Designing better phages

Introduction

Algorithm

Experiment results

Future work

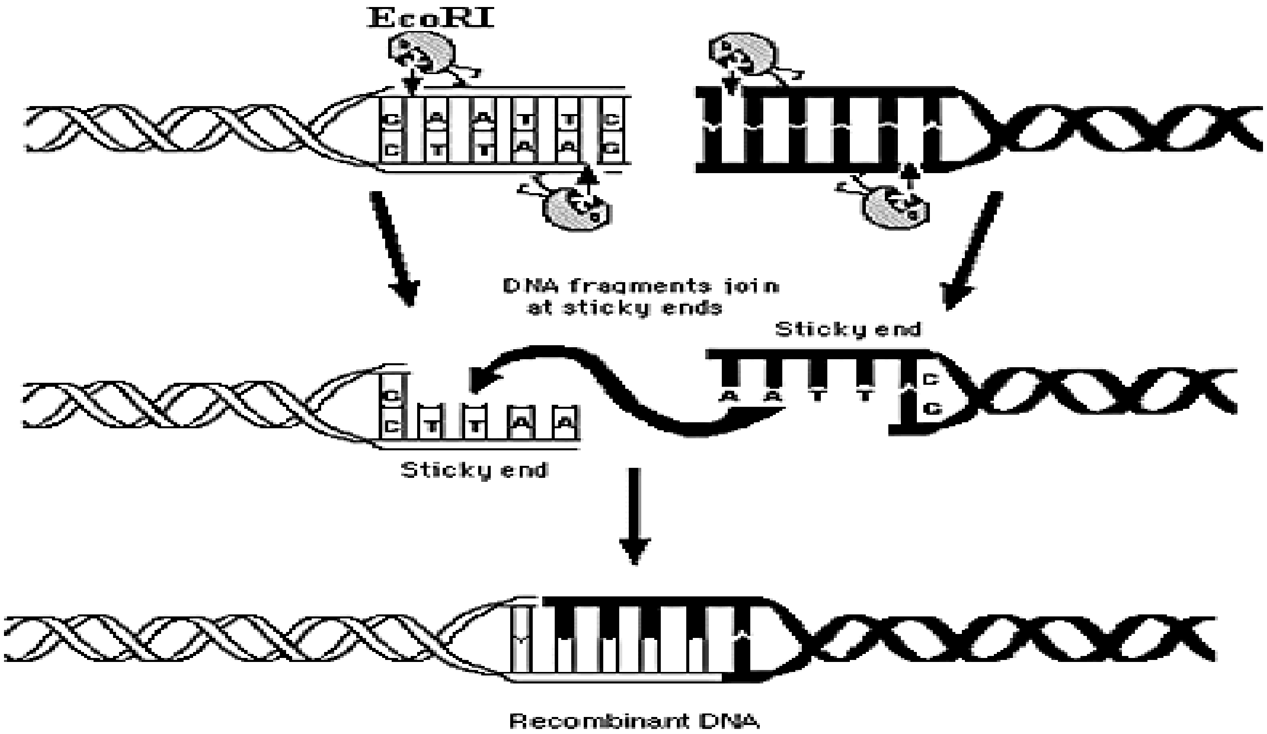
Introduction

- **What is bacteria**? given the proper nutrients, can grow and reproduce on their own
- **What is virus**? Virus cannot "live" or reproduce without entering inside some living cell, whether it's a plant, animal, or bacteria.
- **What is Bacteriophage?** virus that infects bacteria, antibacterial agent.
- **How bacteriophage multiplies**? Attach itself to bacterium, inject DNA, and produce mass of new virus, eventually kill bacteria.
- **How bacteria can destroy phage**? bacteria deploy restriction enzymes to cut phage DNA to make it biologically inactive. More restriction sites, more vulnerable to restriction enzymes.
- **What is our goal?** seek coding sequences which minimize the number of restriction sites while still coding for the same designed protein.

Restriction enzymes

- Name of each enzyme denotes the order of discovery within the host organism (e.g, EcoRI was the first restriction enzyme discovered in E. Coli)
- [2001] 3487 enzymes ,255 distinct cutter sequences found. Cutter sequence range in length from 2 to 15 bases. Most enzymes cut at specific base patterns, some enzymes recognizes multiple sequences by allowing variants at specific base positions. For example, the cutter GCNNGC matches any sequences starting with GC, ending with GC, separated by any sequence of exactly two bases.

The action of restriction enzymes

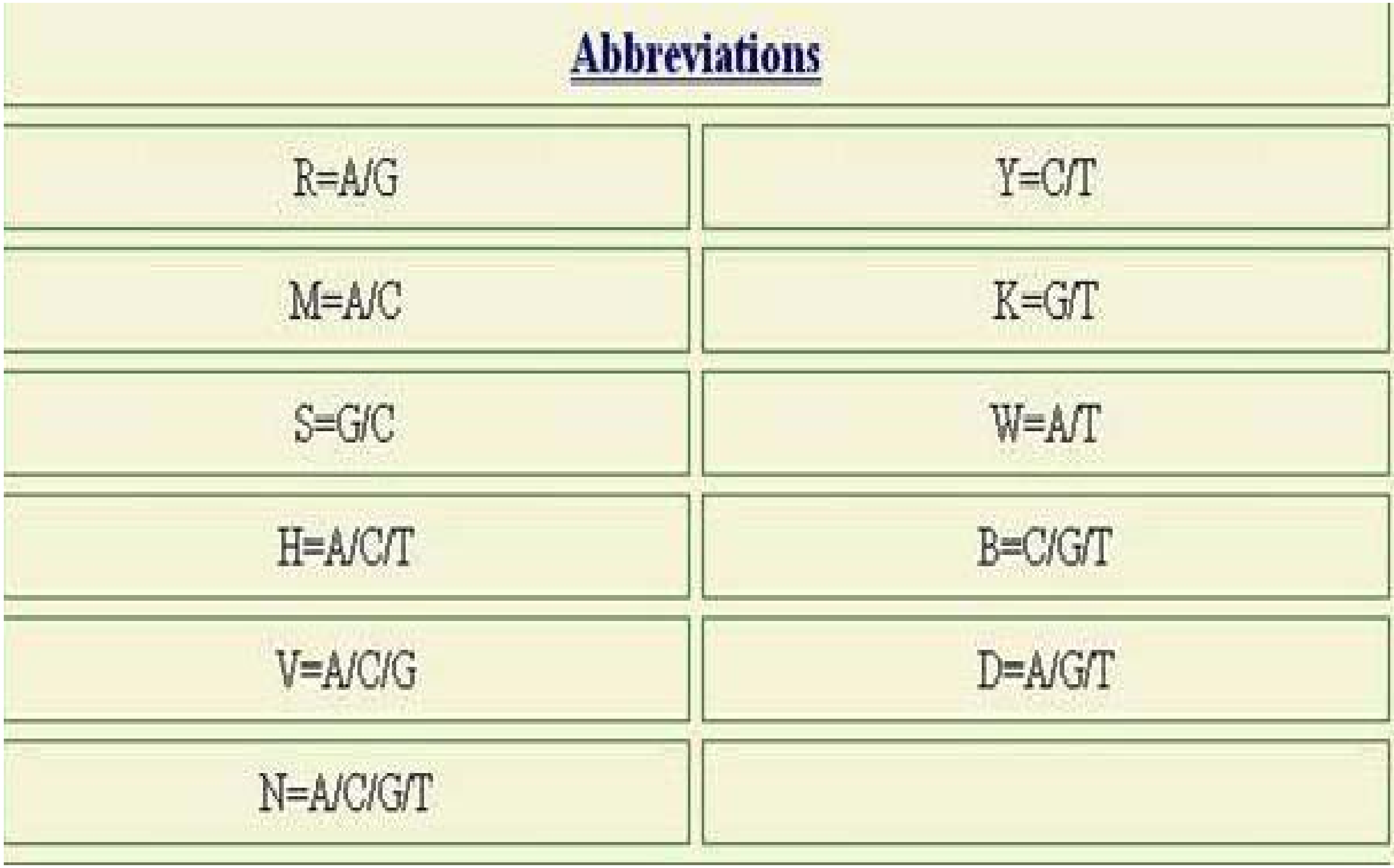


Restriction Enzyme Action of EcoRI

Enzymes tested in the experiment

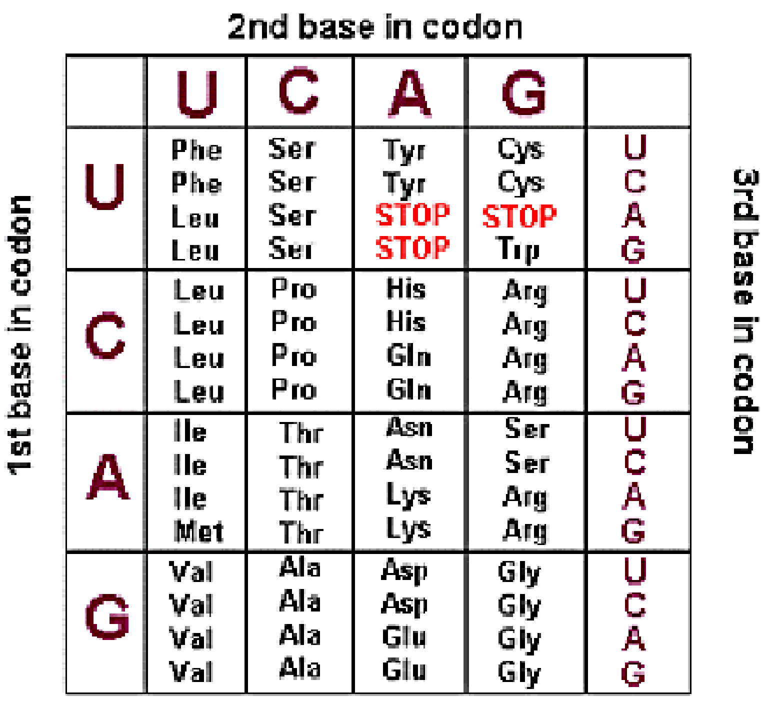
Name	Sequence	Cutting numbers
HaeIII	GGCC	6
Cfr10I	RCCGGY	6
FauI	CCCGC	4
FokI	GGATG	4
HphI	GGTGA	6
AluI	AGCT	7
MaeIII	GTNAC	7
ScrFI	CCNGG	9
SfaNI	GCATC	4
Cac8I	GCNNGC	10
CviRI	TGCA	11
Mn1I	CCTC	1
TauI	GCSGC	11
HhaI	GCGC	15
BbvI	GCAGC	8
HpaII	CCGG	19
TseI	GCWGC	11
AciI	CCGC	12
Fnu4HI	GCNGC	0
CviJI	RGCY	22

Symbols used in restriction enzyme



Abbreviations	
R=A/G	Y=C/T
M=A/C	K=G/T
S=G/C	W=A/T
H=A/C/T	B=C/G/T
V=A/C/G	D=A/G/T
N=A/C/G/T	

Genetic code & Codon-bias Table

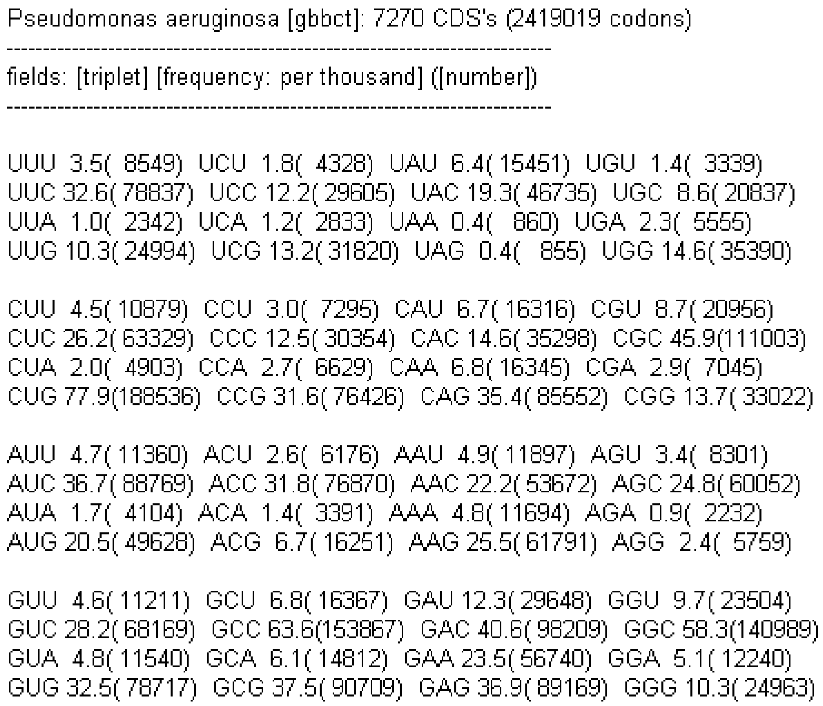


2nd base in codon

1st base in codon

3rd base in codon

	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

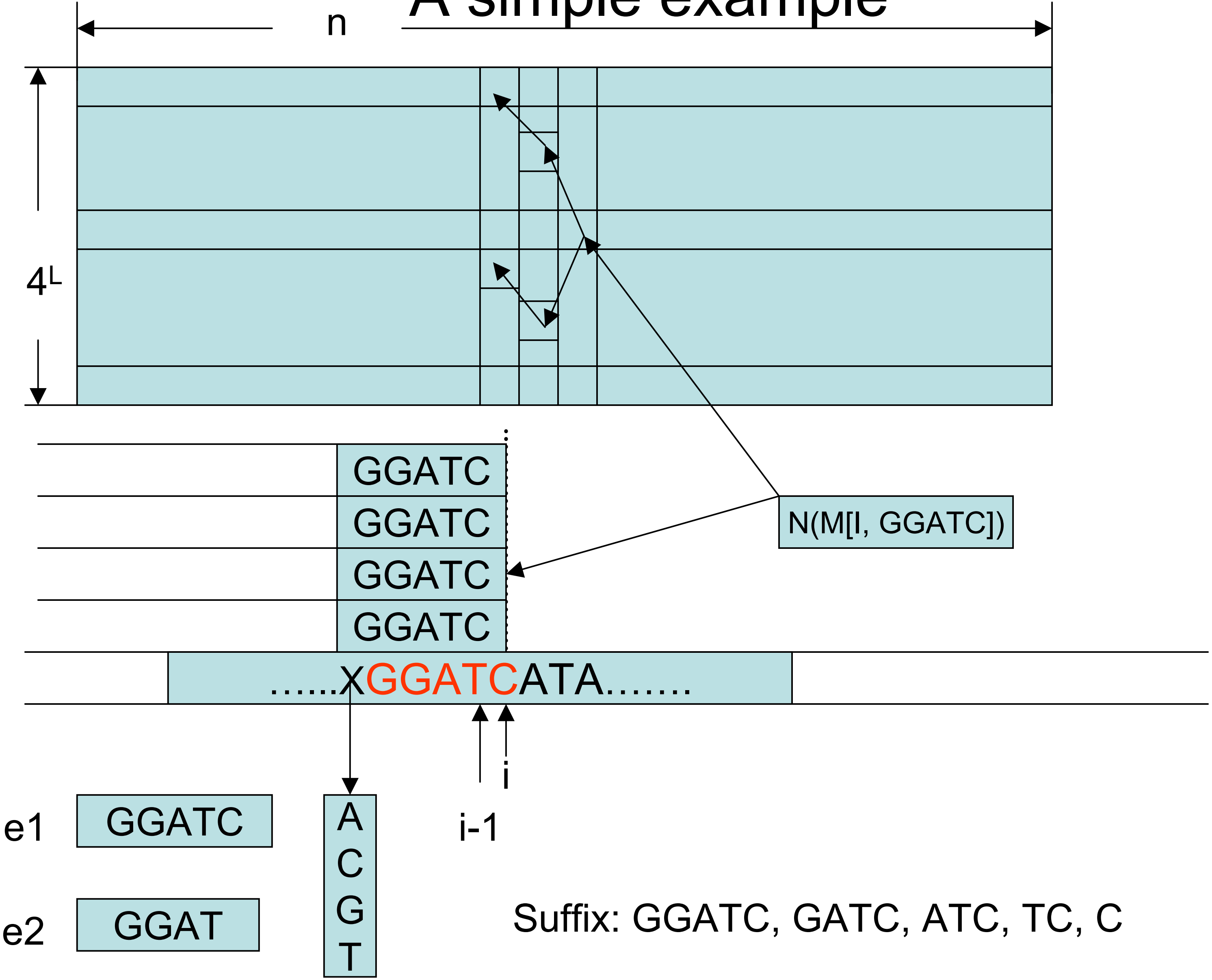


```
Pseudomonas aeruginosa [gbbct]: 7270 CDS's (2419019 codons)
--------------------------------------------------------------------------------
fields: [triplet] [frequency: per thousand] ([number])
--------------------------------------------------------------------------------

UUU  3.5(  8549)  UCU  1.8(  4328)  UAU  6.4( 15451)  UGU  1.4(  3339)
UUC 32.6( 78837)  UCC 12.2( 29605)  UAC 19.3( 46735)  UGC  8.6( 20837)
UUA  1.0(  2342)  UCA  1.2(  2833)  UAA  0.4(   860)  UGA  2.3(  5555)
UUG 10.3( 24994)  UCG 13.2( 31820)  UAG  0.4(   855)  UGG 14.6( 35390)

CUU  4.5( 10879)  CCU  3.0(  7295)  CAU  6.7( 16316)  CGU  8.7( 20956)
CUC 26.2( 63329)  CCC 12.5( 30354)  CAC 14.6( 35298)  CGC 45.9(111003)
CUA  2.0(  4903)  CCA  2.7(  6629)  CAA  6.8( 16345)  CGA  2.9(  7045)
CUG 77.9(188536)  CCG 31.6( 76426)  CAG 35.4( 85552)  CGG 13.7( 33022)

AUU  4.7( 11360)  ACU  2.6(  6176)  AAU  4.9( 11897)  AGU  3.4(  8301)
AUC 36.7( 88769)  ACC 31.8( 76870)  AAC 22.2( 53672)  AGC 24.8( 60052)
AUA  1.7(  4104)  ACA  1.4(  3391)  AAA  4.8( 11694)  AGA  0.9(  2232)
AUG 20.5( 49628)  ACG  6.7( 16251)  AAG 25.5( 61791)  AGG  2.4(  5759)

GUU  4.6( 11211)  GCU  6.8( 16367)  GAU 12.3( 29648)  GGU  9.7( 23504)
GUC 28.2( 68169)  GCC 63.6(153867)  GAC 40.6( 98209)  GGC 58.3(140989)
GUA  4.8( 11540)  GCA  6.1( 14812)  GAA 23.5( 56740)  GGA  5.1( 12240)
GUG 32.5( 78717)  GCG 37.5( 90709)  GAG 36.9( 89169)  GGG 10.3( 24963)
```

Algorithm introduction

- Two functions: for each sequence
  - B(s) = sum of frequency of each codon by codon-bias Table)
  - N(s) = number of the cutter sequences
- Input: DNA sequence S of a given phage. A set of restriction enzymes E = {e1, …em}. A codon-bias Table. L is the length of the longest enzyme
- Output: A DNA sequence G, where |G| = |S|, G and S correspond to same protein. Among all sequences with minimal cutting sites, G has the largest value B.
- Dynamic programming: Let M[i, w] be the output sequence to encode the first i bases of s, where the last L bases are defined by the string w = w1w2….w(L).
  - N(M[i,w]) = min N(M[i-1, xw1w2…w(L-1)]) + cuts(w1w2…w(L)). X is one base from the set of a, c, g and t. Cuts(w) is the number of the cutter sequence that matches a suffix of w. assume Y the set of bases creating same minimal N value in the equation
  - B(M[I, w]) = max B(M[i-1, yw1w2…w(L-1)]), if i = 1, 2(mod3) or B(M[i, w]) = max B(M[i-1, yw1w2…w(L-1)]) + T(w(L-2)w(L-1)w(L)), if i = 0(mod3)

A simple example



Suffix: GGATC, GATC, ATC, TC, C

Algorithm Complexity

- The naïve time analysis is $O(n4^{L}mL)$. There are each sequence position 4^{L} windows will be considered. When calculating cuts(w), we should consider all suffix cuts, which have L ones. for each suffix cut, we should compare it with each enzyme, which has m ones.
- In fact, there can only be at most $3*6^{L/3} = O(1.817^{L})$ distinct legal windows of length L at any single position. The most heavily represented residue is assigned only 6 codons. For each window w with size L, cuts(w) can be stored firstly and used later. It need time $ML4^{L}$, so the total time complexity is $O(n1.817^{L}+ML4^{L})$

Experiment results

- The input DNA sequence is:
atgaaaacgcccaccattcccacccttctggggccggacggcatgacatcgctgcgcgaatatgccggttatcacgg
cggtggcagcggatttggagggcagttgcggtcgtggaacccaccgagtgaaagtgtggatgcagccctgttgccca
actttacccgtggcaatgcccgcgcagacgatctggtacgcaataacggctatgccgccaacgccatccagctgcatc
aggatcatatcgtcgggtcttttttccggctcagtcatcgcccaagctggcgctatctgggcatcggggaggaagaagc
ccgtgccttttcccgcgaggttgaagcggcatggaaagagtttgccgaggatgactgctgctgcattgacgttgagcga
aaacgcacgtttaccatgatgattcgggaaggtgtggccatgcacgcctttaacggtgaactgttcgttcaggccacctg
ggataccagttcgtcgcggcttttccggacacagttccggatggtcagcccgaagcgcatcagcaacccgaacaata
ccggcgacagccggaactgccgtgccggtgtgcagattaatgacagcggtgcggcgctgggatattacgtcagcga
ggacgggtatcctggctggatgccgcagaaatggacatggataccccgtgagttacccggcgggcgcgcctcgttca
ttcacgtttttgaacccgtggaggacgggcagactcgcggtgcaaatgtgttttacagcgtgatggagcagatgaagat
gctcgacacgctgcagaacacgcagctgcagagcgccattgtgaaggcgatgtatgccgccaccattgagagtgag
ctggatacgcagtcagcgatggattttattctgggcgcgaacagtcaggagcagcgggaaaggctgaccggctggat
tggtgaaattgccgcgtattacgccgcagcgccggtccggctgggaggcgcaaaagtaccgcacctgatgccgggt
gactcactgaacctgcagacggctcaggatacggataacggctactccgtgtttgagcagtcactgctgcggtatatcg
ctgccgggctgggtgtctcgtatgagcagctttcccggaattacgcccagatgagctactccacggcacgggccagtg
cgaacgagtcgtgggcgtactttatggggcggcgaaaattcgtcgcatcccgtcaggcgagccagatgtttctgtgctg
gctggaagaggccatcgttcgccgcgtggtgacgttaccttcaaaagcgcgcttcagttttcaggaagcccgcagtgc
ctgggggaactgcgactggataggctccggtcgtatggccatcgatggtctgaaagaagttcaggaagcggtgatgct
gatagaagccggactgagtacctacgagaaagagtgcgcaaaacgcggtgacgactatcaggaaatttttgcccag
caggtccgtgaaacgatggagcgccgtgcagccggtcttaaaccgcccgcctgggcggctgcagcatttgaatccgg
gctgcgacaatcaacagaggaggagaagagtgacagcagagctgcgtaa
- number of the total cutting sites are 173

Experiment results

- The output sequence is:
  atgaagactcctactattcctactcttcttggtagtgatggtatgactagtcttcgtgagtatgctggttatcatggtggtggta
  gtggttttggtggtcaacttcgtagttggaatcctagtagtgagagtgttgatgctgctcttcttcctaattttactcgtggtaat
  gctcgtgctgatgatcttgttcgtaataatggttatgctgctaatgctattcaacttcatcaggatcatattgttggtagttttttc
  gtcttagtcatcgtcctagttggcgttatcttggtattggggaggaggaggcacgtgcttttagtcgtgaggttgaggcggc
  gtggaaggagtttgctgaagatgattgttgttgtattgatgttgagcgtaagcgtacttttactatgatgattcgtgagggtgtt
  gctatgcatgcttttaatggggaactttttgttcaggcgacttgggatactagtagtagtcgtctttttcgtactcagtttcgtat
  ggttagtagtaagcgtattagtaatagtaataatactggggatagtcgtaattgtcgtgctggtgttcagattaatgatagtg
  gtgctgctcttggttattatgttagtgaagatggttatcctggttggatgagtcagaagtggacttggattccccgtgaacttc
  ctggtggtcgtgctagttttattcatgtttttgaacctgttgaagatggtcagactcgtggtgctaatgttttttatagtgttatgga
  gcagatgaagatgcttgatactcttcagaatactcaacttcagagtgctattgttaaggcgatgtatgctgctactattgag
  agtgaacttgatactcagagtgctatggattttattcttggtgctaatagtcaggaacagcgtgagcgtcttactggttggat
  tggggagattgctgcttattatgctgctgctagtgttcgtcttggtggtgctaaggttagtcatcttatgagtggggatagtctt
  aatcttcagactgctcaggatactgataatggttatagtgtttttgagcagagtcttcttcgttatattgctgctggtcttggtgtt
  agttatgagcaacttagtcgtaattatgctcagatgagttatagtactgctcgtgctagtgctaatgagagttgggcgtatttt
  atgggtcgtcgtaagtttgttgctagtcgtcaggcgagtcagatgtttctttgttggttggaggaggcgattgttcgtcgtgttg
  ttactcttcctagtaaggcacgttttagttttcaggaggcacgtagtgcttggggtaattgtgattggattggtagtggtcgtat
  ggcgattgatggtcttaaggaggttcaggaggcggttatgcttattgaggcgggtcttagtacttatgagaaggagtgtgc
  taagcgtggcgatgattatcaggagatttttgctcagcaggttcgtgagactatggagcgtcgtgctgctggtcttaagag
  tcctgcttgggcggcggcggcgtttgagagtggtcttcgtcagagtactgaggaggagaagagtgatagtcgggcagc
  a
- Number of total cutting sites is 20

Experiment results

Name	Sequence	number of cutter sequences before	Number of cutter sequences after
HaeIII	GGCC	6	0
Cfr10I	RCCGGY	6	0
FauI	CCCGC	4	0
FokI	GGATG	4	1
HphI	GGTGA	6	0
AluI	AGCT	7	0
MaeIII	GTNAC	7	1
ScrFI	CCNGG	9	2
SfaNI	GCATC	4	0
Cac8I	GCNNGC	10	1
CviRI	TGCA	11	1
Mn1I	CCTC	1	0
TauI	GCSGC	11	4
HhaI	GCGC	15	0
BbvI	GCAGC	8	1
HpaII	CCGG	19	0
TseI	GCWGC	11	9
AciI	CCGC	12	0
Fnu4HI	GCNGC	0	0
CviJI	RGCY	22	0

Future work

- We can define different function to minimize, such as $F = aN(s) - bB(s)$. A and b are parameter. According to assign different parameters, we can get required result. The algorithm can be easily expanded to resolve such problems.
- If assign $a = 1$, $b = 0$. we select the sequence with minimal restriction sites
- If assign $a = 0$, $b = 1$. we select the sequence with more efficient translation.