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
  – Checksum

• 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

Encoding table

Quaternary Number System

DNA sequence

Compression
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 → 12210111301110131

Use “A, T, C & G” to Represent the Numbers

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

iGEM → 12211011301110131
   → ATCTATTGATTTATGT
Enter your text here: igEM is very interesting

Quaternary Encoding:
12211013101110310200122113030200131212111302132102001221123213101211130212111303131012211232
1213

DNA Encoding:
TCCTTATGTATTTAGTACAATCCTTGAGACAATGTCTCTTTGACTGCTACAATCCTCGCTGTATCTTTGACTCTTTGAGCCTTGATCCTTCCGCTCTG

Original message input
Converted to Quaternary number
Converted to DNA sequence
• **DEFLATE** — a compression algorithm

1. Can reduce the **homopolymer** and repetitive regions

2. Can store more information
The length and repetitive sequence is greatly reduced
• Provide DNA variation
• DNA Shuffling system
• Examples:
  
<table>
<thead>
<tr>
<th>Homologous recombination</th>
</tr>
</thead>
<tbody>
<tr>
<td>RACHITT</td>
</tr>
</tbody>
</table>

Encryption
Simulation Analysis

The United States Declaration of Independence

8074 Characters!!!
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 2\textsuperscript{nd} fragment

Header of 1\textsuperscript{st} fragment

Header of 3\textsuperscript{rd} fragment

Header of 4\textsuperscript{th} fragment
If insertion size per cell is 1kb......

The United States Declaration of Independence

Only 18 cells!!!
Capacity

1 gram of cells consists of ~ 10 Million cells.

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

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

sequencing

Identification of **repeat**, **message**, **checksum**

Checksum system

**Header**
- **AAATAA**
- **TTTATT**
  - Contains: Data fragment order

**Repeats**
- **GCATTA**
- **AACGC**
- **GACGG**
- **CGTAAT**
- **TTGCG**
- **CTGCC**

**Footer**
- **GCGGCG**
- **CGCCGC**
  - Contains: Checksum
Checksum Mechanism

F1 → F2 → F3 → Checksum

Checksum 1:
F2 → F3 → F1

Checksum 2:
F1 → F2 → F3

CRC64
WET LAB TEAM
To prove our concept...

Coding
(extended ASCII code)

Encryption
(message + rci system)

Storage
(in *E.coli* DH5α)

Decryption
(sequencing)
4. The host cell duplicates the data storage vectors, which helps to ensure data integrity by redundancy.
• 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)
• Repeat A sequence in natural shufflon system has the highest inversion frequency

• 19bp
Parts designed

- **Message gene template (438bp)**
  - Synthesized DNA

- **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

Prefix part (from digest with EcoRI & Spel)
Suffix part (from digest with Xbal & PstI)
Destination vector (from digest with EcoRI & PstI)

Three way ligation (no agarose gel purification)

Composite BioBrick part
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

```
5'    BA    3'  
Header Repeat+Message+Repeat Footer 
A'B

Can be amplified and sequenced
```
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

- Coding System
  - Encryption
  - Decryption
- MP Storage System

• Experimental proof
Bio-hard disk

<table>
<thead>
<tr>
<th></th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard disk</td>
<td>2000GB</td>
</tr>
<tr>
<td>1 gram E.coli</td>
<td>900,000GB</td>
</tr>
</tbody>
</table>

Therefore....

1 gram (wet weight) of E.coli = 2 TB hard disk

= 450
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
Further Information


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
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 $\rightarrow$ decrease Rci activity $\rightarrow$ 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)
3A assembly

product---rci system
(promoter + RBS + rci gene + terminator + pSB1C3)

plasmid backbond
(pSB1T3)
(BBa_J04450)

message

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