DNA-based data storage: constrained optimal compression rate and efficient transcoding

Sara Al Sayyed, Aline Roumy, Thomas Maugey
INRIA, Rennes

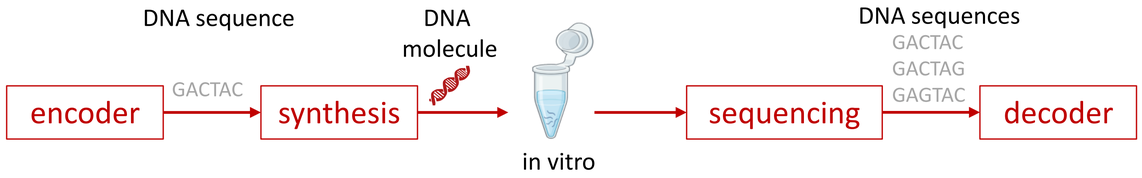
I - DNA-based data storage

Definition: is a technology that stores digital information by encoding the data into synthetic DNA sequences using the nucleotides A, C, G, and T.

Benefits:

- high density
- high durability
- low energy consumption

Workflow:



II - Safety vs Efficiency

Encoding Constraints:

1. **GC content** : 40 to 60 % (for stability)
2. **No Homopolymers** : repetition of the same nucleotide (e.g AAAA)

Safety: Adhering to the constraints
Efficiency: Low compression rate (in nts/pixel)

III - BreakSequence : Perfect Safety - Variable Transcoding Rate

Goal: Allow runs (homopolymers) of length up to m.

Code principle:

1. Mapping: 2 bits $\mapsto$ 1 nucleotide

bit pair	n
00	A
01	T
10	C
11	G

2. If m identical nucleotides occur, break the pattern and map the next 1 bit $\mapsto$ 1 nucleotide

$$\underbrace{b_0}\ \underbrace{b_1}\ \underbrace{b_2}\ \underbrace{b_3}\ \underbrace{b_4}\ \underbrace{b_5}\ \underbrace{b_6}\ \underbrace{b_7}\ \underbrace{b_8}$$

$$\text{if } \ n_0 = n_1 = n_2 \quad n_3 \quad n_4$$

such that:

$$\begin{cases} \textbf{One constraint :} n_3 \neq n_2 (\text{no homopolymer}) \\ \textbf{Two constraints :} n_3 \neq n_2 \,\&\, n_4 \ (\text{ for rate efficiency}) \end{cases}$$

One Constraint:

$$b_6 \in \{0,1\} \quad : 2 \text{ values to encode}$$
$$n_2 = \mathrm{A} \neq n_3 \in \{\mathrm{T}, \mathrm{C}, \mathrm{G}\} \quad : 3 \text{ possible values}$$

but rate sub-optimal

Idea: homopolymer can start in n_4 instead of n_3 : **two constraints**

Rule (Two Constraints):

- If $n_2 \neq n_4$, then $n_3 \in \{A, T, C, G\} \setminus \{n_2, n_4\}$.
  CODE: Choose the dictionary such that $\{n_2, n_4\}$ belong to the same column. Encode using the other option.

	Dic1		Dic2		Dic3	
Bit/Option	O_1	O_2	O_1	O_2	O_1	O_2
0	A	C	A	T	A	C
1	T	G	C	G	G	T

- If $n_2 = n_4$, then $n_3 \in \{A, T, C, G\} \setminus \{n_2\}$.
  CODE: encode with dictionary Dic1 with option $\neq n_2$.

Performance: This method is perfectly safe by construction.

V - Optimal Rate of a constrained code [1][2]

Constrained Transcoding

$$B_i \in \{0,1\} \qquad U_n \in \{A, T, C, G\}$$
$$B^i = (B_1, \ldots, B_i) \xrightarrow{\text{Transcoder } f} U^n = (U_1, \ldots, U_n)$$

Constraint on U^n: maximum **run length** $= m$

Def (Optimal rate for constraint m).

$$R^*(m) \triangleq \lim_{i\to\infty} \min_{f:\, \max\text{run}=m} \frac{1}{i} \mathbb{E}\left[\ell\left(f\left(B^i\right)\right)\right] \quad \text{(nts/bit)}$$

$$\text{Optimal rate (given } n) = \frac{\text{length of quaternary sequence}}{\text{length of binary sequence}} = \frac{n}{\log_2 |\mathcal{U}_n(m)|}$$

$$|\mathcal{U}_n(m)| = N_n(m) = \begin{cases} 4^n & n \leqslant m, \\ (4-1)\sum_{k=1}^{m} N_{n-k}(m), & n > m. \end{cases}$$

where k is the length of the last consecutive run of identical symbols.

k	$U_1, U_2, \ldots, U_{n-m}, \ldots, U_{n-2}, U_{n-1}, U_n$	$N_n(m) = \lvert\mathcal{U}_n(m)\rvert$
1	$\ldots\ A\ \begin{matrix}C\\T\\G\end{matrix}$	$3\,N_{n-1}(m)$
2	$\ldots\ A\ \begin{matrix}C & C\\T & T\\G & G\end{matrix}$	$3\,N_{n-2}(m)$
m	$\ldots\ A\ \begin{matrix}C & C & \ldots & C & C\\T & T & \ldots & T & T\\G & G & \ldots & G & G\end{matrix}$	$3\,N_{n-m}(m)$

Initialization: $N_n(m) = 4^n\ ,\ n \leqslant m$
Growth: as $n \to \infty$, $N_n(m) \approx A(m)\lambda(m)^n$,

where $\lambda(m)$ is the largest real root of the characteristic equation :

$$x^{m+1} - 4x^m + 4 - 1 = 0.$$

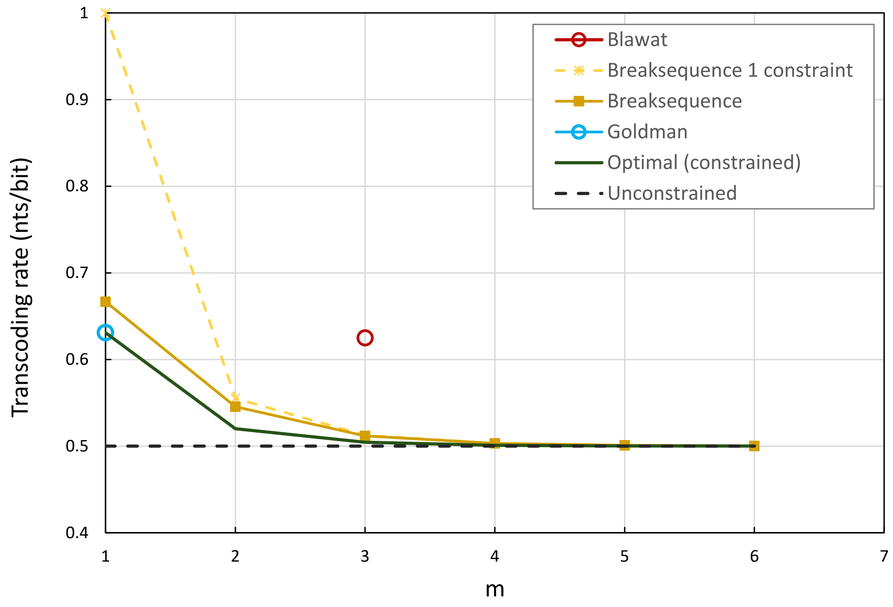
$$R^*(m) = \lim_{n\to\infty} \frac{n}{\log_2 |\mathcal{U}_n(m)|} = \lim_{n\to\infty} \frac{n}{\log_2 \left(A(m)\lambda(m)^n\right)} = \frac{1}{\log_2 \lambda(m)} \text{(nts/bit)} .$$

VI- Results

Goldman[3]: Transcode 1 ternary symbol into one nucleotide (m=1):

$$R_{Goldman} = R^*(1)$$

Blawat[4]: Transcode 8 bits into 5 nucleotide (m=3)



Conclusion

We propose a constrained transcoder (BreakSequence) that maps binary data to the quaternary domain . We then compare its rate to the optimal constrained rate and show that it outperforms the state-of-the-art method (Blawat).

[1] K. Immink, K. Cai, Properties and Constructions of Constrained Codes for DNA-Based Data Storage, 2020.
[2] C. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, 1948.
[3] N. Goldman, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 2013.
[4] M. Blawat et al., "Forward error correction for DNA data storage," Procedia Computer Science, 2016.

sara.al-sayyed@inria.fr, aline.roumy@inria.fr, thomas.maugey@inria.fr

France 2030, PEPR MoleculArXiv, ANR-22-PEXM-0005