable reads
- Assemble consensus of covered regions
- Generate possible splices between neighboring exons
- Build seed table index from unmappable reads
- Map reads to possible splices via seed-and-extend
- *Trapnell et al., Bioinformatics, 2009*

- uses Bowtie,
- problems for pseudogenes.

**Tophat2**

- Transcription alignment (optional)
- Genome alignment
- Spliced alignment
- Segment alignment to genome
- Segments aligned to genome
- Transposome index
- Genome index
- Genome-trapped reads with alignments extending 5' or 3' barcoding sites are re-aligned to transposome

*Kim et al., Gen. Biol., 2013*
Other considerations

**Quality** Current tools provide read quality. Better use it.

**Pair end reads** Most tools have a special “fragment rescue” mode.

**Low quality 3’ ends** Some tools truncate them.

**Binary encoding** A: 00, C: 01, G: 10, T: 11, other: ???

---

### Tool comparison

<table>
<thead>
<tr>
<th>DNA mapping tools</th>
<th>seed-and-extend</th>
<th>pigeon hole</th>
<th>spaced seed</th>
<th>q-gram</th>
<th>suf. tree</th>
<th>B–W</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSAHA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUMmer2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intros</td>
</tr>
<tr>
<td>Eland</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAQ</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PatMaN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMAP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SeqMap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mosaic</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5OC</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZOOM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PASS</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAP2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BWA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHRIMP</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowtie</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mFAST</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RazerS</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CloudBlust</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mpSscan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PerM</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BFfAST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sagemehl</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNU/Map</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>revFfAST</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nognalign</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GASSST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stampy</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOAP3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowtie2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samman</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BaraCUDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sensitivity** map most reads,

**Specificity** mappings are correct,

**Time**

**Memory**

⇒ Balance between the criteria.
“Unbiased” comparisons


⇒ Results are already irrelevant!

RGASP

The RNA-seq Genome Annotation Assessment Project.

(Li Website: http://lh3lh3.users.sourceforge.net/alnROC.shtml)

(Bowtie2 vs BWA)

(Langmead et al., Nat. Meth., 2012)

(RGASP)

(Engström et al., Nat. Meth., 2013)
Which tool should I use?

Common situations
You can use widely-used tools.
- **DNA-Seq**  BWA(-SW/-mem), Bowtie2
- **RNA-Seq**  TopHat2, Star
  - Cannot refuse your paper!
  - Well maintained.
  - Large software suite (Tuxedo).

A niche?
- mrFAST: for high number of copies
- SHRiMP: for color space
- Stampy: for highly divergent sequencing (with BWA)
- CloudBurst: with Map/Reduce
- SOAP3: for GPU

Want to buy France?
Use CRAC.