The Big Data of Cells

**Big numbers** are everywhere and they represent an enormous amount of continuously generated data and information, from bank accounts to supermarket transactions, website accesses and videos uploads into the cloud. Current technologies allow us to measure everything with special units, like bytes, Hertz, centimeters, etc. Now, biologists are coming across huge amounts of information, especially after the “Genomics Revolution” coupled to noteworthy breakthroughs in computer technology. In the past 15-20 years, many high-throughput biological data sets have been published, particularly after the Human Genome Project was completed in 2003. Merely 200,000 genomes would fit in the 700 terabytes of the Netflix’s repository; with a global population expected to reach 9 billion by 2050, at least 5,000 times more data would be produced by genome sequencing. The number of sequenced genomes could reach more than one million within five or ten years. In short, more data has been generated in the past two years than in human history.

Bioinformaticians and computational biological scientists have been struggling with conventional computing technology limitations, and, therefore, they started to wonder about some ideas that sound as if they came out of a sci-fi movie. If we have reached physical limits of computer memory, can we use other type of material for storage? That material happens to be biological and every living being on Earth has it: the DNA.

The DNA is a molecule that stores the genetic information of a certain specimen and can last more than thousands of years. It surpasses by far the information density and energy needed for operation offered by flash memory available today. It may seem odd, but five years ago scientists already successfully used DNA to encode five files, including Shakespeare’s sonnets and a snippet of Martin Luther King’s ‘I have a dream’ speech. In 2016, researchers from Microsoft and the University of Washington produced a DNA archive of 200 megabytes (MB), including a music video, setting a record. Last year, the corporation announced a considerable investment in this new promising technology.

<table>
<thead>
<tr>
<th>Different systems capacities</th>
<th>Hard disk</th>
<th>Flash memory</th>
<th>Bacterial DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read-write speed (µs per bit)</strong></td>
<td>~3,000-5,000</td>
<td>~100</td>
<td>&lt;100</td>
</tr>
<tr>
<td><strong>Data retention (years)</strong></td>
<td>&gt;10</td>
<td>&gt;10</td>
<td>&gt;100</td>
</tr>
<tr>
<td><strong>Power usage (watts per gigabyte)</strong></td>
<td>~0.04</td>
<td>~0.01-0.04</td>
<td>&lt;10^-10</td>
</tr>
<tr>
<td><strong>Data density (bits per cm³)</strong></td>
<td>~10^13</td>
<td>~10^16</td>
<td>~10^19</td>
</tr>
</tbody>
</table>

Unfortunately, some obstacles remain. Despite being a very efficient system, it is also prone to errors, slow and expensive because of the chemical process used to assemble the DNA strands. Microsoft’s project, for instance, used 13,448,372 unique pieces of DNA, costing at least $800,000. Regarding the writing speed, it is estimated in only 400 bytes per second. Data retrieval can be done using a high-speed sequencing machine, but still, its efficiency would have to double to be commercially feasible.
The First Binary DNA String

In 1988, the artist Joe Davis, in association with Harvard Medical School, designed and synthesized an 18 base-pair message encoding the image of the ancient Germanic rune representing the female Earth – Microvenus –, pasted into a vector and transformed into E. coli creating a living work of art. This message was coded under a simple principle, where the lines of the image were translated to ones and zeros in a 5x7 grid, converted to DNA with phase-change values rather than numerical values; each base is used to indicate how many times each binary bit (zero or one) is to be repeated before changing to the other binary bit.

This is a technique used in many forms of computer-compression. For example, the binary code 10101 would be CCCCC, because each binary code occurs only once before switching to the other digit. Following the same logic, 0111000 would be CAA.

Programming in Life’s Language

Many research groups, universities and companies have already developed their own techniques to write and read in the DNA storage system. The most promising ideas are from the University of Illinois research group, which despite its high error rate, allows random access to data, is portable, includes an error detection and rectification step and reaches a net density of 1.72 bits/bp (base pair), the highest achieved until now. Their approach includes a data encoding step and a post-processing step. The first consists of compression of data, removal of certain substrings and balancing of GC content, which is a problematic issue for synthesis and sequencing. Moreover, mathematically constructed addresses for pieces (gBlocks) of data are added through constrained coding for future random access through highly selective PCR reactions. In the post-processing step, address sequence identification and error corrections take place.

Figure 1. In the encoding step the data is first compressed, then the bits are encoded into DNA blocks through constrained coding, which limits occurrence of the address block to one or two predefined positions only and balances the GC-content in each substring. Additional homopolymer checks are added in the string for posterior errors correction.
During the post-processing step the data is read out after nanopore sequencing, which can be performed on a portable MinION device. After this initial reading, estimates are obtained by running several Multiple Sequence Alignment (MSA) algorithms with exact matches of the addresses. MSA consists in aligning three or more biological sequences of similar length and from the output can be used for phylogenetic analysis. However, in this case, the similarity between the DNA blocks are used to define a consensus sequence and eliminate small deletions within the code. Afterwards, several error corrections are run and the file is integrally recovered.

Challenges

Research teams and many companies that are at present investing in this revolutionary technology, are focusing in reducing the costs of DNA synthesis, just as genomics reduced the high cost of sequencing. However, current trends in technological advances are reducing the costs of DNA synthesis at a pace that should make some schemes cost-effective within the next decades.

Some other issues include DNA’s vulnerability to errors and data retrieval. An efficient method to avoid or correct mistakes while the information is being written is essential. Moreover, standard sequencing methods only allows to retrieve data after reading the whole DNA string, there is no way to get to a specific piece of data.

The Microsoft-University of Washington researchers outlined their solution for data retrieval. Through the polymerase chain reaction (PCR) they pinpoint and make more copies of the region they want to extract. This approach is cheaper and more accurate. In their demonstration, they could read the Sydney Opera House image without any mistakes.

In order to make DNA-based data storage commercially competitive, it is critical to reduce synthesis cost, provide random point access, reduce sequencing costs through portable readout systems, and offer extremely low error rates.

Bibliography


