DNA data storage using photolithographic synthesis and error correction

Reinhard Heckel

Technical University of Munich & Rice University

April 9, 2021

Joint work with:

R. Grass, P. L. Antkowiak, J. Lietard, M. Z. Darestani, M. M. Somoza, W. J. Stark

Storing information in DNA

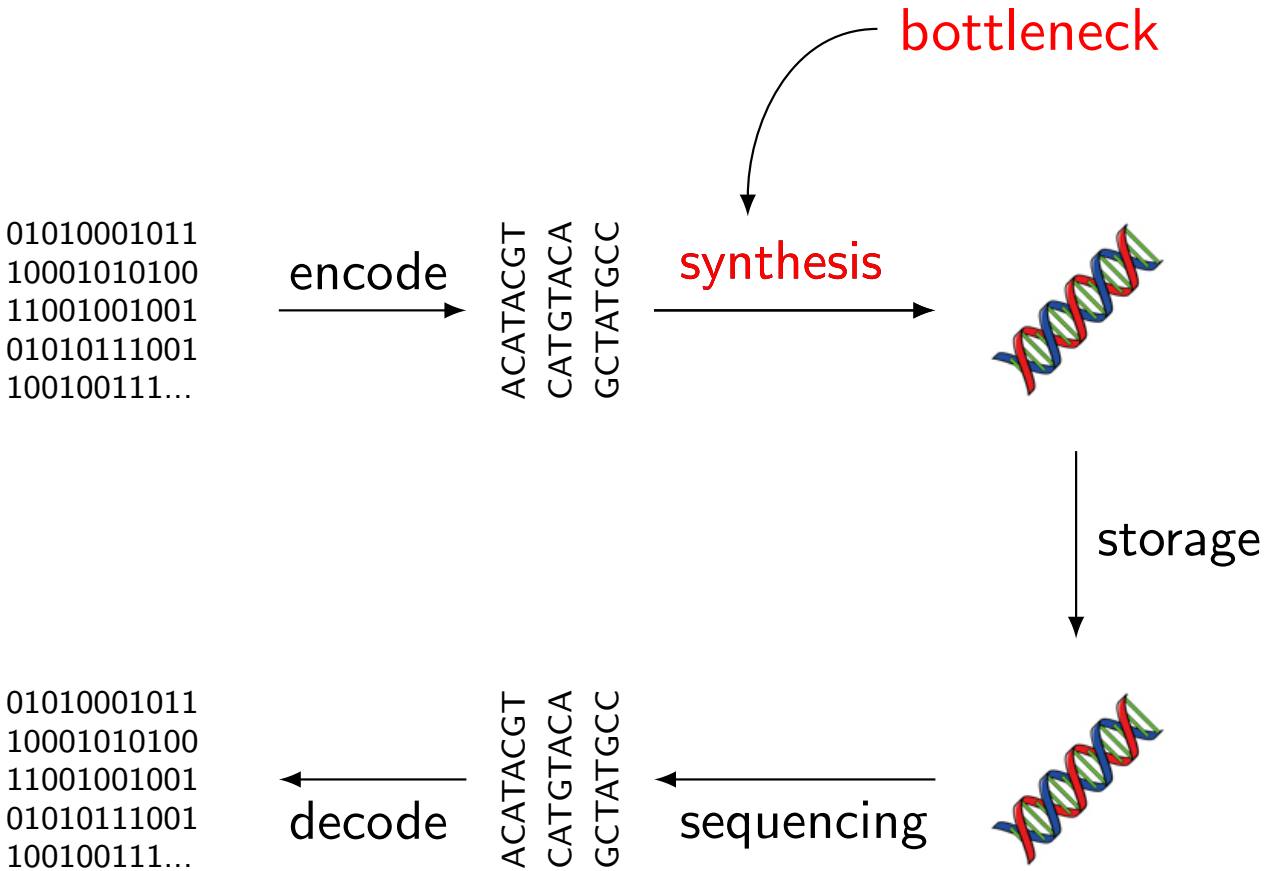
- Excellent longevity & Information density


NETFLIX
BIOHACKERS
STORED IN DNA
BIOHACKERS

Storing information in DNA

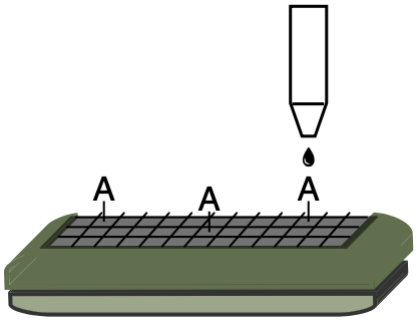
- Excellent longevity & Information density
- But expensive and slow

DNA storage system



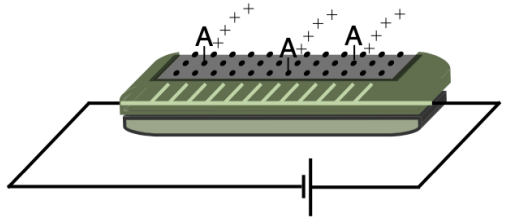
DNA synthesis

printing



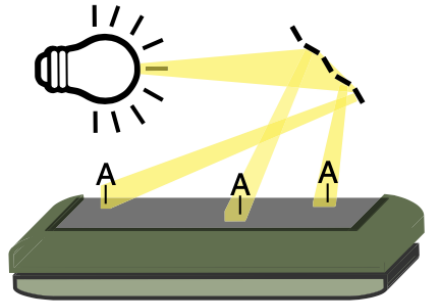
low-error, expensive

electrochemical



medium-error, expensive

light-directed



high-error, cheap

+

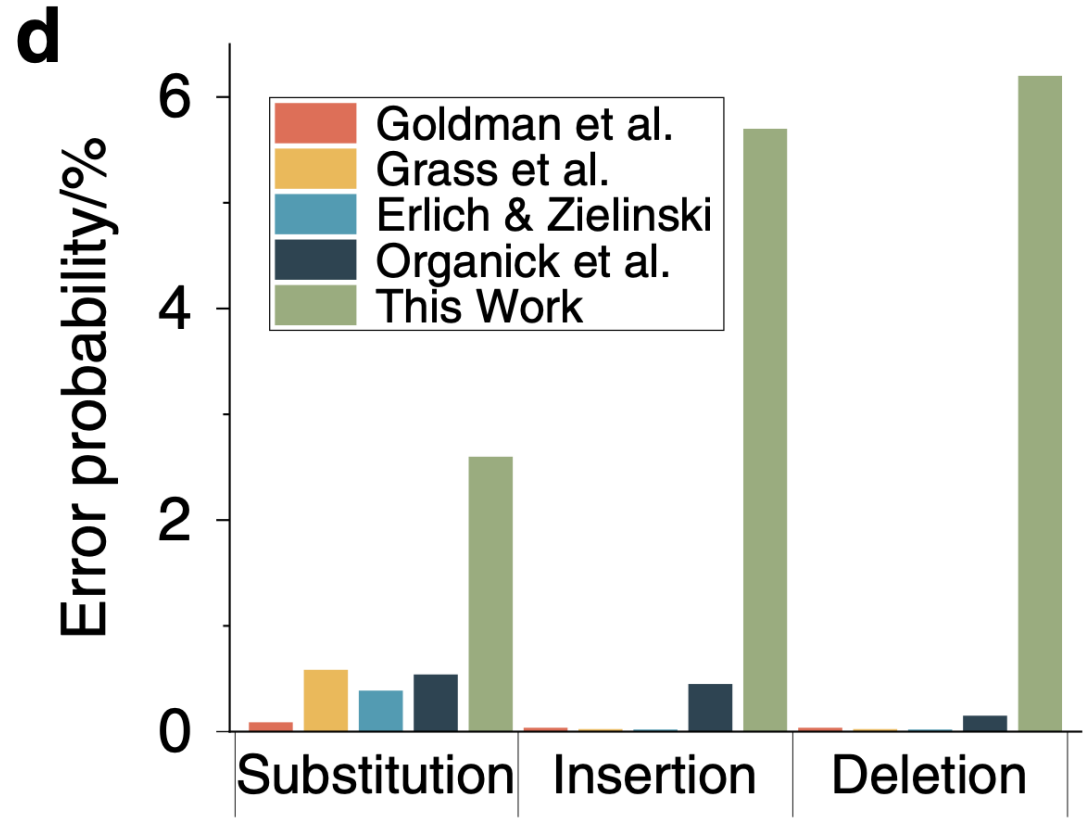
extensive error correction

this work

Errors in DNA storage

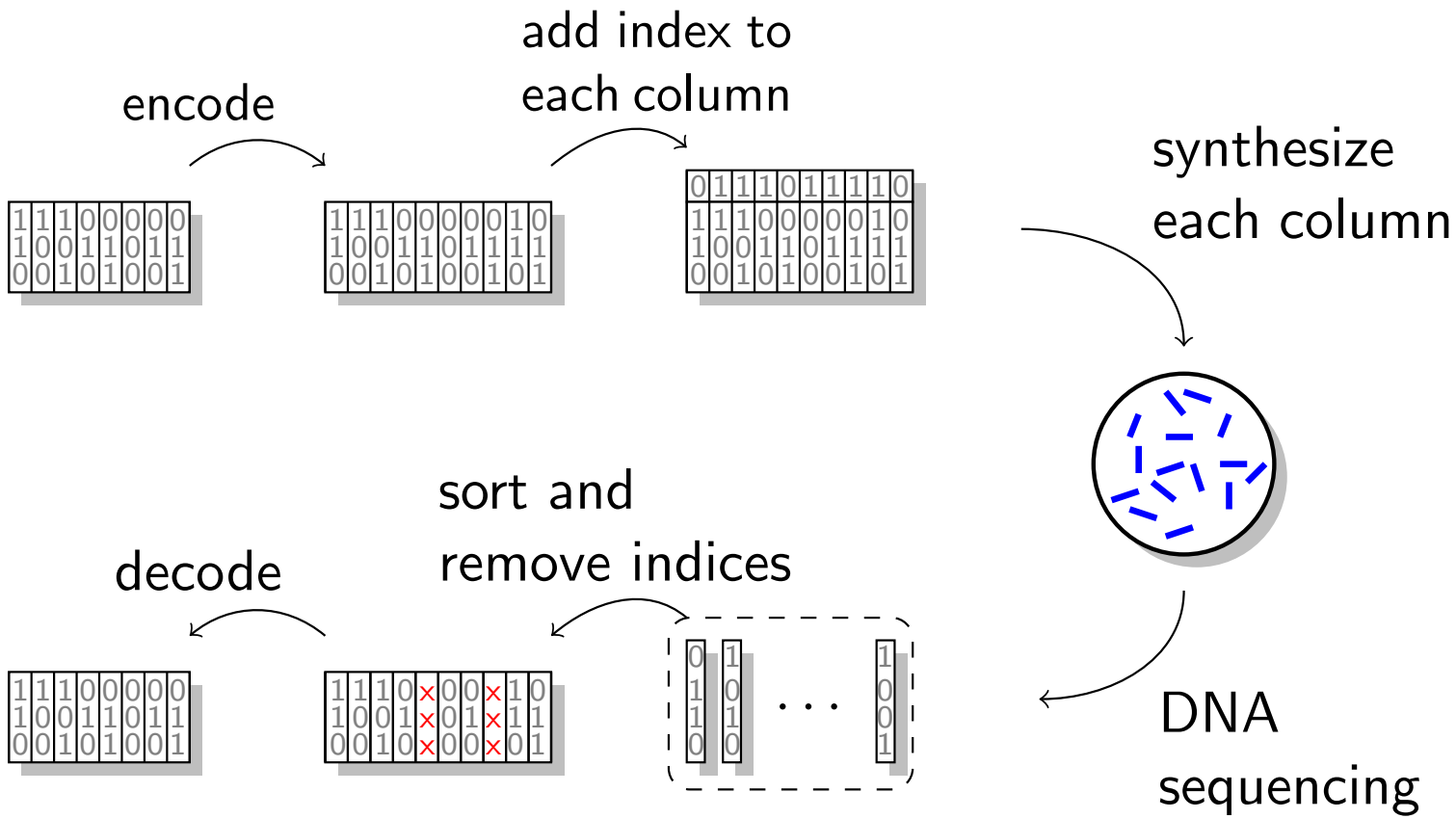
- Loss of sequences
- Substitution errors: 'ATA' $\rightarrow$ 'CTA'
- Deletion and insertion errors: 'ATA' $\rightarrow$ 'ACTA'

Errors in molecules



1/ Correcting for lost sequences

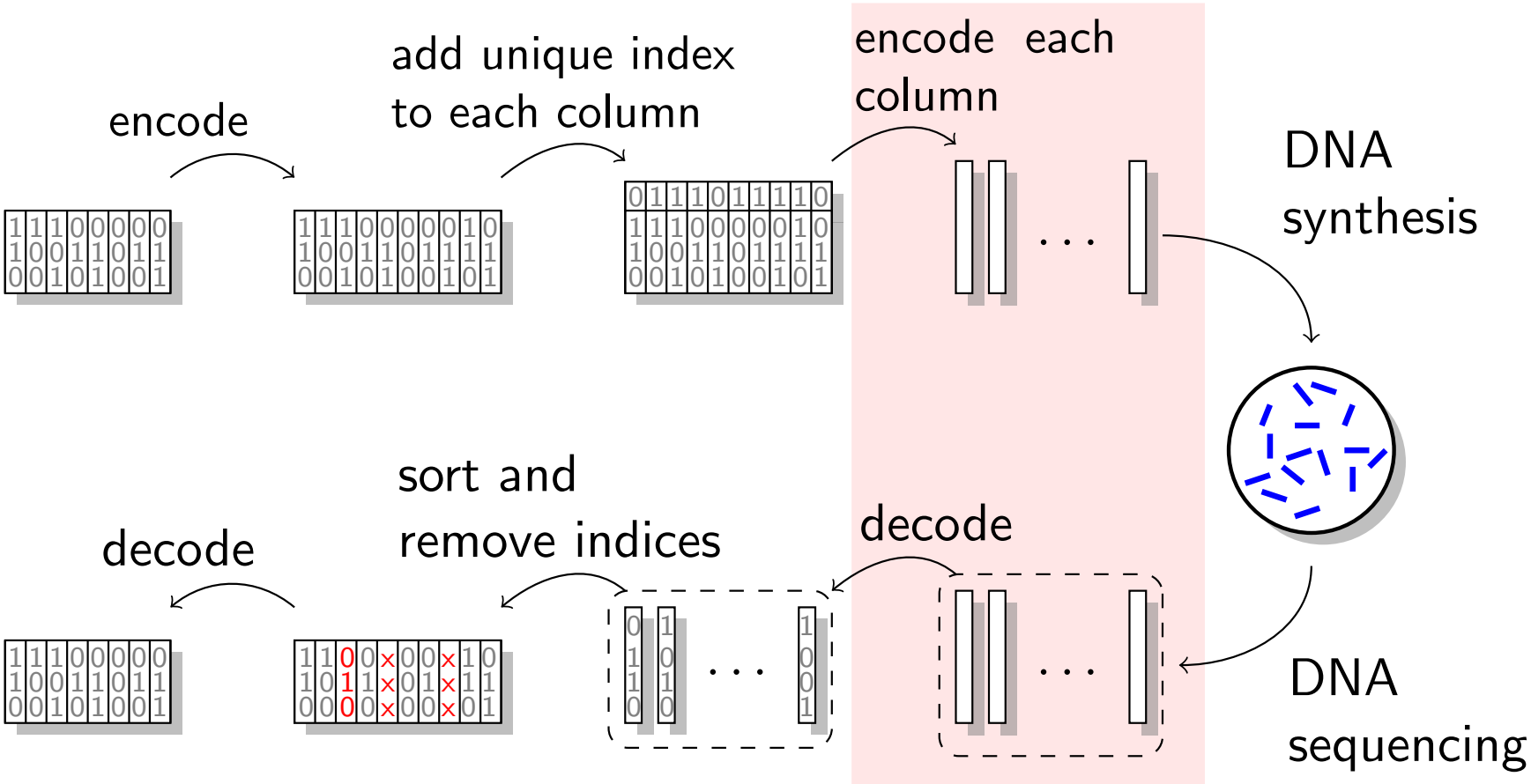
Encode with optimal erasure code



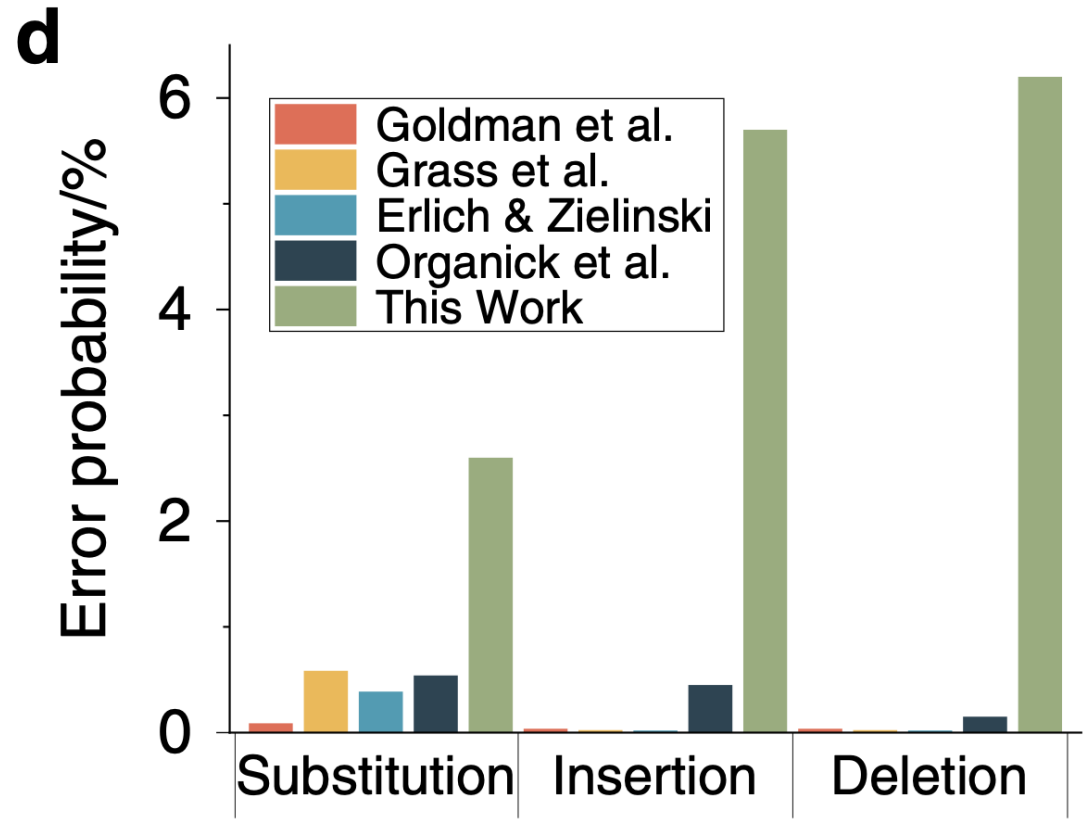
Theorem (H., Shomorony, Ramchandran, Tse, 2017)

This scheme maximizes bits/nucleotide, i.e., is optimal.

2/ Encoding+decoding if few errors in sequences

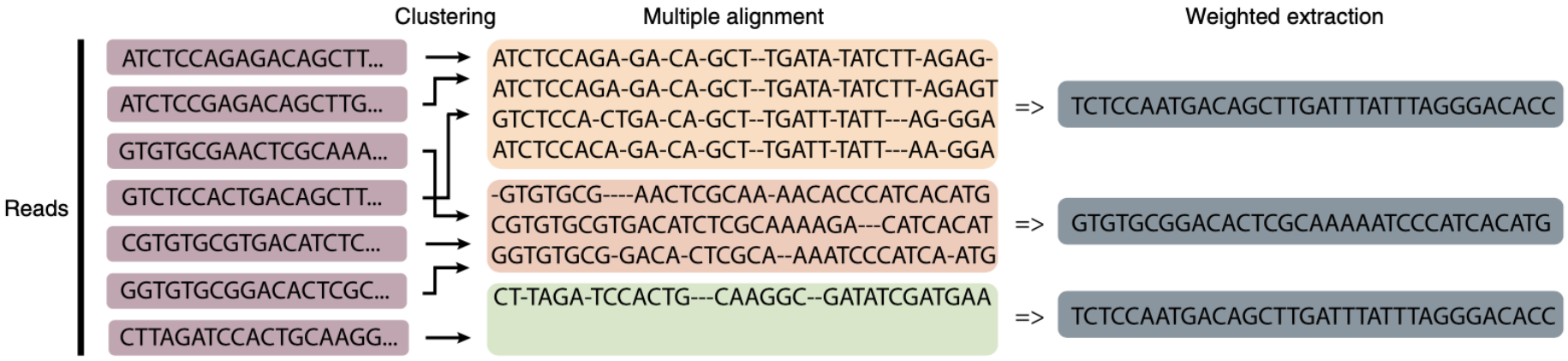


Light directed synthesis induces huge errors



- Only one in 10^7 sequences is error free
- Impossible to correct this with an inner code only

3/ Encoding+decoding if many errors in molecules



* We use locality-sensitive hashing based clustering (Haveliwala et al., 2000); see also Rashtchian et al., 2017 "Clustering Billions of Reads for DNA Data Storage" for an efficient clustering algorithm).

Summary

- Reliably storage with extremely noisy synthesis is possible
- Codes and algorithms enable shifting costs and working with noisy synthesis or sequencing

Long term information storage

- Information theory
- Chemistry

Prof. Robert Grass


MATT BLACK
MEZZANINE
DNA
CONTAINS
1 million copies of the
Mezzanine album encoded*
in 901'065 DNA sequences,
each 146 basepairs long
and encapsulated in silicone
particles for long-term
storage stability*
* Grass et al. Angew. Chem.
Int. Ed. 54, 2552 (2015).
* Paunescu et al. Nat. Protoc. 8,
2440 (2013).

- Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction, P. L. Antkowiak, J. Lietard, M. Z. Darestani, M. M. Somoza, W. J. Stark, R. Heckel and R. Grass, 2020.
- Robust chemical preservation of digital information on DNA in silica with error-correcting codes, R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, 2015.
- Fundamental limits of DNA storage systems, R. Heckel, I. Shomorony, K. Ramchandran, and D. Tse, 2017.

Thank you!