An Approach To Measure Quality among Multiple Alignments and Align Oxford Nanopore Single Molecule Sequence Reads

By

Faisal Ahmed

Reg. No.: 2011331047

Supervisor:

Biswapriyo Chakrabarty

Lecturer,

Dept. of CSE, SUST

Co-Advisor:
Md. Ruhul Amin Shajib
Assistant Professor,
Dept. of CSE, SUST

What is Sequence Alignment

Sequence alignment is a way to arrange several sequences of DNA or RNA to identify similar regions that may have consequences on relationships among the sequences.

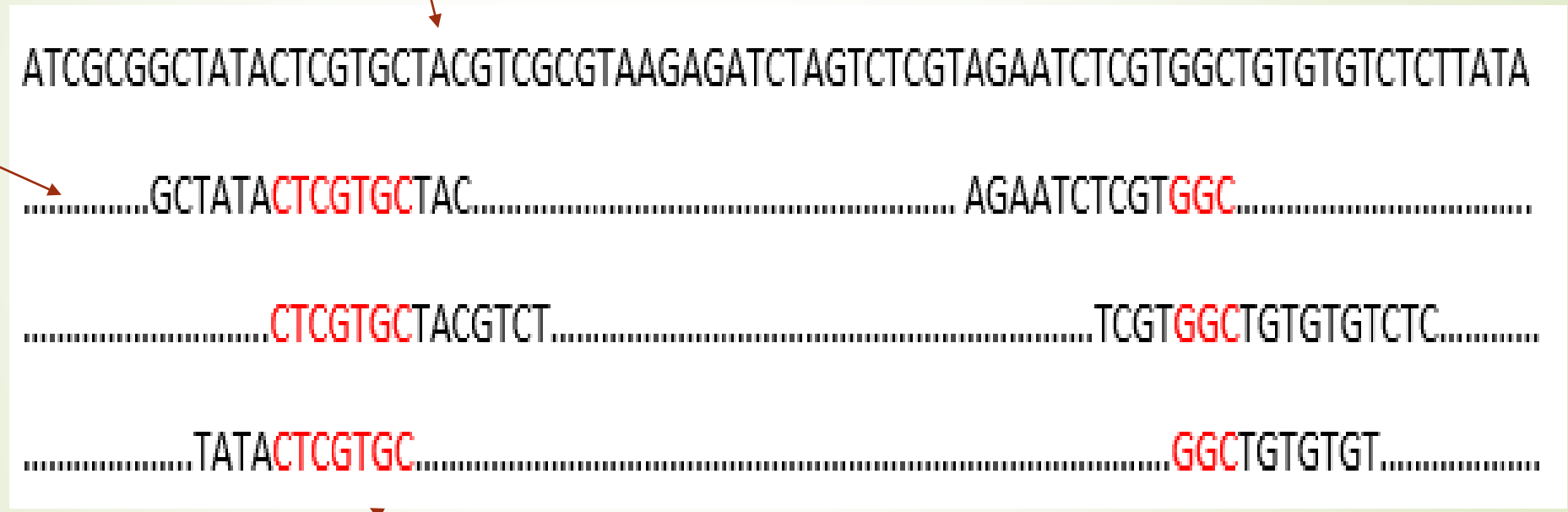
Sequence Alignment

- Global Alignment
- Local Alignment

Our task on Sequence Alignment

- Here we are working on finding similarity between a reference genome and some reads.

Reference Genome



Short Reads

Common Area
Coverages

Sequence Alignment Mapping

- Sequence Alignment Mapping (SAM) is a standard format to show how a tool makes alignment on several reads with any other genome reference.

Sequence Alignment Mapping File Format Specifications

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

channel_93_read_44_template_/dev/shm/downloads/LomanLabz_E.coli_MG1655_3311_1_ch93_file45_strand.fast5
0
chrMG1655
4164525
255
1M3I9M1I2M3X1M1X2D7M1X1I7M2X1M1D1M2D7M1D1M1X1D1M2D2M3X1D2M1I1M1I2M1I2M1I2M1X2M1I3M5D5M1X5M1X3M1D1M1D2M1D1M1D1M1D2M1X1M1X1I2M1D2M1D2M2D2M3D1M1X1M3X2M1X1D1X1M2I1M1I2M1X4
1I1M2D2M1X1D2M1I1M1X4M1X1I1M3I1M1I2M1X1I1M2X1M1I2M1D1M1I2M1X1M2X1M1X1I2M1X2M1X6M3X1M2D5M1I2M1X2M2X1M2D2M1D1M1X2M1X1D3M1X1M1D6M1X1M1X1I2M1D2M1D2M1X1M2I2M1D2M1I1M1X1I4M1
X3M1I1M2X1I2M2I1M1I1M1I3M1X1M1X1M1X1M1D2M4X2M2I2M1X2M1X2M3I3M1X1M2X1M1X2M2X2M1I1M1I3M1X2M3I2M1I1M1I3M2X4M1X5M3I4M1D2M1I4M2X2I2M1I1M1X1M2X1M1I2M1X1M2X3M1X2D1M1D1M2D4M1D
3M3I7M1D1M2D2M1X1M1D3M1I2M1X1I2M1X1M2X1I5M2D5M1X1D3M2X1M2D3M1X3D2M1I1M1X5M2X1M2D3M1X2M1X1I3M2I2M3X4M1X1M3I2M1I2M1X1M1D1M1D2M1D5M1D3M1X2D2M1X2M2I3M2X1D6M1X3I3M2X1D6M1X1
D2M1D5M2I3M1X3M1X1M1D1M1X1M1X2M2X1M1I2M1X1D1M2X1M1X1M1D5M1X2I5M2X1D1M1D5M2X7M1D1M1D2M3I4M1X1M1X3M2X1M1X2D14M1X6M1X6M1X1M1X1M2X1M1D1X3M1D4M1X2M1D2M1X2M3D1X1M2X4M1D2M1X1
1X2M1X3M1X3M1D2X6M1D2M2X3M3I2M3X1M3X5M1X1M1X4M1D1M2X2M2X2M4X2M5D1M1D1M1D1X6M2D1M2D1M1D3X2M1D1M1X1M1X3M1D3M1I2M1X1M1X2M1X2M1I3M4X1M1D2X1M2X1M1I2M2D2M1D1X2M3X1M2X4M3X2M2
4M1I2M2I1M2I2M3I1M2I2X3M1D2M2X2M1I1M1I2X3M5I2M1I1M1I1M1I4M2X2M1D2X2M1X2M3I1X4M1I1X2M2X3M1X2M1X3M1I1M2D4M2D1M2D1M1D1X3M2I2M1D1M1X2M1D2M3D1M1X1M1D1X2M1X2M1I4M1D5M1I3M1I2
*
0
0
ATGCCTACGATCCGCCCGAAATCGTTATTTATTCGTCTCCACTCGATGGCGGCGGGCAGCGCGTTCGCATTGCGGCATGTCCTCGTTTGTCCGTGATGTTCTCGGTTCCTATATTGAGGAAAGCGGAGGCCGACGCGTACCCATTTATTTGGGGGCTGGGACGTAAG
CTCGTTTCCCGTCTGAGCGAGAGGGCAGCCGTTCCTTTGCTCGAATGTGCGATGTCGCCCGCCGTTGCACGTTCTCCGATTCAAATGGTTCCTCGTGAACTAGTCGAATCGAAGTTTCATTCAATCGATCACGAGTTCACTTCGCATTTTTGTTGGTCAGCCATGCT
TTGGAAAGTCAATGGAGCTGGAAGACGGTACCCCCAGCCGAGCAAGCGTCGCAGCAGGAGGCTCTTCAAGTCACTGCCAACGATGTTCCCGAGATGATTAGCCGCGTAGTCACCGGTAGCCCCGGGCTGCCGAGCTGTTTGGATACAGAGCCAGAGCGGGCGGACTG
GCATCACGAGTCGTCGTAGTCCGTGGGGTCTGCGACTAAGCGGGGAGTCCGTCGAAACCTAGTAATTCGTGAGAAATCCGGTTGCGGCGTTCCCGTCAATCCCCGCACGCGTCATCGCCGAAATGAGTGGTCGGGGGAAGTGAACAACCTCGTCGGGCGGCCACATC
GGAGGCGGAGCGAAACAGGGCGTGAGGAGCGGCAACGATTTGCTTCCCGGGGTGGGGGCCAATTAGGCGTTGGGGCGTACACTTTTGCCCCGTAGCGTAGTTATTGAGTTACCAAGCGGGACCGAGGAGAACGAGTCCTCTGCTTCAGTCAACGGGGCGGGTGAACG
ACCAGCAGTCCGGGAGGCGAGAACGTCCTATCTCCTTCGGGAGGCACATGATAATCGGTAGGTCCCGCTATGGGGTGAATAGGGGGTCTGAGAGCGAAGCGTAACCTACTTGCAACCCAGCGCATATTTTCAGGGCGAGGTCGTACCGTTGCCCTGCCCGTTTCTAG

Tools To Evaluate SAM Files

- Qualimap
- Picard tools
- RNA-SeQC
- RSeQC

	Picard tools	RNA-SeQC	RSeQC	**Qualimap**
Scope	Various NGS applications	RNA-seq	RNA-seq	Various NGS applications
User interface	Command-line	GUI + Command-line	Command-line	GUI + Command-line
Input formats	Alignment: BAM; annotations: RefSeq; reference: FASTA	Alignment: BAM; annotations: GTF; reference: FASTA	Alignment: BAM; annotations: BED.	Alignment: BAM; annotations: GFF/GTF or BED
Output	Raw data and plots. Different for each tool	Summary report including both html reports and raw data	Raw data and plots. Different for each tool	Summary report (HTML or PDF) and raw data
Alignment statistics	Yes	Partially	Partially	Yes
Overall Summary	No	Yes	No	Yes
Coverage analysis	Genome-wide or genes	Genes and transcripts	Genes and transcripts	Genome-wide or arbitrary regions
Average gene 5'-3' coverage plot	No	Yes	Yes	No
Counts computation	No	Yes	Yes	Yes
Insert size estimation	Yes	Yes	Yes	Yes
Sequence quality	No	Yes	Yes	Yes
Multiple samples support	No	Yes	No	2 samples

Analysis On NanoBLASTer

- Evaluation of corresponding SAM file

Analysis On NanoBLASTer

- Evaluation of corresponding SAM file
- Analysis on results

Analysis On NanoBLASTer

- Evaluation of corresponding SAM file
- Analysis on results
- Experimenting new method on tool's used algorithm

Analysis On NanoBLASTer

- Evaluation of corresponding SAM file
- Analysis on results
- Experimenting variations on tool's used algorithm
- Analysis results on experimental changes

Data Set

- We took 1000 reads from Oxford Nanopore Single Molecule Reads with minimum 100 length and maximum of 1000 length.

Data Set

- We took 1000 reads from Oxford Nanopore Single Molecule Reads with minimum 100 length and maximum of 1000 length.
- Data Set is highly noisy.

Scale Measurement Definition

- Percentage of Identity means Percentage of Matches considering all penalties (insertion, deletion, mismatches)

Scale Measurement Definition

- Percentage of Identity means Percentage of Matches considering all penalties (insertion, deletion, mismatches)
- POI = Total number of matches/ (Total Matches + Total Penalties)

Scale Measurement Definition

- Counted-
- Total Matches
- Total Mismatches
- Total Insertions
- Total Deletions

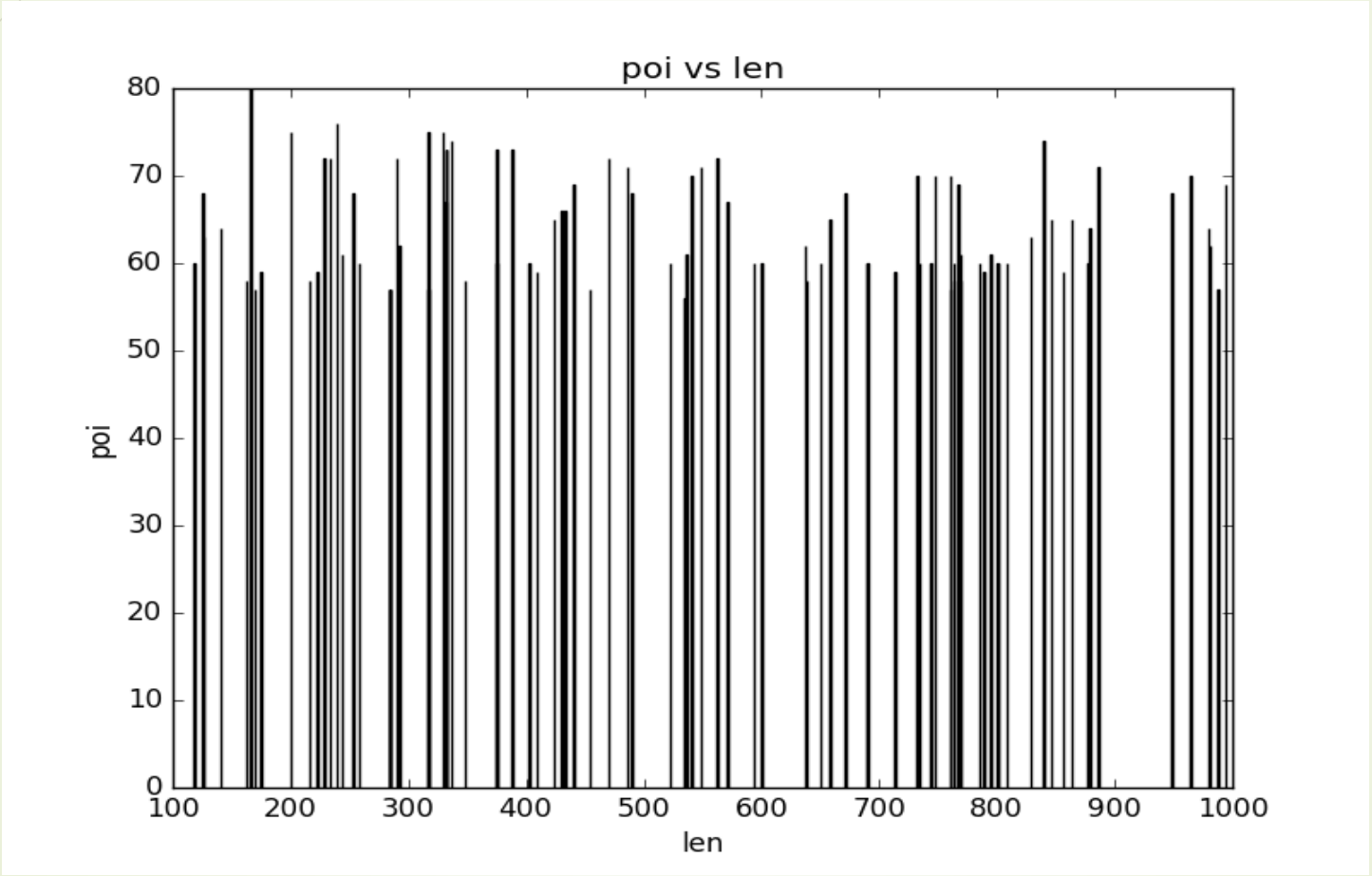
Measuring Quality

- Measured POI of sample data and compared with implemented alignment POI
- If POI (from implemented alignment) > POI (from sample data) then quality falls otherwise it will be called better in this scale

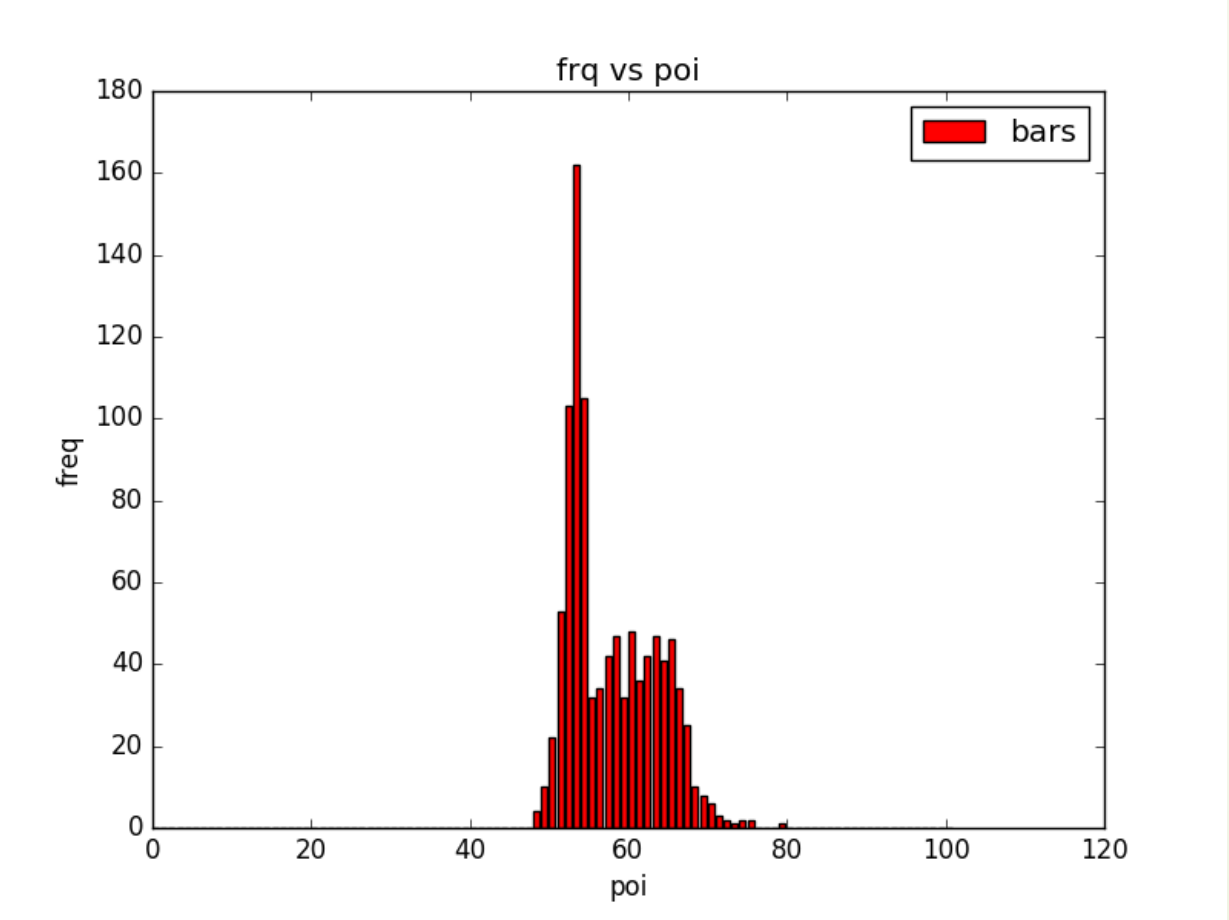
Graph Plotting

- X - axis is POI from 1 to 100
- Y - axis is frequency of POIs'

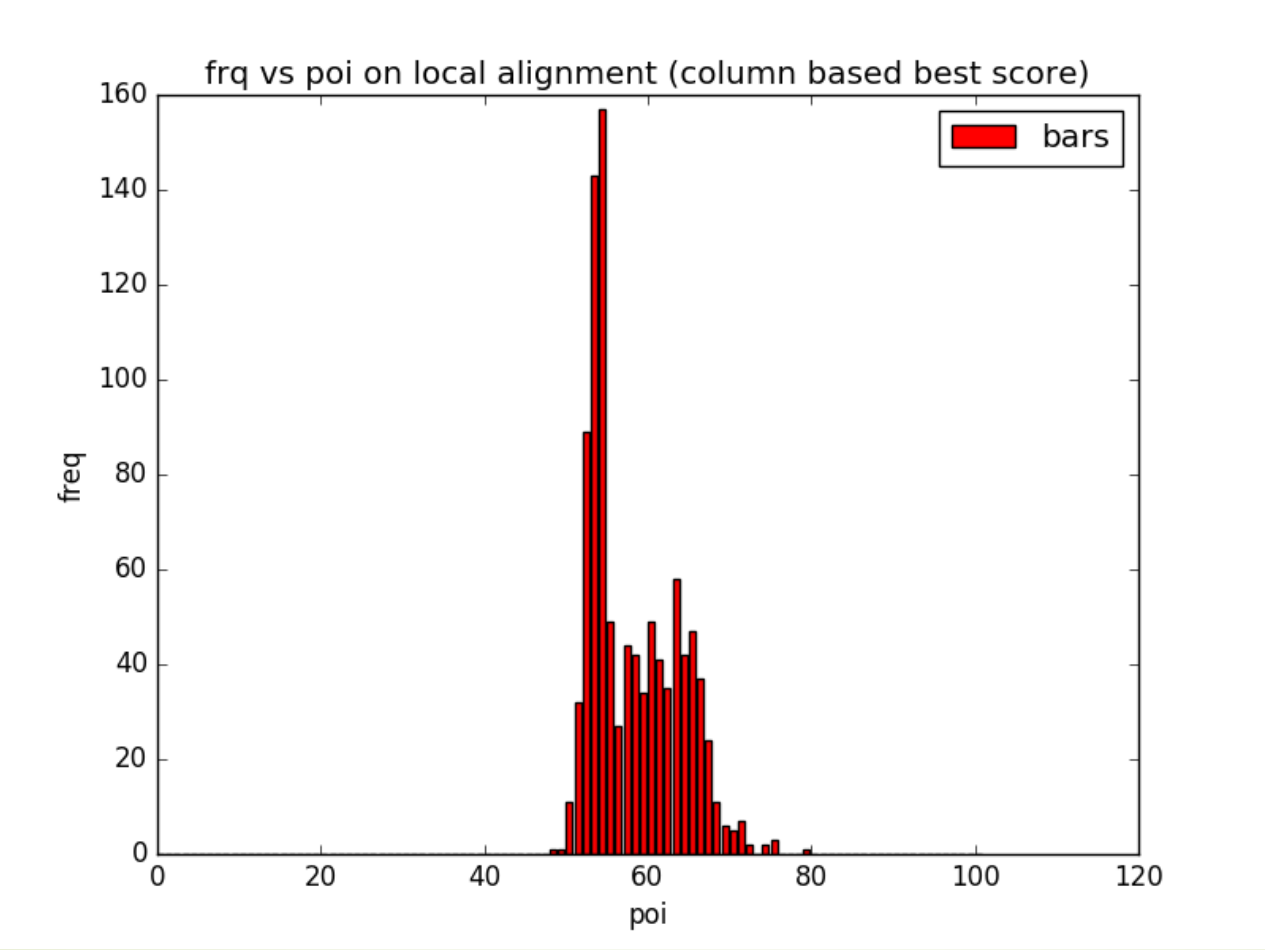
Max POI vs. Length



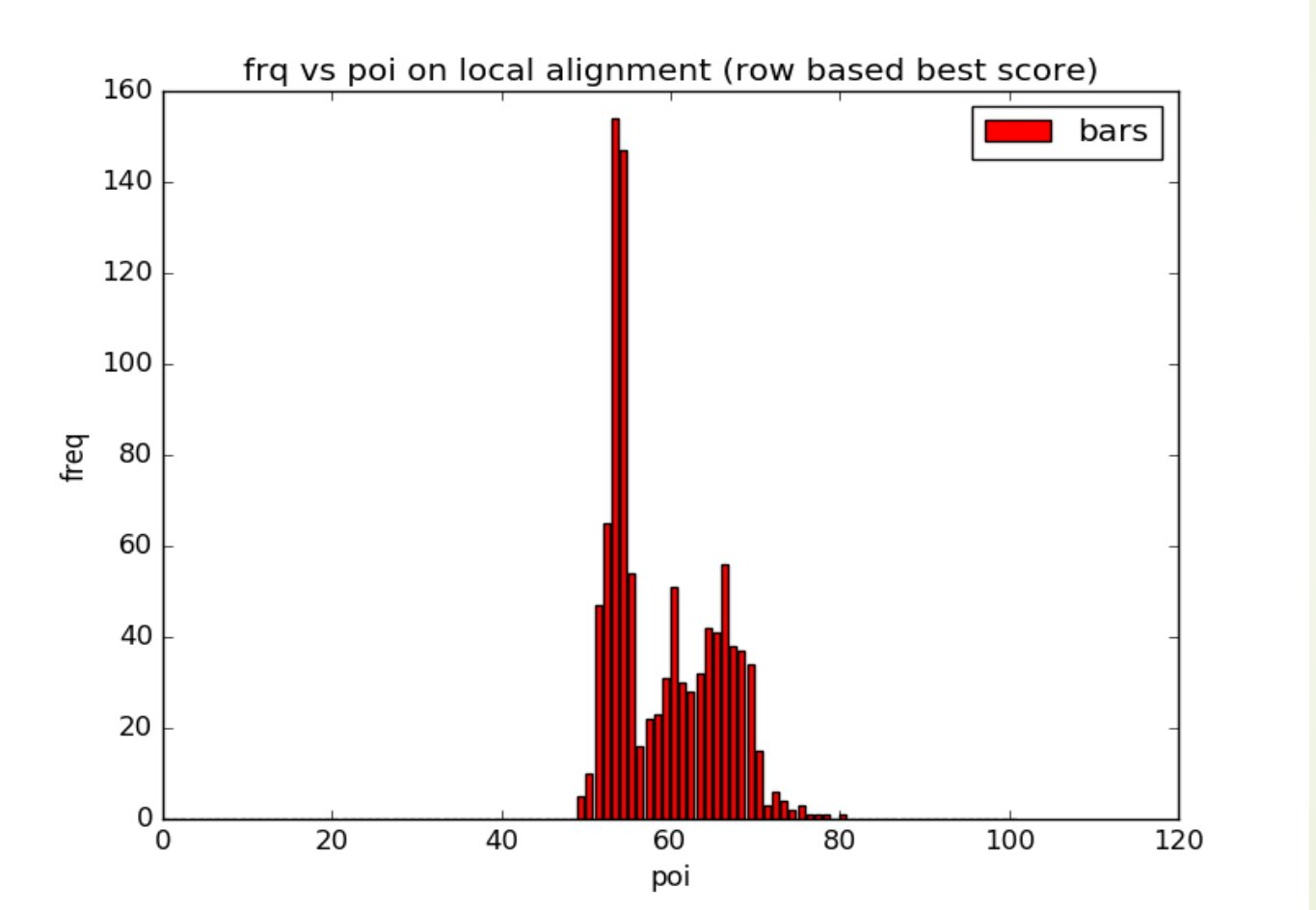
Frequency vs. POI (Global Alignment)



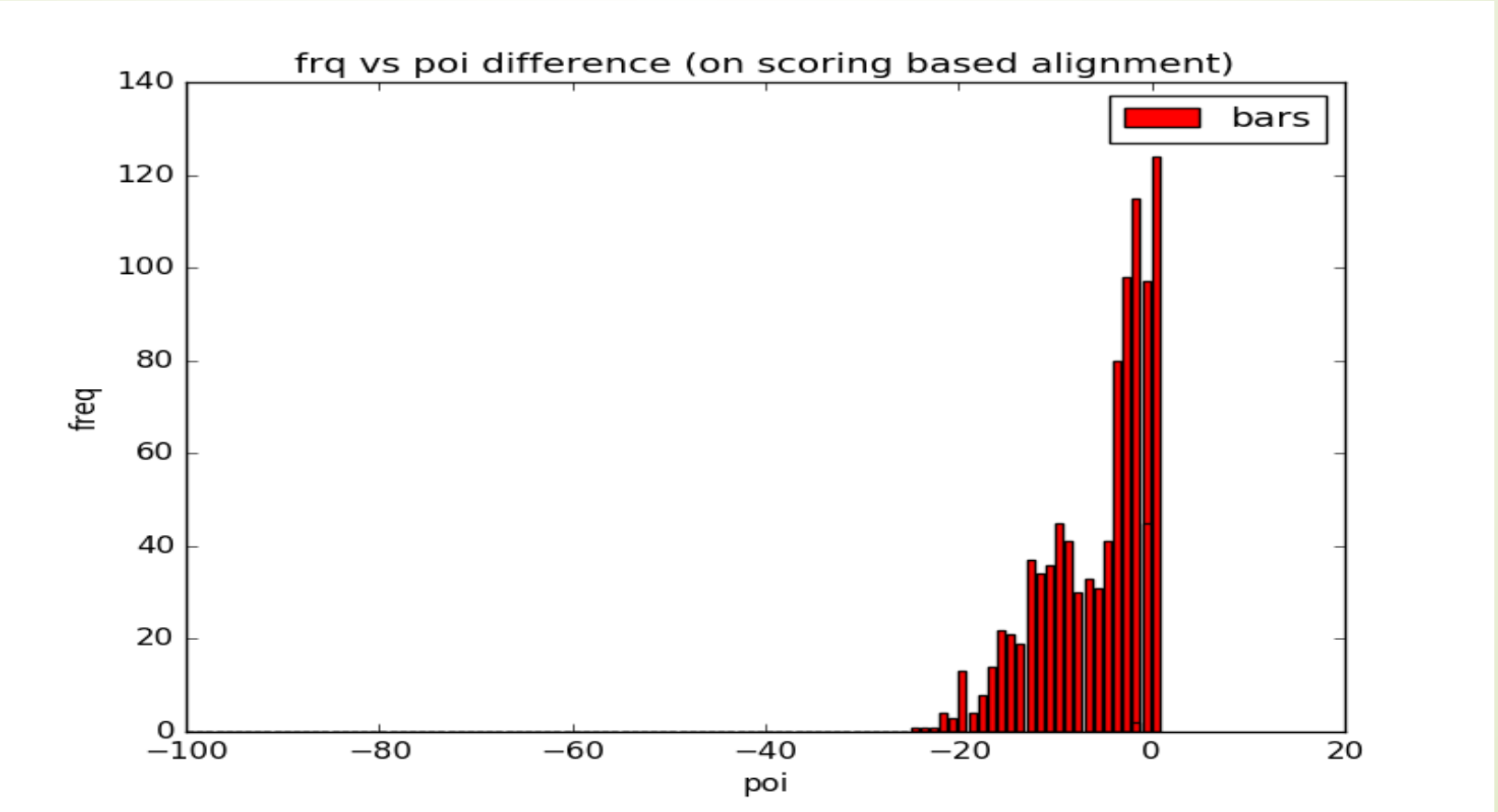
Local Alignment (Considering Column)



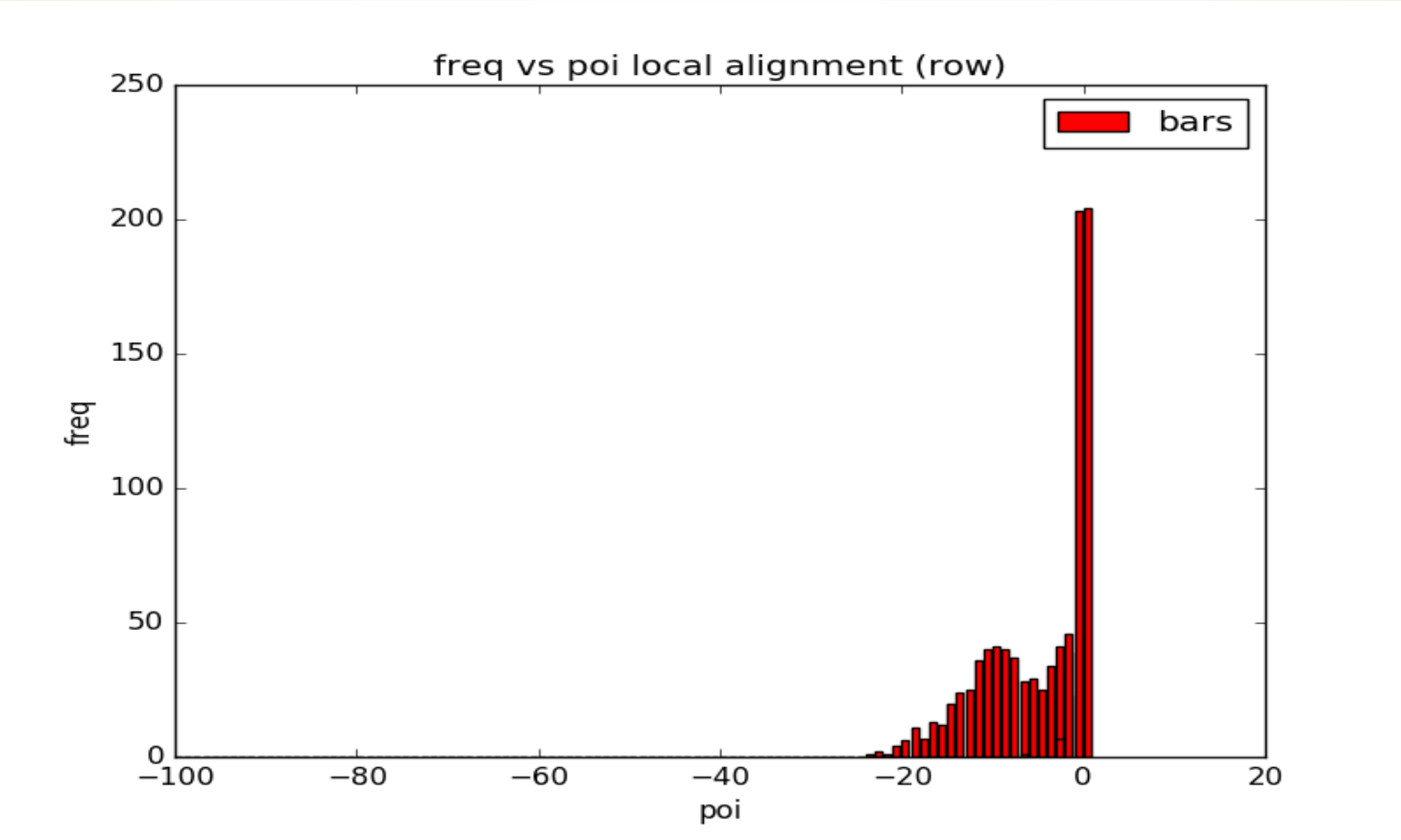
Local Alignment (Considering Row)



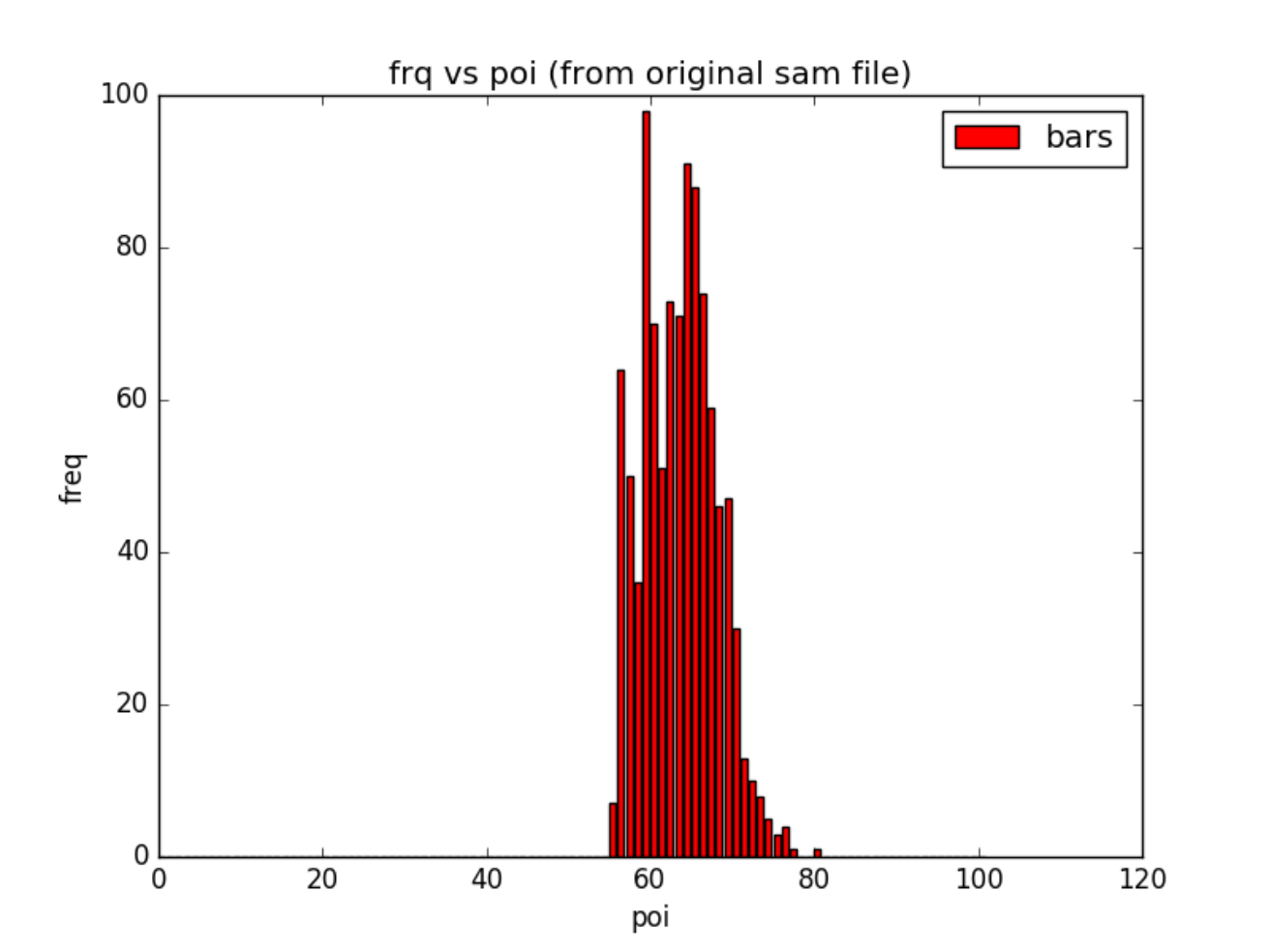
Difference with Implemented Global Alignment



Difference with Implemented Local Alignment



Freq. vs. POI (from original Input Data)



Experimented Method

- Mapping reference genome into several K-mers

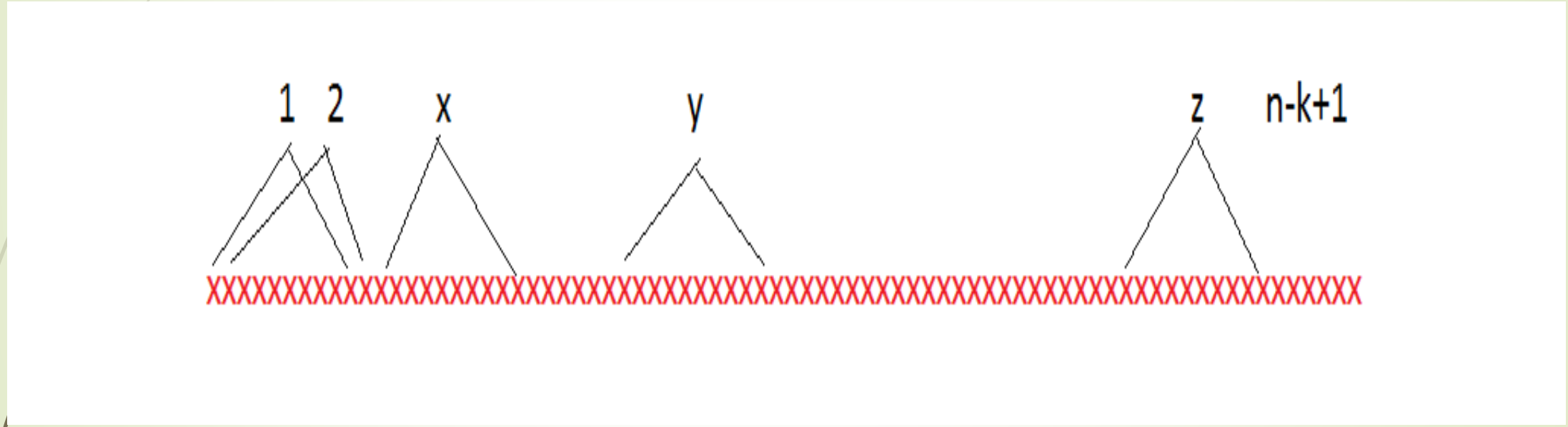
Experimented Method

- Mapping reference genome into several K-mers
- Mapping reads into several K-mers

Experimented Method

- Mapping reference genome into several K-mers
- Mapping reads into several K-mers
- Preparing 2-D vector with the mapped values with matching considering threshold values.

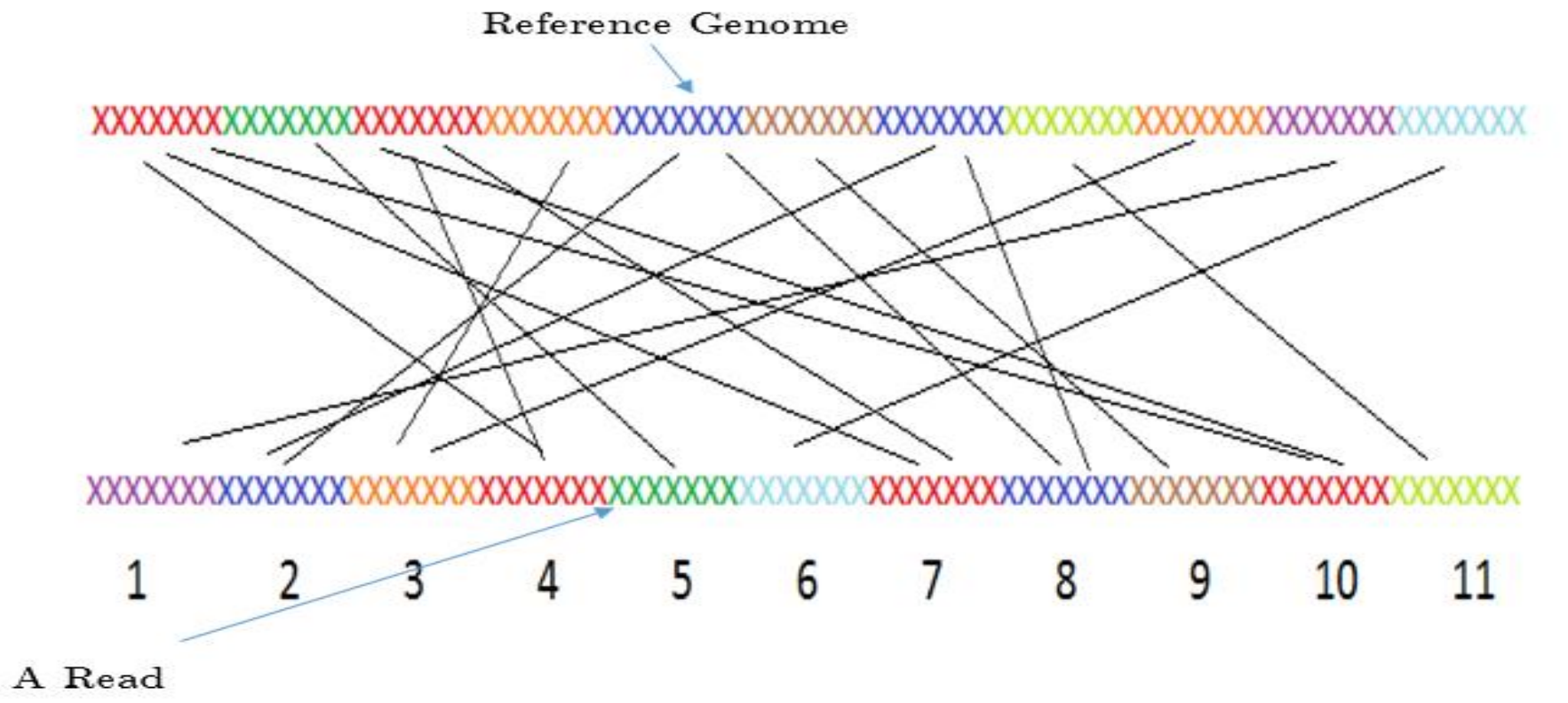
Mapping Reference Genome



Mapping Reference Genome

1	xxxxxxx
2	xxxxxxx
3	xxxxxxx
4	xxxxxxx
5	xxxxxxx
n-k-1	xxxxxxx
n-k	xxxxxxx
n-k+1	xxxxxxx

Preparing 2-D Vector

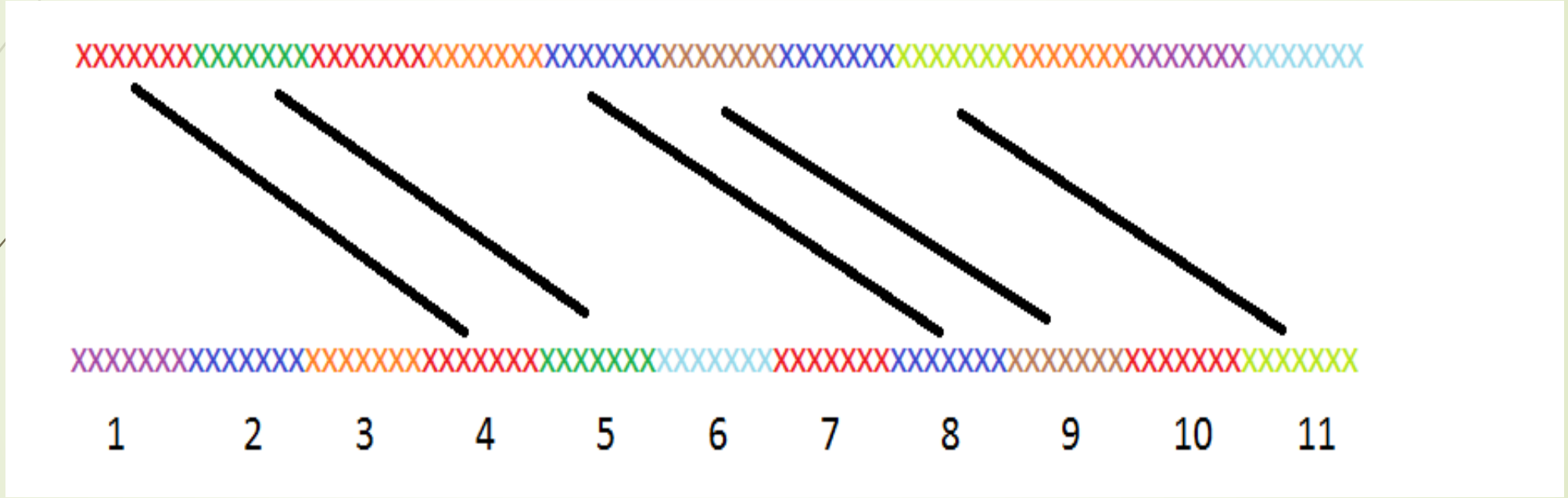


K-Mer starting positon from Reference Genome	K-Mer starting position from a Read
1	4
1	7
1	10
2	5
3	4
3	7
3	10
4	3
5	2
5	8
6	9
7	2
7	8
8	11
9	3
10	1
11	6

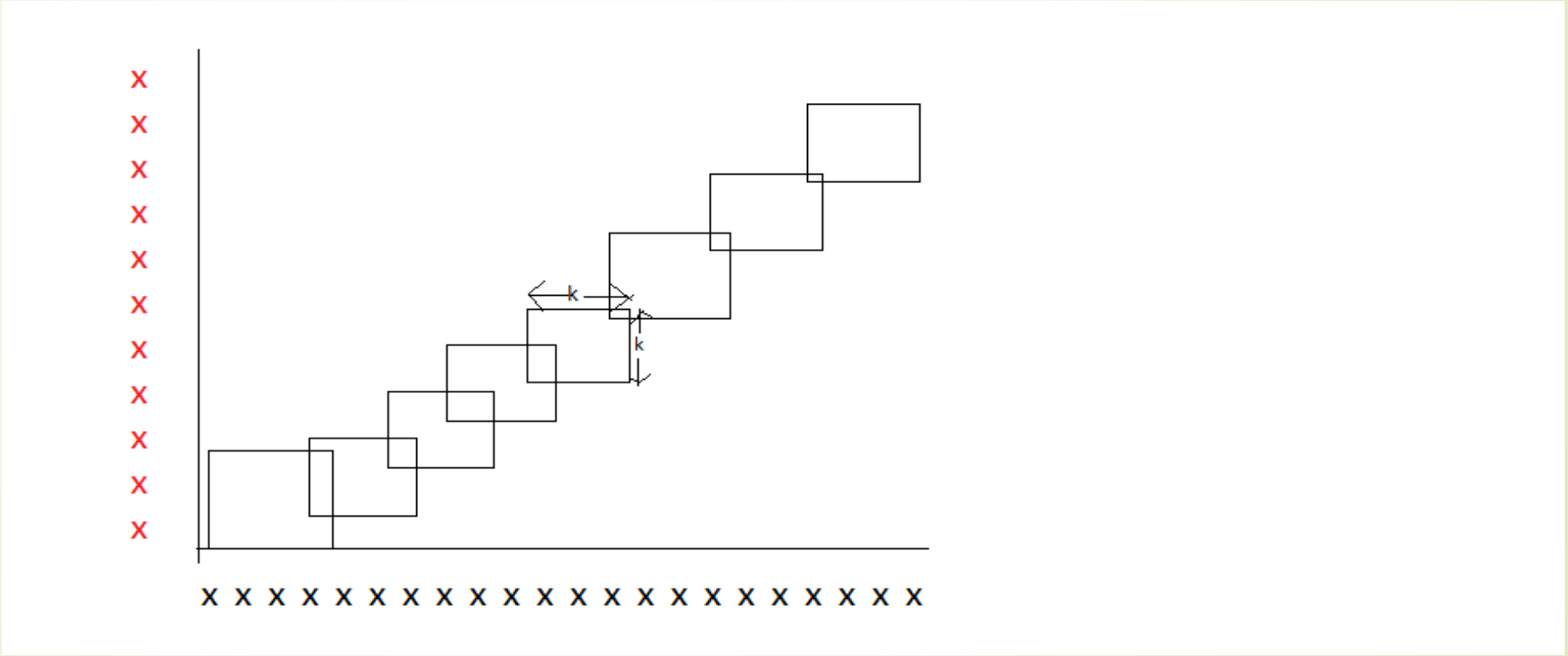
Result After Running Longest Increasing Subsequence Algorithm on the Vector

K-Mer starting positon from Reference Genome	K-Mer starting position from a Read
1	4
2	5
5	8
6	9
8	11

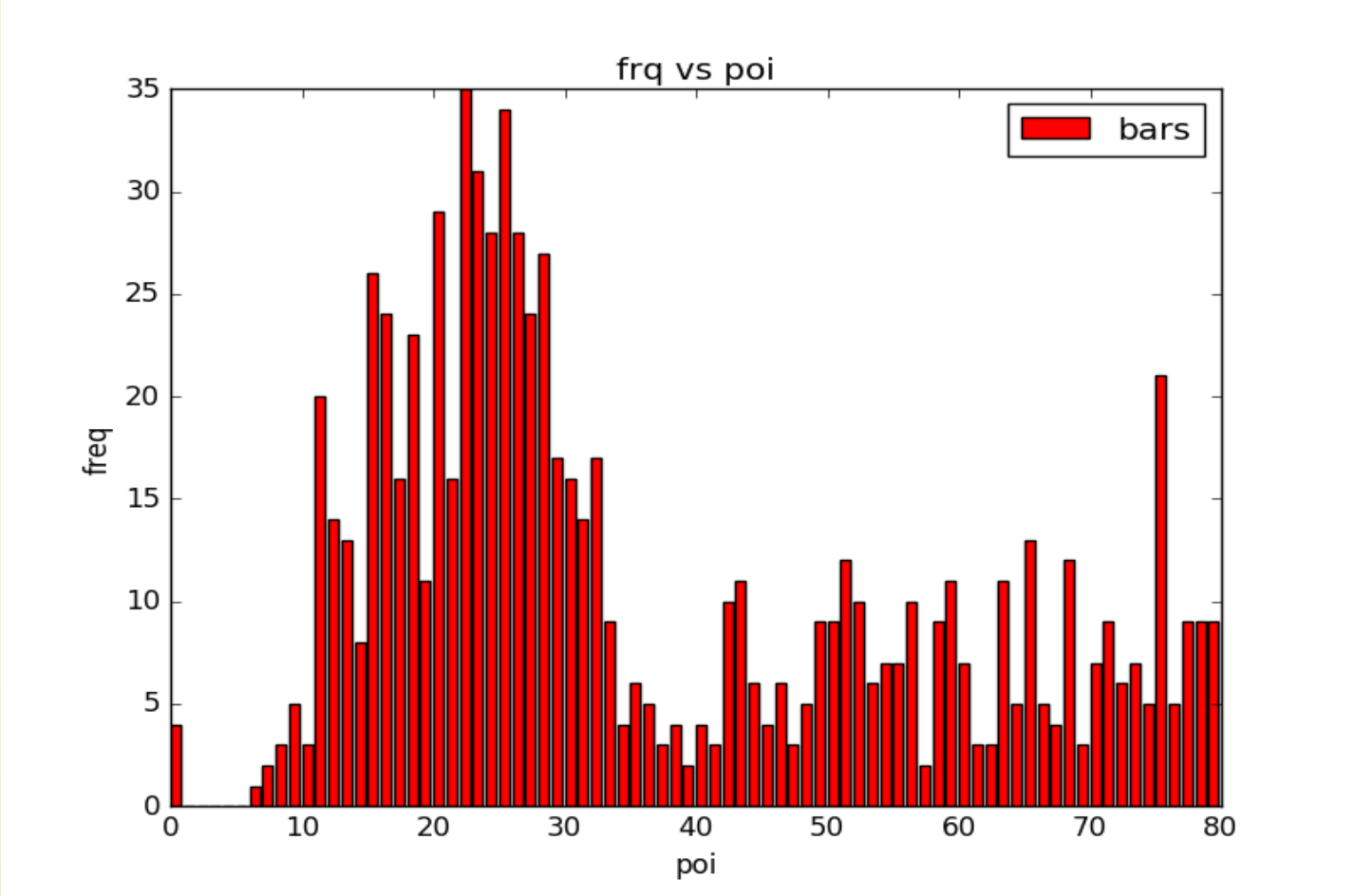
Visualizing LIS result



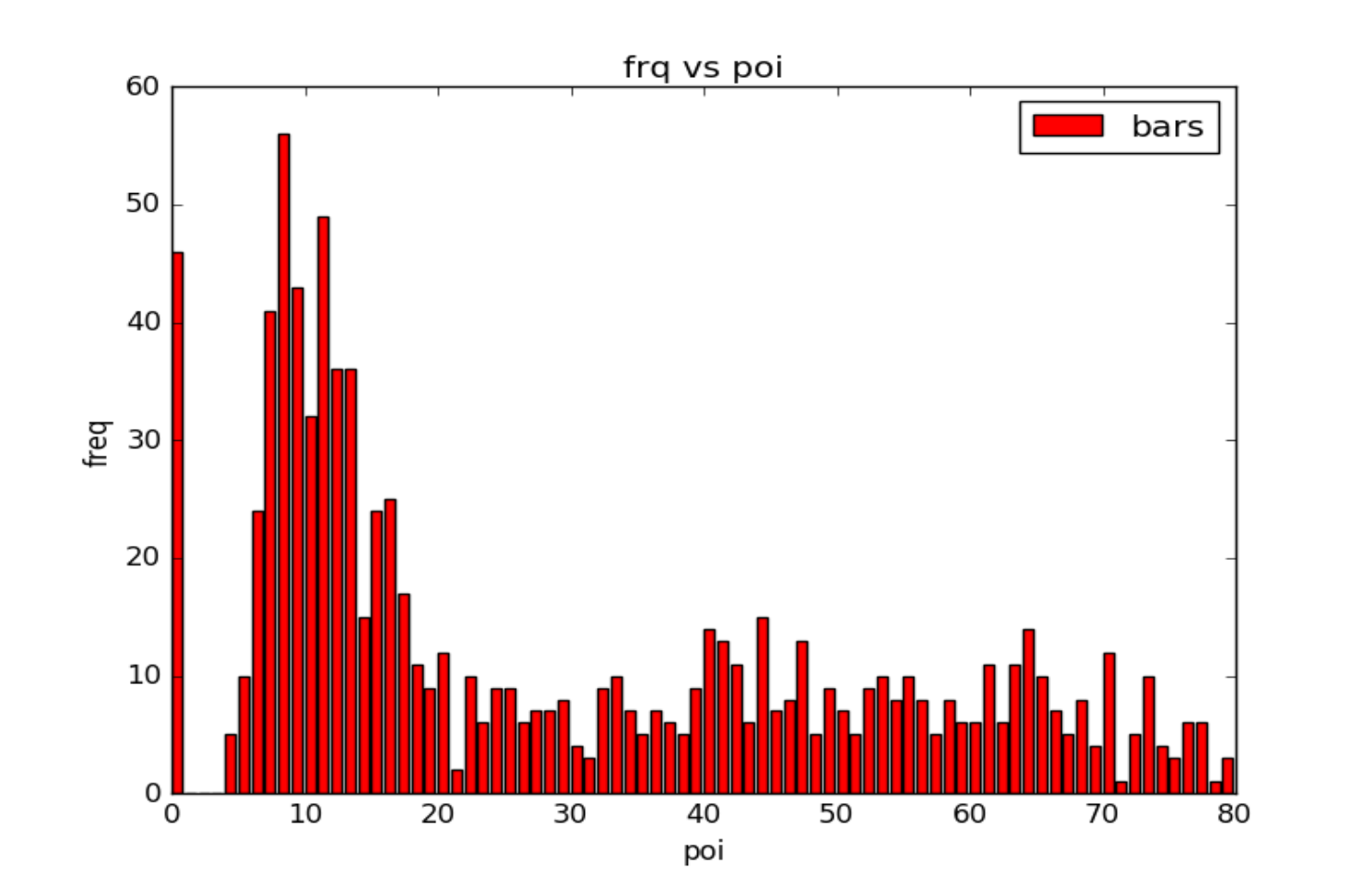
Block Alignment On Unaligned Areas



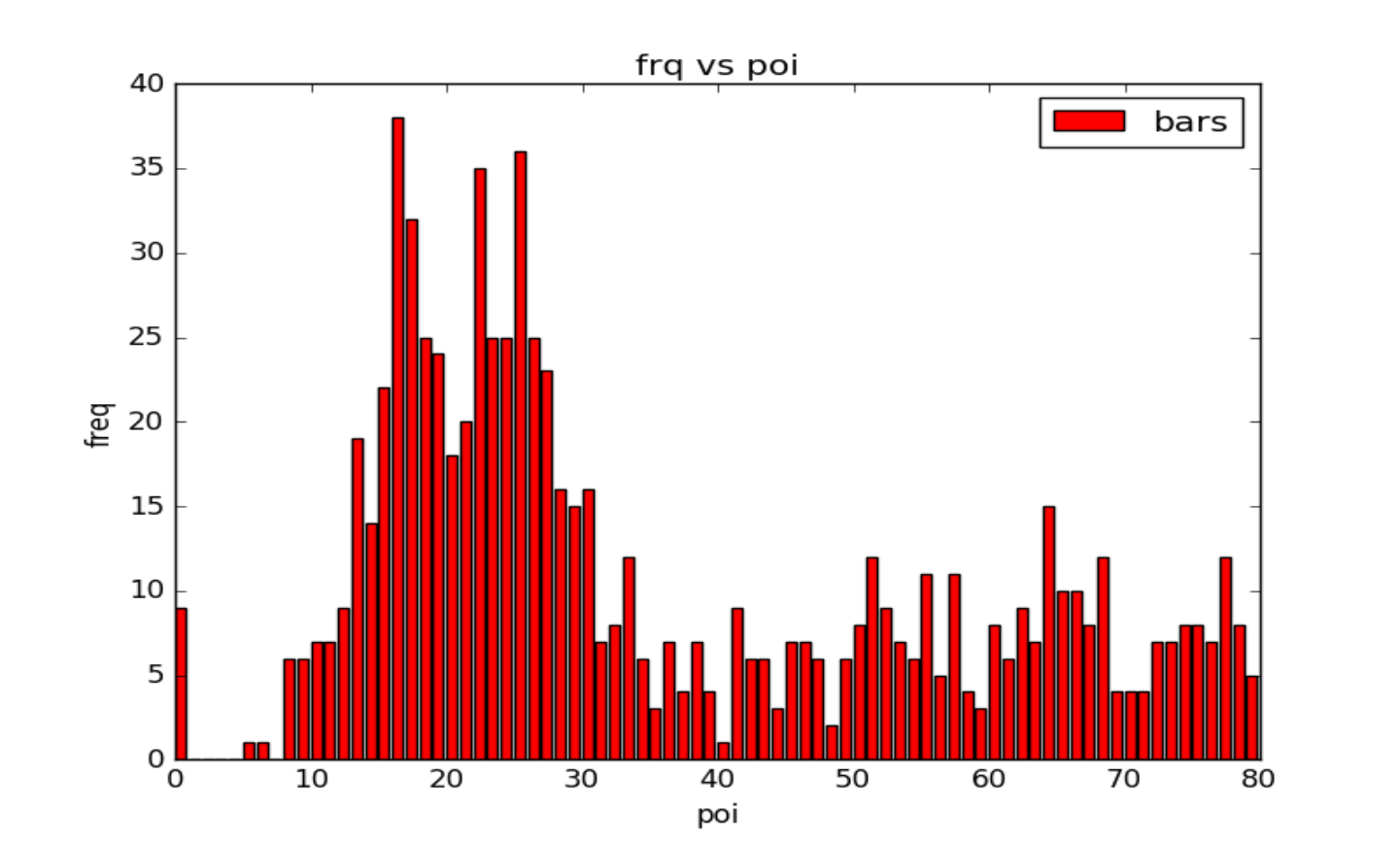
For k=11


frq vs poi
bars
freq
poi
0
5
10
15
20
25
30
35
0
10
20
30
40
50
60
70
80

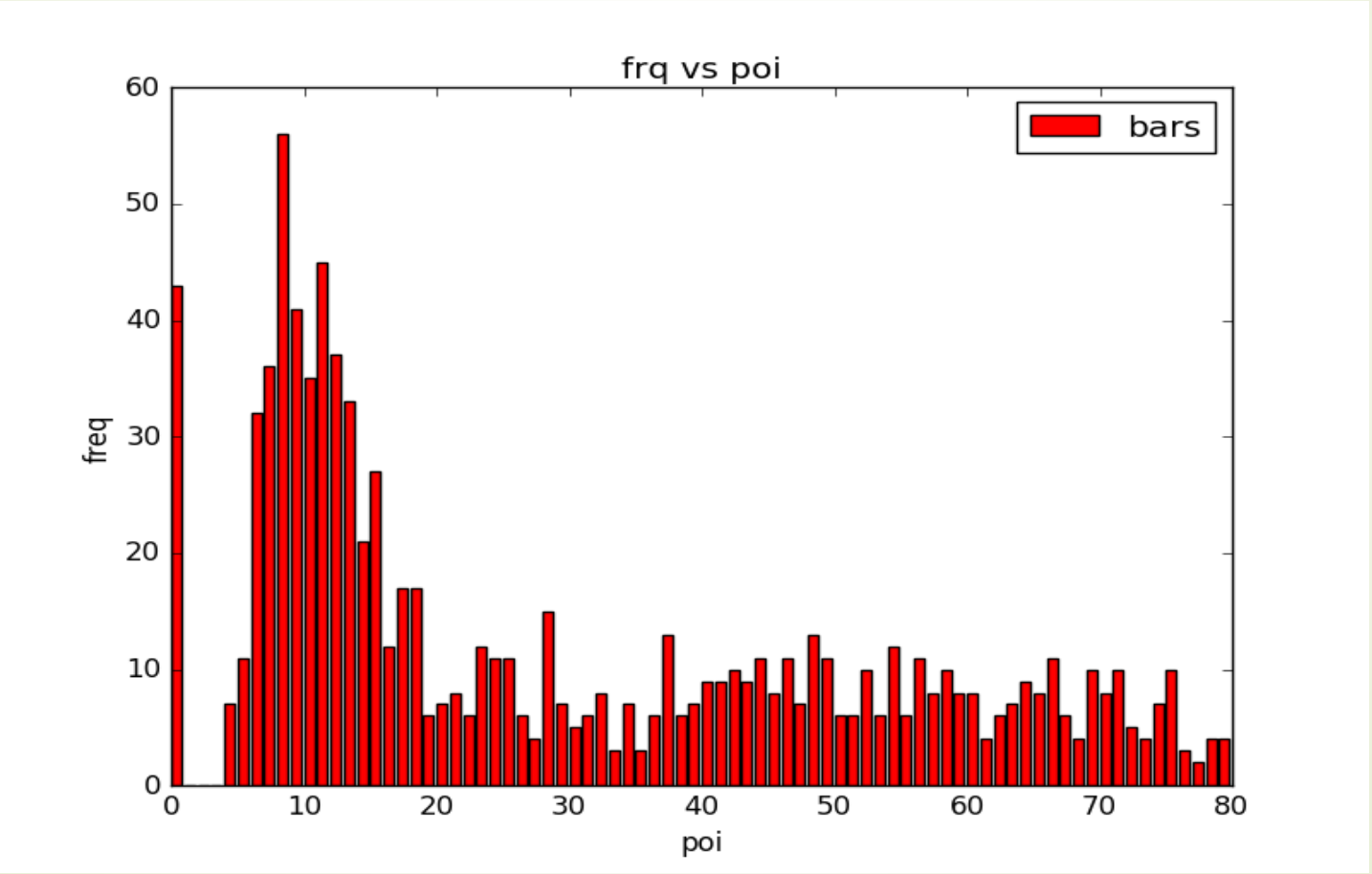
For k=12



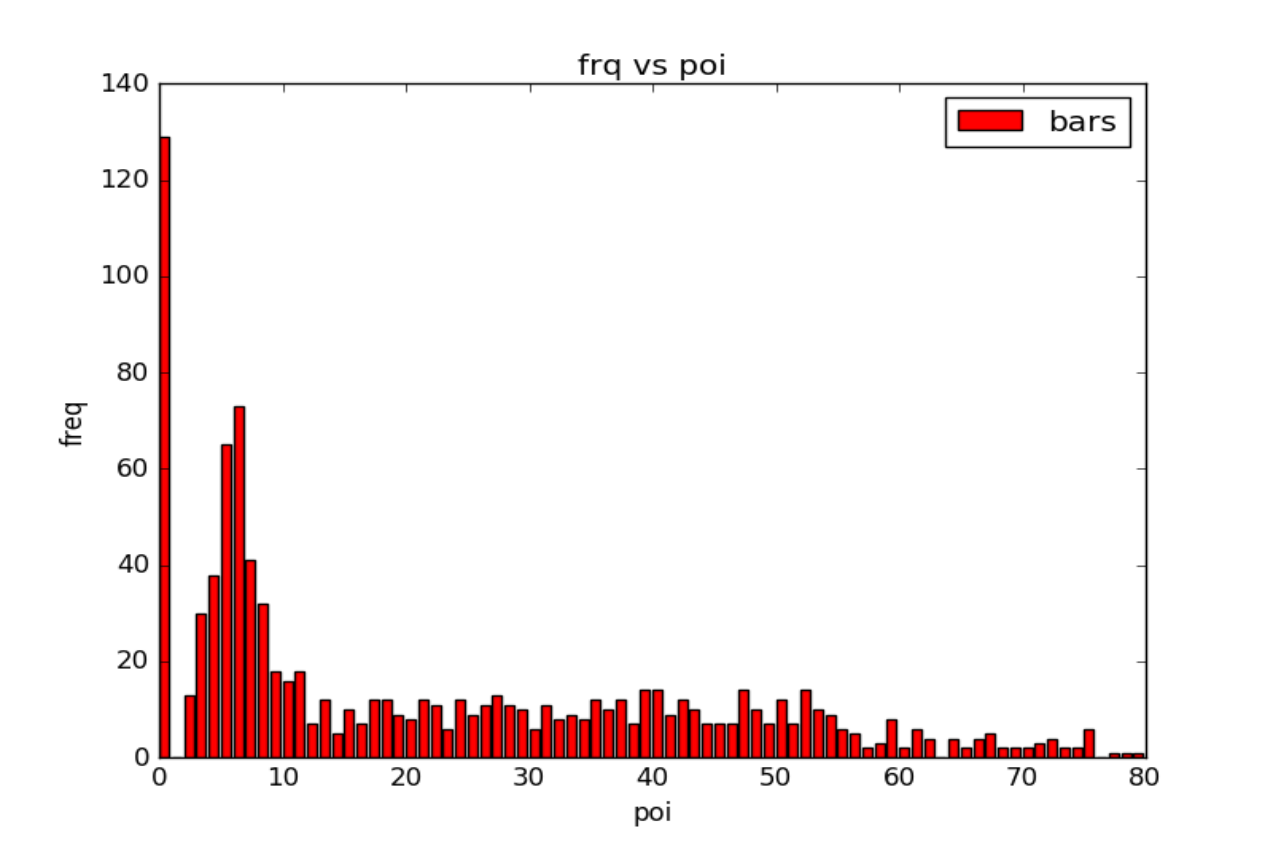
For k=13


frq vs poi
bars
freq
poi
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70
80

For k=14



For k = 15


frq vs poi
bars
freq
poi
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70
80

Observation

- Result has been getting worse with incrementing values of k

Observation

- Result has been getting worse with incrementing values of k
- Best result got with k=11 and k=13

Questions…..