DNA data storage method

Jakub Dobrík
University of Tartu
2016
We will go through:

- Brief history of DNA storage method
- Why there is a need for this method
- Basic elements of DNA storage method
- Writing and reading processes
- Data representation
- Homework
Me listening to some biological problems
• Creator Nick Goldman
• The idea started as a joke
• Among first things stored in this form were Shakespeare's sonnets and a snippet of Martin Luther King's 'I have a dream' speech
• Storing data the traditional way will become too expensive
• Solution for this issue is storing the data in DNA

![STORAGE LIMITS](source: http://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496)

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today’s storage technology.

<table>
<thead>
<tr>
<th></th>
<th>Hard disk</th>
<th>Flash memory</th>
<th>Bacterial DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read–write speed</strong> (µs per bit)</td>
<td>~3,000–5,000</td>
<td>~100</td>
<td>&lt;100</td>
</tr>
<tr>
<td><strong>Data retention</strong> (years)</td>
<td>&gt;10</td>
<td>&gt;10</td>
<td>&gt;100</td>
</tr>
<tr>
<td><strong>Power usage</strong> (watts per gigabyte)</td>
<td>~0.04</td>
<td>~0.01–0.04</td>
<td>&lt;10(^{-10})</td>
</tr>
<tr>
<td><strong>Data density</strong> (bits per cm(^3))</td>
<td>~10(^{13})</td>
<td>~10(^{16})</td>
<td>~10(^{19})</td>
</tr>
</tbody>
</table>
DNA storage system

Figure 3. Overview of a DNA storage system.

• Basic unit is DNA strand (100-200 nucleotides) – capable of storing 50-100 bits of data
• The DNA strands are stored in so called „pools“
• These pools have „stochastic spatial organization“ – not structured addressing
• The address of the data must be written in the strand data itself (so the data’s value can be retrieved)
Writing process

- The input is key and a value
- The key is used to obtain the PCR primer sequences, calculating the high part of the address, then finding a pool in the DNA library where the results will be stored
- The low part of the address indexes the strands
- The data is encoded + the primary target sequence is attached
- The synthesizer manufactures the final DNA sequence

(a) The write process performs `put(key, value)`, generating a DNA library.

Input is a key
Identification of molecules associated with the key
Physical extraction of a sample from a DNA pool containing the data (lots of unrelated data attached)
Amplification of the desired strands
Finally, DNA sequencer produces the digital data

Data representation

- Huffman code is used to map binary byte into ternary digits.
- The rotating nucleotide encoding maps this string to the DNA sequence CTCTG.

Useful reference 😊

Questions?

Please don't
Homework

• Short test with 4 possible answers for each question
• Only one answer is correct
• Please send the results (sequence of correct answers in txt or pdf with highlighted correct answers) to kubodobrik@azet.sk until 2th od December, 10:00.
Thank you for your attention

- References:
Test:

Only one answer is correct!

1. Why was the idea behind DNA data storage created?
   a) Just for fun
   b) Current storage media is becoming insufficient
   c) Better duration during space travels
   d) CEO of Microsoft placed an order on the web

2. What makes DNA better solution than conventional solutions?
   a) Data can be written into molecules more densely than the basic elements of conventional storage technologies can
   b) The European Union plans to ban the conventional hard drives as it is bad for the environment
   c) Nothing
   d) It is cheaper technology and easily recoverable

3. A DNA storage system consists of:
   a) A balloon, hot water and a nail
   b) a DNA synthesizer, storage container and a DNA sequencer
   c) a DNA synthesizer, storage container, a DNA sequencer and two hard drives
   d) a DNA synthesizer and a storage container

4. How many bits can a single DNA strand store?
   a) 100 MB
   b) 60-400 KB
   c) 50-100 bits
   d) 1 bit

5. How do we get rid of unrelated data while reading from a DNA?
   a) Use a “clean” function
   b) Use Pearson correlation
   c) Brush it with a broom and tell it to leave immediately
   d) Send the sample and the PCR primers to the PCR thermocycler

6. Find some project/tool/interesting articles related to the topic and please provide a link (just to find me some further reading 😊)