Imperial College London

15th September 2021

Encoding Information into DNA

MatBio 2021, King's College London

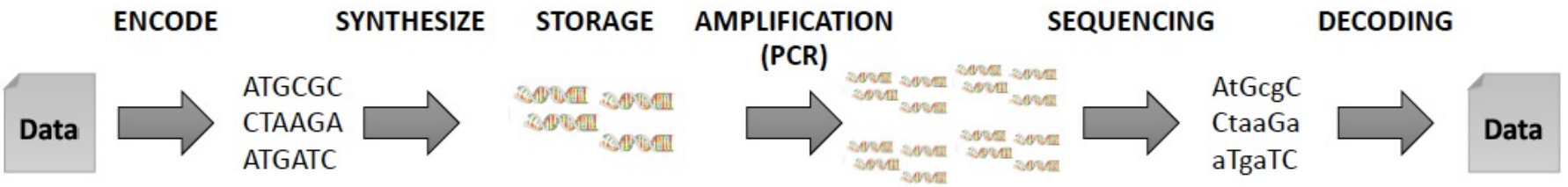
Dr. Jamie Alnasir (Post-doc.) & Dr. Thomas Heinis (P.I.)

Why Encode Information into DNA?

- **Yearly global production of data growing exponentially**
  – outpacing our ability to store this data
  – hardware and software obsolescence, storage medium decay
  – risk is that information currently stored on magnetic or optical media will not be recoverable in a century or less

- **Benefits of DNA**
  – theoretical possibility of storing 455 EB/g
  – un-treated, DNA has a half-life of approximately 520 years
  – when treated (desiccated, chilled), can be stable for thousands of years
  – currently could be used for high-latency storage, i.e. cold storage, archiving
  – numerous encodings for mapping data to DNA's Base 4 (A,T,G,C)
  – reagents (Adenine, Thymine, Guanosine and Cytosine) are readily available and abundant

What's involved in the end-to-end workflow?



Phase	
Writing	**Encoding the data** • apply **error-correction** codes • ensure sequence meets **biological constraints** **Synthesise the DNA**
Storage	**Store DNA in/on appropriate medium**
Reading	**Sequence the DNA** • raw sequence trace undergoes **base-calling** **Decoding the data** • **apply error detection, error correction** to recover any corrupted blocks

What are the considerations?

Constraint	Reason
Biological Constraints	**Homopolymers (i.e. runs of same base letter, i.e. GGGGG)** – difficult to synthesise and often recalcitrant to sequencing **GC extremes** – unstable and difficult to sequence
Error detection & Error Correction	**Synthesis and Sequencing** – both prone to errors i.e. insertions, deletions, substitutions Error detection alone insufficient, require **error correction**
Information Density	**Synthesis cost.** Metadata (**primers, error correction**) costs nucleotides and thus reduce information density
Data retrieval method	**Sequencing whole DNA pool** inefficient **vs Random access**
Storage medium	**In vivo (in cells)** – not ideal for storage (has been done) **Microplate / wells / immobilised** (more practical)

Early approaches and encodings

- **Mostly focused on the novelty, artistic projects**

Microvenus (1996) [1]	Encoded Germanic Rune for Life and the Female Earth (5 x 7, or 7 x 5 figure, totaling 35 bits). Stored in E. coli (transfected)
Genesis Project (1999) [2]	Encoded sentence from Genesis, using morse code mapping
Hiding data in DNA microdot (1999) [3]	Ternary code. Stored in microdot
Organic memory using DNA (2003) [4]	Ternary code. Stored in E. coli and extremophile Deinococcus Radiodurans

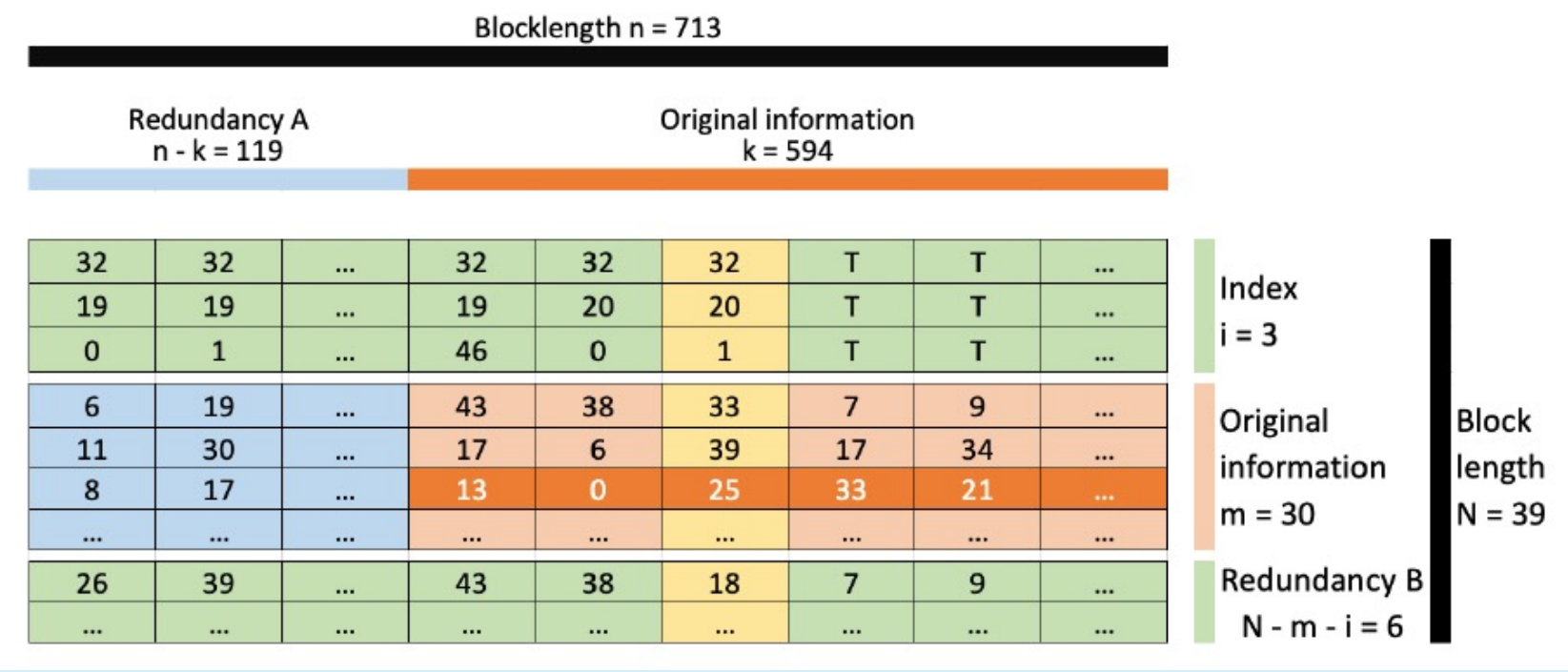
- **Used trivial encoding schemes**
  - **Morse code –** mapping Dash → T, Dot → C, and space → A.
  - **Phase-change encoding –** codes bit repeats until a change occurs
  - **Ternary → Quaternary mapping –** A → 0000 → AAAA, B → 0001 → AAAC, C → 0002 → AAAT, …
  - **Ignorant of biological constraints i.e. homopolymers, extreme GC**

Huffman, Comma and Alternating codes [5]

Huffman code	**Most frequent character is encoded with the least number of symbols**, and the least frequent character the most symbols, such that **e → T, z → CCCTG** • As the frequency decreases, the codeword increases in size • Varying codeword length makes it difficult to distinguish between artificial and natural sequences • No error-detection or error-correction is incorporated
Comma code	Uses **G nucleotide as a comma to separate all other codewords of length 5**. G is never used in any codeword • Proposed code uses G as the comma every six nucleotides encoding is easily identified as synthetic. 5 nucleotide codeword uses A,T,C but not G. Balances GC content, more efficient amplification process. • General **format of codeword is CBBBC where B in {A,T}, with B's C's in any permutation** (80 possible)
Alternating code	**64 codewords, six nucleotides long** • In the form **XYXYXY, where X in {A,G}, Y in {C,T}** • XYXYXY / YXYXYX avoids homopolymers of three • Encoding used has suboptimal GC content, could be improved using: X in {C,G}, Y in {A,T}

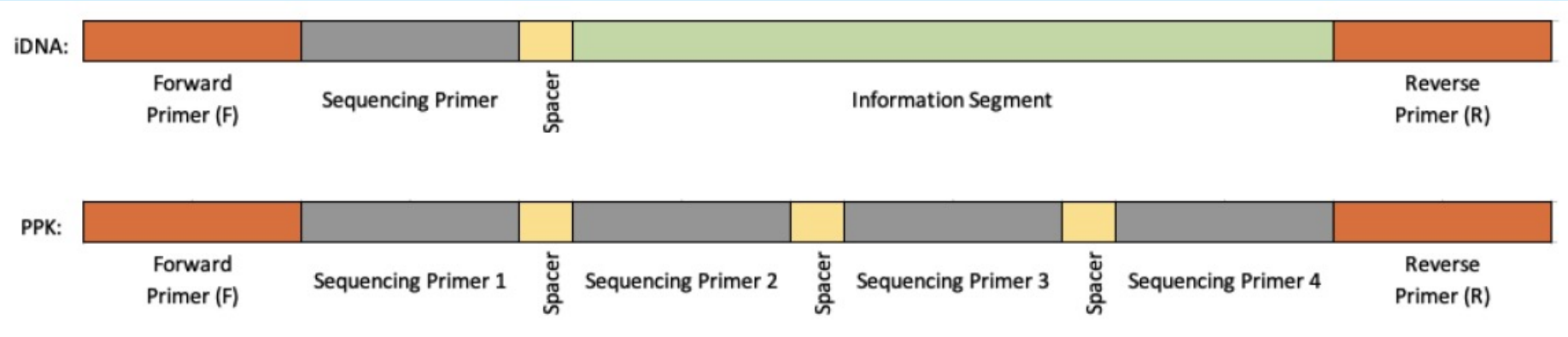
Reed-Solomon error-correcting codes [6]

- Most newer approaches implement Reed-Solomon codes
- An inner and outer code add error-checking/correction, redundancy
- Corrupted blocks can be recovered 😄



Random access using Primers [7]

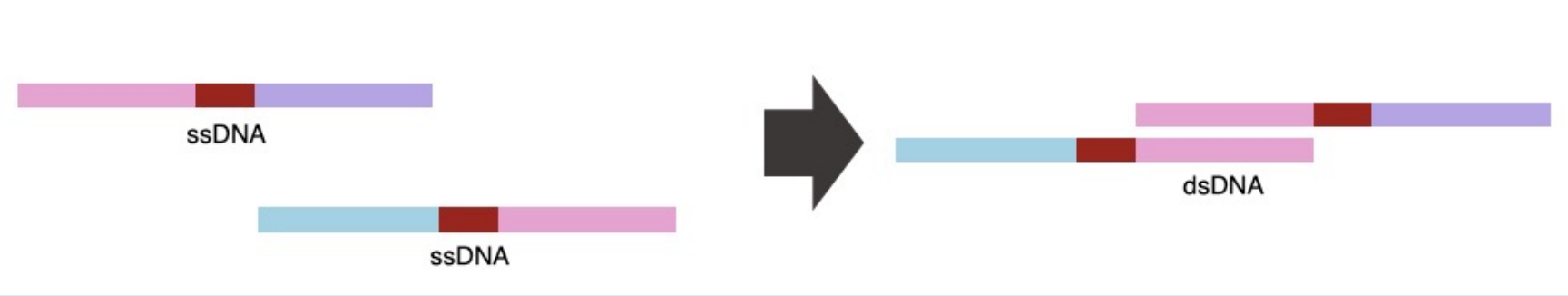
- Sequential access requires that the whole DNA pool of Oligonucleotides is read – this is inefficient and time consuming
- **Solution:** Use a separate Oligo (PPK – Poly-Primer Key) comprising multiple Primers to index blocks (sequences) on data-bearing Oligonucleotides (iDNA – information DNA)
- This method allows for a *subset* of the information (i.e. sequences in a subset of iDNAs) to be retrieved from the DNA Oligo pool



Inter-molecular computation (DB join example) [8]

- OligoArchive is an archival tier built on a Postgre SQL relational DB
- Performs schema-aware encoding & decoding of relational data
- Object Identifier (Primer) uniquely identifies subset of Oligos (similar to PPK and iDNA we saw earlier)
- *get* operation uses Primer to filter and extract all oligos for the Primer

- *Inter-molecular (in-vitro) join* operation leverages molecular biology
  - the annealing of two complementary single-stranded (ssDNA) sequences to form a double-stranded oligo (dsDNA)
  - resulting oligos are then sequenced to retrieve the data



Challenges and relevant areas to explore

- DNA Synthesis is still more expensive than sequencing
  - sequencing costs driven down by clinical diagnostics, R&D etc not so with synthesis
- Choice of Sequencing technology
  - Illumina vs Nanopore (very different profiles)
  - optimisation of base-calling using ML
- Standardisation of encoding
- Miniturisation of the synthesis/sequencing technologies
- Xeno nucleic acids (XNA) – synthetic (distinguishable from DNA)
- Biosecurity
  - prevention/flagging of generation of pathogenic sequences
    - requires a sequence check against a database

References

1. Joe Davis. 1996. Microvenus. Art Journal 55, 1 (1996). https://doi.org/10.1080/00043249.1996.10791743

2. Eduardo Kac. 1999. GENESIS. http://www.ekac.org/geninfo.html. (1999)

3. Catherine Taylor Clelland, Viviana Risca, and Carter Bancroft. 1999. Hiding Messages in DNA Microdots. Nature 399, 6736 (1999). https://doi.org/10.1038/21092

4. Pak Chung Wong, Kwong-kwok Wong, and Harlan Foote. 2003. Organic Data Memory Using the DNA Approach. Commun. ACM 46, 1 (Jan. 2003), 4. https://doi.org/10.1145/602421.602426

5. Geoff C. Smith, Ceridwyn C. Fiddes, Jonathan P. Hawkins, and Jonathan P.L. Cox. 2003. Some Possible Codes for Encrypting Data in DNA. Biotechnology Letters 25, 14 (01 Jul 2003). https://doi.org/10.1023/A:1024539608706

6. Robert N. Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, andWendelin J. Stark. 2015. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. Angewandte Chemie International Edition 54, 8 (2015). https://doi.org/10.1002/anie.201411378

7. Carter Bancroft, Timothy Bowler, Brian Bloom, and Catherine Taylor Clelland. 2001. Long-Term Storage of Information in DNA. Science 293, 5536 (2001). https://doi.org/10.1126/science.293.5536.1763c

8. Raja Appuswamy, Kevin Le Brigand, Pascal Barbry, Marc Antonini, Olivier Madderson, Paul Freemont, James McDonald, and Thomas Heinis. 2019. OligoArchive: Using DNA in the DBMS Storage Hierarchy. In Proceedings of the 9th Biennial Conference on Innovative Data Systems Research CIDR. http://cidrdb.org/cidr2019/papers/p98-appuswamy-cidr19.pdf

DNA Storage Team @ SCALE Lab

Thomas Heinis

Lab Director

Jamie J. Alnasir

Research Associate

Omer S. Sella

Research Assistant

William Hunter

Doctorate Research

Jasmine Quah

M.Sc. Student

Chandler Low

M.Sc. Student

Samantha Kwok

M.Sc. Student

We are open to collaboration

Data Science Department, Eurecom, Biot, France; Université Côte d'Azur & CNRS, Institut de Pharmacologie Moléculaire et Cellulaire; Université Côte d'Azur & I3S, France; Department of Medicine and Department of Computing, Imperial College London, London, UK