EMBL-EBI

UNIV. OF CAMBRIDGE
Dept. of Engineering
SigProC Lab

Codes for efficient data storage on DNA molecules

Jossy Sayir

Information, Inference and Entropy Symposium, March 2016

Your mission: save the world!

- Mankind is self-destructing
- We will lose all technology and knowledge
- Archive our information so it remains accessible when all current technology is lost

Your mission: save the world!

- Mankind is self-destructing
- We will lose all technology and knowledge
- Archive our information so it remains accessible when all current technology is lost

My partners

David MacKay

Nick Goldman

Emily Hesketh

Roland Schwarz

Ewan Birney

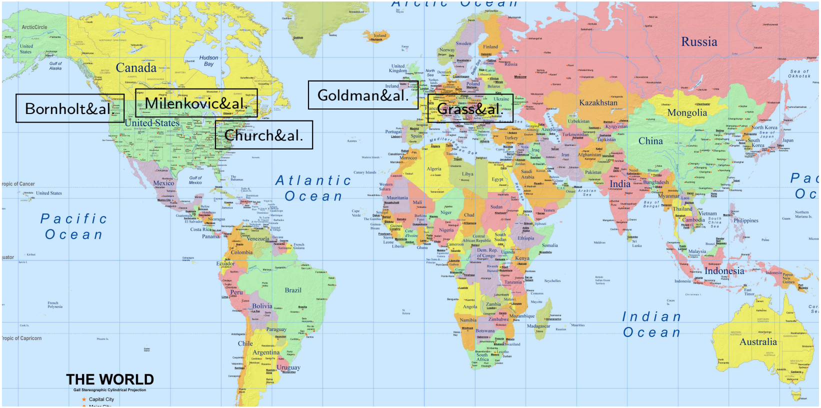
DNA Storage

When the planet finally emerges from the dark ages, whatever intelligent life from develops will eventually re-discover DNA and learn how to read it.

Store information on DNA!

DNA Storage

When the planet finally emerges from the dark ages, whatever intelligent life from develops will eventually re-discover DNA and learn how to read it.
Store information on DNA!



Bornholt&al.

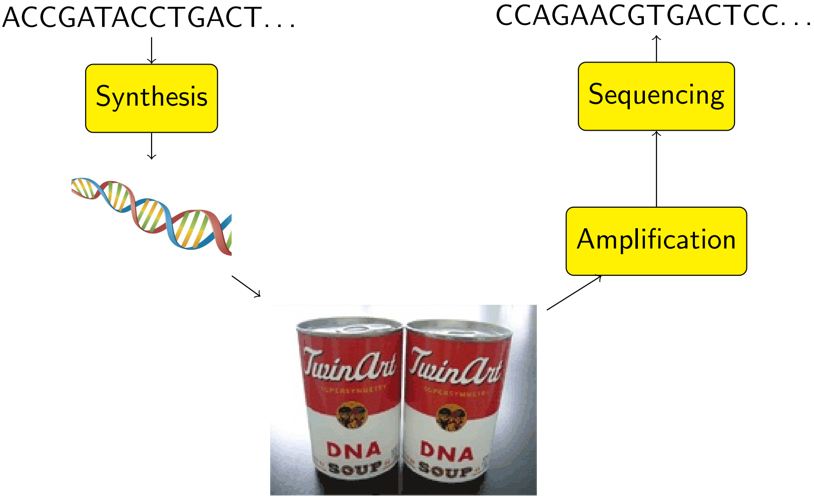
Milenkovic&al.

Goldman&al.

Grass&al.

Church&al.

DNA Storage



ACCGATACCTGACT. . .

Synthesis

Sequencing

Amplification

CCAGAACGTGACTCC. . .

DNA Storage

- Synthesis: order from a handful of companies, slow, expensive
- Amplification: any lab can do it, cheap
- Sequencing: specialised lab equipment, getting cheaper

DNA Storage

- Synthesis: order from a handful of companies, slow, expensive
- Amplification: any lab can do it, cheap
- Sequencing: specialised lab equipment, getting cheaper

Past work:

- Feasibility studies (order of kiloBytes)
- Grass&al. glass coated DNA tested robustness (10,000 years)

DNA Storage

- Synthesis: order from a handful of companies, slow, expensive
- Amplification: any lab can do it, cheap
- Sequencing: specialised lab equipment, getting cheaper

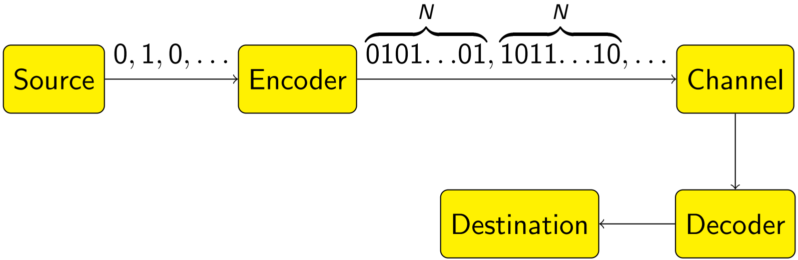
Past work:

- Feasibility studies (order of kiloBytes)
- Grass&al. glass coated DNA tested robustness (10,000 years)

Current and future work:

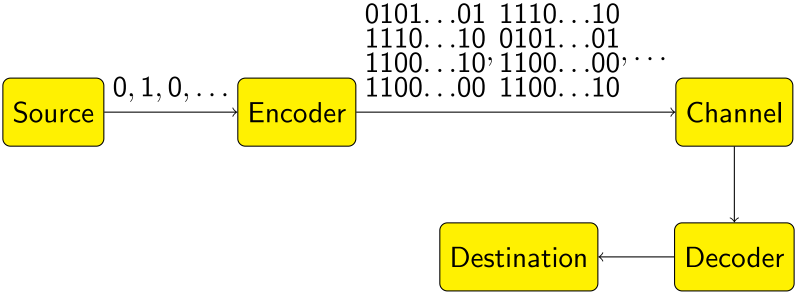
- Probabilistic characterisation of the storage channel
- Dedicated coding techniques
- Optimised data rates (kbit per \$) and reliability (10^{-n} error probabilities for $n = 6, 7, 8, \ldots$)

“Normal” symbol coding


Source
$0, 1, 0, \ldots$
Encoder
$\overbrace{0101\ldots01}^{N}, \overbrace{1011\ldots10}^{N}, \ldots$
Channel
Decoder
Destination

$$\boxed{N \gtrsim 10,000}$$

Packet coding

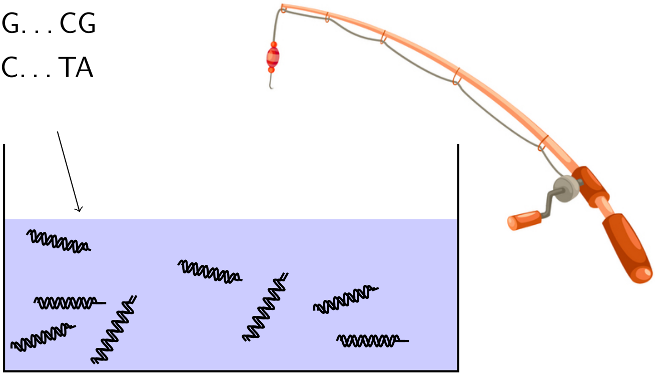

Source
$0, 1, 0, \ldots$
Encoder
$\begin{matrix}0101\ldots01\\1110\ldots10\\1100\ldots10\\1100\ldots00\end{matrix}, \begin{matrix}1110\ldots10\\0101\ldots01\\1100\ldots00\\1100\ldots10\end{matrix}, \ldots$
Channel
Decoder
Destination

$$\boxed{\text{Packet size} \simeq 8 - 10,000, \ \text{Codeword length} \simeq 100 - 10,000}$$

The DNA Soup Channel

$$\overset{N \simeq 200}{\longleftrightarrow}$$

1. ACGCA. . . AT
2. GGACT. . . TG
3. ATCTG. . . GA
4. TTACG. . . CG
5. GCTAC. . . TA
6. . . .



The DNA Soup Channel

- DNA is quaternary
- Synthesis/sequencing constraints dictate DNA strand lengths in the 100s
- Too short for proper coding
- Difficult but feasible size for “packet” coding
- But:

- Packet order is lost in the soup
- Identical packets (out of order) are indistinguishable

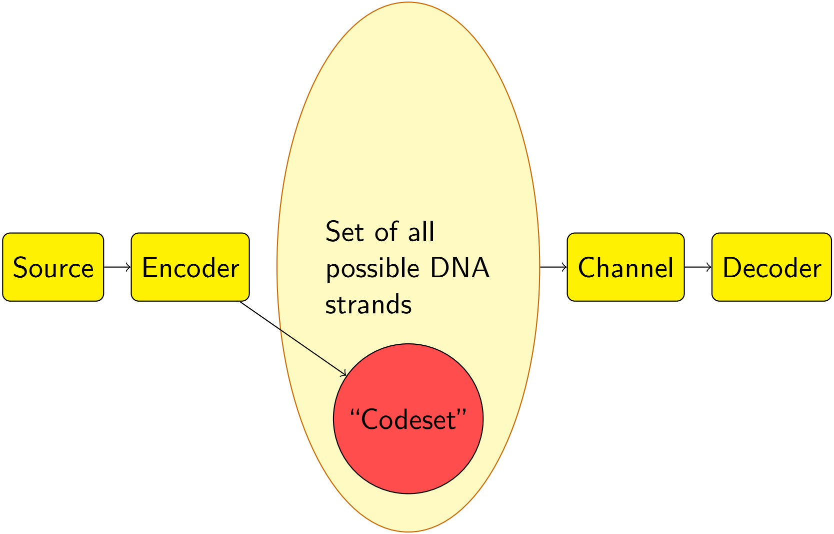
The DNA Soup Channel

- DNA is quaternary
- Synthesis/sequencing constraints dictate DNA strand lengths in the 100s
- Too short for proper coding
- Difficult but feasible size for “packet” coding
- But:

- Packet order is lost in the soup
- Identical packets (out of order) are indistinguishable

What are the theoretical limits for storage in the DNA soup channel?

Subset coding à la MacKay



Subset coding à la MacKay

- The decoder obtains (possibly repeated) noisy observations of the elements in the codeset.
- Its role is to determine which codeset the encoder selected.
- The “codesets” take up the role of “codewords” in traditional coding.

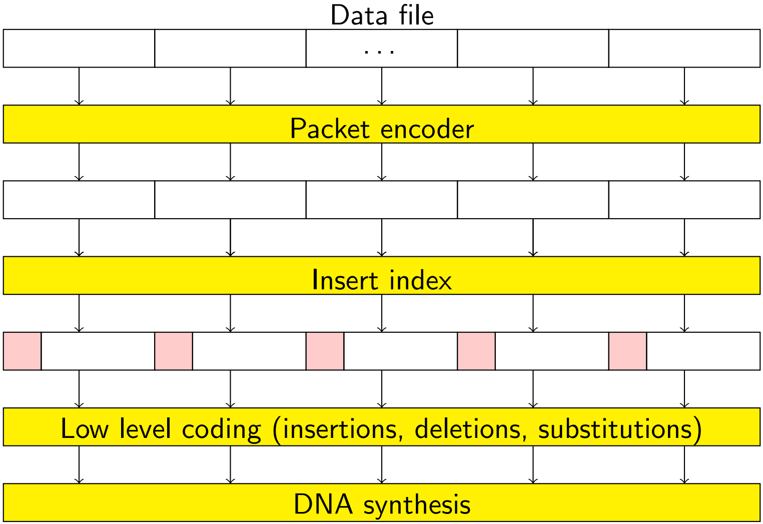
Subset coding à la MacKay

- The decoder obtains (possibly repeated) noisy observations of the elements in the codeset.
- Its role is to determine which codeset the encoder selected.
- The "codesets" take up the role of "codewords" in traditional coding.

For the noiseless channel, optimal construction:

- Prefix of DNA packet runs through an index sequence 0,1,2,3,. . .
- Remaining portion of DNA packet determined by traditional encoder where the index maps the position of the symbol in the codeword
- This is equivalent to fountain coding, indexed Reed Solomon coding and indexed random linear coding.

Current system



Current system

- “Noisy fountain coding” through low-level high rate intra-packet coding (similar to [Venkiah, Poulliat Declercq] papers
- The “index” portion needs perfect protection or the system fails

This is not optimal and the “graal” of noisy DNA soup channel coding would be to invent a true codeset coding system.

Current work

- Channel measurement and estimation
- Evaluate low-level codes (Marker codes [Ratzer&MacKay], watermark codes [Davey&MacKay], convolutional codes)
- Evaluate packet encoders (Fountain, RS codes)
- Unequal error protection for the index
- Direct codeset coding for the noisy DNA soup channel

Ethical questions

- If "Mankind version 1.0" is so terrible, should we store our knowledge at all?
- Will future intelligent life know about Reed Solomon codes?
- Who says we are version 1.0?