NanoMapper: An Efficient Read Mapping Algorithm Using Gapped Minimizer and FM-indexing for Oxford Nanopore Long Noisy Reads

Supervisor:

Eamin Rahman
Assistant Professor,
Department of Computer Science and Engineering
Shahjalal University of Science and Technology

Presenters

Enamul Hassan
Reg. No.: 2011331051
4th year, 2nd Semester

Md. Khairullah Gaurab
Reg. No.: 2011331063
4th year, 2nd Semester

Department of Computer Science and Engineering
Shahjalal University of Science and Technology

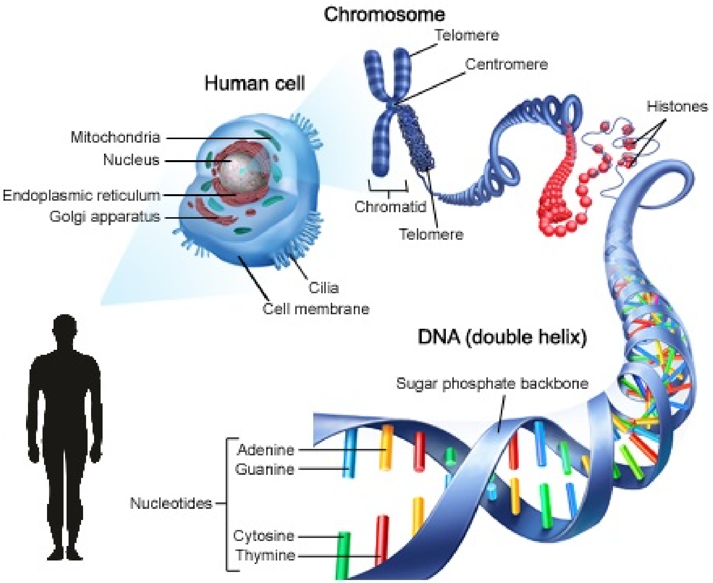
Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

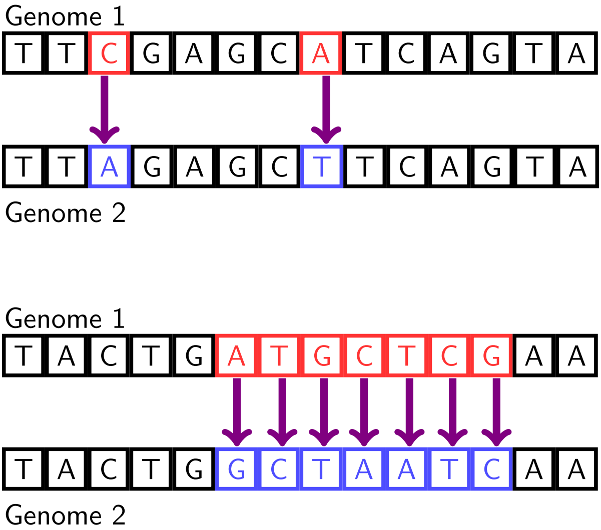
Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Human Cell To DNA Sequence



Mutation



Genome 1

T T C G A G C A T C A G T A

T T A G A G C T T C A G T A

Genome 2

Genome 1

T A C T G A T G C T C G A A

T A C T G G C T A A T C A A

Genome 2

Figure: Random mutations and Segment mutation.

Mapper

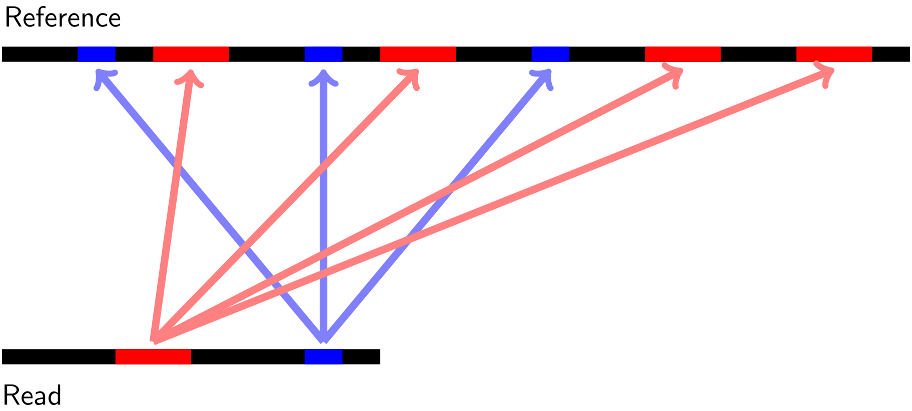


Figure: Mapping is just indicating the clusters of a large segment of the read in reference.

Aligner

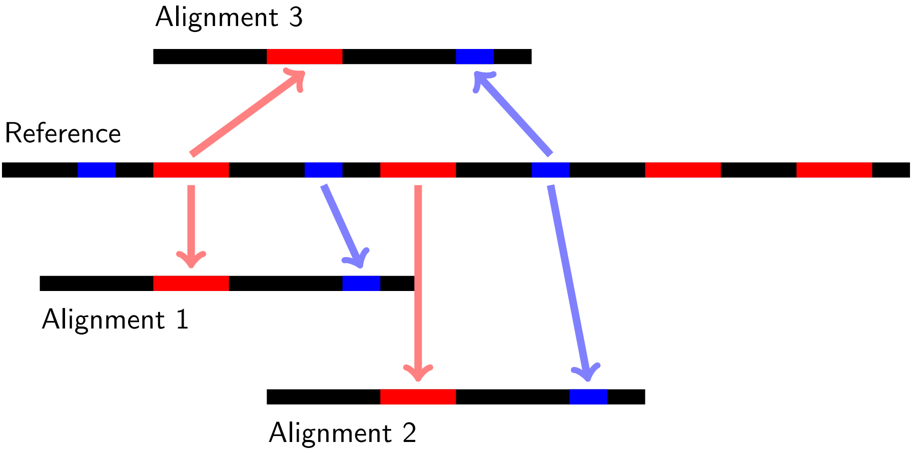


Figure: Three possible alignments based on the read mapping.

Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Alignment Needs Mapping

- Variant Calling
  - Identification of causative genes , candidate genes, passenger and driver genes in many complex diseases, disorders and cancers.

Alignment Needs Mapping

- Variant Calling
  - Identification of causative genes , candidate genes, passenger and driver genes in many complex diseases, disorders and cancers.
- DNA Binding
  - Finding DNA-Binding sites on specific reference genome sequence.

Alignment Needs Mapping

- Variant Calling
  - Identification of causative genes , candidate genes, passenger and driver genes in many complex diseases, disorders and cancers.
- DNA Binding
  - Finding DNA-Binding sites on specific reference genome sequence.
- Gene Expression
  - Classification of human tumors, profiling breast cancer, Ontological analysis.

Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Challenges Developing Mapper

- Illumina/Solexa Technologies Produces Short Reads – 30 - 300 BP

Challenge

Very Repetitive for Long Reference Like 3 billion BP

Challenges Developing Mapper

- Illumina/Solexa Technologies Produces Short Reads – 30 - 300 BP

Challenge

Very Repetitive for Long Reference Like 3 billion BP

- PacBio's Oxford Nanopore Technology Produces Long Reads – 10K - 60K

Challenge

The More The Length, The More The Noise

Challenges Developing Mapper

- Illumina/Solexa Technologies Produces Short Reads – 30 - 300 BP

Challenge

Very Repetitive for Long Reference Like 3 billion BP

- PacBio's Oxford Nanopore Technology Produces Long Reads – 10K - 60K

Challenge

The More The Length, The More The Noise

- Needs Huge Memory To Index Reference Sequence

Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. **Existing Tools**
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Minimap

- Used Minimizer and Min-sketch
- Tweaked for Miniasm
- Multi-threaded
- Not Tested Enough Yet

Output Format

The output format is PFA which is different than other tools.

BWA

- Two versions:
  - BWA / BWA-MEM for Illumina/Solexa reads
  - BWA-SW for Oxford Nanopore Reads of MinION instrument
- Used BWT
- Used Prefix-Trie Traversing Top-Down Fashion
- Applied Smith–Waterman-like Dynamic Programming
- Multi-threaded

Bowtie

- Two versions:
  - Bowtie
  - Bowtie 2
- Used BWT FM-index
- Quality-aware Backtracking Algorithms Enabling Mismatches
- Double Indexing – To Avoid the Excessive Use of Backtracking

MUMmer

- Three versions: 1, 2, 3
- Used Suffix-Tree

NanoBLASTer

- Used "seed-and-extend" Technique
- Dynamic Programming Based Extension Mechanism
- Faster than Leading Alignment Tools
- Less False Positive Rate

Limitation

Could not Handle Long Insertion or Deletion

Long Insertion

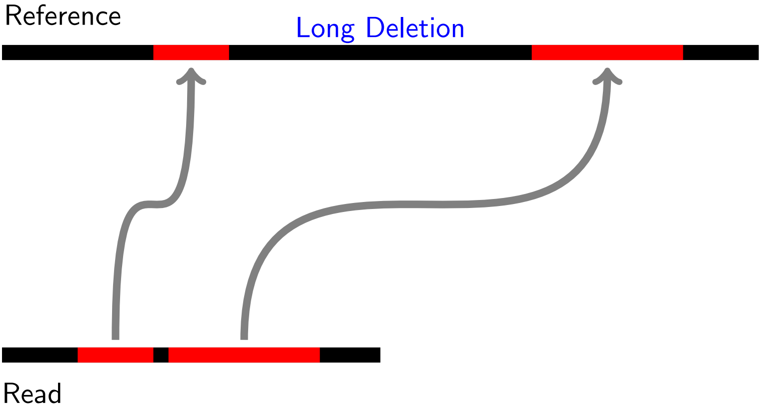


Figure: Mapping Two Long K-mers with Long Deletion in Read.

Long Deletion

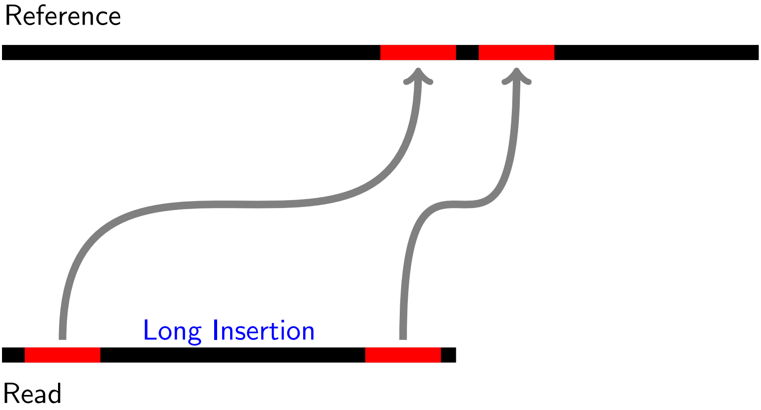


Figure: Mapping Two Long K-mer with Long Insertion in Read.

Table of Contents

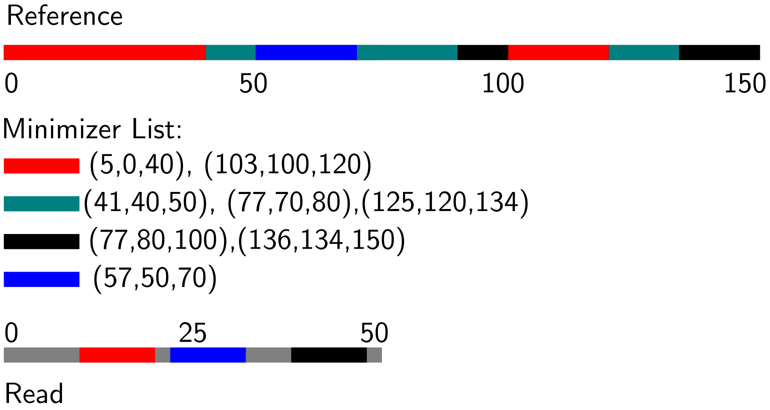
1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. **Our Approaches**
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Minimizer Approach

There are two versions of this approach:

- Naive Window Technique
- Efficient Window Traversing Technique

Minimizer Approach



Minimizer List:

🟥 (5,0,40), (103,100,120)

🟩 (41,40,50), (77,70,80),(125,120,134)

⬛ (77,80,100),(136,134,150)

🟦 (57,50,70)

Figure: Mapping Minimizers.

Minimizer Approach : Result

Table: Minimizer Hits For Increasing Error Rate.

Error (%)	Found In Reference (%)	Found In Range (%)	Found Both In and Out of Range (%)
0	99.96	69.2	30.7
5	71.8	62.4	27.2
10	53.8	52.8	25.5
15	43.5	47.4	21.7
20	34.8	39.8	17.9
45	0.2	40	40

Gapped Minimizer

Problem

The More The Error, The More The Mismatches

Gapped Minimizer

Problem

The More The Error, The More The Mismatches

Solution

Insert Gap in Both Reads and Reference To Neutralize Mismatches At Every Certain Number of Bases.

Gapped Minimizer

Problem

The More The Error, The More The Mismatches

Solution

Insert Gap in Both Reads and Reference To Neutralize Mismatches At Every Certain Number of Bases.

Remark

In NanoMapper, Gaps are Added Every Third Base Assuming 33% Error.

Example: `ATCTGGTAATCATAGCGTAC`

With Gap: `AT_TG_TA_TC_TA_CG_AC`

BWT FM-index Approach

This approach also has two versions:

- Naive Approach
- Enhanced Approach

BWT FM-index Approach : Method

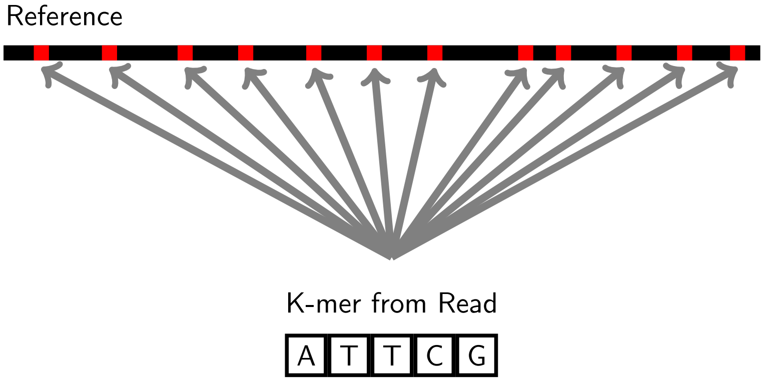
1. Index the Reference Genome
2. Take a Read
3. Take the Next K-mer. If There is not Any, Go to Step 9

9. Exit

BWT FM-index Approach : Method

1. Index the Reference Genome
2. Take a Read
3. Take the Next K-mer. If There is not Any, Go to Step 9
4. If it is not in the Reference, Go to Step 3
5. Let $i = 1$
6. If $(K + i)$-mer does not Exist, Go to Step 8
7. Do $i = i + 1$ and Go to Step 6.
8. Write the Locations of $(K + i - 1)$-mer To the Output and Go to Step 3
9. Exit

BWT FM-index Approach : Method



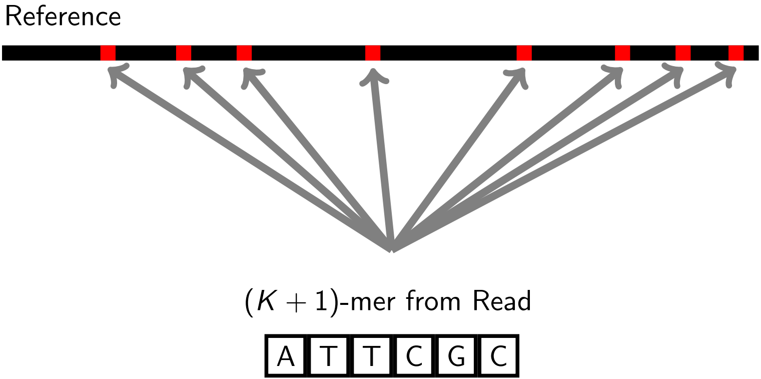
Reference

K-mer from Read

A T T C G

Figure: A K-mer with Value $K = K_{min}$ is Picked Up from Read and Indicated Where The K-mer is Found in the Reference.

BWT FM-index Approach : Method



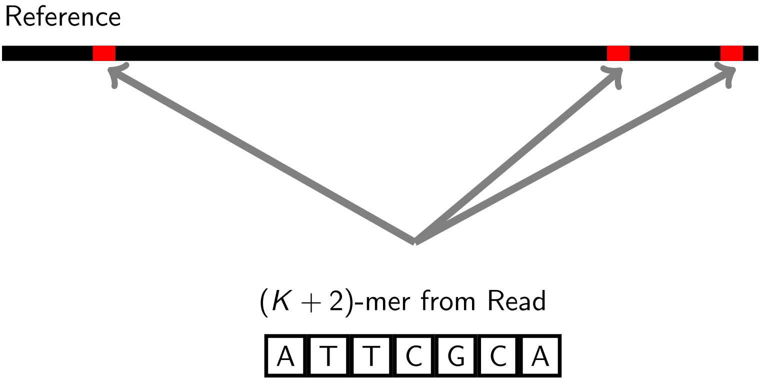
Reference

$(K+1)$-mer from Read

A T T C G C

Figure: Extending One Base in K-mer, The Locations of $(K+1)$-mer in the Reference is Reduced.

BWT FM-index Approach : Method



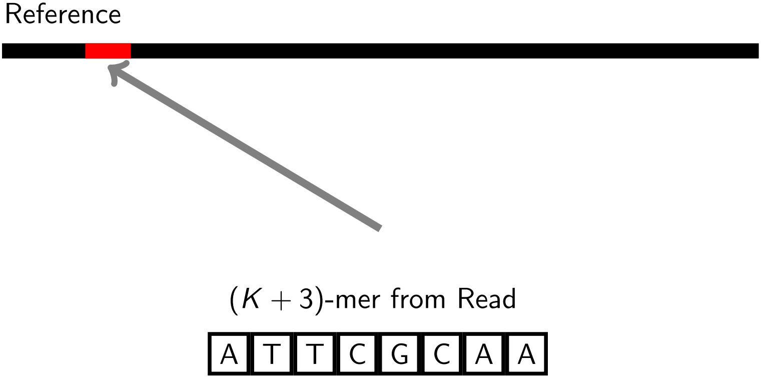
Reference

$(K + 2)$-mer from Read

A	T	T	C	G	C	A

Figure: Extending One More Base , The Locations of $(K + 2)$-mer in the Reference is Reduced And Now It is Only 3.

BWT FM-index Approach : Method



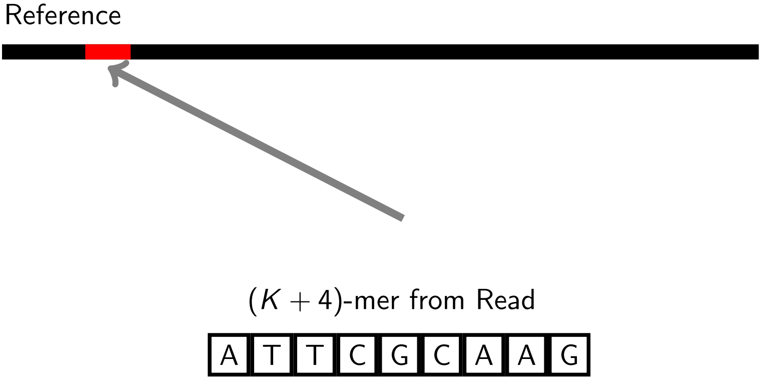
Reference

$(K+3)$-mer from Read

A	T	T	C	G	C	A	A

Figure: Continuing the Extension, $(K+3)$-mer is Created. The Count in Reference is Only One.

BWT FM-index Approach : Method



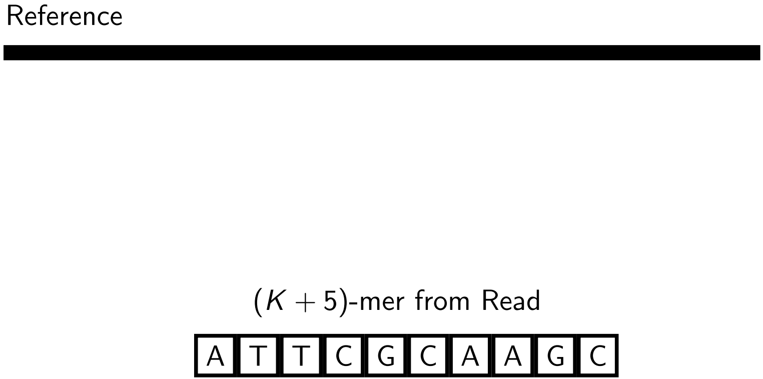
Reference

$(K+4)$-mer from Read

A T T C G C A A G

Figure: $(K+4)$-mer is Made By Appending One Base From Read. It has No Consequence in The Count in Reference.

BWT FM-index Approach : Method



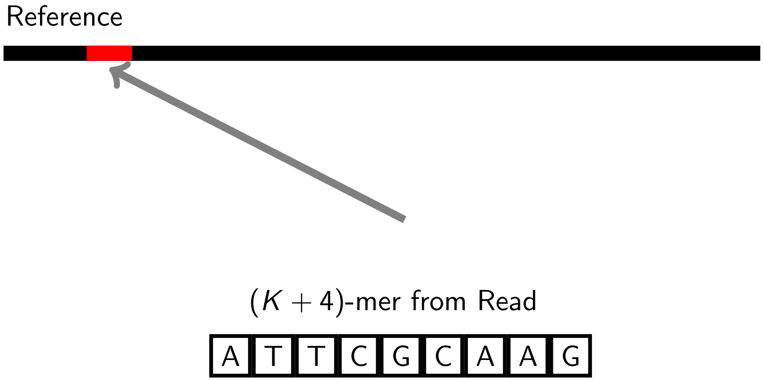
Reference

$(K+5)$-mer from Read

A	T	T	C	G	C	A	A	G	C

Figure: One Base Extension in $(K+4)$-mer, There is No Existence of $(K+5)$-mer in Reference. So, the Locations Got From $(K+4)$-mer Would be Considered as Final.

BWT FM-index Approach : Method



Reference

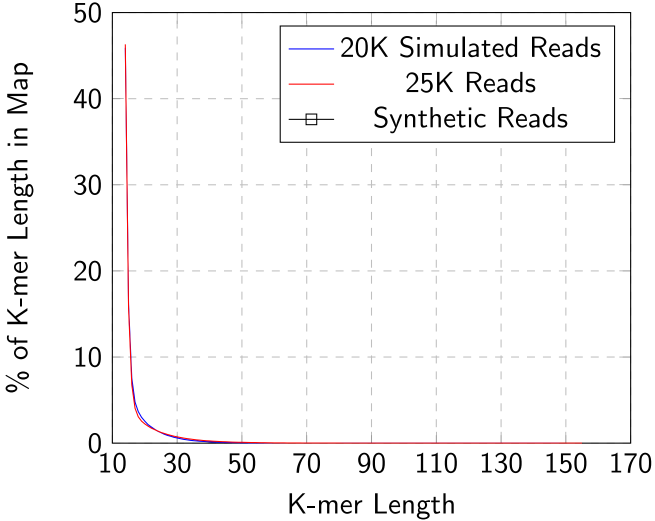
$(K + 4)$-mer from Read

A	T	T	C	G	C	A	A	G

Figure: $(K + 4)$-mer is Made By Appending One Base From Read. It has No Consequence in The Count in Reference.

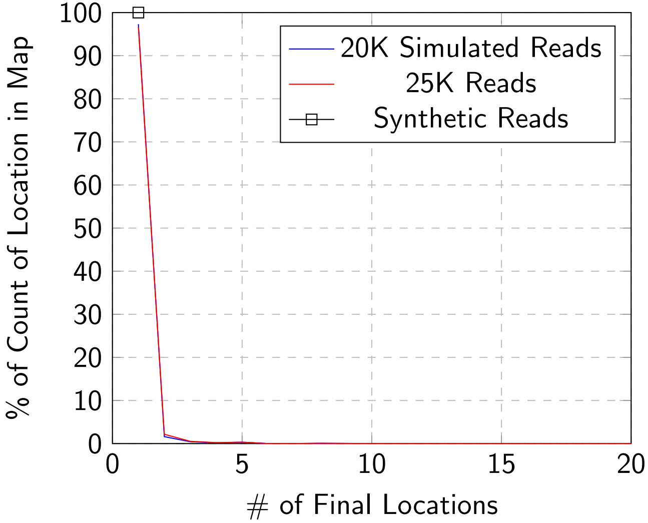
BWT FM-index Approach : Result

Which Length of K-mer Dominates the Mapping by What Percentage



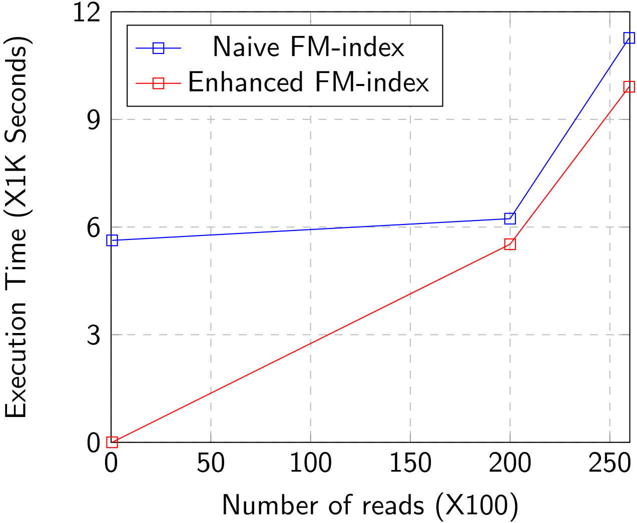
BWT FM-index Approach : Result

# of Locations Where the Final K-mers are Found VS Their Percentage



BWT FM-index Approach : Naive vs Enhanced

Exec. Time Comparison Between Naive vs Enhanced FM-index Approach



BWT FM-index Approach vs Minimap

Memory Requirement for Indexing: FM-index Approach vs Minimap

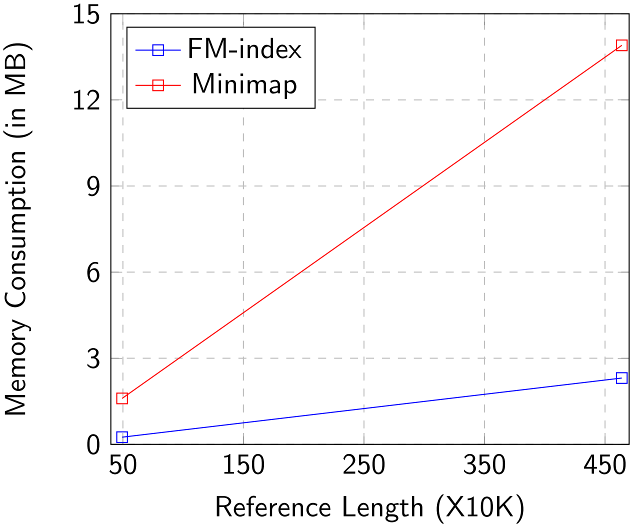


Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Hotchpotch Recipe

- Minimizer
- Window Technique

Hotchpotch Recipe

- Minimizer
- Window Technique
- Gap Insertion
- Enhanced BWT FM-index

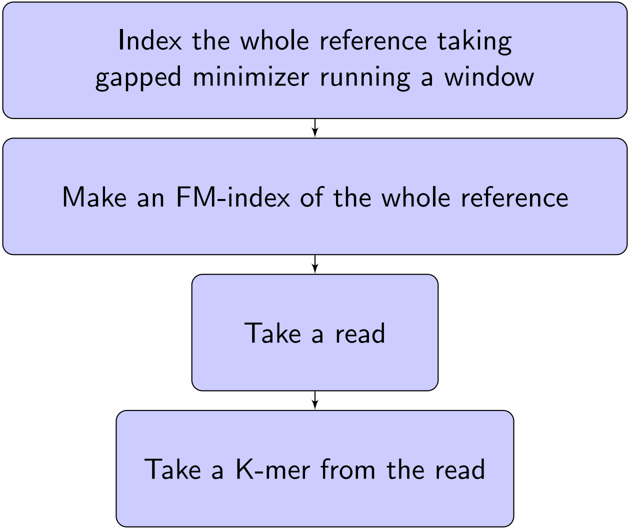
Hotchpotch Recipe

- Minimizer
- Window Technique
- Gap Insertion
- Enhanced BWT FM-index

NanoMapper

A Hot Hotchpotch having All Above Characteristics

NanoMapper



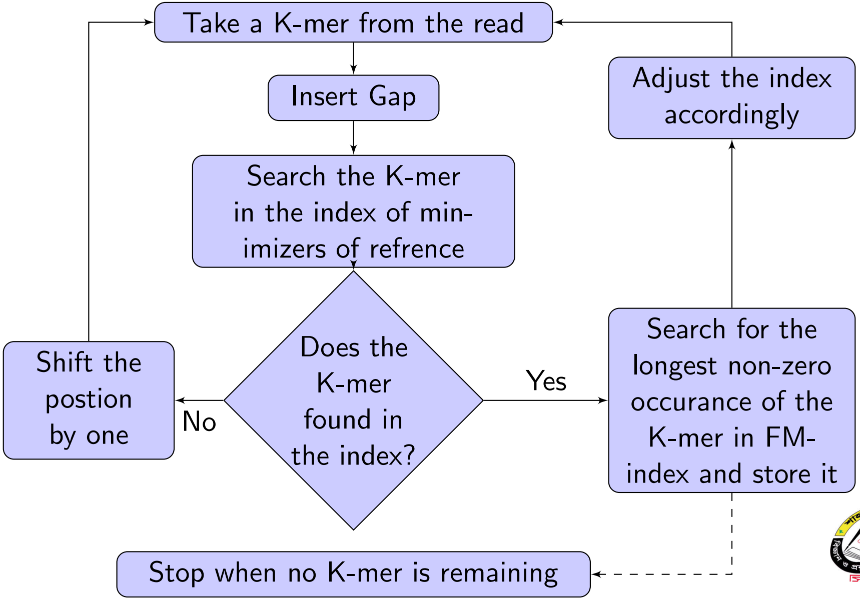
Index the whole reference taking gapped minimizer running a window

Make an FM-index of the whole reference

Take a read

Take a K-mer from the read

NanoMapper



Take a K-mer from the read

Insert Gap

Search the K-mer in the index of minimizers of refrence

Does the K-mer found in the index?

No

Shift the postion by one

Yes

Search for the longest non-zero occurance of the K-mer in FM-index and store it

Adjust the index accordingly

Stop when no K-mer is remaining

NanoMapper : Result

Table: Summary of Processing Reference genome

No	Reference Name	Reference Length	# of Minimizer	Indexing Time (Mini.)	Indexing Time (FM)	Memory Usage (MB)
1	E.Coli	4639211	682246	0.46	*1.76*	1.67
2	Synthetic	493290	12225	0.04	*0.16*	0.18

NanoMapper : Result

Table: Summary of Processing Read Sequences

No	Name of Data Set	Total Length of Reads	Time to Map (Naive)	Time to Map (Enhanced)
1	20K Simulated Reads	118335765	*19538.9 (5h 26m)*	*17051 (4h 44m)*
2	25K Reads	216906558	*9408.94 (2h 37m)*	*9869.78 (2h 45m)*
3	Synthetic Reads	500000	*3867.97 (1h 5m)*	*0.90*

Table of Contents

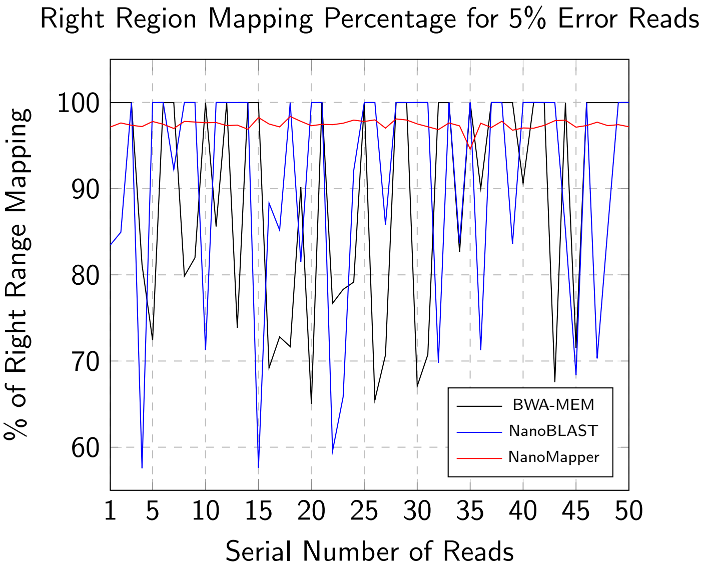
1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

NanoMapper : Accuracy

Table: Mapping Comparison While 5% Error Added in Read Data

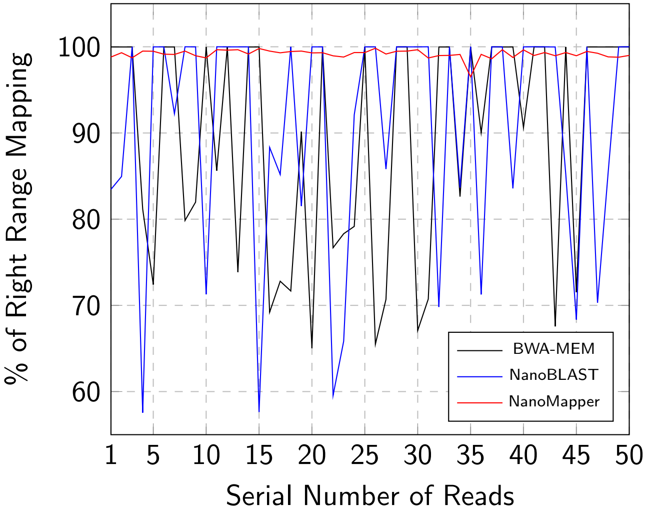
Name of the Tool	Right Position (%)	Wrong Position (%)
BWA-MEM	88.4	11.6
NanoBLAST	86.55	13.45
NanoMapper	97.42	2.58

NanoMapper : Accuracy



NanoMapper : Accuracy

Eliminating K-mers Having Length $\leq$ 15



NanoMapper : Accuracy

Table: Mapping Comparison While 10% Error Added in Read Data Eliminating K-mer Having Length ≤ 15.

Name of the Tool	Right Position (%)	Wrong Position (%)
BWA-MEM	88.82	11.18
NanoBLAST	88.72	11.28
NanoMapper	99.02	0.98

NanoMapper : Accuracy

10% Error Reads Eliminating K-mers Having Length ≤ 15

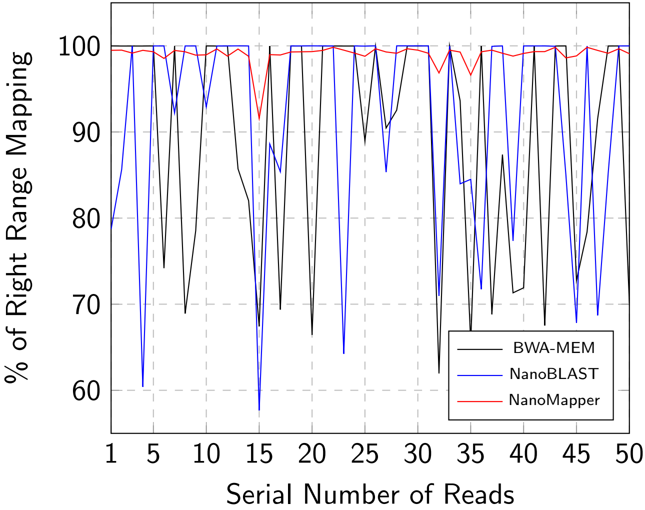


Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Future Work

- API Enhancement

Future Work

- API Enhancement
- Integration with NanoBLASTer

Future Work

- API Enhancement
- Integration with NanoBLASTer
- Testing

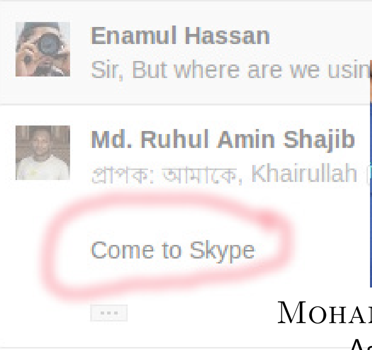
Future Work

- API Enhancement
- Integration with NanoBLASTer
- Testing
- Developing New Aligner

Table of Contents

1. Human Cell To Mapper
2. Why Mapper Is Needed?
3. What are the challenges?
4. Existing Tools
5. Our Approaches
6. Gapped Minimizer and BWT FM-index Merged Approach
7. Comparison
8. Future Work
9. Conclusion

Special Thanks



Enamul Hassan

Sir, But where are we usin n of vectors? Are not we taking t...

Md. Ruhul Amin Shajib

প্রাপক: আমাকে, Khairullah

Come to Skype

...

Mohammad Ruhul Amin

Assistant Professor,

Department of Computer Science and Engineering

Shahjalal University of Science and Technology

Thank You