A leading institution at the heart of the digital society

EURECOM
Sophia Antipolis

Storing and reviving databases on synthetic DNA

Raja Appuswamy

EURECOM

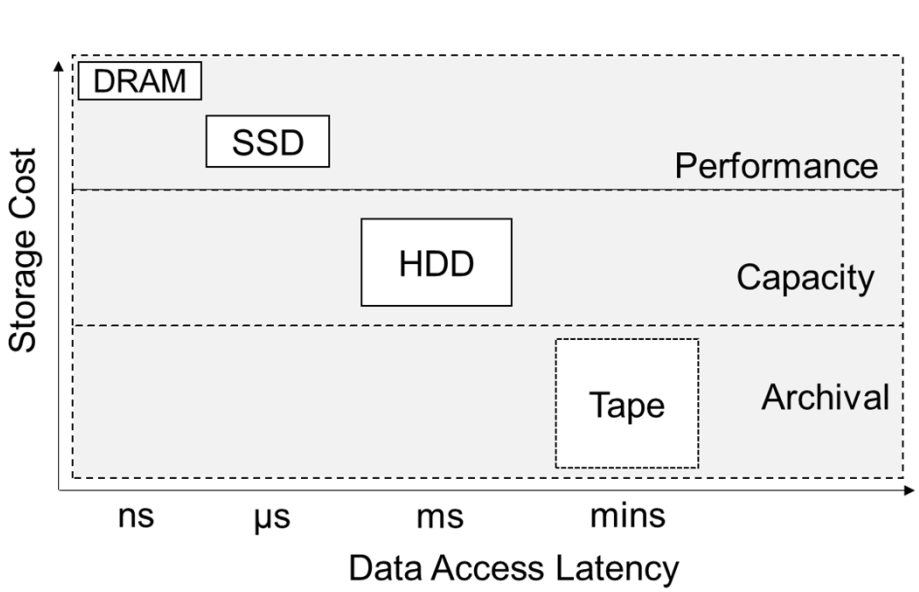
(in collaboration with members of EU FET OligoArchive and Danish National Archive)

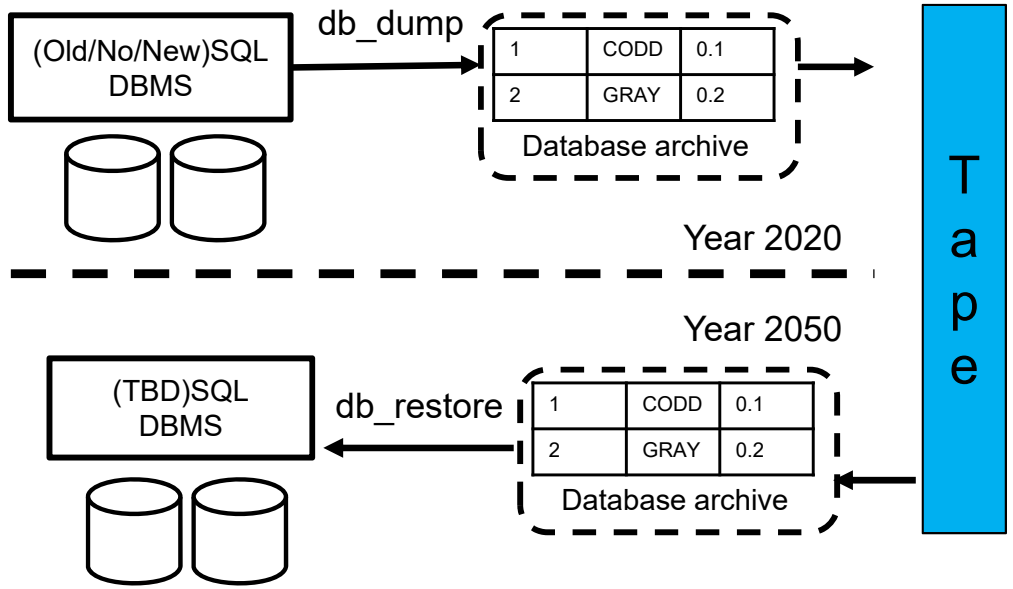
Cti
Commission des Titres d'Ingénieur

Growth of archival data

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

“80% data is cold, and increasing at 60% CAGR” [Horison]





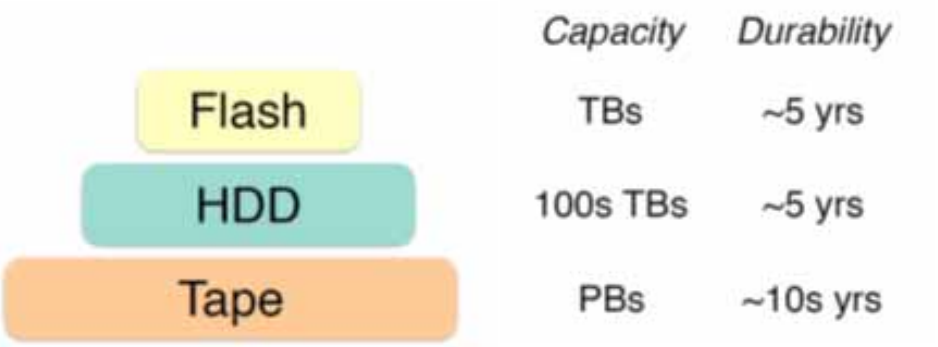
Current tape-based archival suffers from fundamental limitations

Continuous data migration with tape

“60% of archival data stored longer than 20 years”

[SNIA 100 Year Archive]

Media decay



Media obsolescence

	Tape Drives				
Version	LTO-6	LTO-5	LTO-4	LTO-3	LTO-2
LTO6	Read/Write				
LTO6 WORM	Read/Write				
LTO5	Read/Write	Read/Write			
LTO5 WORM	Read/Write	Read/Write			
LTO4	Read	Read/Write	Read/Write		
LTO4 WORM	Read	Read/Write	Read/Write		
LTO3		Read	Read/Write	Read/Write	
LTO3 WORM		Read	Read/Write	Read/Write	
LTO2			Read	Read/Write	Read/Write
LTO1				Read	Read/Write
Cleaning Tape	Supported	Supported	Supported	Supported	Supported

Net effect: Media Obsolescence

28 Apr 2017 | 15:00 GMT

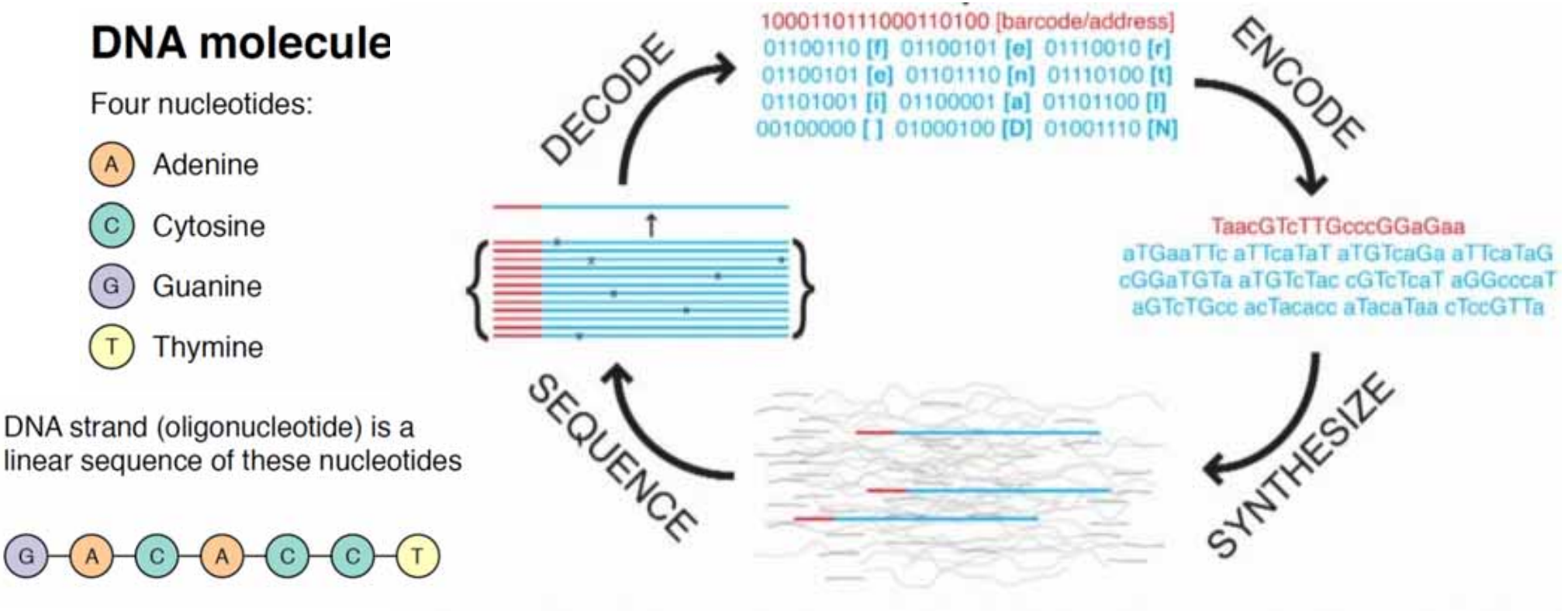
The Lost Picture Show: Hollywood Archivists Can't Outpace Obsolescence

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

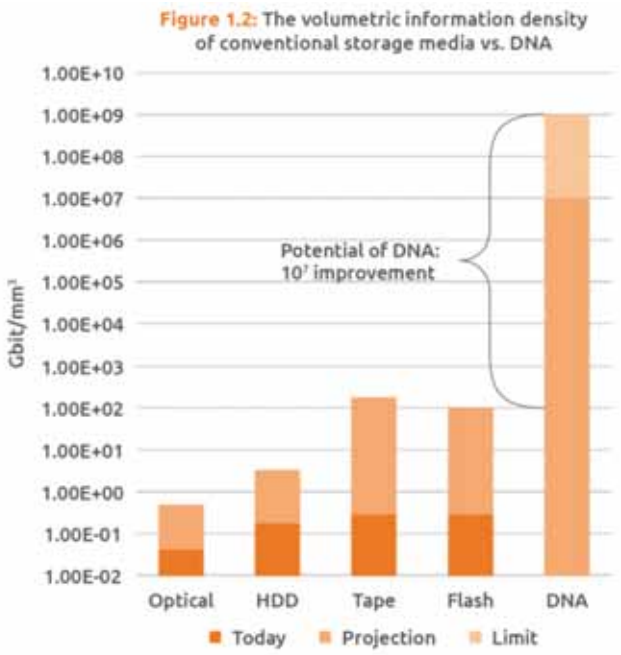
By **Marty Perlmutter**

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

DNA as a digital storage media



Why DNA

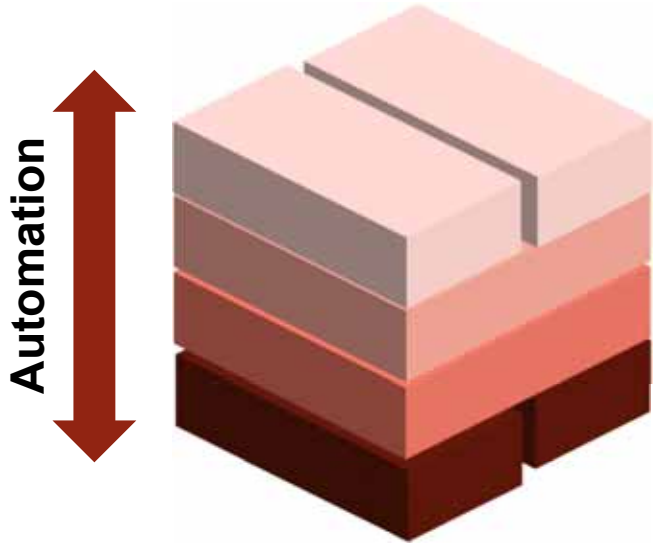



Woolly mammoth on verge of resurrection, scientists reveal

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

Project Oligoarchive focuses on using DNA as an intelligent storage medium

OLIGO ARCHIVE



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

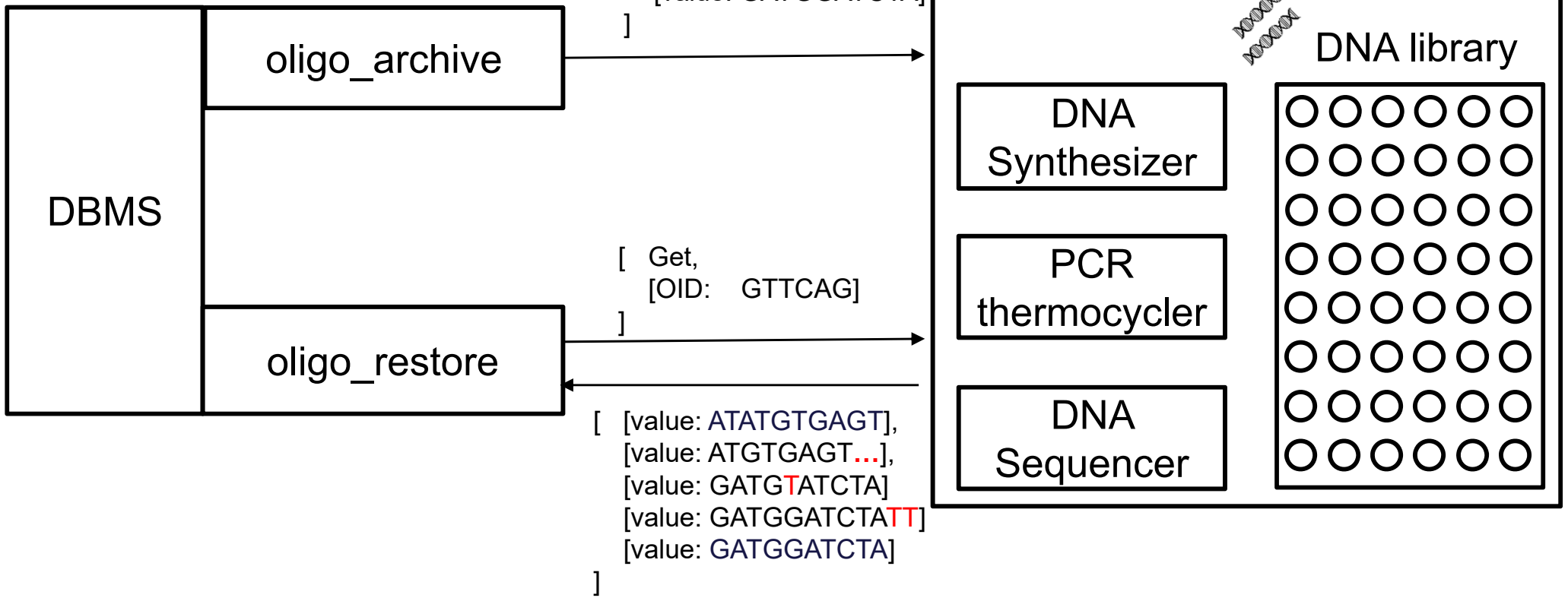
OS Layer
Advanced access paths (block, fs, …)

Controller Layer
Near-molecule query processing

Media Layer
Synthesis and Sequencing

Database archival/preservation with DNA

1	CODD	0.1
2	GRAY	0.2

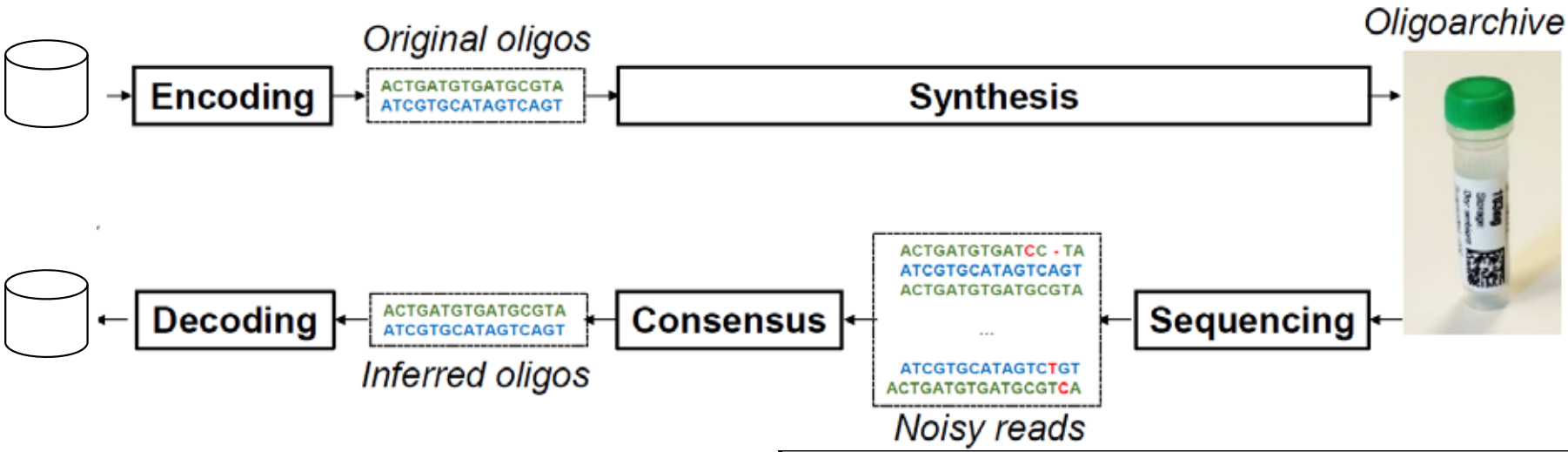


Challenges in using DNA as a storage medium

- **Conceptual mapping: 00-A, 01-T, 10-C, 11-G**
  - 2 bits per nucleotide
  - Cannot be used in practice
- **Each DNA is limited to a few hundred nucleotides**
  - For 300nt DNA strand, 600 bits can be stored
  - Data spread out across millions of DNA strands
- **Not all DNA are created equal**
  - G-C content limitations, homopolymers (AAAA)….
- **DNA has no addressing**
  - Need to reserve some nucleotides for ordering information
- **Synthesis (writing) and sequencing (reading) errors**
  - Biochemical steps introduce substitution, insertion, deletion errors

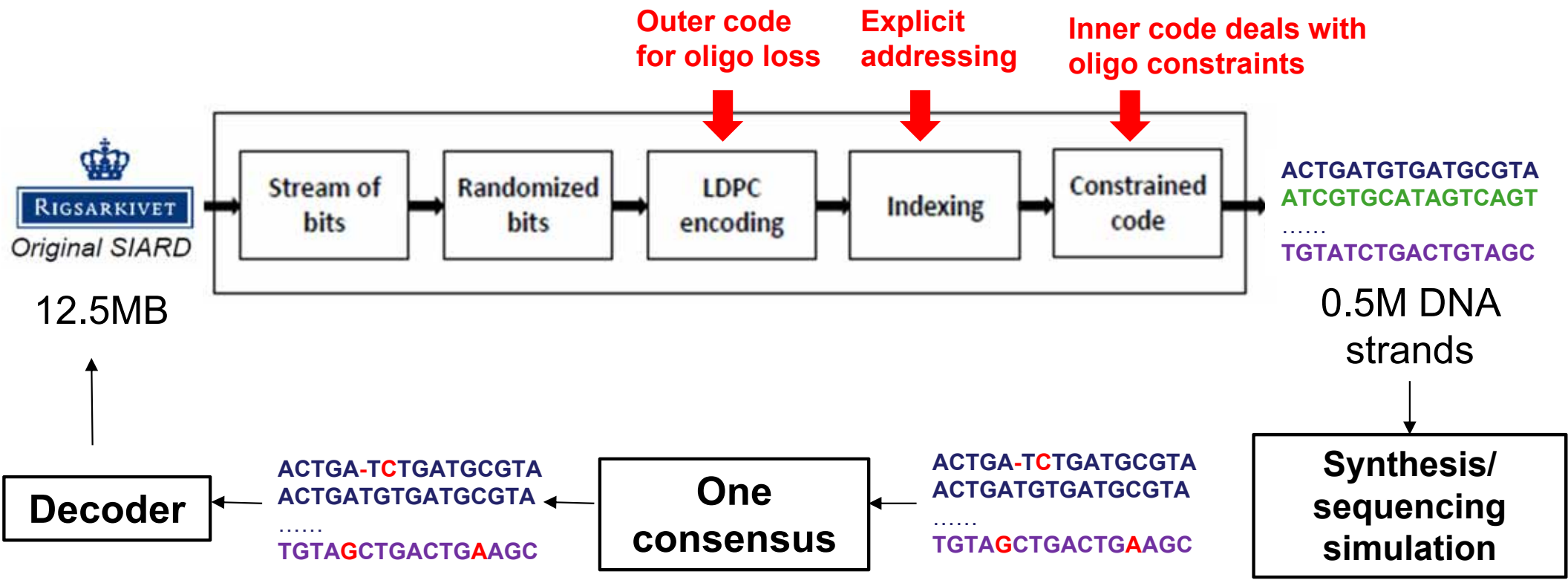
DNA archival & restoration: Challenges

- **Each DNA is limited to a few hundred nucleotides**
  - Data spread out across millions of DNA
- **Not all DNA are created equal**
  - G-C content limitations, homopolymers
- **DNA has no addressing**
  - Need to add ordering information in DNA



Biochemical errors
- substitution, insertions, deletions,
- Bias & duplication

DNA4DNA Collaboration: Synthetic DNA and the Danish National Archives



- **Third largest academic DNA storage experiment**
  - 200MB (Microsoft/UW), 22MB (Blawat et al.)
  - Storage density :1.73 bits/nt

Devil's Advocate: Open Problems

- **Price: DNA synthesis is 10^7 times more expensive than tape (10$/TB for tape vs 100M$/TB for DNA)**
  - Novel synthesis techniques under research
- **Automation, performance of synthesis and sequencing**
  - Synthesis/sequencing is labor intensive and slow
  - DNA throughput O(Kb/s) compared to tape's MB/s
- **DNA does not solve media/format obsolescence**
  - SIARD helps with format obsolescence
    - What about non-standard formats?
  - Who preserves the decoder (media obsolescence)?
    - Ongoing collab. with EUPALIA on emulation + DNA storage

Conclusion

- **Contemporary magnetic media suffers from limited lifetime**
  - Continuous migration complicates long-term archival
- **DNA provides a biological alternative**
  - Dense, durable, eternal relevance
- **DNA does not solve all problems**
  - Need cheap synthesis & scalable, end-to-end automation
  - Can synergistically combine {standards + emulation + DNA storage} for an end-to-end solution
- **Reach out for collaboration on database archival**