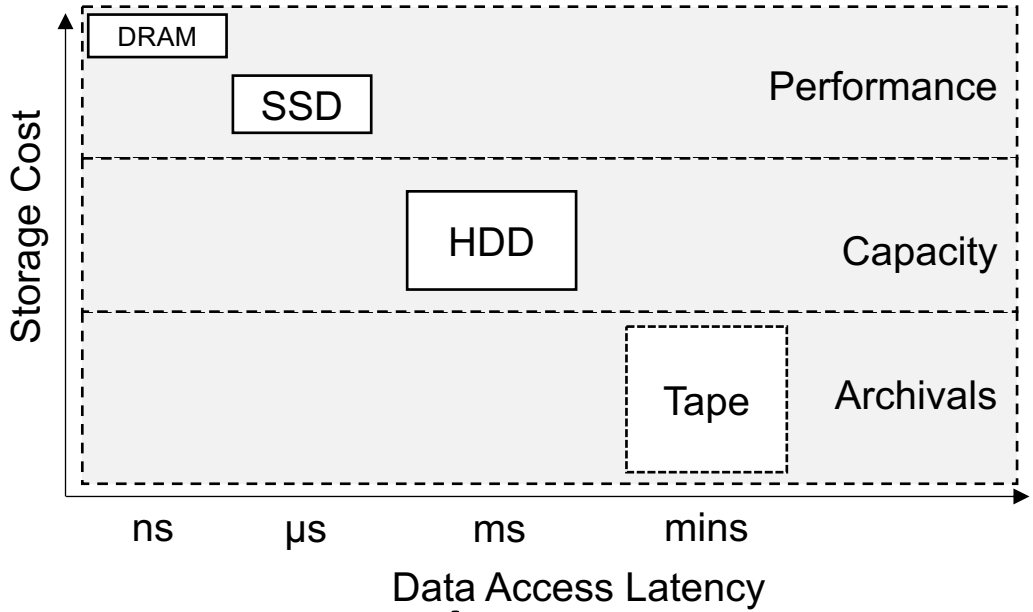
Using Synthetic DNA to Store and Process Data

Growth of Archival Data

"50% of 175ZB global datasphere will be enterprise data in 2025" *[IDC]*

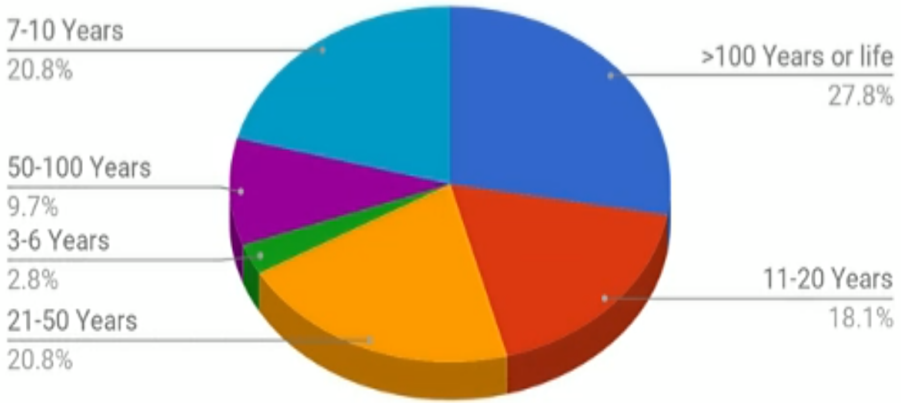
"80% enterprise data is cold, and increasing at 60% CAGR" [Horison]



Tape provides lowest cost/GB for data archives today

Problems with Tape (and Storage in general)

"60% of archival data stored longer than 20 years" *[SNIA]*



Tape lifetime is 10-20 years - continuous data migration

Net effect: Media Obsolescence

28 Apr 2017 | 15:00 GMT

The Lost Picture Show: Hollywood Archivists Can't Outpace Obsolescence

Studios invested heavily in magnetic-tape storage for film archiving but now struggle to keep up with the technology

By **Marty Perlmutter**

"There's going to be a large dead period," he told me, "from the late '90s through 2020, where most media will be lost."

Enterprise DBMS archives might soon face obsolescence

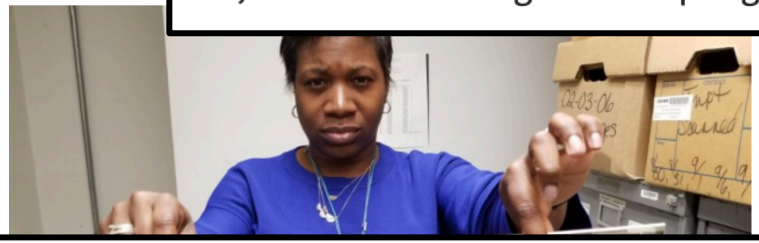
Microfilm to the Rescue…

Vinegar Syndrome Is Eating Away Cook County History

Kristen Schorsch
June 12, 2019

PLAY 4 MIN

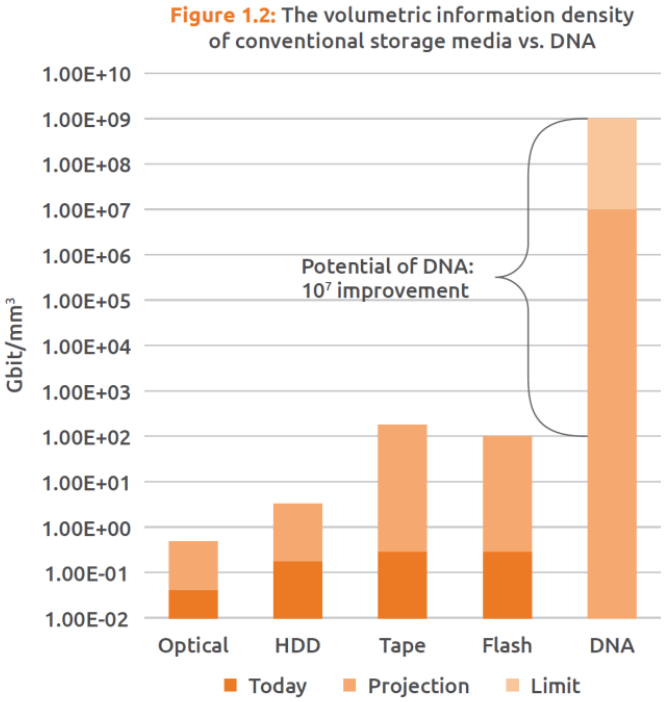
Potentially millions of records on microfilm have what's called vinegar syndrome. That happens when acetate film, like in old movie reels or government microfilm, is stored in hot, humid rooms. It gives off a pungent odor.



The vault's problem with vinegar syndrome is spreading. Gleffe said the oldest microfilm — potentially up to 35 million records from 1871 to 1959 — will likely have to be destroyed. A small percentage of the 24 million images that cover the following 25 years or so likely will have to go too.

Why DNA?

Dense



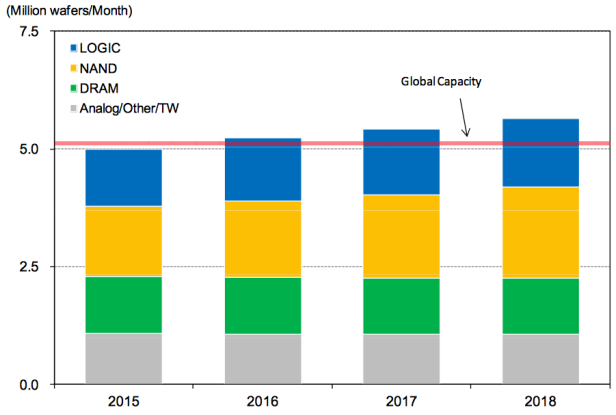
Durable

Woolly mammoth on verge of resurrection, scientists reveal

Scientist leading 'de-extinction' effort says Harvard team could create hybrid mammoth-elephant embryo in two years

Why DNA?

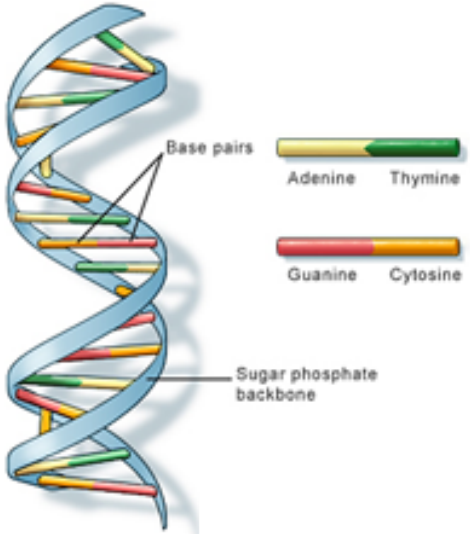
Scarcity of Silicon Supply

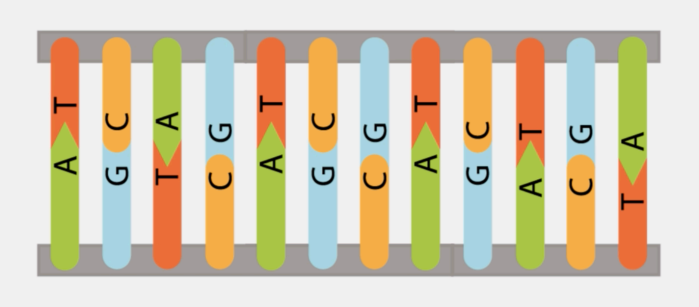


Parallelism

DNA

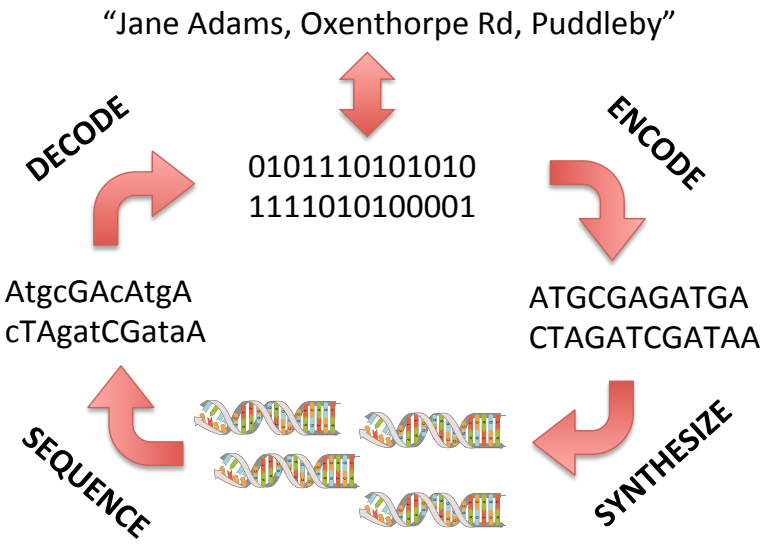
- Carrier of genetic instructions
- Double, long chain of molecules called nucleotides
- Four different nucleotides: A, T, C & G
- Complementarity (A & T, C & G) provides stability



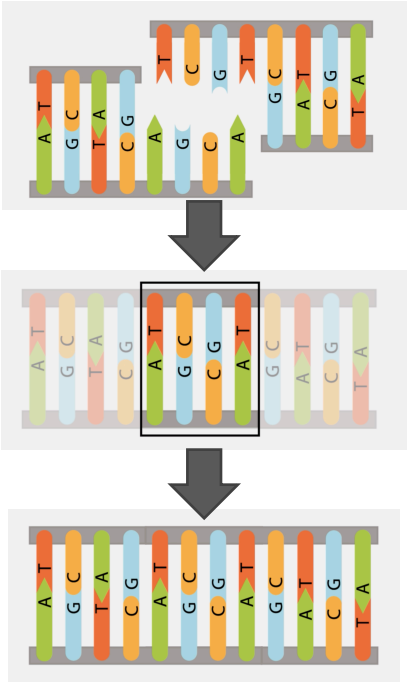


Storing & Processing Information in DNA

Storage

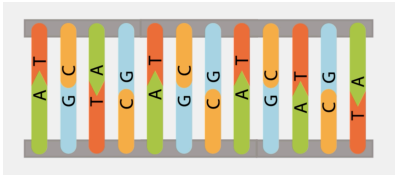


Processing



Encoding Information in DNA

Translate arbitrary data to nucleotides (A, T, C, G)

010101110101 ➡ 

Goal: compact code! Minimize the number of nucleotides per bit. E.g.:

00	→	A
01	→	T
10	→	C
11	→	G

010101110110

T T T G T C

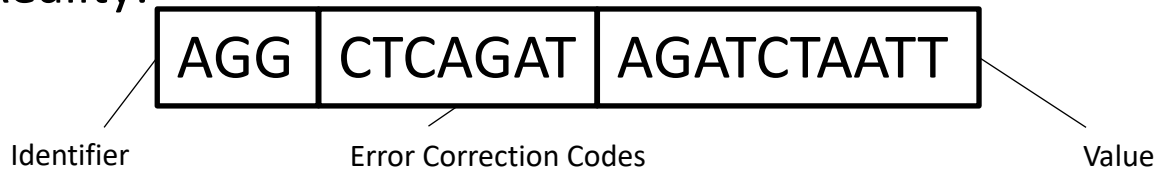
Encoding Information in DNA

Theory: 0.5 nucleotides per bit

Challenges:

- Biological constraints
- Error-prone synthesis & sequencing

Reality:



Typically around 1 nucleotide per bit

Information Processing in DNA

Exploit chemical processes:

- Annealing of complementary nucleotides
- Polymerase chain reaction (PCR) to replicate/amplify DNA sequences
- Loop-mediated isothermal amplification

Purposes:

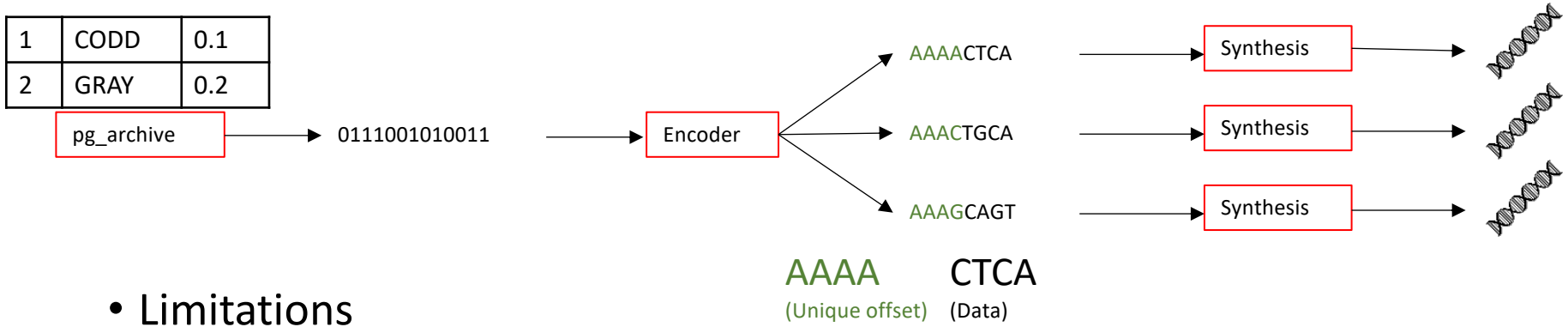
- Content detection
- Content retrieval through amplification:

AGG…ATACTCAGA…TAGATCTAATT…TGC

- Solving combinatorial problems

Writing Data to DNA (1): The Unstructured Way

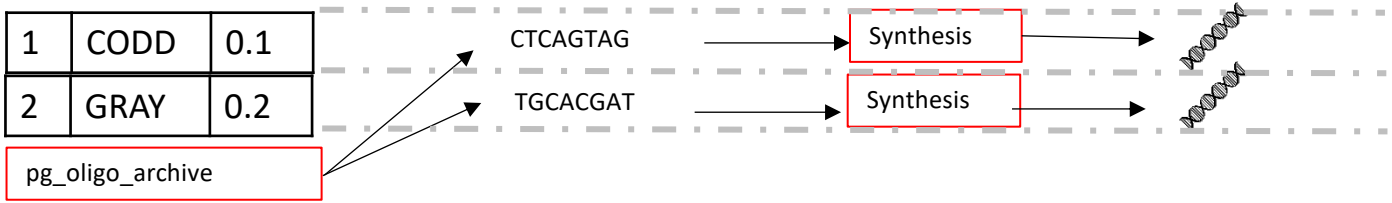
- Issues using DNA as a storage media
  - Limited DNA(oligo) length, homopolymer/G-C constraint, indels/subst. errors
- Approach-1: Dump database to a binary archive file and encode



- Limitations
  - $\log_4$(#segments) nucleotides reserved for offset (1TB => 17 nts in 150nt oligo)
  - No point queries supported
  - Cannot perform near-molecule data processing

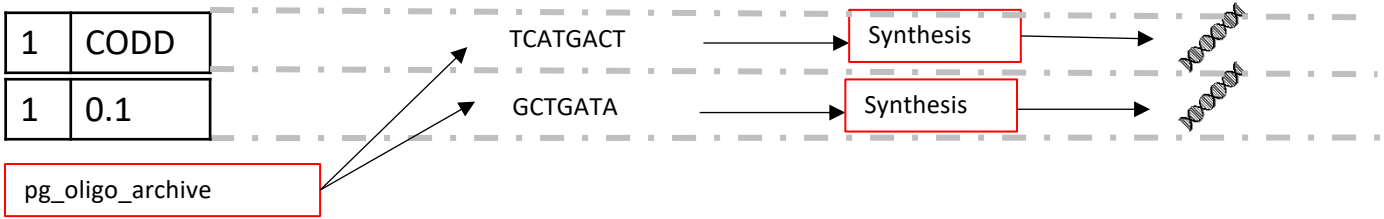
Writing Data to DNA (2): Structured Data Layout

- NSM on DNA: one row per oligo



Use unique primary key to avoid additional indexing

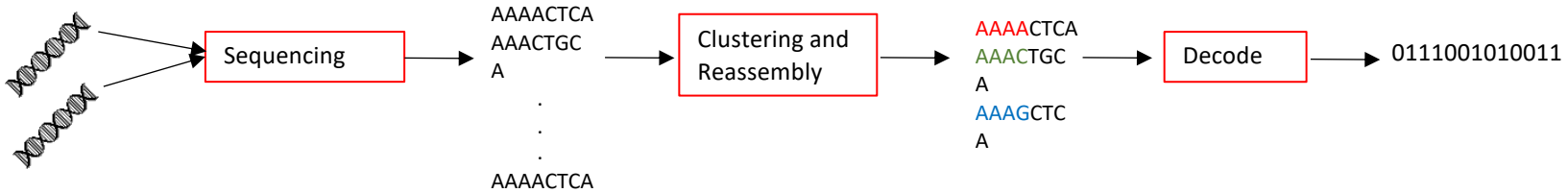
- DSM on DNA: columnset partitioning for “large” rows



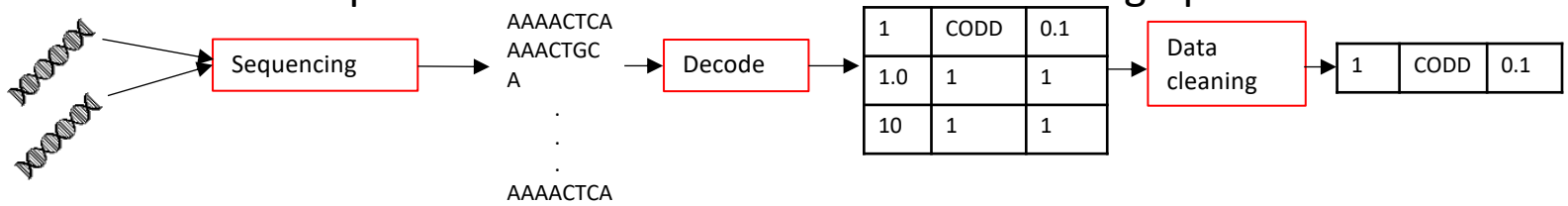
Reduces overhead from $\log_4$(#segments) to $\log_4$(cardinality)
(#segs. >> Card.)

Reading Data from DNA: Data Cleaning

- Read path for restoring unstructured data



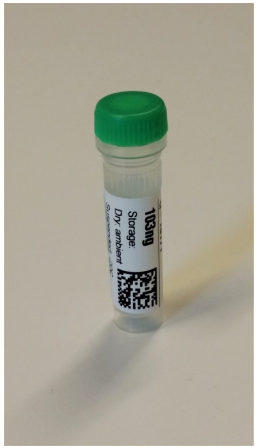
  - Clustering and reassembly time-consuming, necessary step before decoding
- But, our approach performs structure preserving encoding
  - Can map DNA read restoration to a data cleaning operation



Our approach uses schema information to restore data

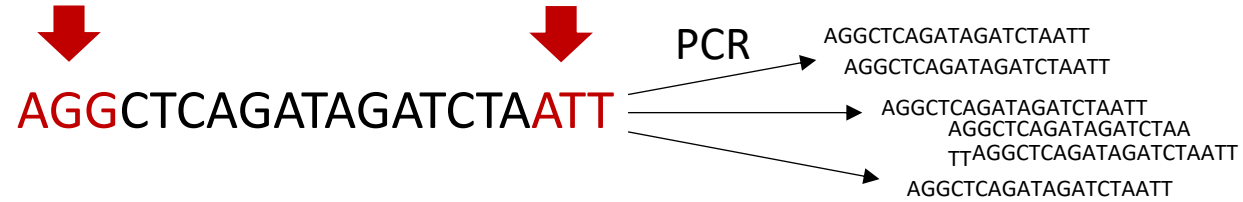
Evaluation: DNA Archival and Restoration

- PostgreSQL TPC-H SF-10^{-5}
  - 36 records across 8 tables, size 12KB
- pg_oligo_archive to archive database to DNA
  - 404, 150nt oligos synthesized with Twist Bioscience
- Sequencing with Illumina NextSeq 500
  - Deep sequencing provided very high coverage
- pg_oligo_restore performed automated restoration

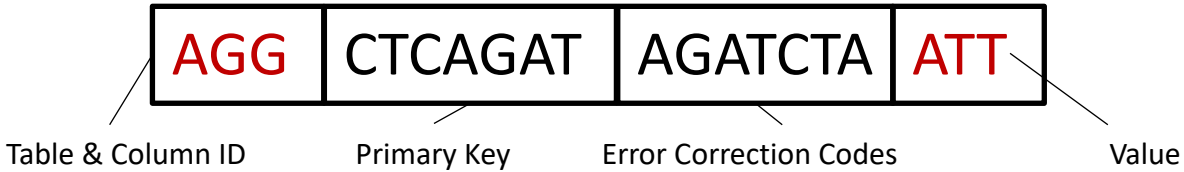


Near-molecule Query Processing: Selection

- Selection: find an oligo encoding an attribute with a particular value
- Key technique: Polymerase Chain Reaction (PCR)
  - amplify, i.e., copy ‘matching’ oligo countless times
  - need to know start and end sequences of matching oligo



  - Encoding, i.e., mapping column to oligo:

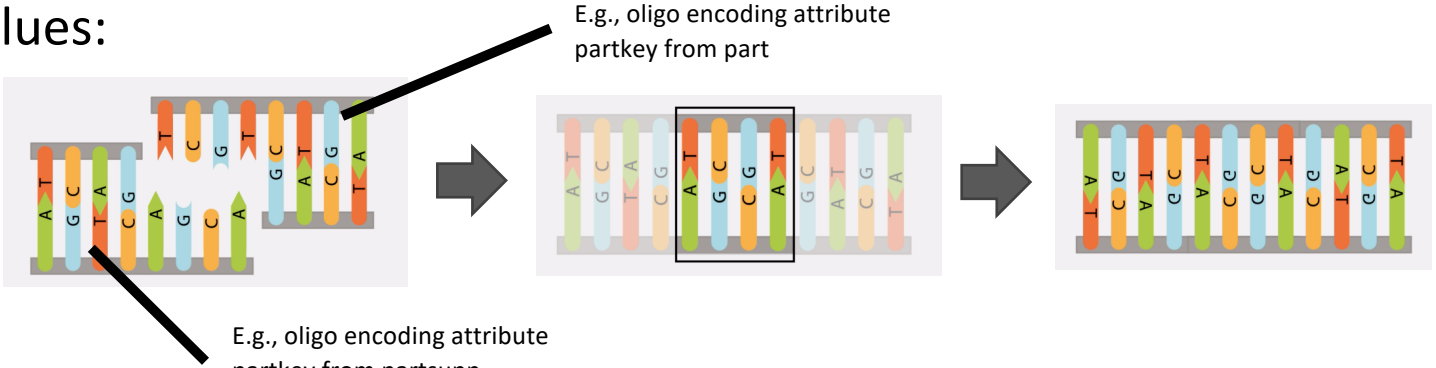


- Sequence using nanopore sequencing (Oxford Nanopore)

Near-molecule Query Processing: Join (1/2)

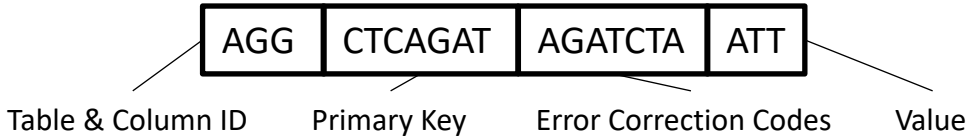
- Goal: join records/attributes with equal value
- Key technique: annealing of complementary single stranded oligos
- Complementarity – matching base pairs:

- Matching records/attributes have complementary encodings of the values:



Near-molecule Query Processing: Join (2/2)

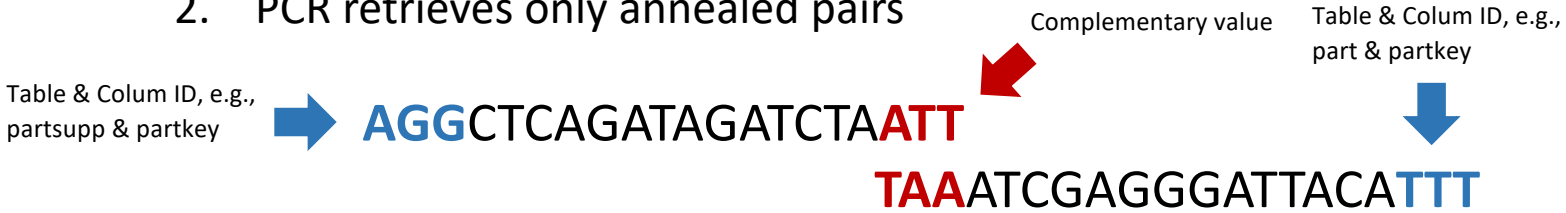
- Each attribute encoded as before:



but an additional reversed oligo with the value complemented:

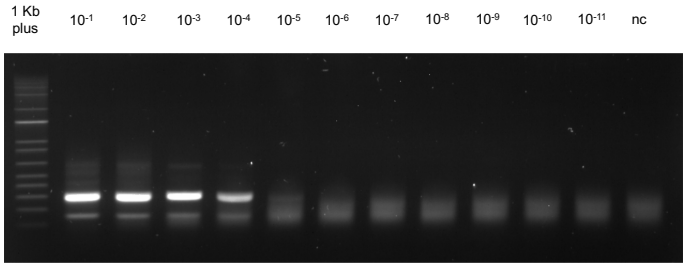
TAA AGATCTA CTCAGAT AGG

- Process:
  1. Annealing binds together matching/equal attributes
  2. PCR retrieves only annealed pairs



Evaluation: Near-molecule Query Processing (Join)

- Proof of concept experiment
- Encode matching records (only value attribute) from the TPC-H part and partsupp tables using two oligos
- Perform join between the matching records in increasing background of random oligos
- Gel electrophoresis after PCR
- Nanopore sequencing to retrieve resulting annealed oligos



OligoArchive

OligoArchive

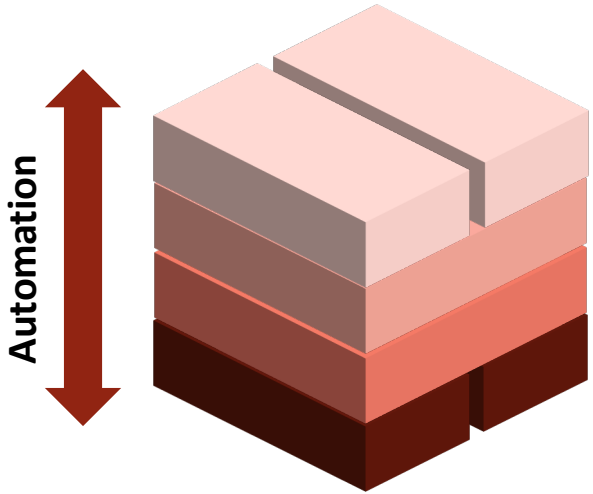
OligoArchive (https://oligoarchive.eu) is a €3M European Commission funded research effort to deliver the building blocks need to make DNA storage a reality. It involves six partners across four countries.

Research will be carried out a long the following directions:

- Near-molecule data analysis for content detection directly in DNA storage
- Efficient encoding for arbitrary data in general and structured (database) and unstructured (images) data
- Accelerated sequencing and thus reading for DNA storage
- Novel, cost-effective synthesis technology for DNA storage
- End-to-end automation, storing and reading arbitrary data from DNA storage through robotic equipment

Overview

The project will implement all the layers of the storage stack along with automation:



Application Layer
Encoding structured (database) and unstructured (imaging) data

OS Layer
File system abstraction

Controller Layer
Data processing capabilities

Media Layer
Synthesis and Sequencing

Trends & Outlook

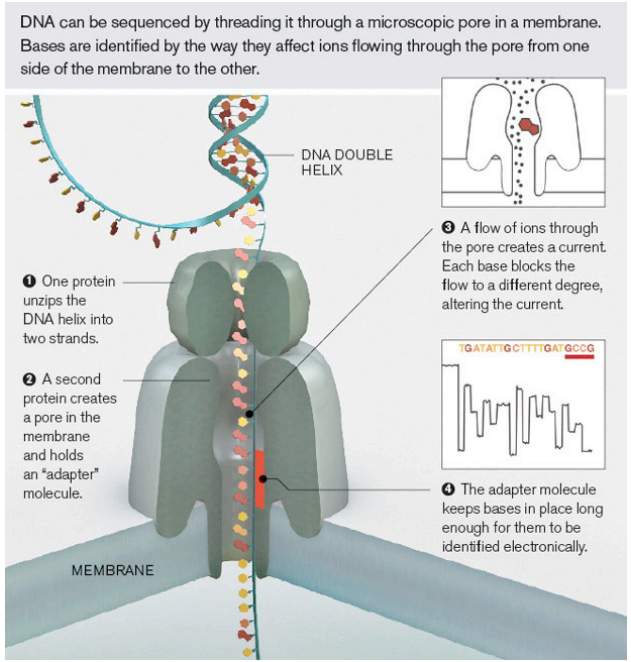
Sequencing Breakthrough

Oxford Nanopore became the first company to provide a commercially available nanopore sequencer in 2015 (available to community in 2012)

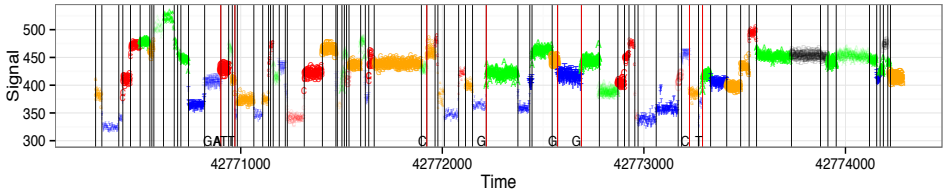
Nanopore is a disruptive technology:

- Sequencer Size
- Read Length

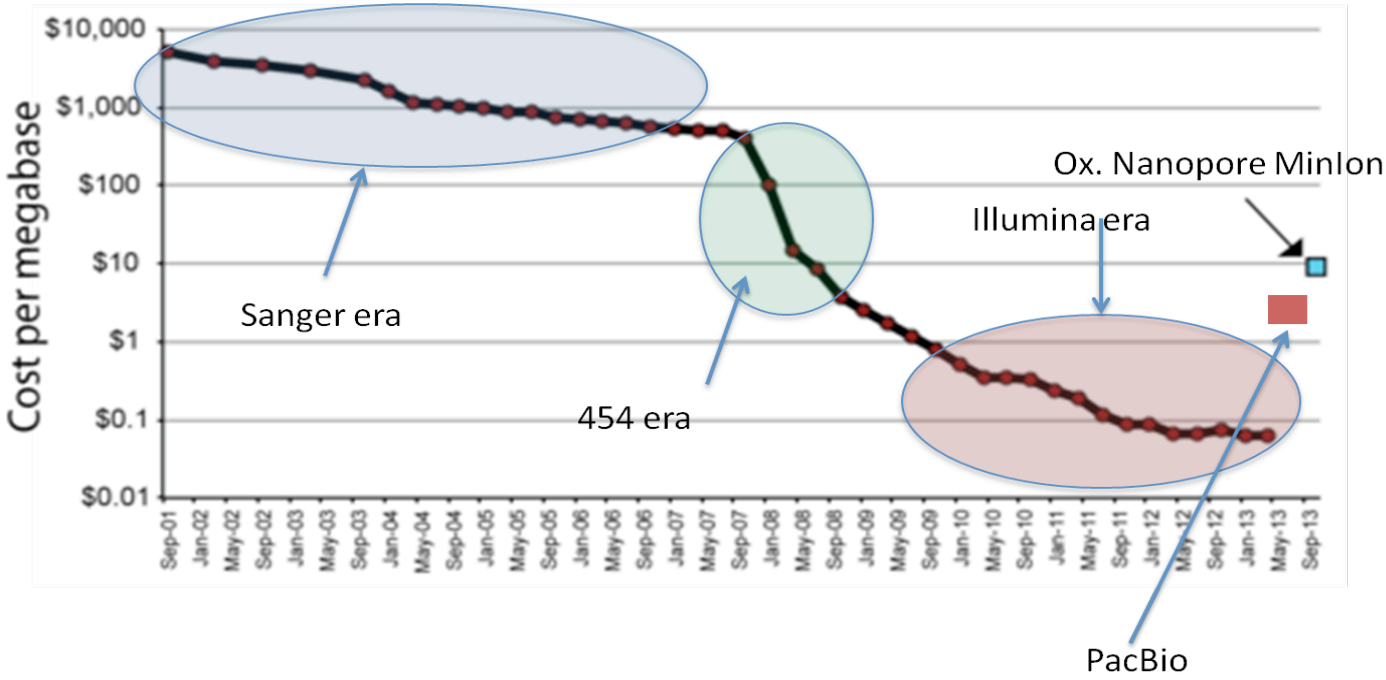
(Oxford) Nanopore Sequencing



Determine the sequence of DNA fragments by passing DNA through a protein (or other) pore in a membrane

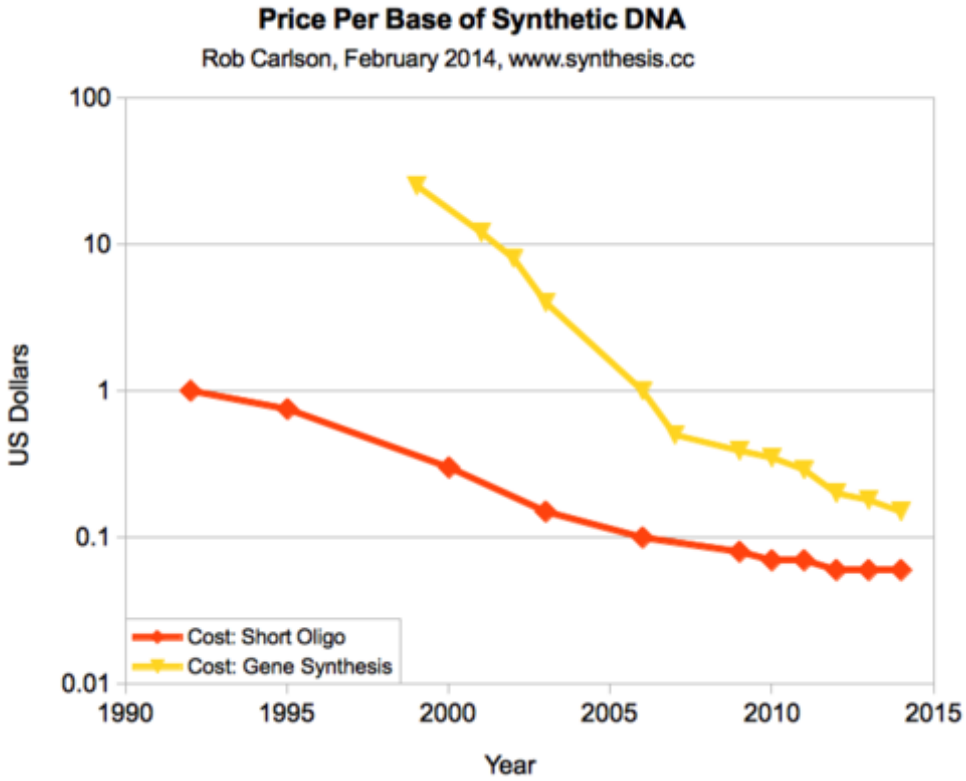


Trends Sequencing Cost



Solid-state nanopores will further lower cost of sequencing

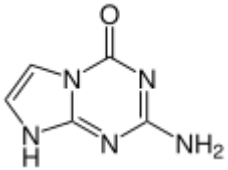
Synthesis Trends

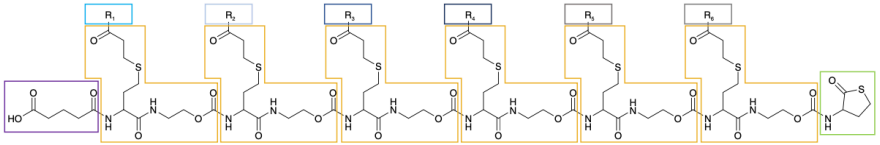


Synthesis Cost Major Roadblock

Novel synthesis techniques for DNA Storage:

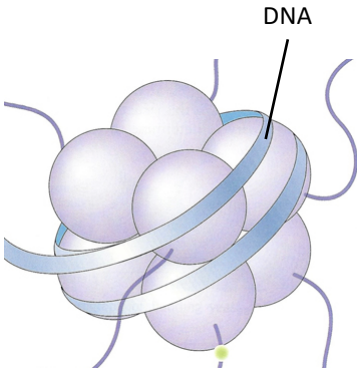
- Tolerance for errors in DNA storage
- Longer sequences
- Longer alphabets (Hachimoji synthesis: P, Z, B, S)
- Synthetic molecules





Current Research Directions

- Encoding
  - Compact error correction codes for synthesis & sequencing technology
- Synthesis Cost:
  - 3D structures
  - Imprecise synthesis
  - Prefabricated sequences
- Synthetic macromolecules
- Automation:
  - Automate synthesis, storage and sequencing
- Packaging/storage: wrap long DNA sequences around synthetic histones



Conclusions

- Huge potential for long-term storage up to thousands of years
- Simple in-storage analysis efficiently possible
- Simple storage - no copying needed

- Clearly not available tomorrow
  - Economically viable in 2-10 years
  - Speed competitive earlier
- Other long-term storage technology also under development

More info on https://oligoarchive.eu