Introduction

Episomes are plasmid present in bacterial cells that can exist independently in the cell or integrated to the main chromosome.
Escherichia coli strain ATCC 43889 has a plasmid, Escherichia coli O157:H7 str. Sakai plasmid pOSAK1.

This report will focus on aligning this plasmid with the main chromosome. This gives an idea where plasmid DNA would be found inside the genome, if any.

For this purpose, it is sugguested to use a traditional sequence alignment methord. But, as both the plasmid and the chromosome are long sequencences, basic traditional sequence alignment is not efficient. Therefore it is suggested to breakdown the long seqences in to shorter segments.

Breaking up long sequences to small segments

Then by using traditional sequence alignment on short segments, we can come to a conclusion.

The length of the chromosome and plasmid are as follows,

Escherichia coli strain ATCC 43889	- 5567434 bases
Escherichia coli O157:H7 str. Sakai plasmid pOSAK1	- 3306 bases

Segmentation of the long sequences were done

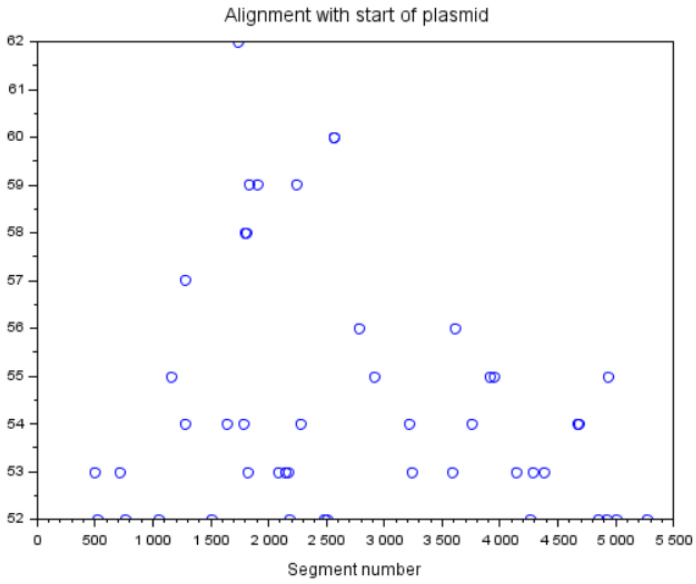
	Total number of bases	Length of a segment	Number of segments
Chromosome	5567434	1000	5567
Plasmid	3306	200	16

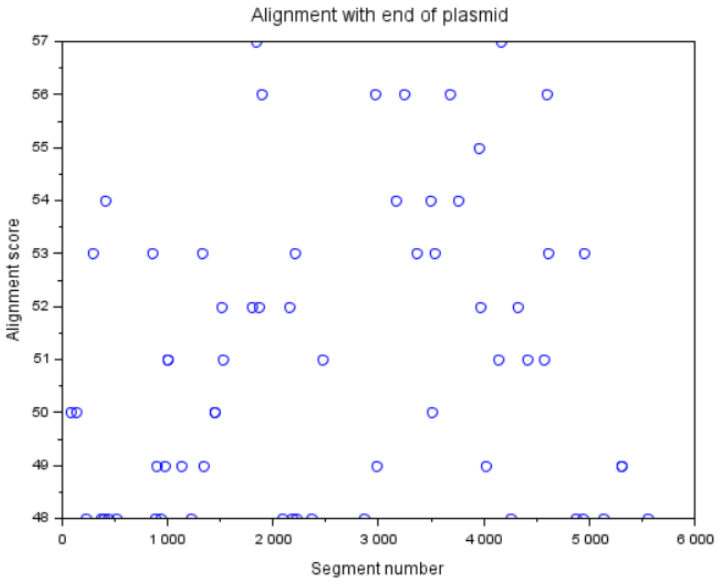
Aligning sequences using local alignment

As the length of plasmid segments are smaller than the lengths of chromosome segments, local alignment is the most suitable methord of aligning.

Performing local alignmet to all pairs of plasmid and chromosome segments is computatinally difficult. Hence, alignment with non-intact query of the start and end of the plasmid was performed with all chromosome segments.
The score used to value the alignment is the maximum value found in the local alignment matrix. The maximum scores obtained are as follows.

Results of local alignment





Best alinged start segment indexes
494. 515. 717. 765. 1048. 1159. 1275. 1277. 1510. 1632. 1739. 1782. 1792. 1801. 1820. 1836. 1907. 2078. 2137. 2171. 2179. 2239. 2278. 2475. 2505. 2561. 2569. 2775. 2915. 3213. 3231. 3580. 3614. 3757. 3916. 3944. 4138. 4258. 4287. 4377. 4671. 4684. 4842. 4919. 4934. 5006. 5268.

Best alinged end segment indexes
83. 129. 229. 298. 377. 401. 409. 441. 512. 850. 883. 893. 929. 976. 1004. 1006. 1130. 1226. 1326. 1345. 1441. 1444. 1518. 1521. 1795. 1836. 1869. 1888. 2094. 2149. 2178. 2213. 2222. 2359. 2473. 2863. 2971. 2984. 3164. 3240. 3367. 3486. 3502. 3536. 3678. 3756. 3949. 3968. 4010. 4130. 4165. 4258. 4324. 4408. 4568. 4589. 4610. 4866. 4928. 4949. 5130. 5305. 5306. 5552.

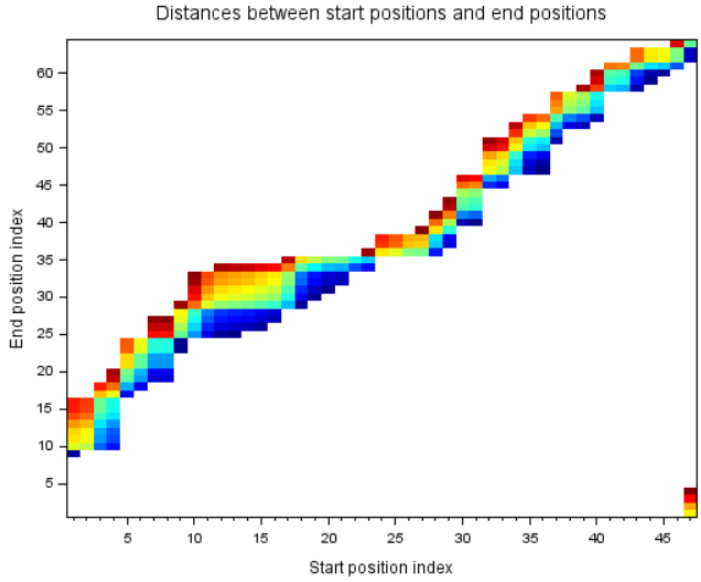
Calculating distances between start and end points

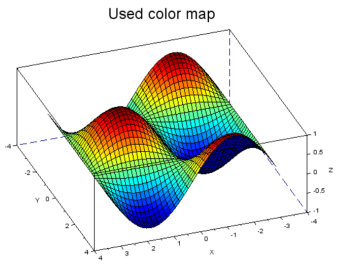
A general idea of where the two ends of the plasmid is most likely to be seen in the chromosome has been obtained.

If the plasmid is to be aligned with the chromosome, the aligned length would be close to the original length of the plasmid.

Finding best start alignment and end alignment

Distances between the filtered start position segements and end position segments were obtained. The results were graphed in a Matplot to visualize and filter the most suitable start and end position pairs.



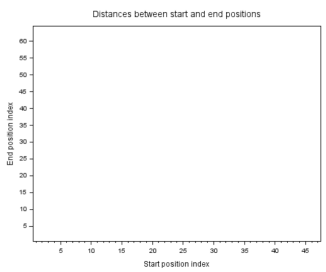


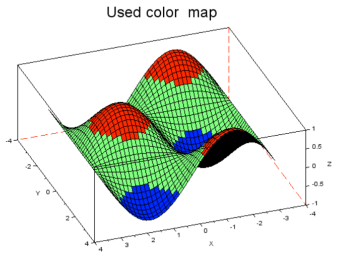
The white area represents data points out of bounds to the color map. Only distances that can occupy a plasmid need to be considered. It is unlikely that a plasmid is spread out over very long reigions. So, only the lowest possible distances between the start alignment and end alignment are considered.

Checking the lower limits

The distance between the start and end position is not prefered to be much less than the actual plasmid length. A segment of chromosome is 1000 bases in length. As the plasmid is 3306 bases in length, if the distance between the start and end alignments is less than 3 segments, such an alignment is unlikely.

By reducing the color map size to 3, such segment pairs can be identified.

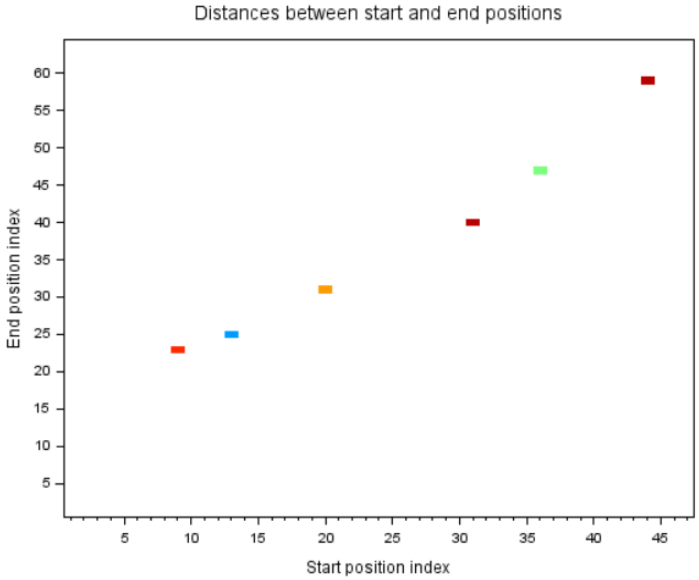


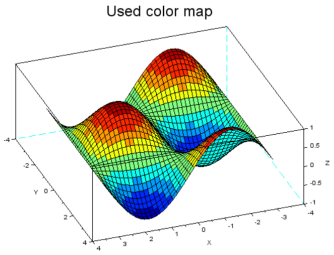


From the Matplot we can observe that no such distances are seen.

Finding 5 best alignments

Increasing the colormap size untill 5 points are visible.





Out of the 6 points, points 4 and 6 are both maximums of the set of distances, hence point 4 is arbitrarily chosen.
The indices of (start position,end position) can be chosen as,

(9,23) (13,25) (20,31) (31,40) (36,47)

Which corresponds to the segment sections,

1510-2213, 1792-1795, 2171-2178, 3231-3240, 3916-3449

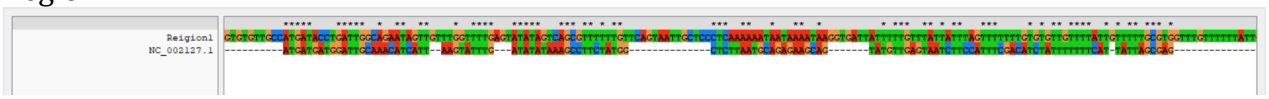
The reigions of interest can be found accordingly to be,

1509001-1518000
1791001-1795000
2170001-2178000
3230001-3240000
3943001-3949000

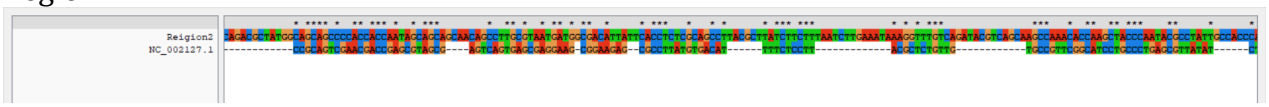
Performing alignment with ClustalX

FASTA files were created for the above reigions and aligned with the plasmid using clustaX.

Region 1



Region 2



Region 3

Region 4

Region5