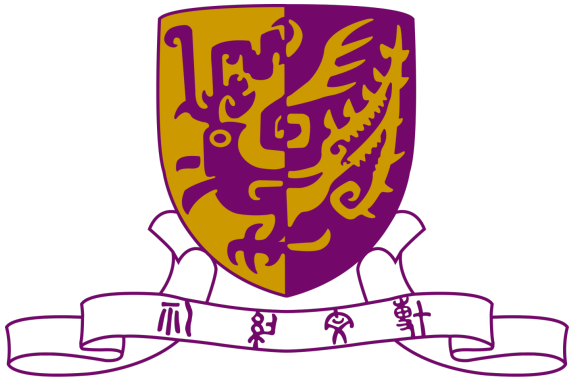
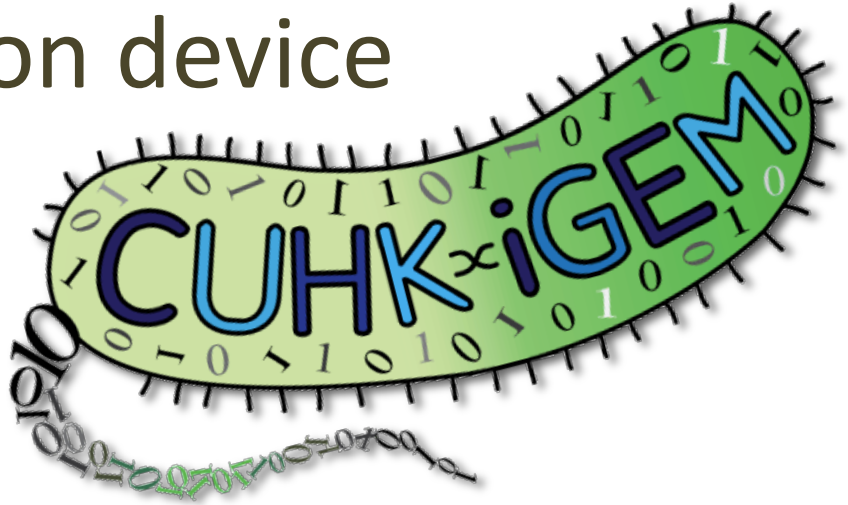
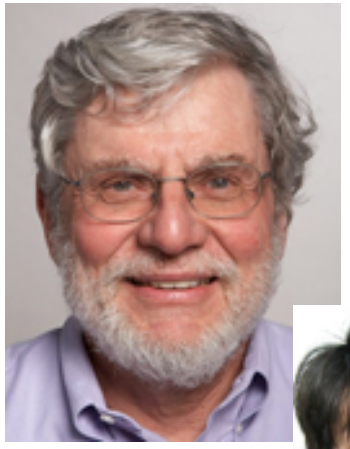
The Chinese University of Hong Kong – iGEM 2010

Bacterial based storage and encryption device





Bacterial based information storage device



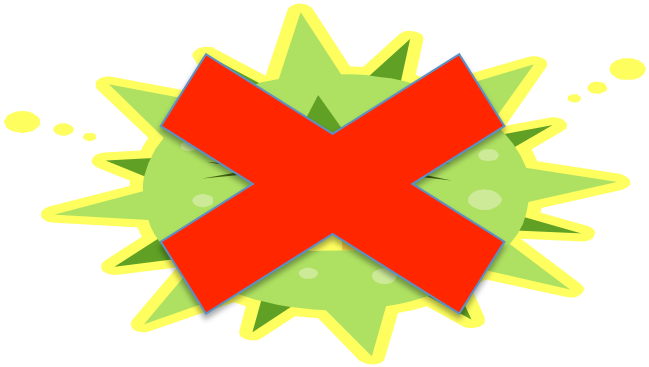
- Bancroft's group (2001)

  Mount Sinai School of Meducube

- Yachie's group (2007)

  Keio University

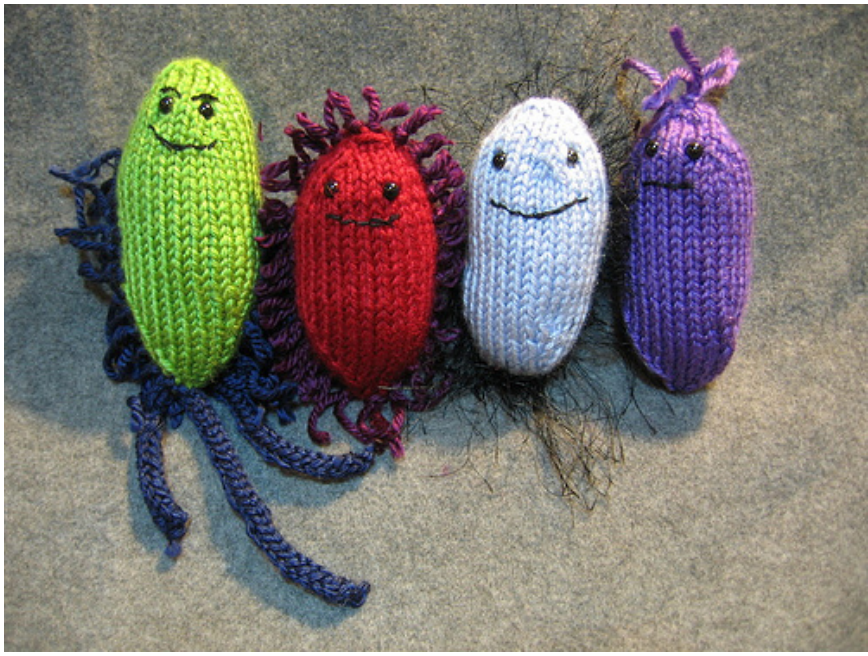
However..



This year, The CUHK…

- True, massively parallel bacterial storage system

It is not the only thing we did..!!



In Addition…

- Encryption module with DNA shuffling system
  - Rci system

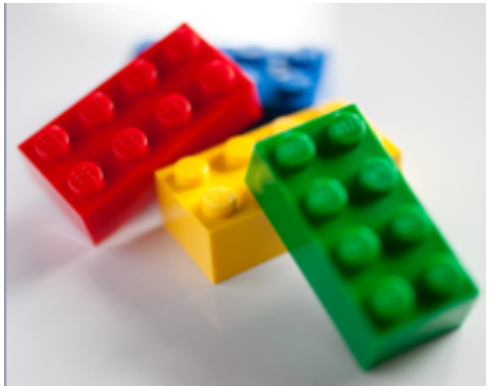
- The data proof-read
  - Chechksum

- Strategy deal with synthesis/sequencing difficulties
  - Homopolymer, repetitive sequence

A LIVING DATA STORAGE SYSTEM

Basic infrastructure of the system

Module-based



Coding System

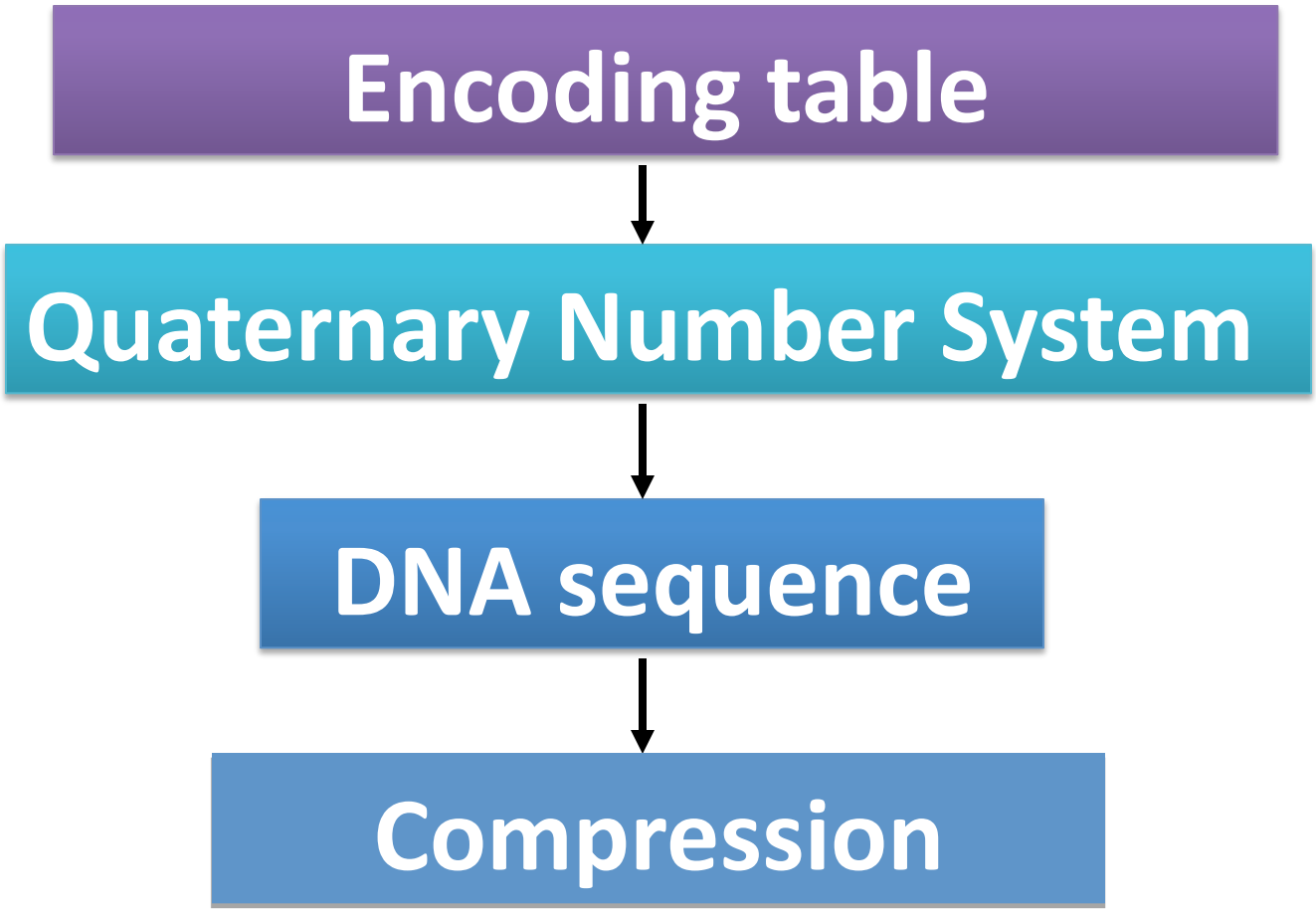
Encryption

Decryption

MP Storage System

Processing

Coding System



Use Numbers to Represent the Letters

From ASCII Table

- i = 105
- G = 71
- E = 69
- M = 77

Change to Quaternary Numbers

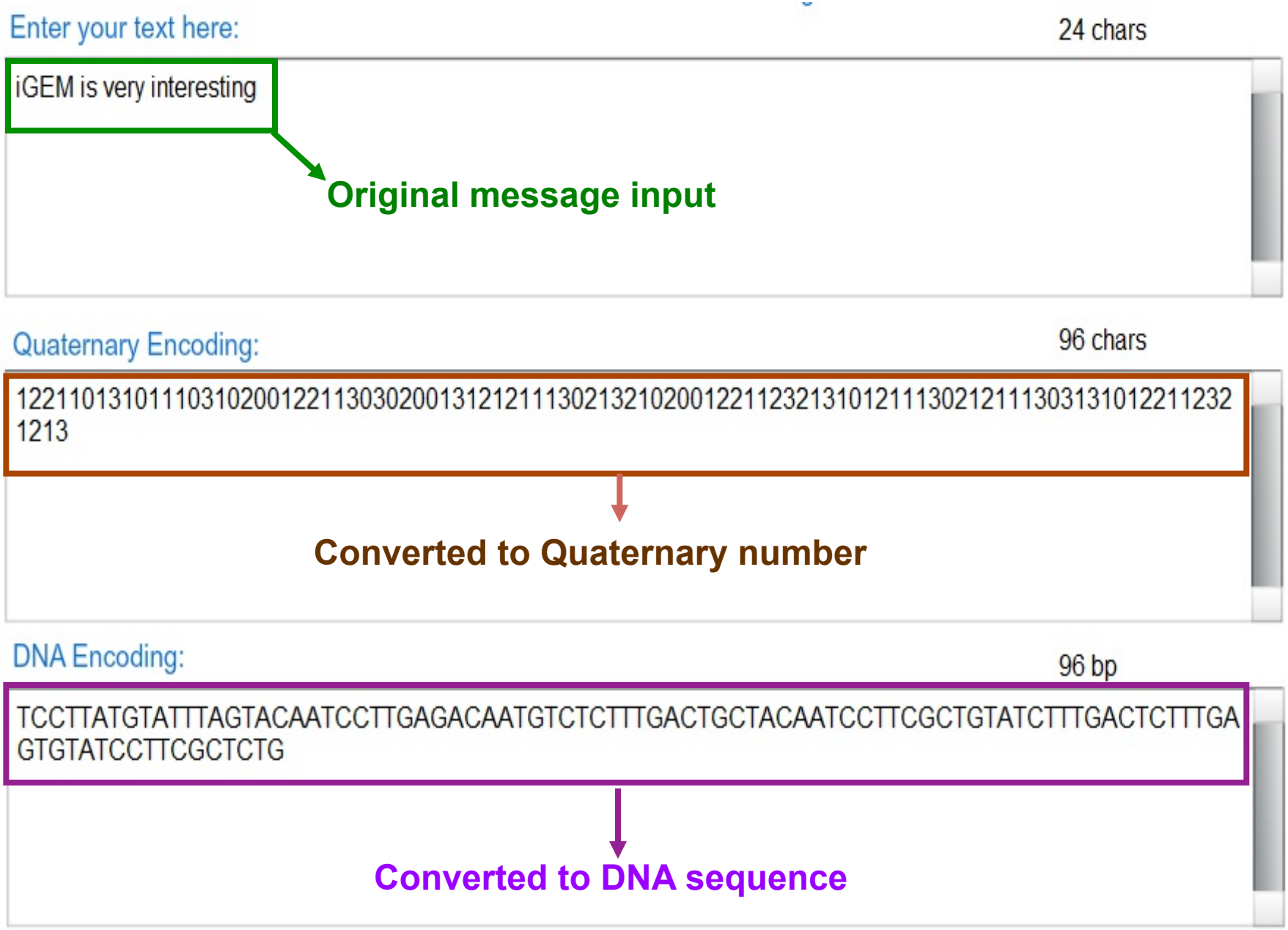
- 105 → 1221
- 71 → 0113
- 69 → 0111
- 77 → 0131

iGEM → 12210113011110131

Use "A, T, C & G" to Represent the Numbers

- 0 = A
- 1 = T
- 2 = C
- 3 = G

iGEM → 12211011301110131
→ ATCTATTGATTTATGT


Enter your text here:
24 chars
iGEM is very interesting
Original message input
Quaternary Encoding:
96 chars
1221101310111031020012211303020013121211130213210200122112321310121113021211130313101221123212213
Converted to Quaternary number
DNA Encoding:
96 bp
TCCTTATGTATTTAGTACAATCCTTGAGACAATGTCTCTTTGACTGCTACAATCCTTCGCTGTATCTTTGACTCTTTGAGTGTATCCTTCGCTCTG
Converted to DNA sequence

Coding System

Encryption | Decryption

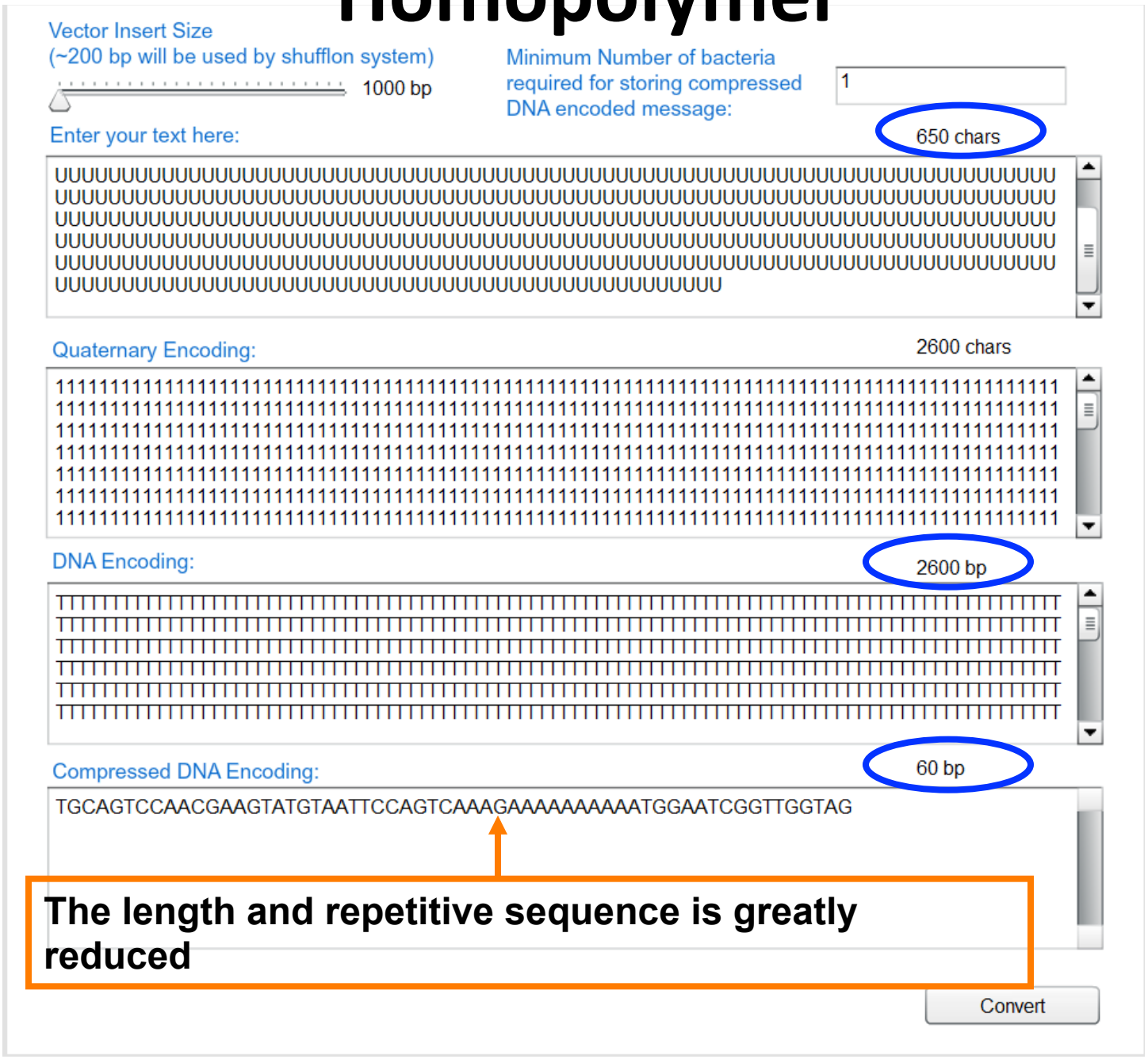
MP Storage System

Coding - Compression

- **DEFLATE — a compression algorithm**

1. Can reduce the **homopolymer** and **repetitive regions**
2. Can store more information

Homopolymer



Repetitive Regions

Vector Insert Size
(~200 bp will be used by shufflon system)

1000 bp

Minimum Number of bacteria required for storing compressed DNA encoded message: 1

Enter your text here: 672 chars

CUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCU
HKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHK
CUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCU
HKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHK
CUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCUHKCU
HKCUHK

Quaternary Encoding: 2688 chars

10031111102010231003111110201023100311111020102310031111102010231003111110201023100311111020
10231003111110201023100311111020102310031111102010231003111110201023100311111020102310031111
10201023100311111020102310031111102010231003111110201023100311111020102310031111102010231003
11111020102310031111102010231003111110201023100311111020102310031111102010231003111110201023
10031111102010231003111110201023100311111020102310031111102010231003111110201023100311111020
10231003111110201023100311111020102310031111102010231003111110201023100311111020102310031111
10201023100311111020102310031111102010231003111110201023100311111020102310031111102010231003

DNA Encoding: 2688 bp

TAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACG
TAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACG
TAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACG
TAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACG
TAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACG
TAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACGTAAGTTTTTACATACG

Compressed DNA Encoding: 69 bp

TGCAGTCCTGAGAAGCGGTTGGAATGTCATGCTTCAGTGCACTAATCACAAAAGGCTCCGGGATAAGCT

Convert

Encryption

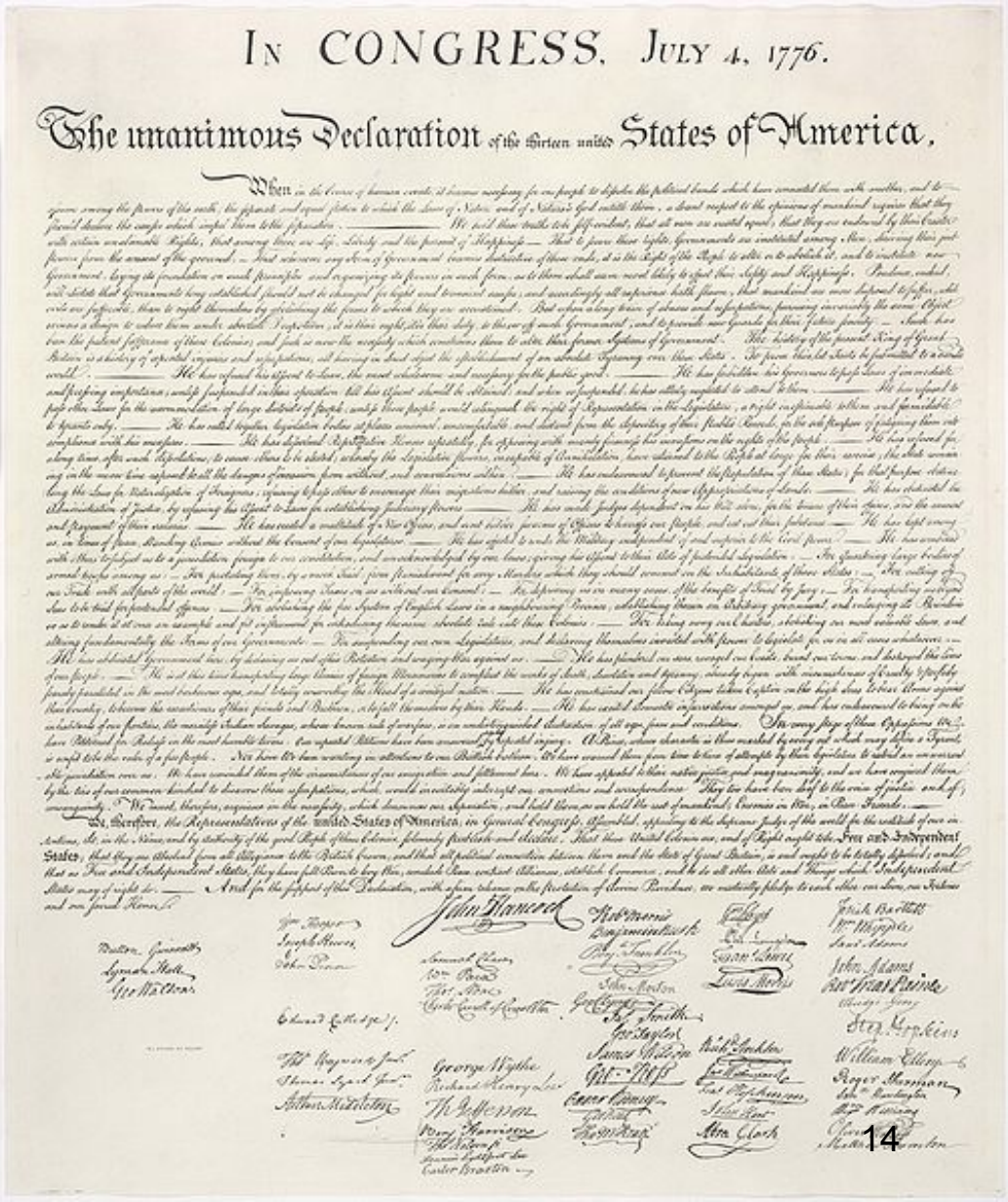
- **Provide DNA variation**
- **DNA Shuffling system**
- Examples:

Homologous recombination
RACHITT

Simulation Analysis

The United States Declaration of Independence

8074 Characters!!!


IN CONGRESS, JULY 4, 1776.

The unanimous Declaration of the thirteen united States of America.

Fragmentation of message

- Larger than the maximum vector insertion size
- Limitation of current DNA synthesis technology

→ Split the message into different parts

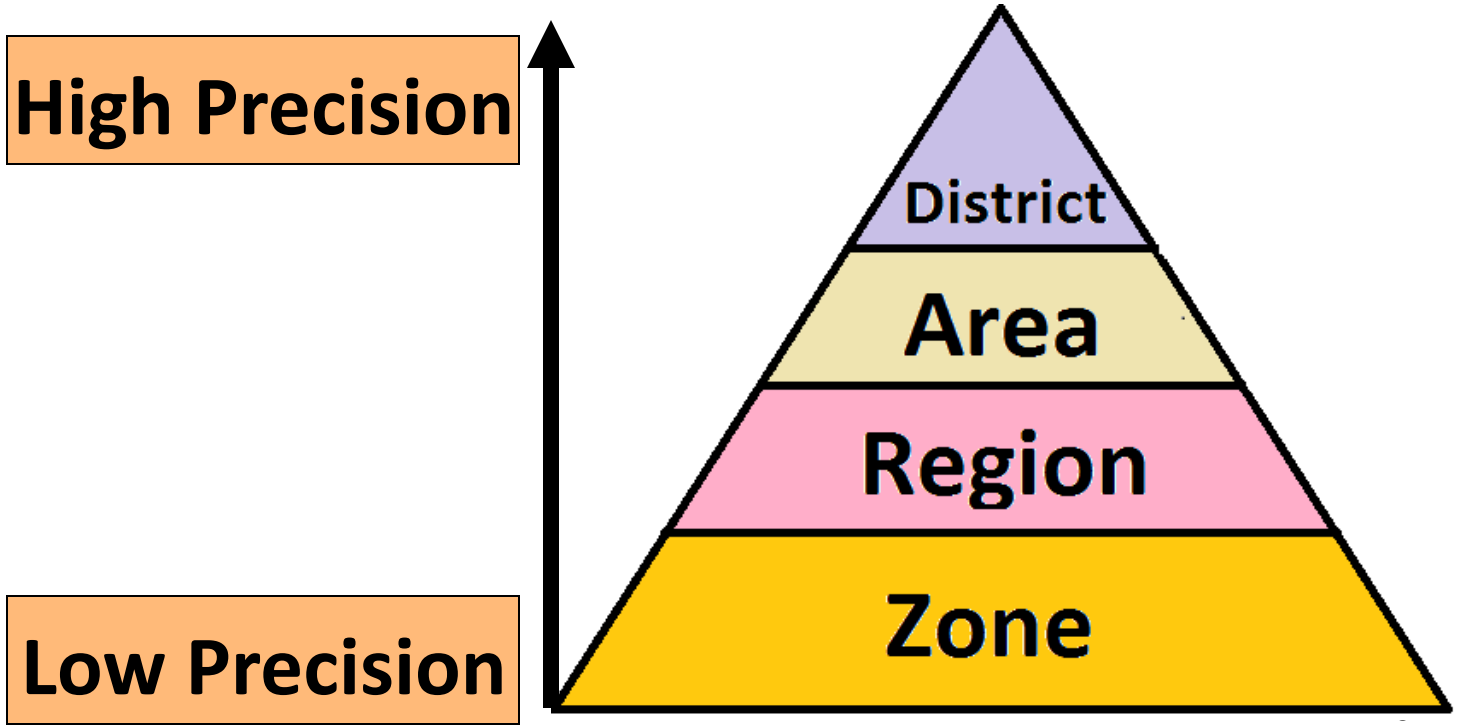
How do you deal with the problem of positioning?

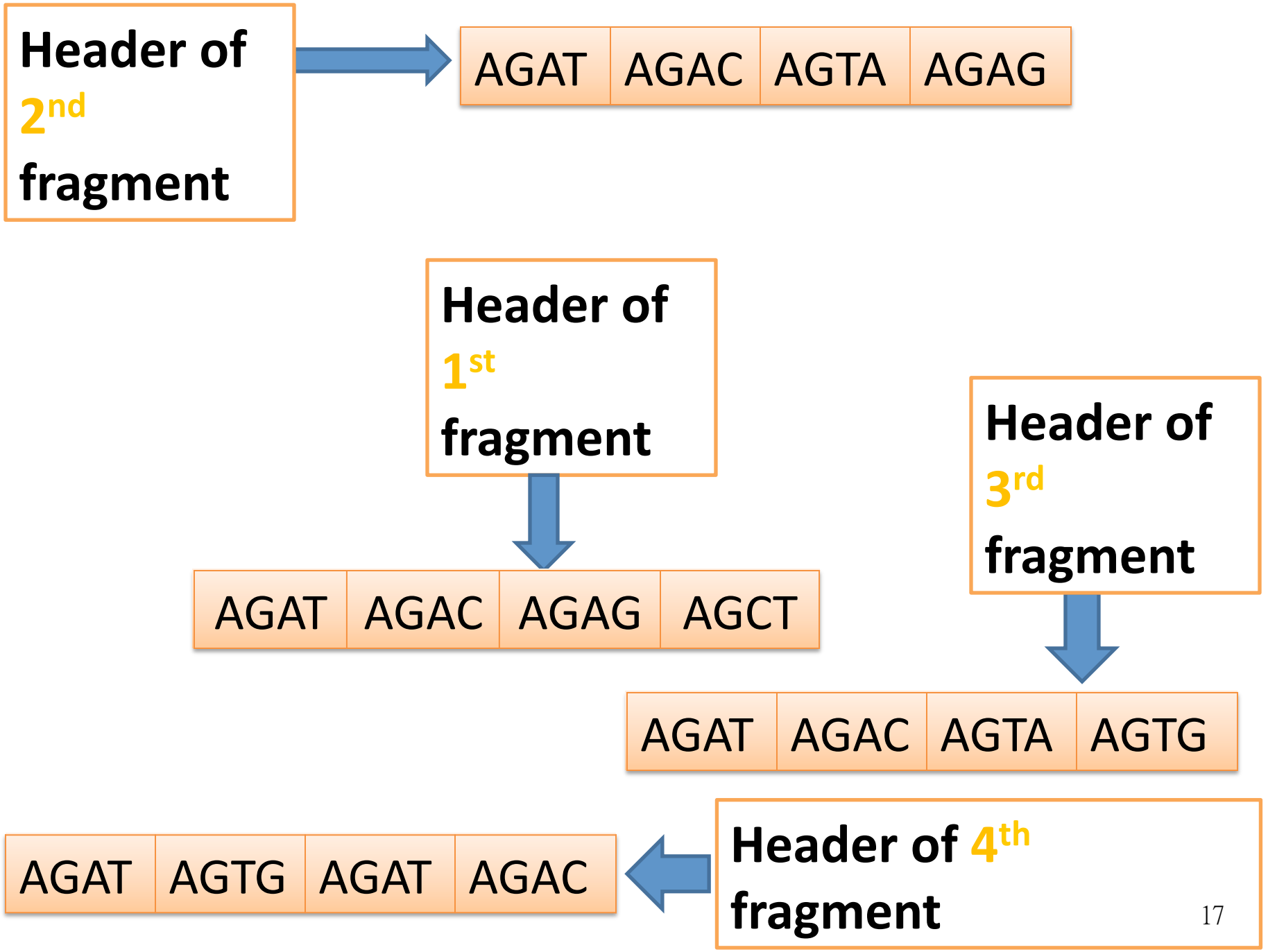
Postal Code

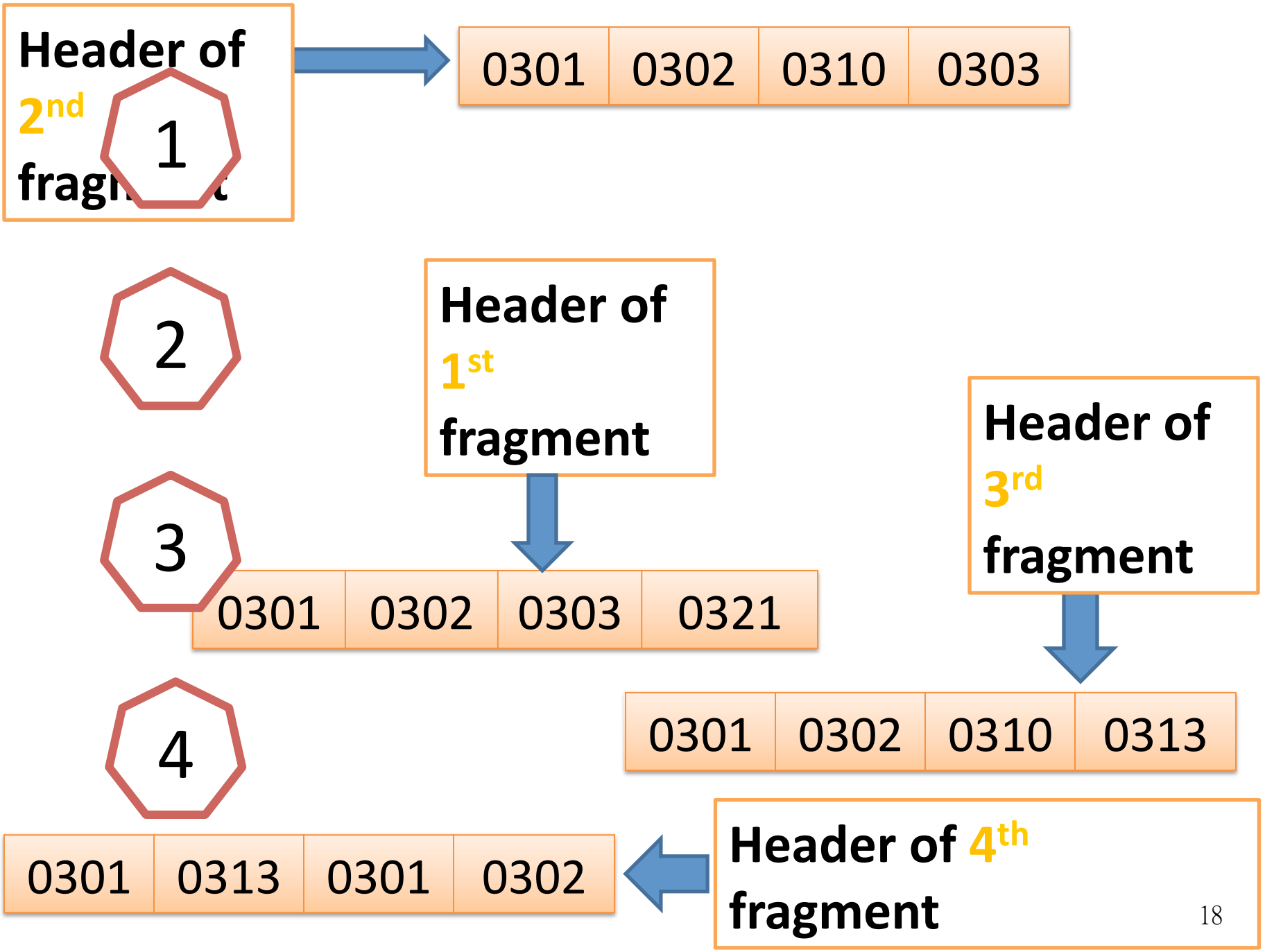
Storage – Massively parallel

Header – Locate particular data fragment of the message

Analogy to the hard disk : 4 address units




Header of 2nd fragment
AGAT
AGAC
AGTA
AGAG
Header of 1st fragment
AGAT
AGAC
AGAG
AGCT
Header of 3rd fragment
AGAT
AGAC
AGTA
AGTG
Header of 4th fragment
AGAT
AGTG
AGAT
AGAC

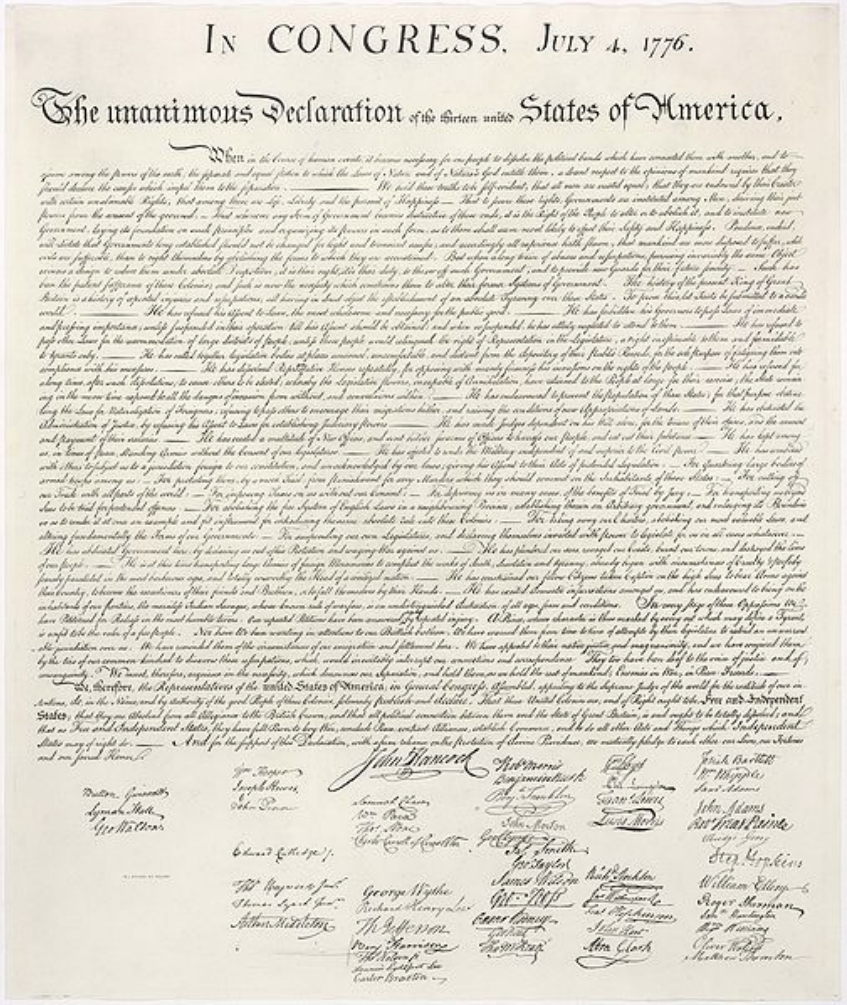

Header of 2nd fragment
1
0301 0302 0310 0303
2
Header of 1st fragment
3
0301 0302 0303 0321
Header of 3rd fragment
0301 0302 0310 0313
4
0301 0313 0301 0302
Header of 4th fragment

Capacity

If insertion size per cell is 1kb……

The United States Declaration of Independence

Only **18 cells!!!**


IN CONGRESS, JULY 4, 1776.

The unanimous Declaration of the thirteen united States of America.

Capacity

1 gram of cells consists of~ 10 Million cells.

The United States Declaration of Independence requires 18cells…..

Each fragment will have at least 500,000 copies!!!

Coding System

Encryption | Decryption

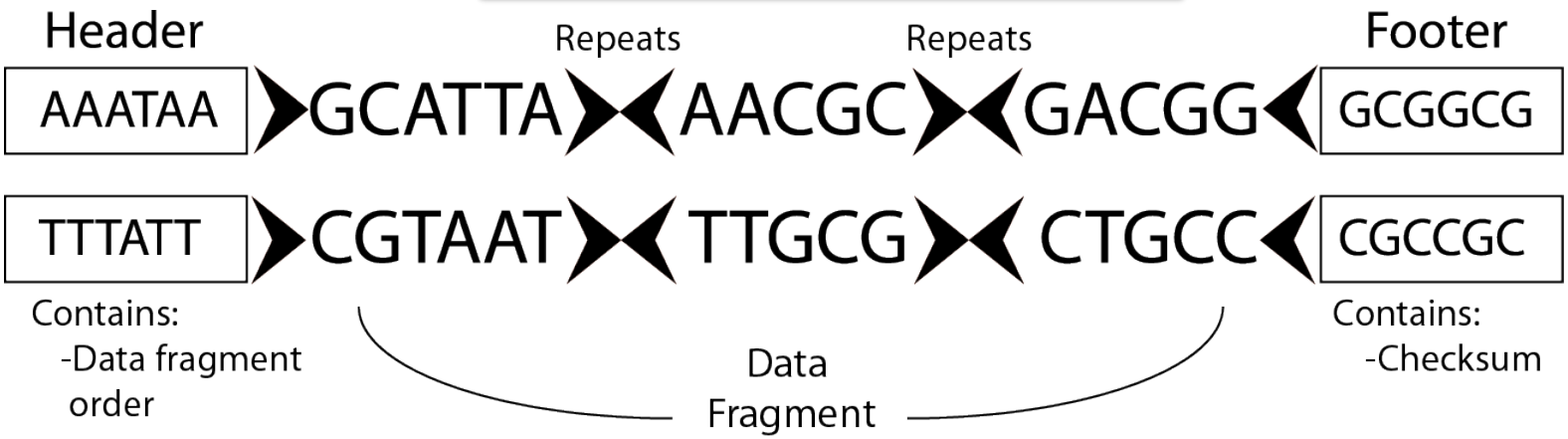
MP Storage System

Decryption

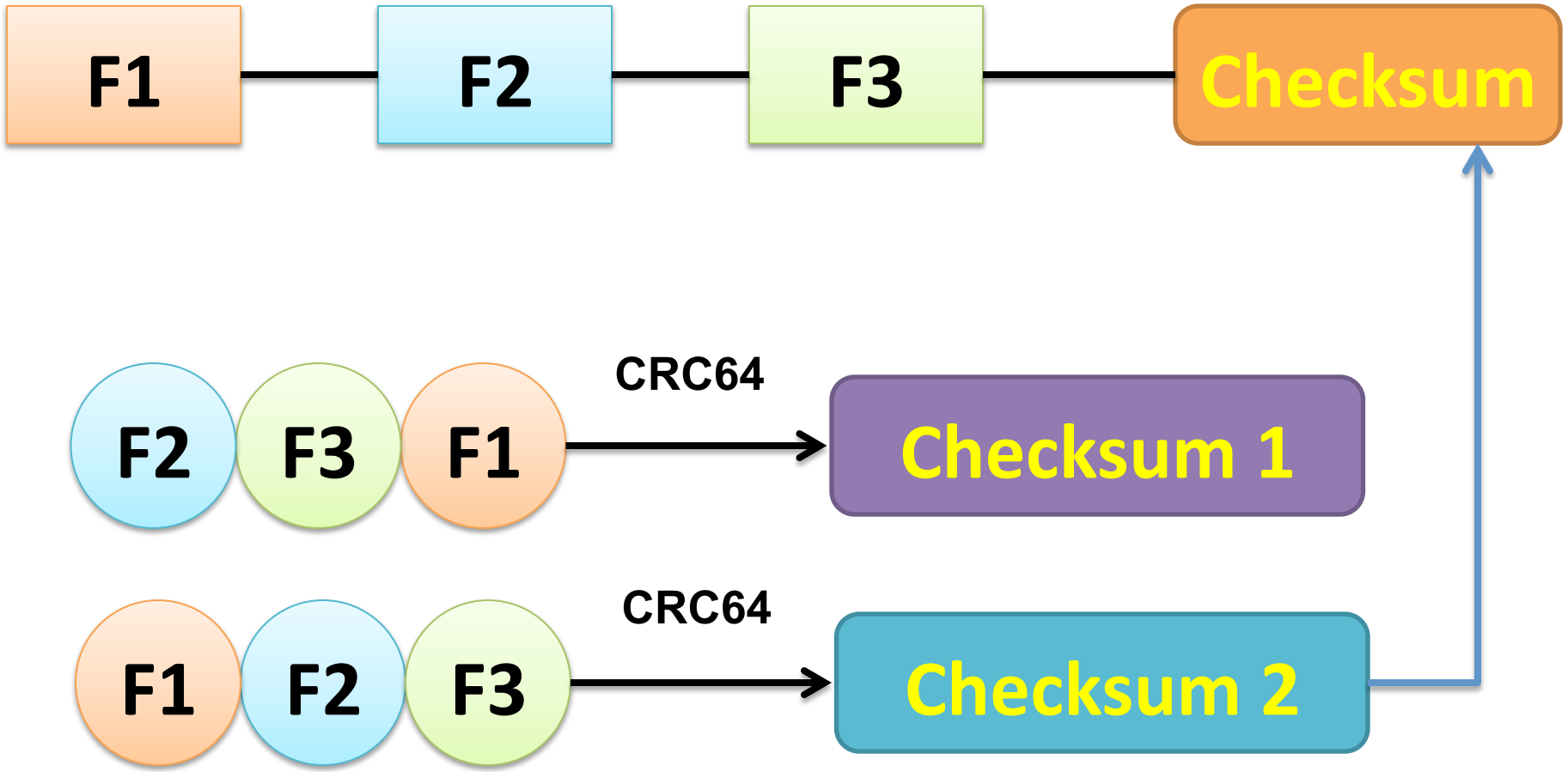
sequencing

Identification of repeat, message, checksum

Checksum system

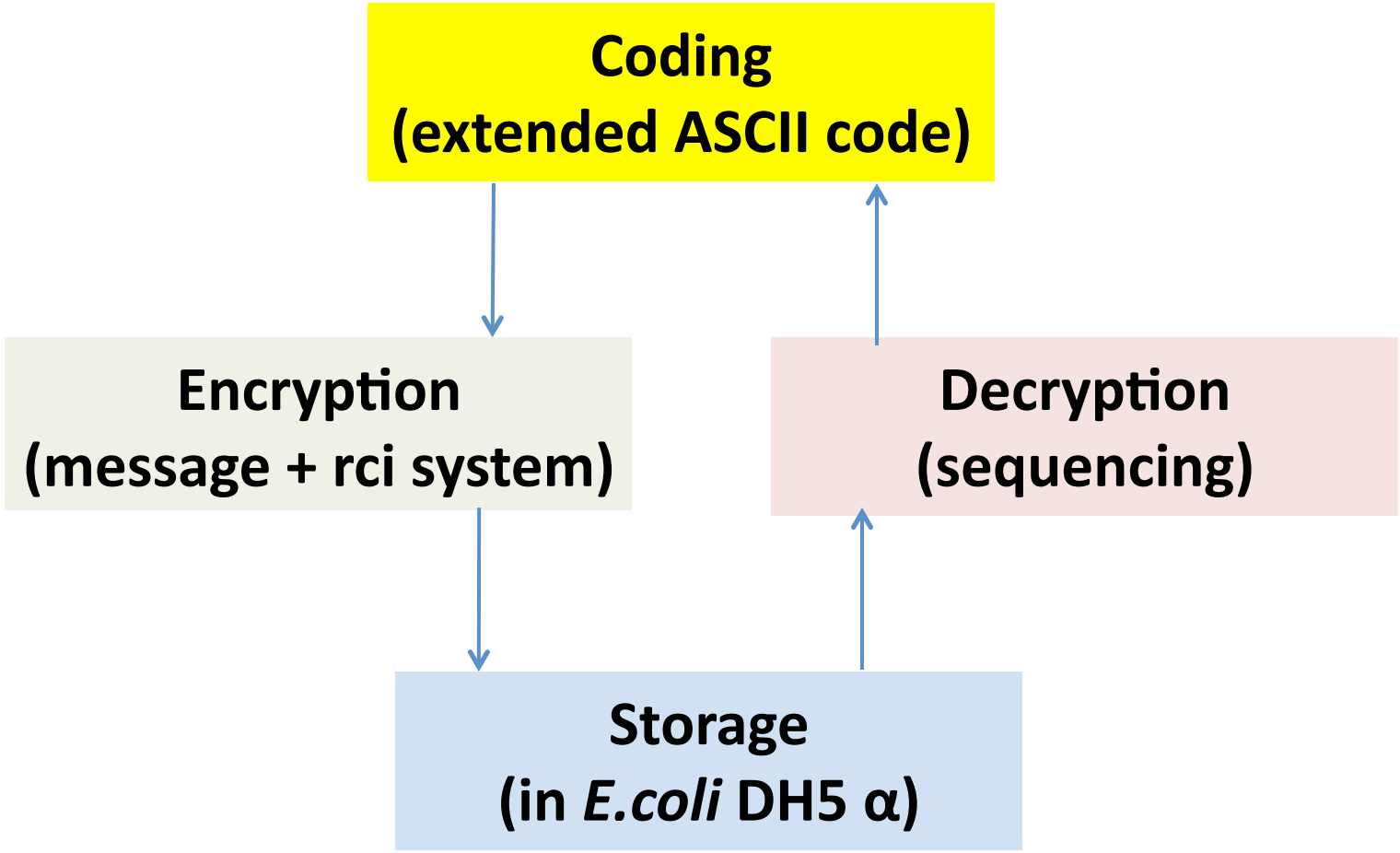


Checksum Mechanism

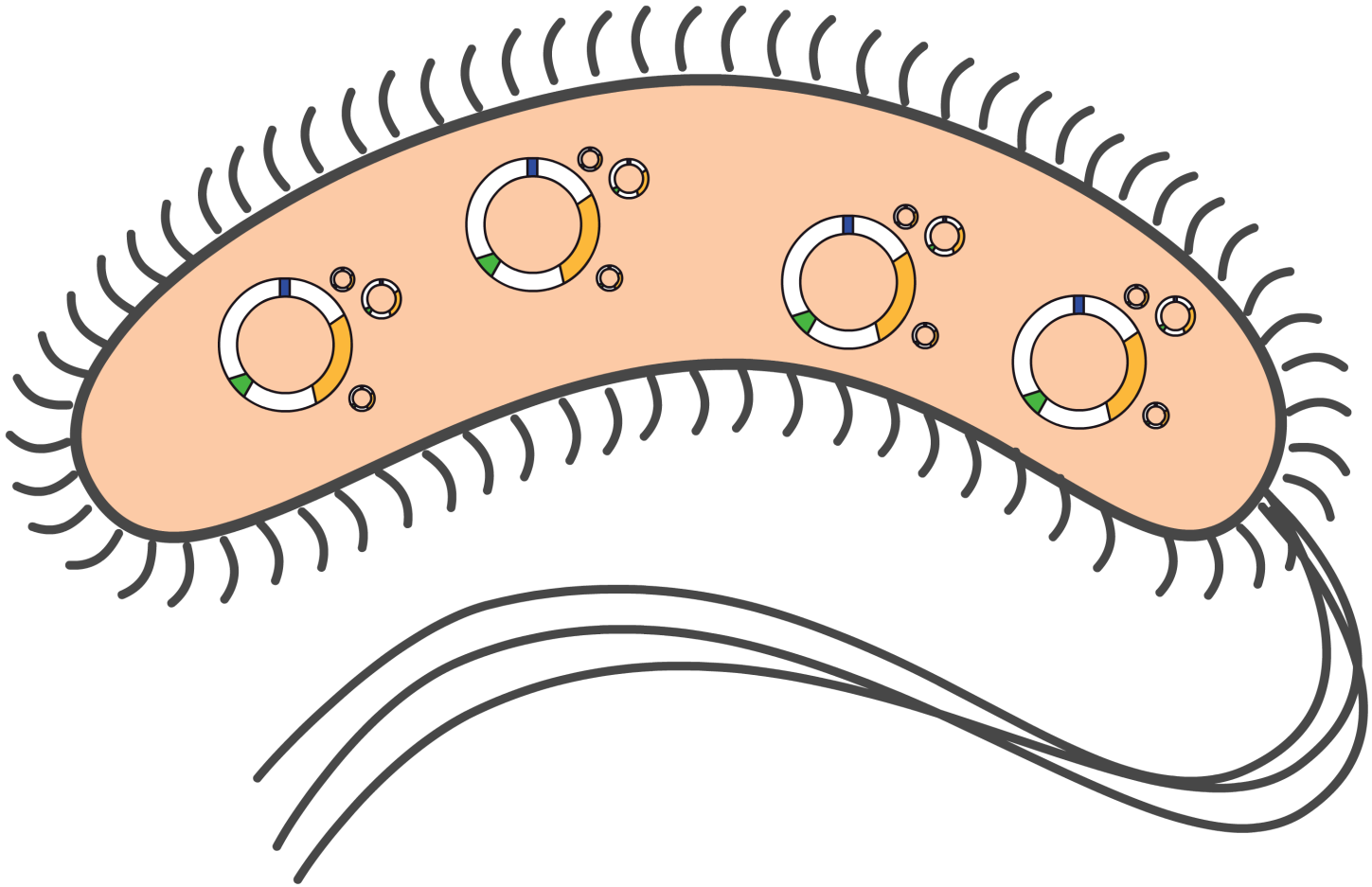


WET LAB TEAM

To prove our concept…



4. The host cell duplicates the data storage vectors, which helps to ensure data integrity by redundancy.



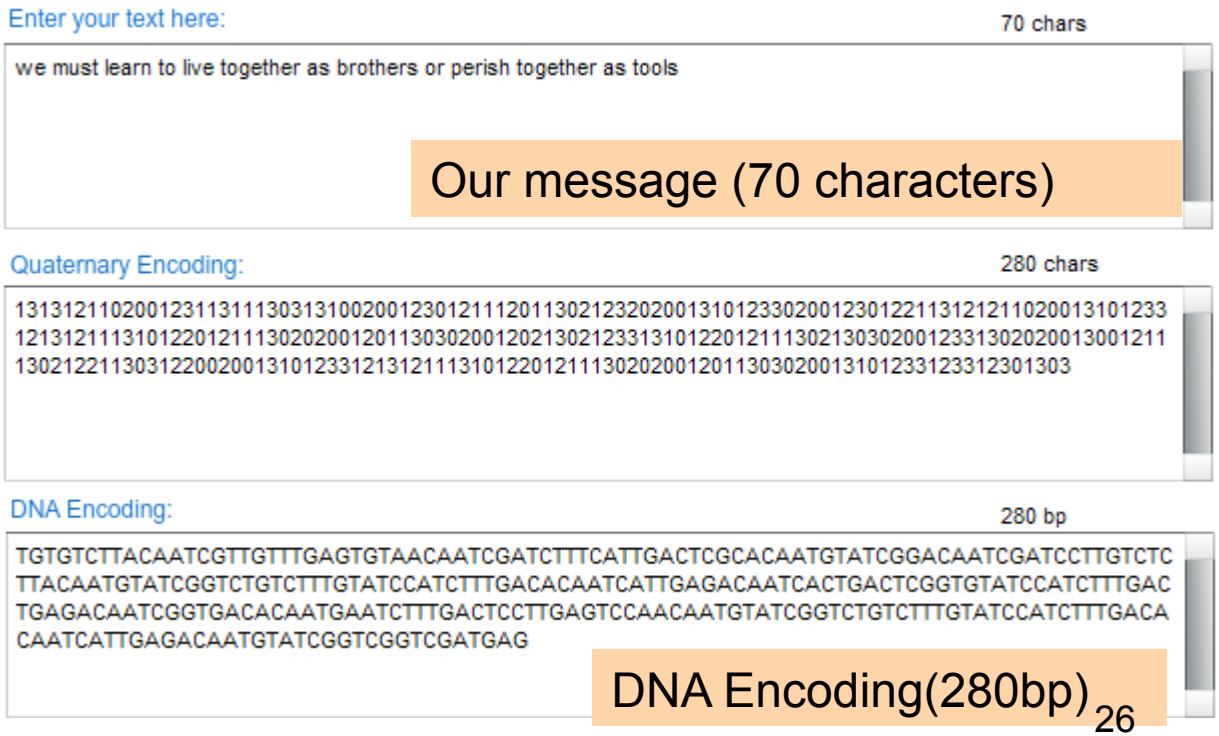
Message

- we must learn to live together as brothers or perish together as tools

  <<code form Dr. Martin Luther King, Jr., a prominent leader in the African American civil rights movement >>

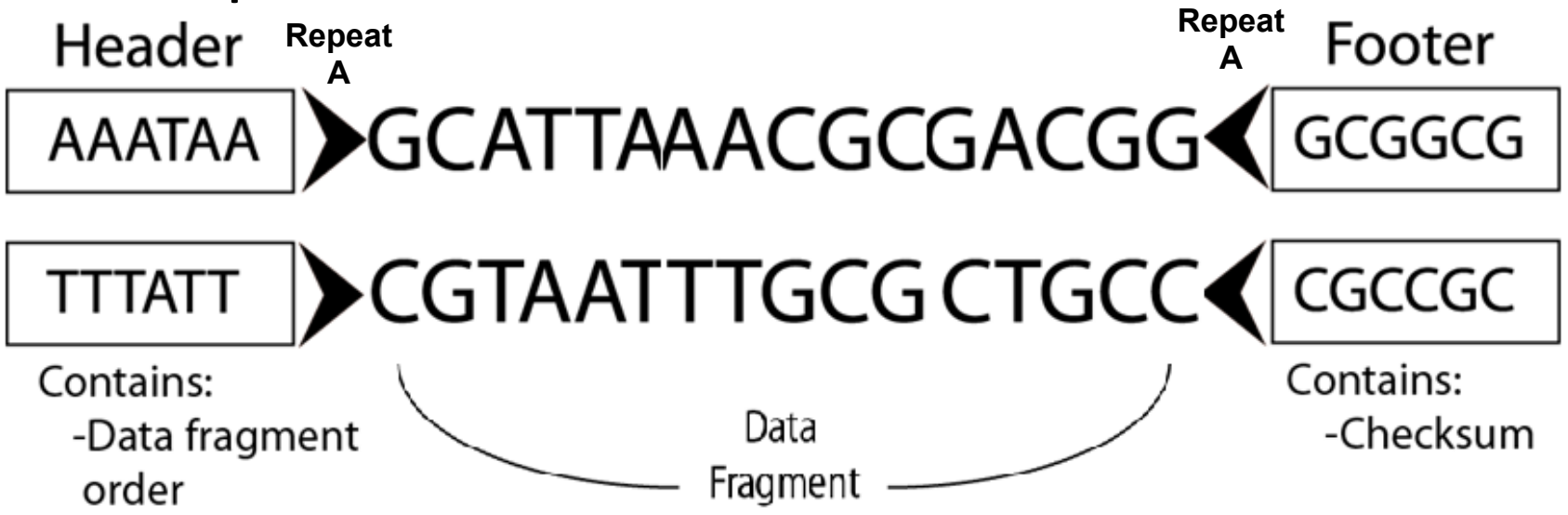
eg. “tools”

DNA encoding:
TGTATCGGTC
GGTCGATGAG
(20bp)



Structure of message

- Repeat A sequence in natural shufflon system has the highest inversion frequency
- 19bp



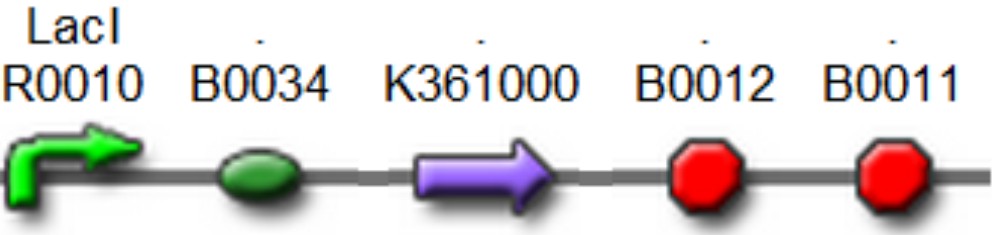
Parts designed

DNA
Message

Message gene template (438bp)
Synthesized DNA

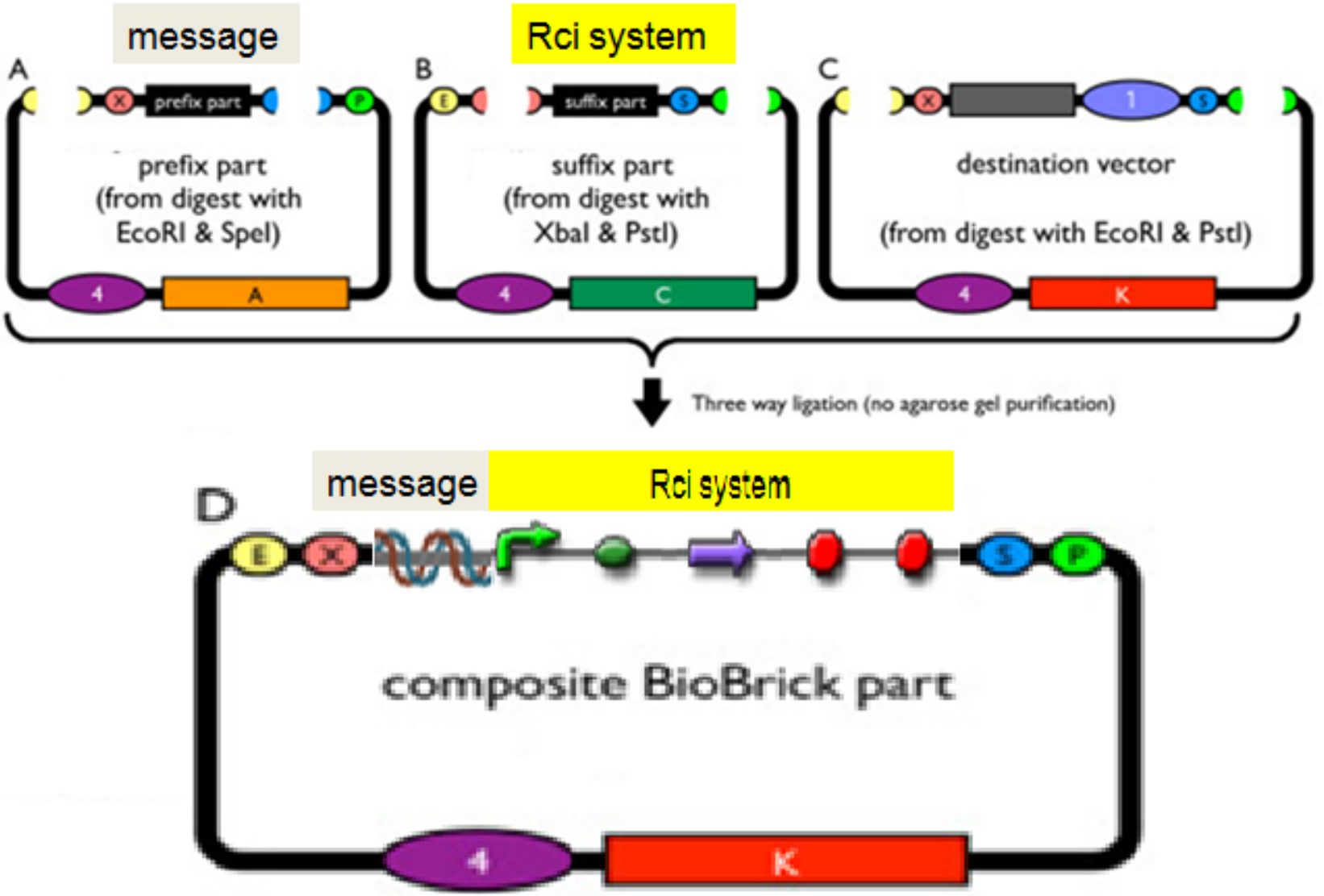
K361000

Rci site-specific recombinase (1155bp)
Synthesized DNA (rci gene sequence of *E. coli* (strain: K-12))



Rci system (1484bp)
- lac promoter
- ribosome binding site
- rci gene
- double terminator

Integration of message to rci system

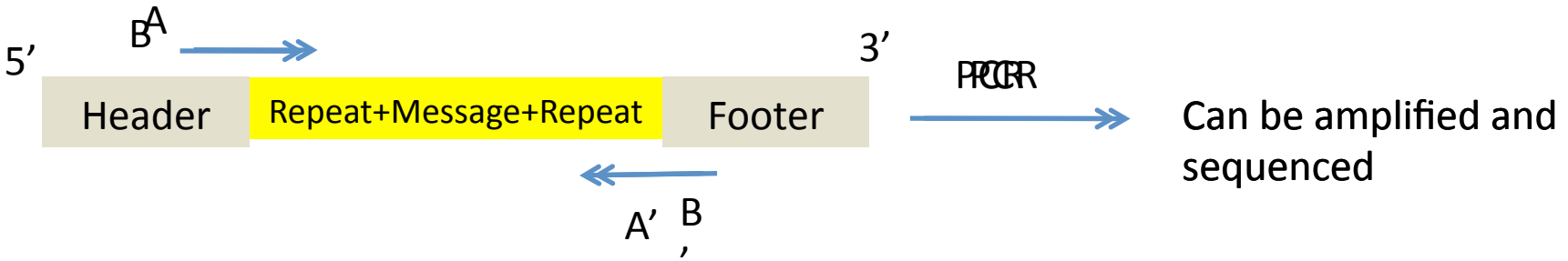


Expectation

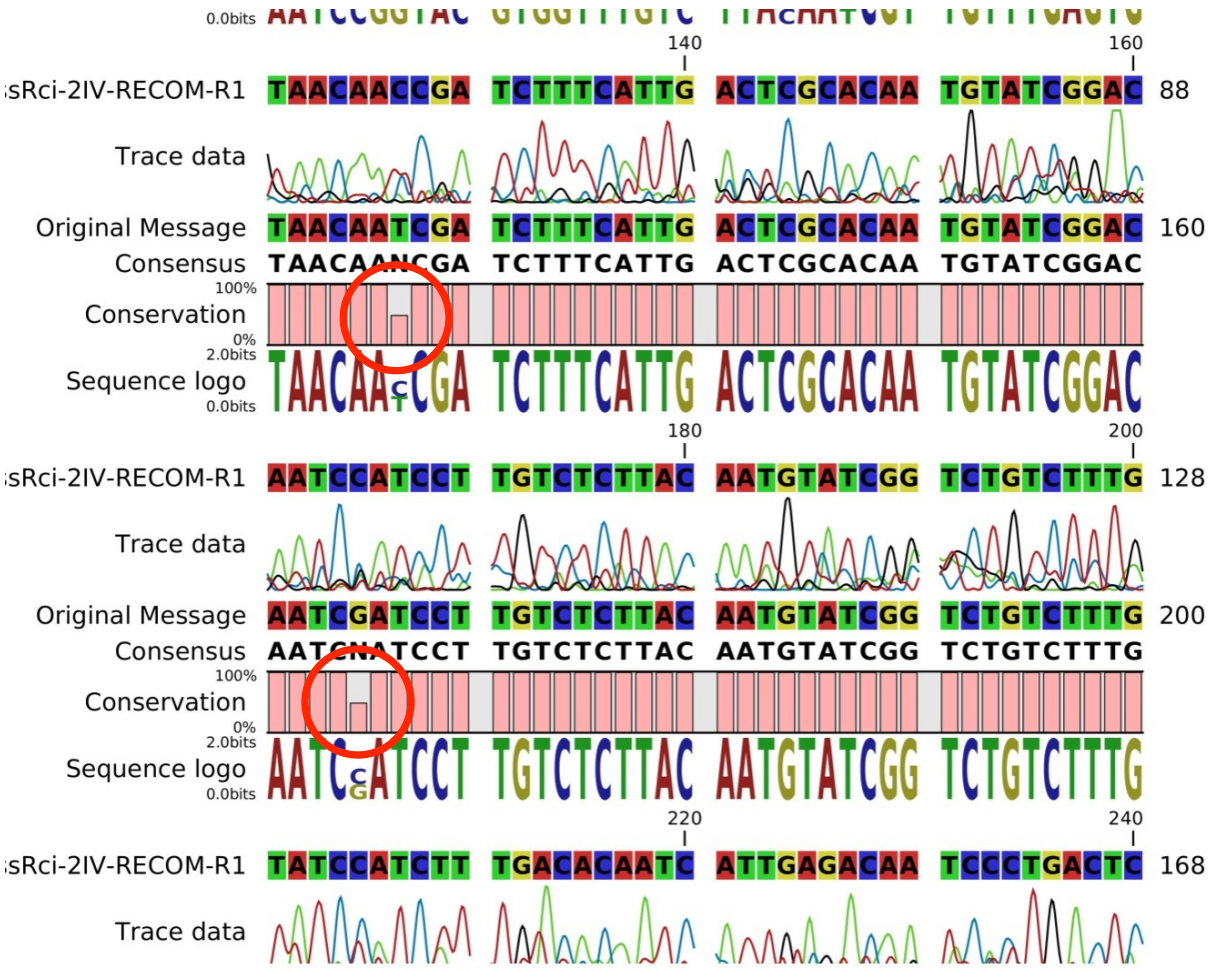
- Repeat sequence + Message + Repeat sequence
- There should be two scenarios:

1. Inversion of message
2. No change of original message

- Two sets of primers are used



Results

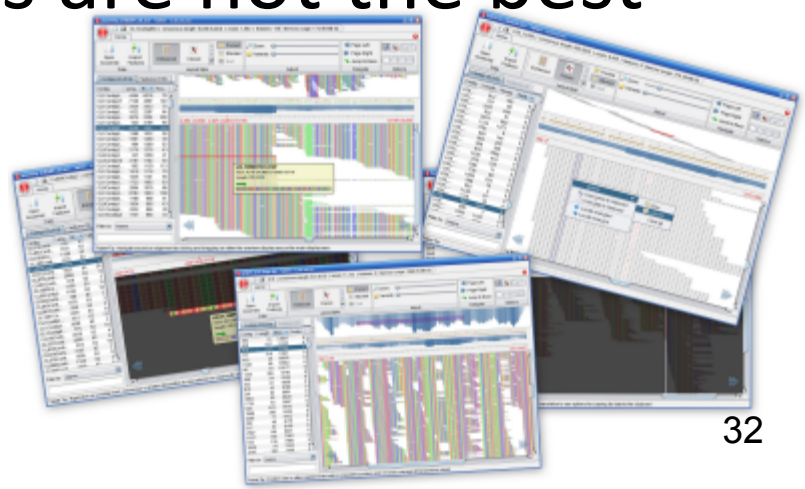


- Inverted and original message were found
- No loss of DNA

Checksum and high throughput sequencing!!!

High throughput sequencing

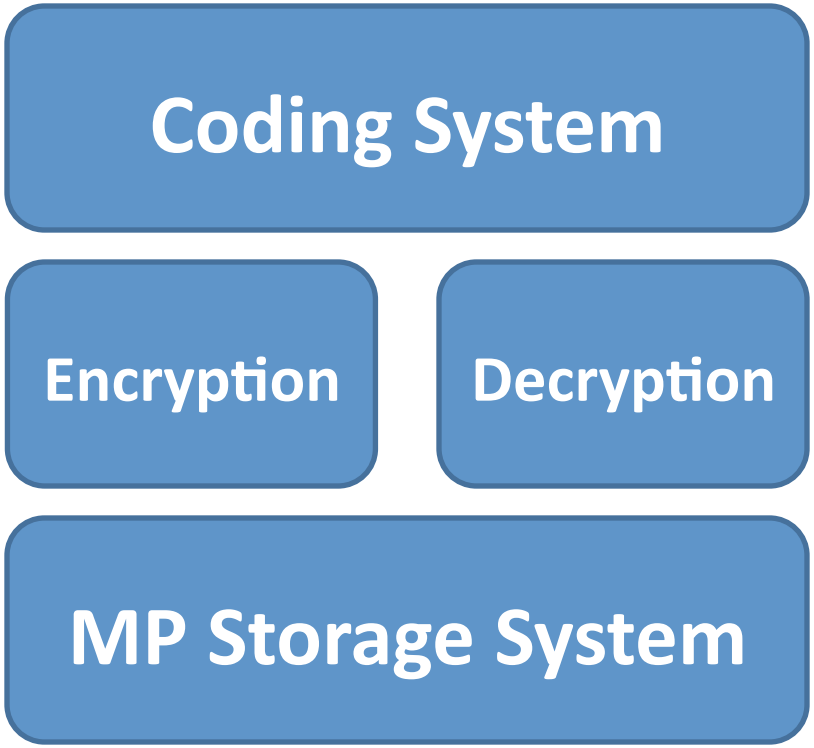
- Massively parallel sequencing process
- Multiple copies of sequencing products (reads) that can cover a particular message stored within the DNA
- Enable us to perform a majority voting on bases for which qualities are not the best



FUTURE PERSPECTIVES

To summarize…

- Infrastructure of our system



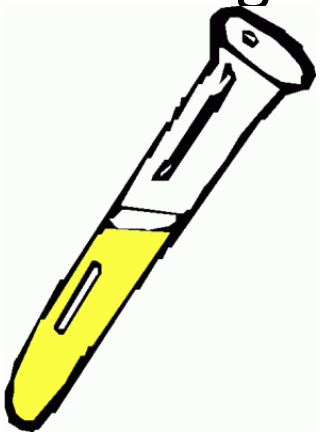
- Experimental proof

Bio-hard disk

	Storage
Hard disk	2000GB
1 gram E.coli	900,000GB

Therefore....

1 gram(wet weight) of E.coli



= ***450***

2 TB hard disk

Rapid & Specific access

- Parallel storage system

Insert *Header & Footer* in every message fragment

↓

Design *specific probe* corresponding to Header

↓

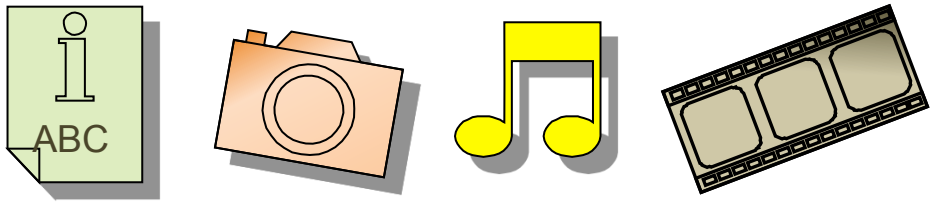
Pick up particular message from pool of data

↓

Targeted sequencing

Future Application

- Can store text, images, music, movies……



- Insertion of barcodes *into synthetic organisms* as a part of current safety protocols to distinguish between synthetic & natural organisms
- Store additional information:

Copyrights **Safety protocols**

Designers of the organisms

Acknowledgement

Eastwin

IDT INTEGRATED DNA TECHNOLOGIES

CuGen

生物化學課程 Biochemistry Programme

Further Information

- Gyohda, A. & Komano, T. (2000). Purification and Characterization of the R64 Shufflon-Specific Recombinase. J. Bacteriol., 182 (10), 2787-92.
- Gyohda, A., Zhu, S., Furuya, N. & Komano, T. (2005). Asymmetry of Shufflon-specific Recombination Sites in Plasmid R65 Inhibits Recombination between Direct sfx Sequences. J. Biol. Chem., 281 (30), 20772-9.

If you would like to know more about our project, you are welcome to visit our Wiki page:

http://2010.igem.org/Team:Hong_Kong-CUHK


FIRE
BLANKET
F01
F02
F03

Q&A

Inversion frequency

1. types of 19-bp repeat sequences

(repeat-a> repeat-d> repeat-b or repeat-c)

2. distance between repeat sequences

(distance increases, frequency increases)

3. DNA sequences surrounding the repeat sequences(symmetric repeat sequence increase frequency)

Inversion frequency

4. presence of HU protein

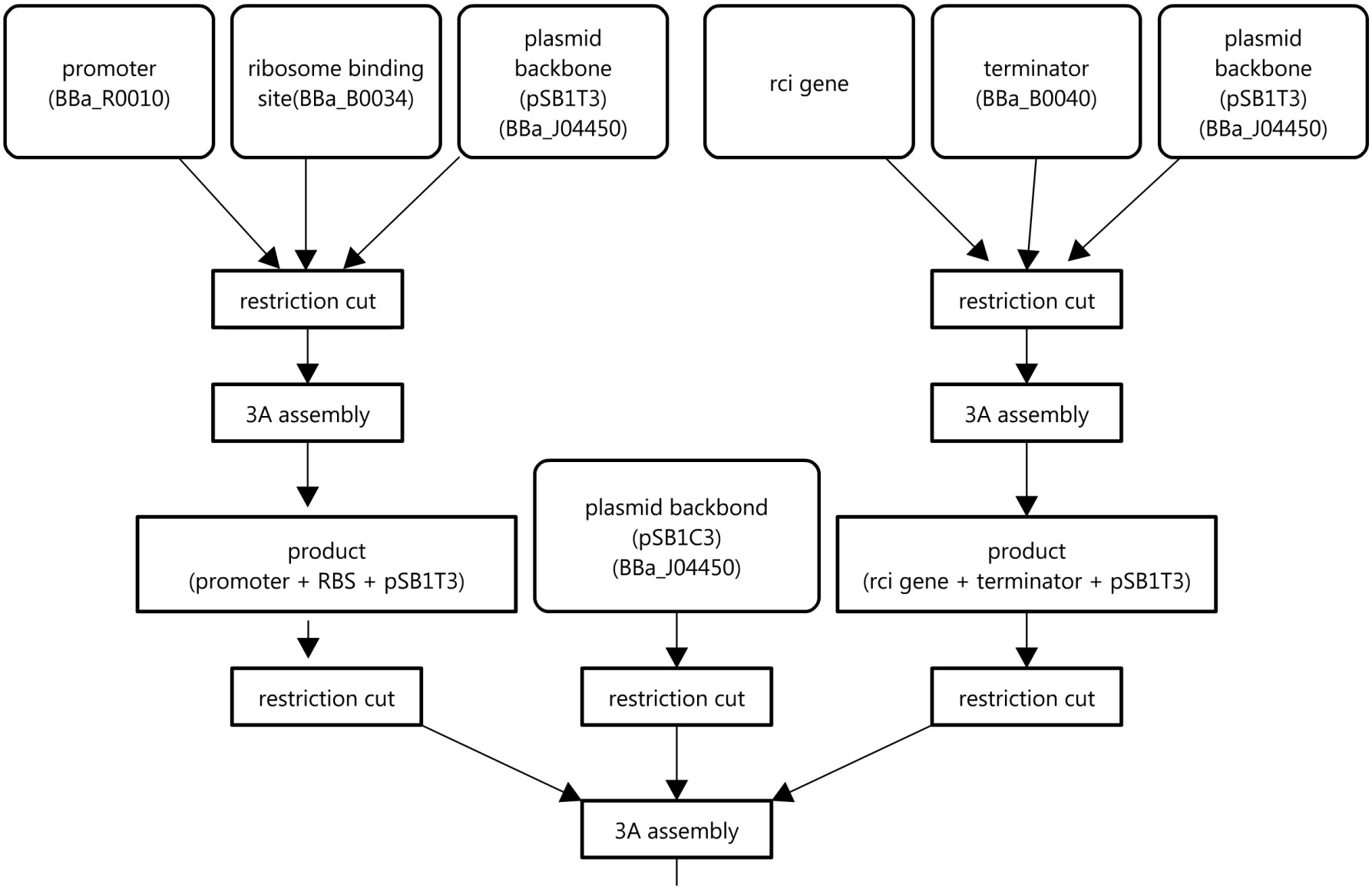
   (binding of HU protein to DNA might facilitate assembly and/or stabilization of the Rci-DNA complex at the recombination sites, increases frequency)

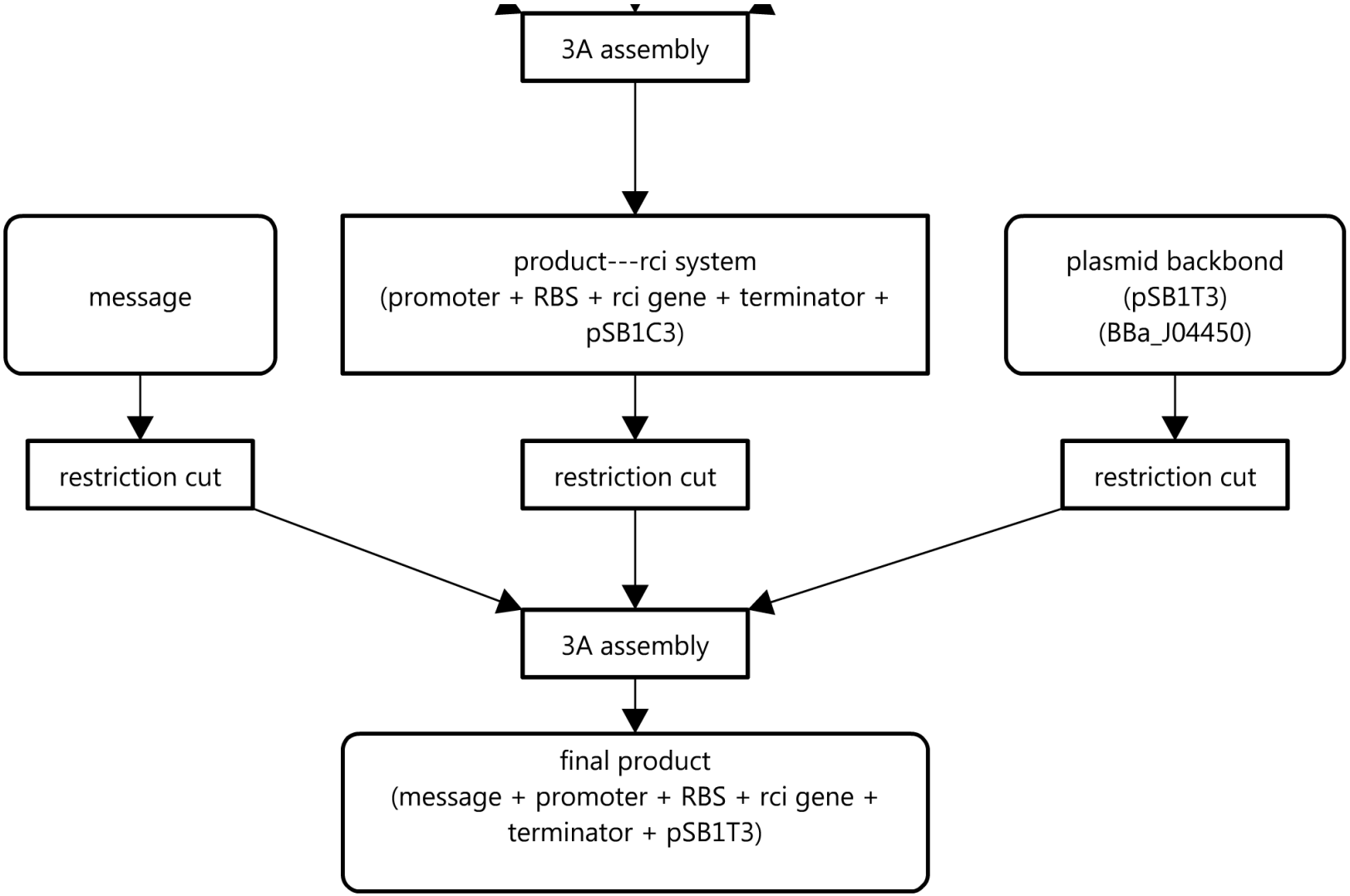
5. extent of DNA supercoiling

   (Inhibition of DNA supercoiling→ decrease Rci activity→ decrease inversion frequency)

To avoid mutation

- Reduce reproductive cycle
- Provide favorable condition
- Move on to eukaryotes, make use of eukaryotes' proofreading system(more sophisticated DNA repair system)


promoter
(BBa_R0010)
ribosome binding
site(BBa_B0034)
plasmid
backbone
(pSB1T3)
(BBa_J04450)
rci gene
terminator
(BBa_B0040)
plasmid
backbone
(pSB1T3)
(BBa_J04450)
restriction cut
restriction cut
3A assembly
3A assembly
plasmid backbond
(pSB1C3)
(BBa_J04450)
product
(promoter + RBS + pSB1T3)
product
(rci gene + terminator + pSB1T3)
restriction cut
restriction cut
restriction cut
3A assembly


3A assembly
message
product---rci system
(promoter + RBS + rci gene + terminator +
pSB1C3)
plasmid backbond
(pSB1T3)
(BBa_J04450)
restriction cut
restriction cut
restriction cut
3A assembly
final product
(message + promoter + RBS + rci gene +
terminator + pSB1T3)

Pacific Bioscience

- Real time
- Read Length : 1000 - 10000bp
- Single Molecule Sequencing
- 30 minutes sequencing process