DNA Data Storage Alliance: Building a DNA Data Storage Ecosystem

Dave Landsman
Senior Director Industry Standards
Distinguished Engineer
Western Digital
What is the problem?

There is too much data to save

And value of saved data is growing
Why DNA?

DNA bits are very small

Volume of an LTO cassette filled with DNA bits would hold ~2,000,000 TB; this would fill ~115,000 LTO-9 tapes

1mm³ holds ~9TB of encoded DNA bits (1/2 LTO-9 tape capacity)

Single DNA Base fits in ~1nm³

… and they last a long time, don’t need much care and feeding, and don’t need migration (= TCO)
… and they’ll benefit from, and accelerate, established investment in DNA technology
But there are questions

- Is there really a need for a medium as dense as DNA?
- Can we scale the underlying technologies?
- How do we create an interoperable DNA data storage ecosystem?

DNA Data Storage Track

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker(s)</th>
<th>Affiliation</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30am-9:20am</td>
<td>Dave Landsman</td>
<td>Distinguished Engineer Western Digital</td>
<td>Creating a DNA Data Storage Ecosystem</td>
</tr>
<tr>
<td>9:30am-10:20am</td>
<td>Joel Christner Alessia Marelli Mark Wilcox</td>
<td>Distinguished Engineer - Dell Technologies CTO - DNA Algo CEO - 21e8</td>
<td>Rosetta Stone: Standard to enable discovery and decode info in a DNA data archive</td>
</tr>
<tr>
<td>10:35am-11:25am</td>
<td>Alessia Marelli Rino Micheloni</td>
<td>CTO and COO DNAAlgo</td>
<td>DNAassim: A Full System Simulator for DNA Storage</td>
</tr>
<tr>
<td>11:35am-12:25pm</td>
<td>João Reis Marília Menossi</td>
<td>Researchers Lenovo / Instituto de Pesquisas Tecnologicas</td>
<td>End-to-End DNA data storage system study</td>
</tr>
<tr>
<td>1:30pm-2:30pm</td>
<td>João Gervasio Adriano Galindo Leal</td>
<td>Researchers Lenovo / Instituto de Pesquisas Tecnologicas</td>
<td>DNA Coding Overview</td>
</tr>
<tr>
<td>2:30pm-3:20pm</td>
<td>Luca Piantanida</td>
<td>Research Scholar Boise State University</td>
<td>Nucleic Acid Memory</td>
</tr>
<tr>
<td>3:35pm-4:25pm</td>
<td>Informal Q&amp;A w/ Presenters</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Oligonucleotides, or oligos, are short single strands of synthetic DNA or RNA.

“A” ≡ “T”  
“C” ≡ “G”
The DNA Data Storage “Channel”

Some (hopefully) illustrative analogies between Electrical and DNA

<table>
<thead>
<tr>
<th>Electrical Channel</th>
<th>DNA “Channel”</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 1’s and 0’s converted to analog wave forms at transmitter, back to 1’s and 0’s at receiver</td>
<td>• 1’s and 0’s converted to DNA bases; that is, “wave forms” in electrical case become DNA bases in DNA case</td>
</tr>
<tr>
<td>• ECC bits added to digital bit stream before transmission, checked/stripped at receiver to check/correct data errors</td>
<td>• ECC bits added to digital bit stream (by codec) before synthesis (transmitter) and checked/stripped after sequencing (receiver)</td>
</tr>
<tr>
<td></td>
<td>• DNA errors: insertions/deletions (indels), substitutions (SNVs), ...</td>
</tr>
<tr>
<td>• Scrambling patterns (reordering 1’s, 0’s) added at transmitter to avoid analog effects which can cause errors on wire</td>
<td>• Some patterns of bases problematic to synthesize/sequence, so we may alter bases (ACCG...) after ECC/metadata encoded</td>
</tr>
</tbody>
</table>

- As with electrical, the “line protocol” for DNA data storage is critical to overall channel efficiency
- There are also “logical” protocol layers above line protocol, e.g., file tagging, packetization
Synthesis Techniques - Snapshot

- **Base-by-base**
  - Two underlying base-by-base techniques: Phosphoramidite and Enzymatic
  - Both methods use similar cyclic process
  - Limit of 200-300 bases per oligo today

- **Ligation**
  - Can enable strands of many hundreds of bases or longer
  - More bits in payload means, in general, less protocol overhead
Synthesis - Scaling

Scaling DNA data storage w/ nanoscale electrode wells

- Study results
  - Electrochemical DNA synthesis w/ 650nm wide electrodes in 200nm wells
  - Acid containment at 2μm pitch with these feature sizes
  - 100-base long DNA data storage with these feature sizes

- Some throughput implications from the study
  - The chip in this study reached synthesis density of 25M synthesis sites/cm²; nearly 3 orders of magnitude greater than previous work
  - An array with this synthesis density could achieve write speed of >2.8 KB/s/cm²* which could be practical minimum for some archival uses
  - But at this density, it would require a 360cm² chip (not manufacturable), or many chips, to achieve, say, 1MB/s, and it would use lots of reagents

* assuming each unique oligo encodes 10 bytes of data and is written over 24 hrs
There is continued progress however:

- Twist Bioscience just announced chip w/ synthesis density of 100M synthesis sites/cm² and ability to write 1GB per run
- Moving toward IARPA Molecular Information Storage (MIST) synthesis goals on capacity, scale, and cost

Much more scaling needed, but foundations are established.
Storage of DNA based bits

- Many preservation methods being explored
- As well as verification methods to enable comparison of data retention capabilities

<table>
<thead>
<tr>
<th>Principle</th>
<th>Procedure</th>
<th>Source (References)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical encapsulation</td>
<td>Moha bones</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>Fox teeth</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>Encapsulation in salts</td>
<td>[12, 16]</td>
</tr>
<tr>
<td>Physical encapsulation</td>
<td>Silica nanoparticles</td>
<td>[1]</td>
</tr>
<tr>
<td></td>
<td>Stainless steel capsules</td>
<td>[3]</td>
</tr>
<tr>
<td></td>
<td>Magnetic silica nanoparticles</td>
<td>[13]</td>
</tr>
<tr>
<td>Inclusion in a matrix</td>
<td>DNASTABLE</td>
<td>[1, 21]</td>
</tr>
<tr>
<td></td>
<td>Gentegra DNA</td>
<td>[1, 22]</td>
</tr>
<tr>
<td></td>
<td>Pullulan</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>Silk</td>
<td>[15]</td>
</tr>
<tr>
<td></td>
<td>300K matrix inclusion</td>
<td>[25]</td>
</tr>
<tr>
<td>Absorption on paper</td>
<td>FTA paper</td>
<td>[1, 23, 24]</td>
</tr>
<tr>
<td></td>
<td>Chitosan treated paper</td>
<td>[17]</td>
</tr>
<tr>
<td>Dehydration on solid supports</td>
<td>Capillaries</td>
<td>[20]</td>
</tr>
<tr>
<td></td>
<td>Glass</td>
<td>[26, 27]</td>
</tr>
<tr>
<td></td>
<td>Tube walls</td>
<td>[28]</td>
</tr>
<tr>
<td>Dissolution in liquid salts</td>
<td>Imidazolium cations</td>
<td>[18]</td>
</tr>
<tr>
<td></td>
<td>Imidazolium cations</td>
<td>[19]</td>
</tr>
<tr>
<td>Living organism</td>
<td>Bacteria</td>
<td>[29]</td>
</tr>
</tbody>
</table>

Sequencing - Snapshot

- **Sequencing by Synthesis (SBS)**
  - Start with a ssDNA (template)
  - Build a complementary strand (synthesis)
  - Each binding event is detectable

- **Nanopore**
  - Guide DNA strand through very small channel: nanopore
  - As strand traverses nanopore, ionic current or tunneling current are detected and the bases are directly read

- In general, today, per base, SBS more accurate/slower and nanopore less accurate/faster
- Throughput battles in commercial systems

Source: Illumina

Sequencing - Scaling

### Analysis

- **Throughput**: We are now at ~10s of GBytes/day (per machine); probably need to get to 100s of TBytes/day
- **Cost**: Using the $/base numbers (assume 1-bit/base)
  - at $48000/TB ($6/Gbit) now
  - with **direct line of site** to $8000/TB ($1/Gbit)
  - and **no conceptual hurdles** to $800/TBytes ($0.1/Gbit)
- This is based on list pricing and requirements for medical/genomic markets
- DNA storage can tolerate higher error rates

### Conclusion

- We have an apparent 3 orders of magnitude to go on both cost/price & throughput for sequencing
- That said, there are many ways to manipulate all phases (codecs, synthesis, sequencing) to balance error tolerance and performance in the whole pipeline
Selective retrieval of digital data encoded in DNA molecules


Also see Scalable and Dynamic File Operations for DNA-based Data Storage (James Tuck, North Carolina State, DNAli Data Technologies) from SDC 2021.
In conclusion: DNA data storage is now resting on a solid foundation
If we build it, who wants to use it and why?

- ADAS
- Hyperscale
- Governmental
- Media/Entertainment
- Genomics/Omics
- Digital Art
- Preservation
- Smart Video
Data Retention - Key to discovery, monetization, …

- Healthcare, astronomy, climate science, sports, smart cities and vehicles, governments/municipalities, etc. seeking to save ever larger data sets

- We cannot know today what data will become relevant to new discoveries or required information tomorrow but storing too much data imposes undue costs

- If we can store more data for a lower total cost, the tradeoff between saving or discarding data can be shifted in favor of saving vs. discarding

Source: Karl G. Jansky Very Large Array - NRAO/AUI/NSF
And data retention puts emphasis on TCO

- Because DNA-based storage promises no data migration and nominal-to-no fixity checks, TCO looks better vs. traditional storage as retention times lengthen.

- DNA-based storage minimizes energy consumption and improves sustainability.

This trend example holds even for one copy of the database; there will be multiple copies.
So how do we build the ecosystem?

DNA DATA STORAGE ALLIANCE
## DNA Data Storage Alliance - At a Glance

### History
- Formed on October 12th, 2020 by Illumina, Microsoft, Twist and Western Digital
- Climbed to 60+ members by 2Q-2022
- Joined SNIA as a Technology Affiliate group as of Jun-2022

### Mission
- Create and promote an interoperable storage ecosystem based on DNA as a data storage medium
- Educate the market to create awareness and adoption of DNA data storage

### Scope
- Develop a DNA data storage industry technology roadmap to drive R&D and funding
- Develop standards and/or specifications as needed by ecosystem
DNA Data Storage Alliance

2022 Activities

- Industry technology roadmap
- Start workgroups for potential standardization
- White Paper #2 - Market segments/use cases
- Newsletter
- Events
Industry Technology Roadmap

Roadmap for how DNA data storage can scale to commercial viability

- Key technologies and challenges in the pipeline
- Success metrics: capacity, transfer rates, cost, …

Guide for academic/industry research and investment
DNA Data Storage Alliance: TWG sub-groups

**DNA Archive Rosetta Stone (DARS)**
Create universal identifier describing how to decode/read rest of archive

**Interoperable Interfaces**
Ensure physical compatibility of synthesis, storage, and sequencing modules to ensure “plug-and-play” integration and avoid vendor lock in:
- plug and play swaps of instruments
- recovery of molecules for read, irrespective of supplier being existent at read time
- issues of fluidics in data centers

**DNA Digital Data Retention**
Define standard metrics and verification methods to enable comparison of data retention in DNA-based data storage solutions (retention properties, complexity of retrieval, etc.)

An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage
A disclaimer before closing

- DNA is not like other media; it needs some explaining
  - “Does this mean I’m going to be storing my music collection in my dog?”

- It is important to clarify

  Storing digital data in synthetic DNA molecules in no way requires the use or creation of any cells, organisms, or life!

  We are ‘simply’ using chemical mechanisms to store digital bits in DNA molecules instead of using electromagnetic or optical mechanisms to store bits in silicon, magnetic, or other materials.
THANK YOU

Enjoy rest of the track and come join the efforts to create a DNA data storage ecosystem!!

Subscribe to our newsletter on our website
https://dnastoragealliance.org

And make sure to check out:
Preserving our Digital Legacy: An Introduction to DNA Data Storage