

IMPERIAL

DNA data storage: A generative tool for Motif-based DNA storage

Samira Brunmayr, Omer S. Sella, Thomas Heinis

23rd USENIX Conference on File and Storage Technologies

February 27, 2025

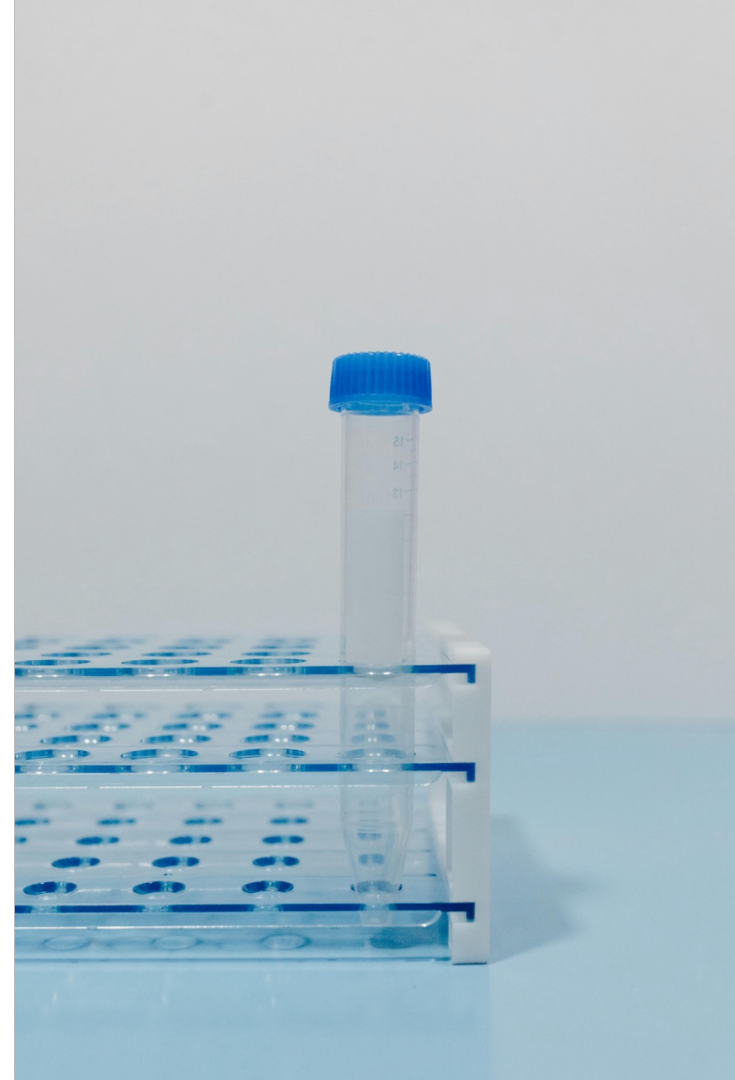
Data Storage

- High demand
- Storage devices deteriorate over time
- Require considerable space and energy

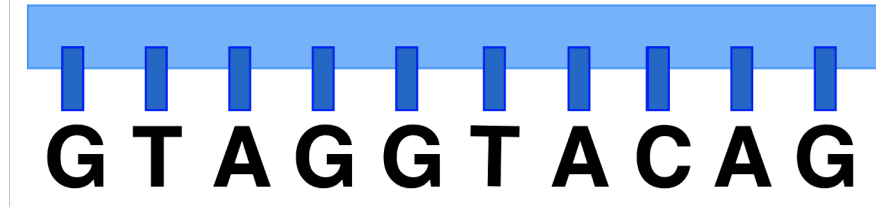


DNA as a Storage Medium

- 10^{18} bytes per mm^3
- Store data for 2000 years
- Store Wikipedia in test tube



How to store data in DNA?

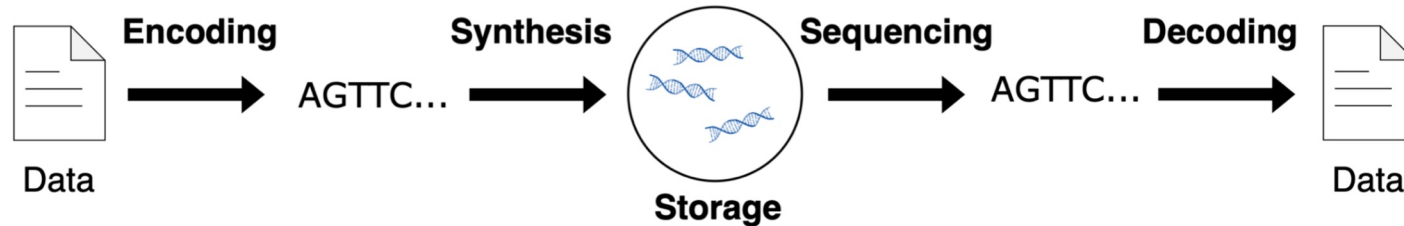


Bases	Binary
A	00
T	01
C	10
G	11

Encoding example

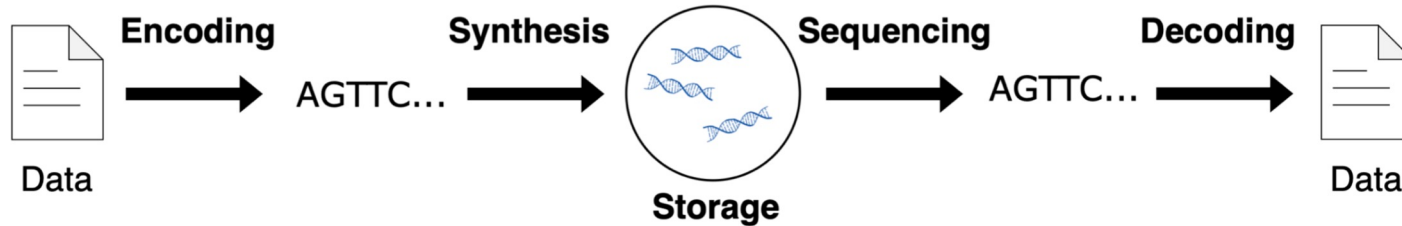
How to store data in DNA?

1. Encoding
2. Synthesis: construct DNA strands
3. Storage of DNA
4. Sequencing: access data in DNA
5. Decoding



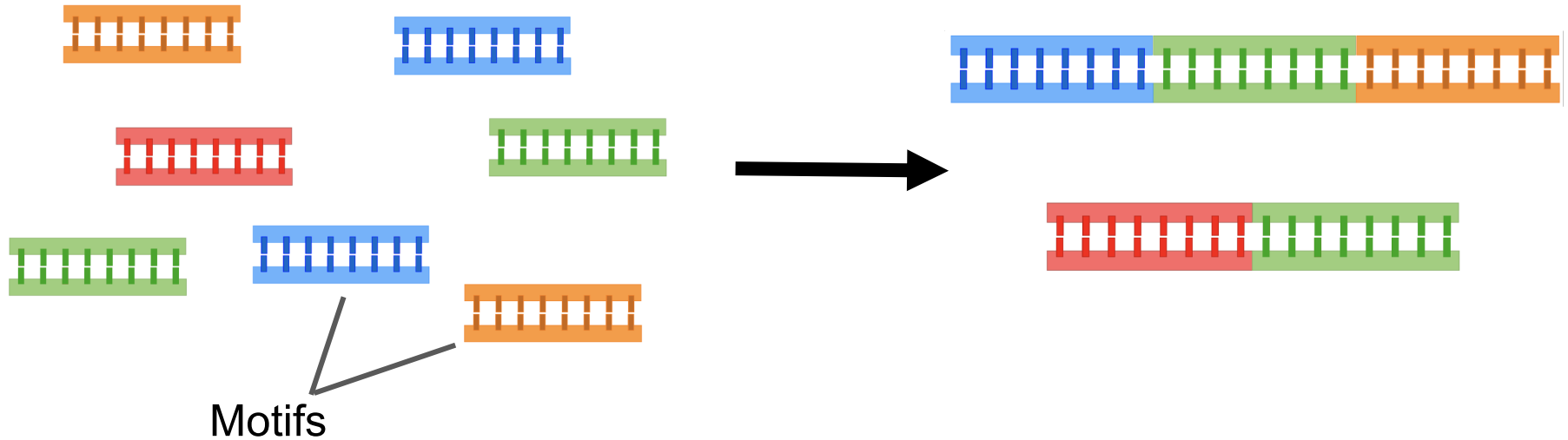
Current Technological Limitations

- Strand length: <150-200 bases
- Synthesis: \$400 million/Terabyte



Solution: Motif-based DNA Data Storage

- Motif: Short DNA sequence



Gibson Assembly Method

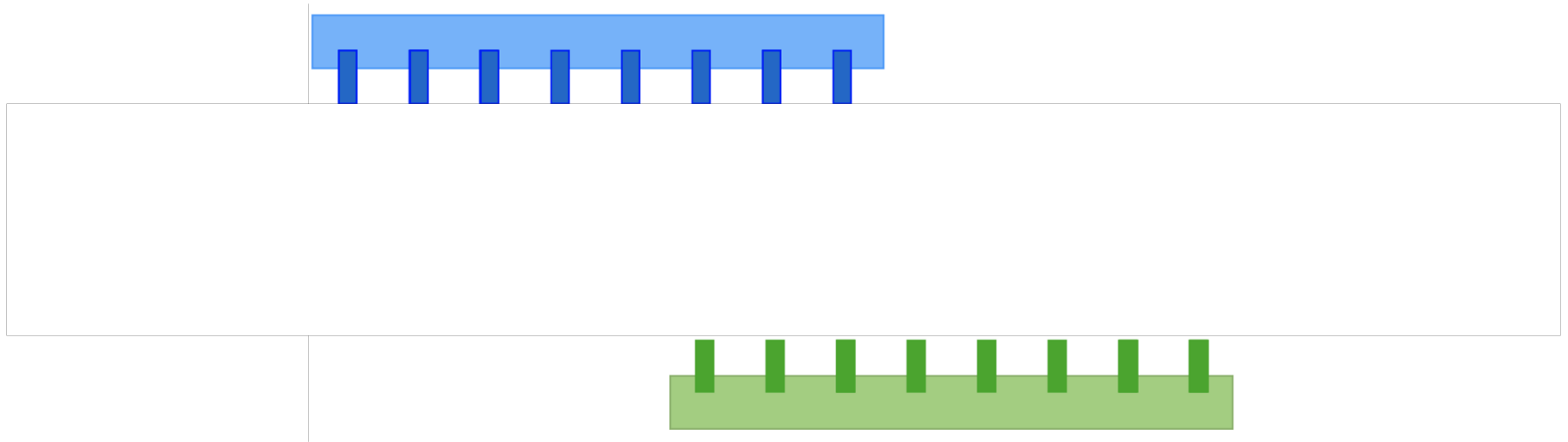
Motif 1



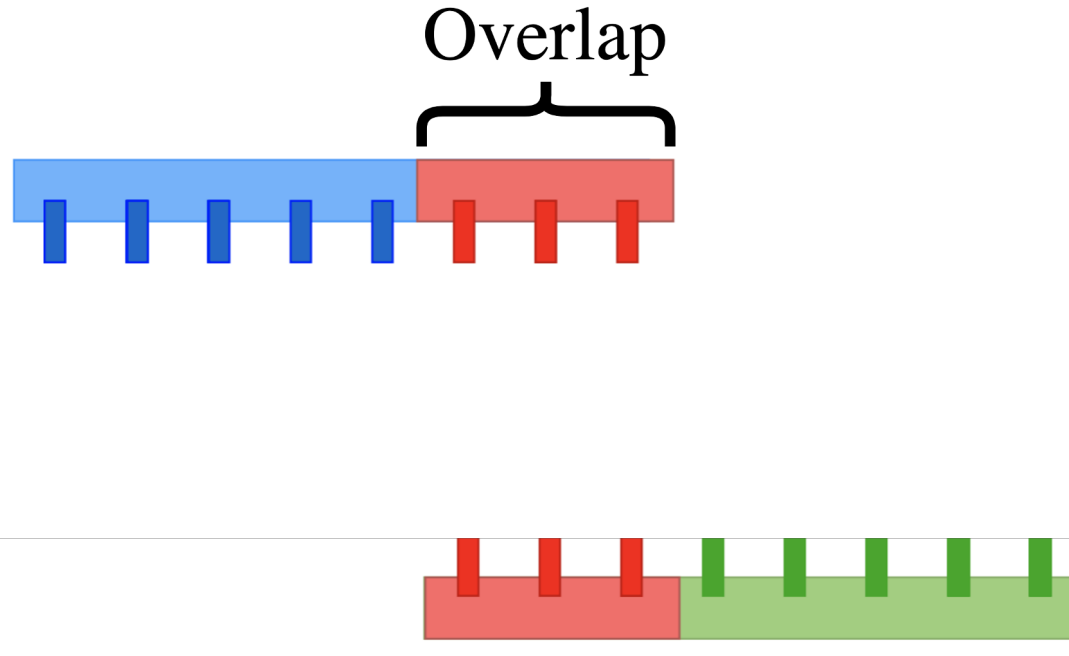
Motif 2



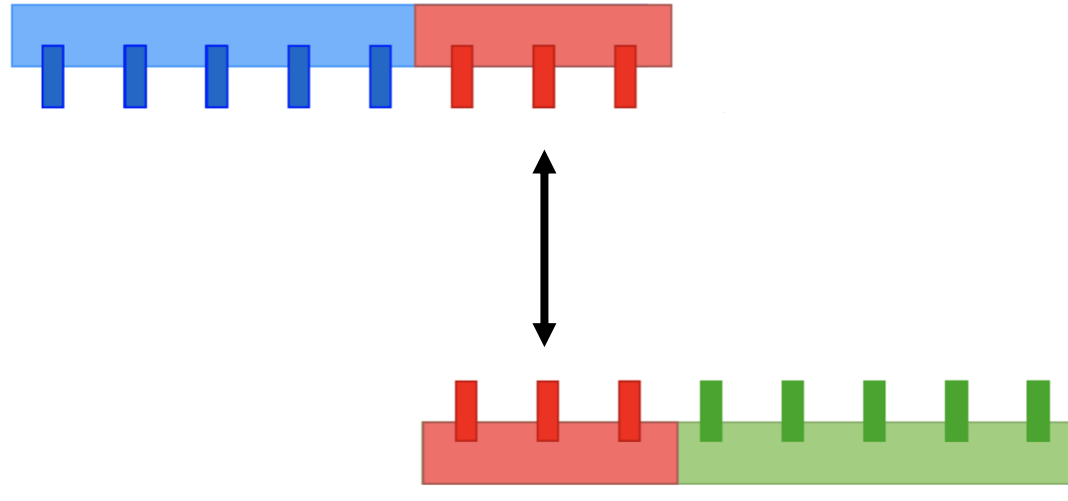
Gibson Assembly Method



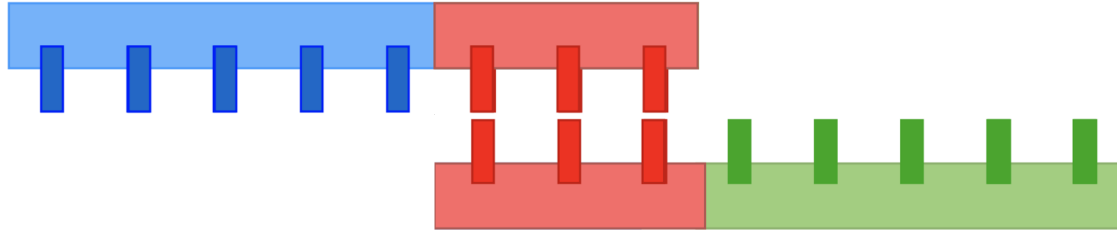
Gibson Assembly Method



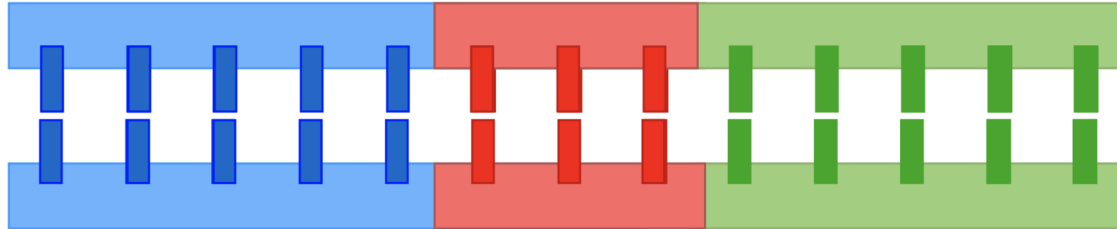
Gibson Assembly Method



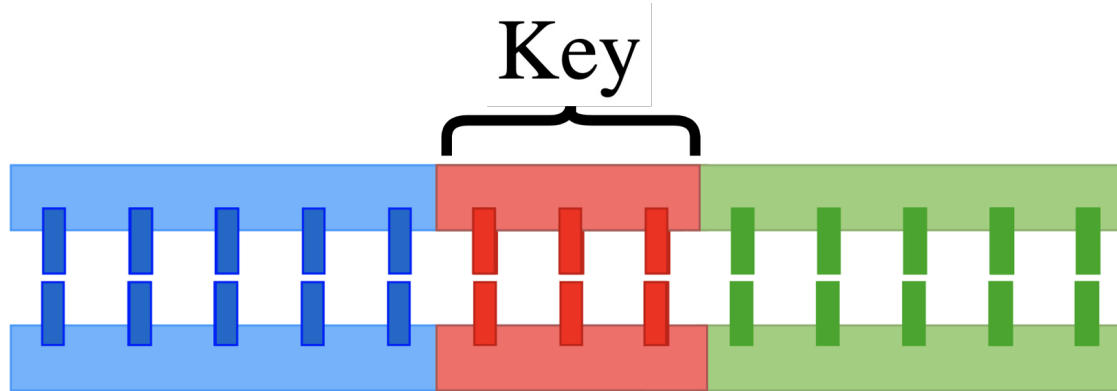
Gibson Assembly Method



Gibson Assembly Method

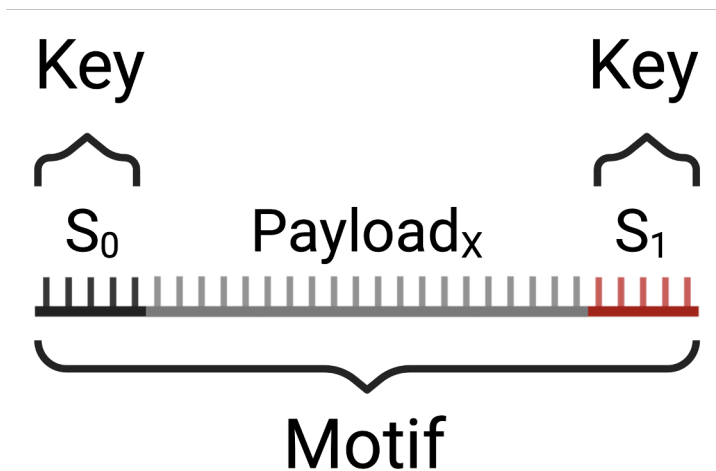


Gibson Assembly Method

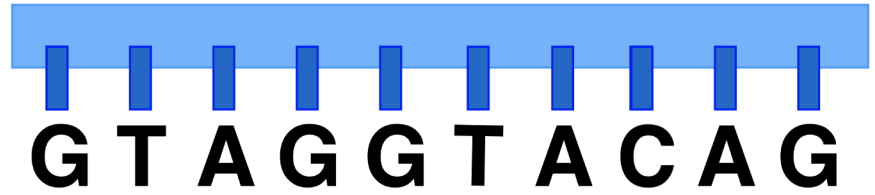


Motif Structure

- Payload: Actual Information
- Key:
 - Used for Assembly
 - Access Data

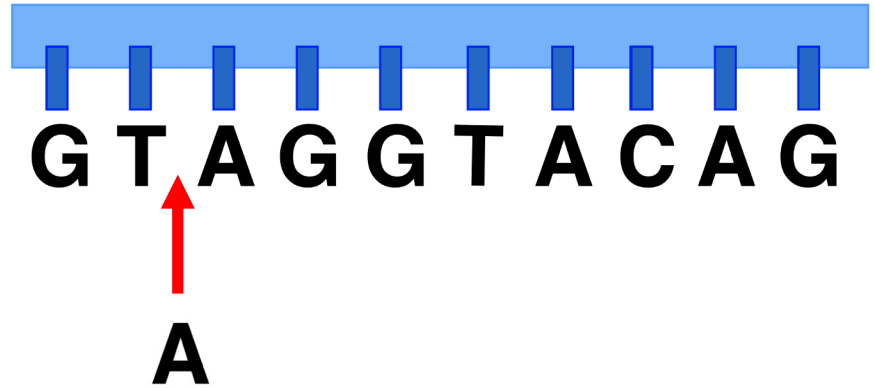


DNA is Error-Prone



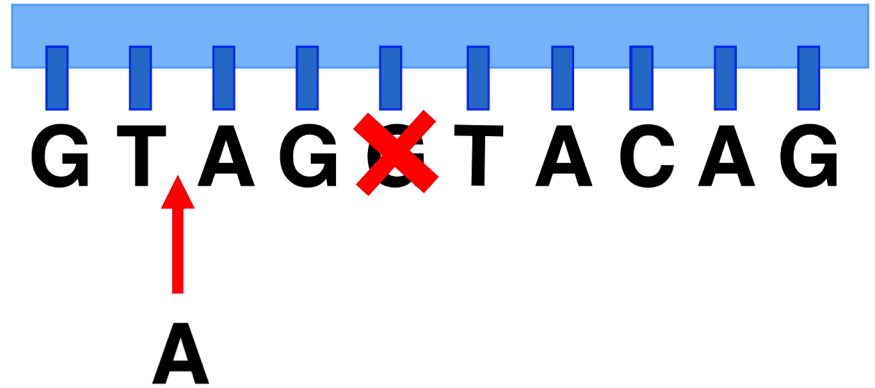
DNA is Error-Prone

- Insertion



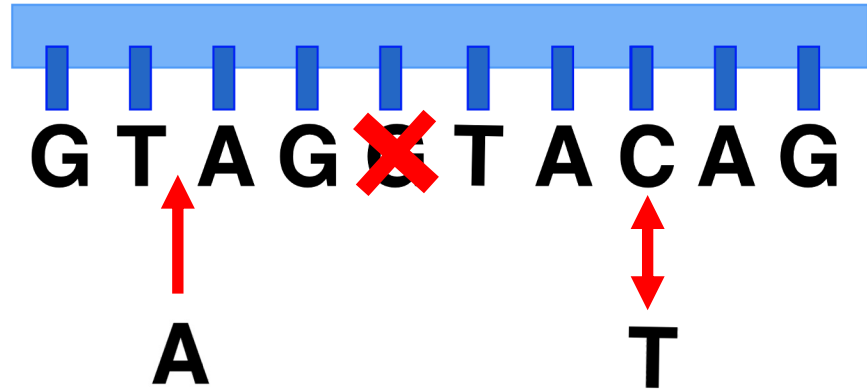
DNA is Error-Prone

- Insertion
- **Deletion**



DNA is Error-Prone

- Insertion
- Deletion
- **Substitution**



Biological Constraints

- Homopolymer
- Hairpins
- GC-Content

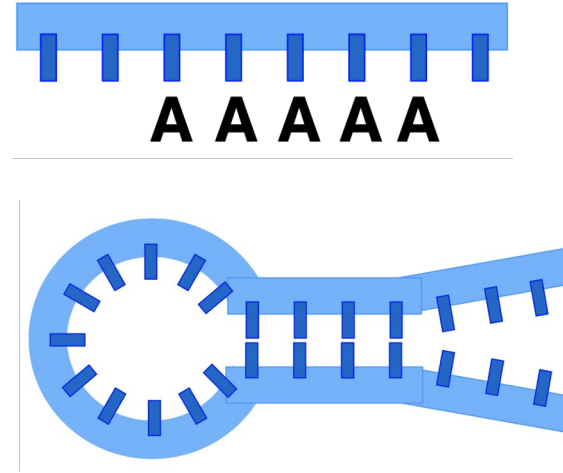
Biological Constraints

- **Homopolymer**
- Hairpins
- GC-Content



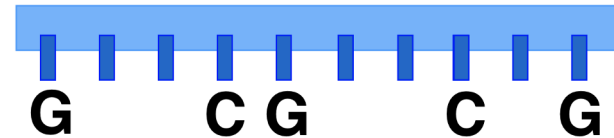
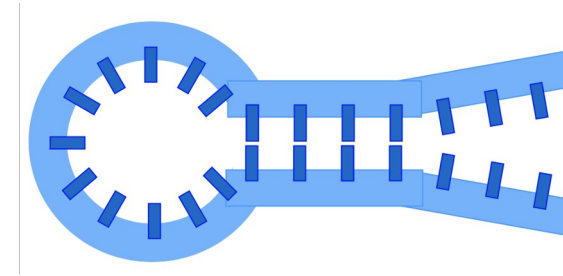
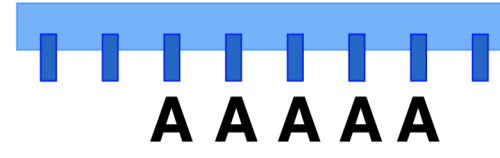
Biological Constraints

- Homopolymer
- **Hairpins**
- GC-Content



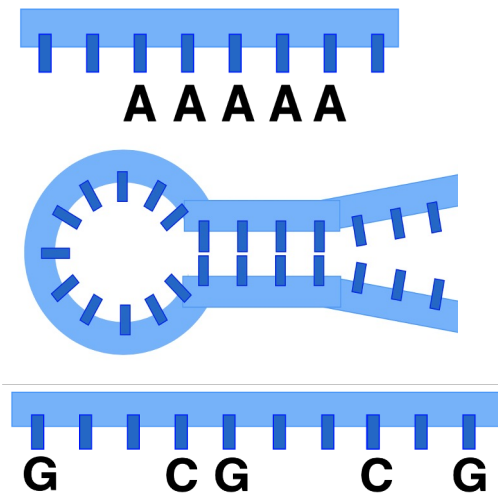
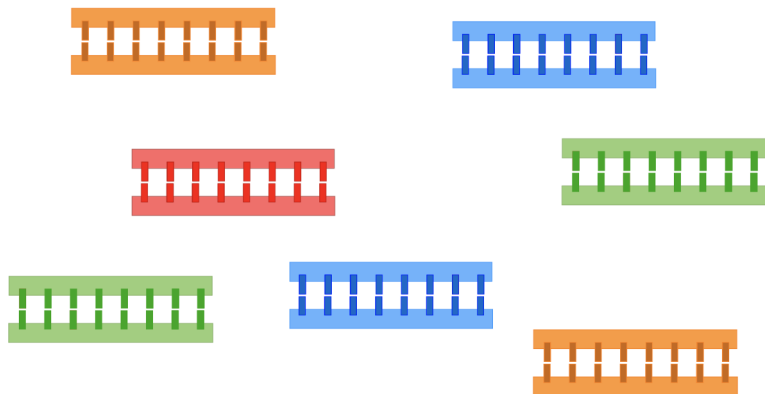
Biological Constraints

- Homopolymer
- Hairpins
- **GC-Content**



What we want:

- **Cost-efficient:** Motif-based storage
- **Reliable:** conform to biological constraints



The **Motif** Generation Tool

Motif Generation Tool

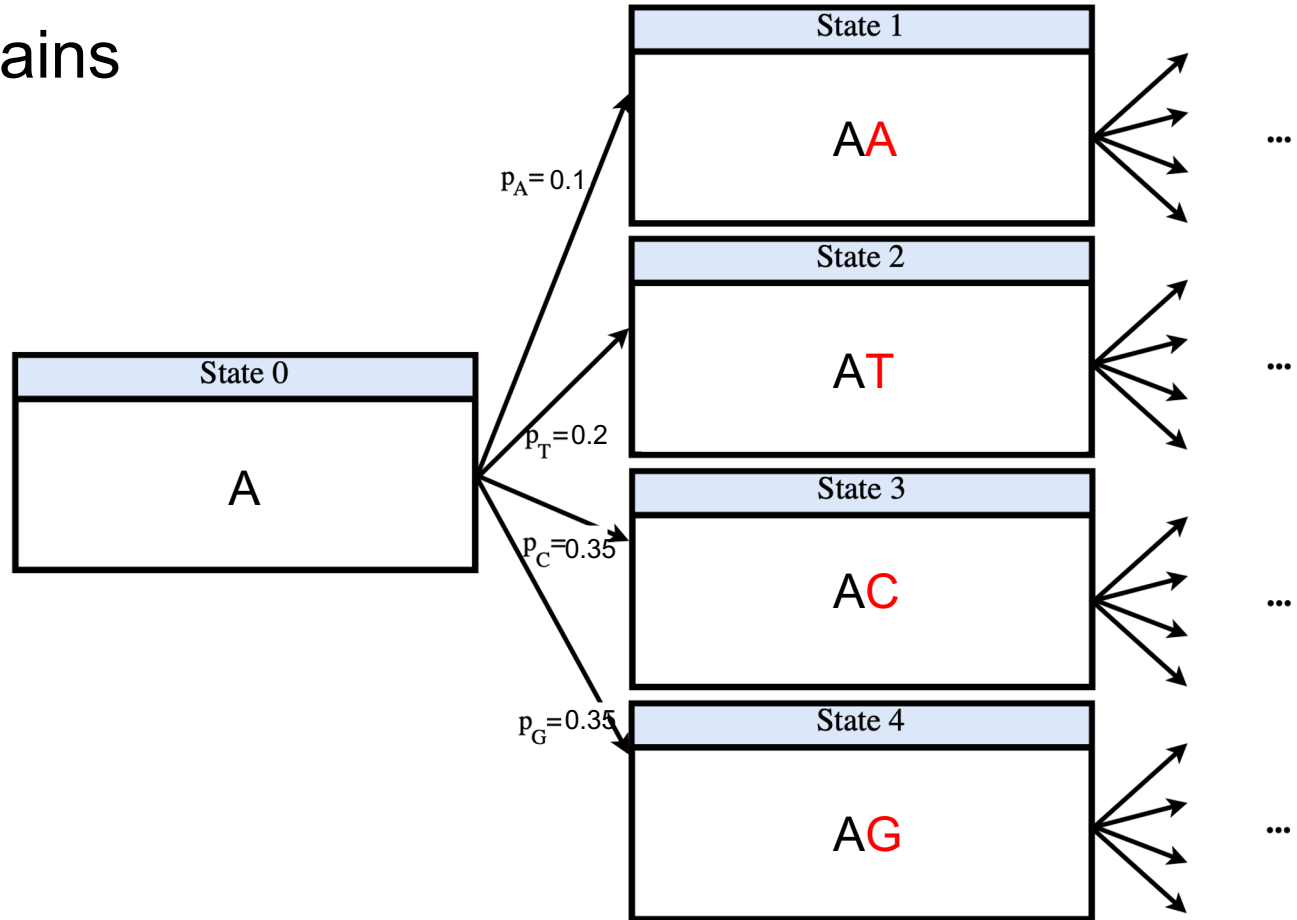
- **Input:**

- Information on Motifs: payload length, ...
- Constraints on:
 - Homopolymer
 - Hairpins
 - GC-Content

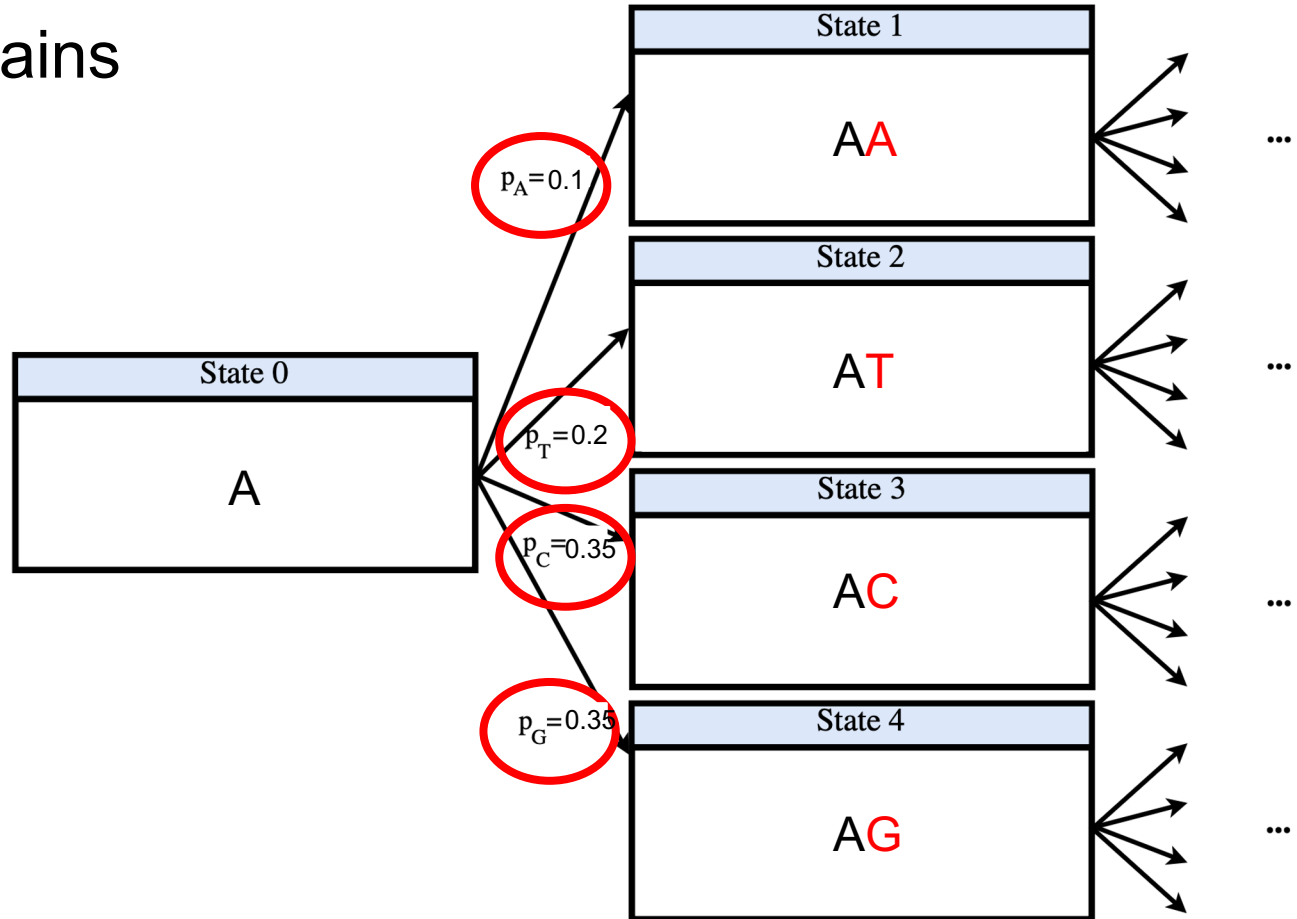
- **Output:**

- Set of motifs conforming to the constraints

Markov Chains



Markov Chains



Transition Probabilities

$$score(i) = score_{hom}(i) \times score_{hairpin}(i) \times score_{gc}(i) \times score_{noKeyInPayload}(i)$$

$$score(i) = e^{\log Score(i)}$$

$$p_i = \frac{score(i)}{\sum_{j \in \{A, T, C, G\}} score(j)}$$

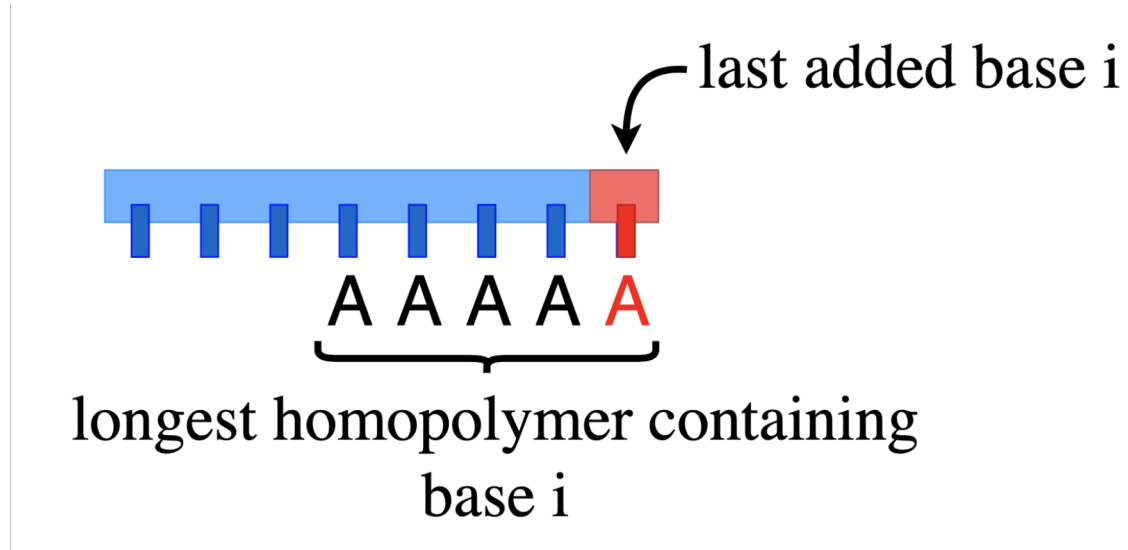
Transition Probabilities

$$score(i) = score_{hom}(i) \times score_{hairpin}(i) \times score_{gc}(i) \times score_{noKeyInPayload}(i)$$

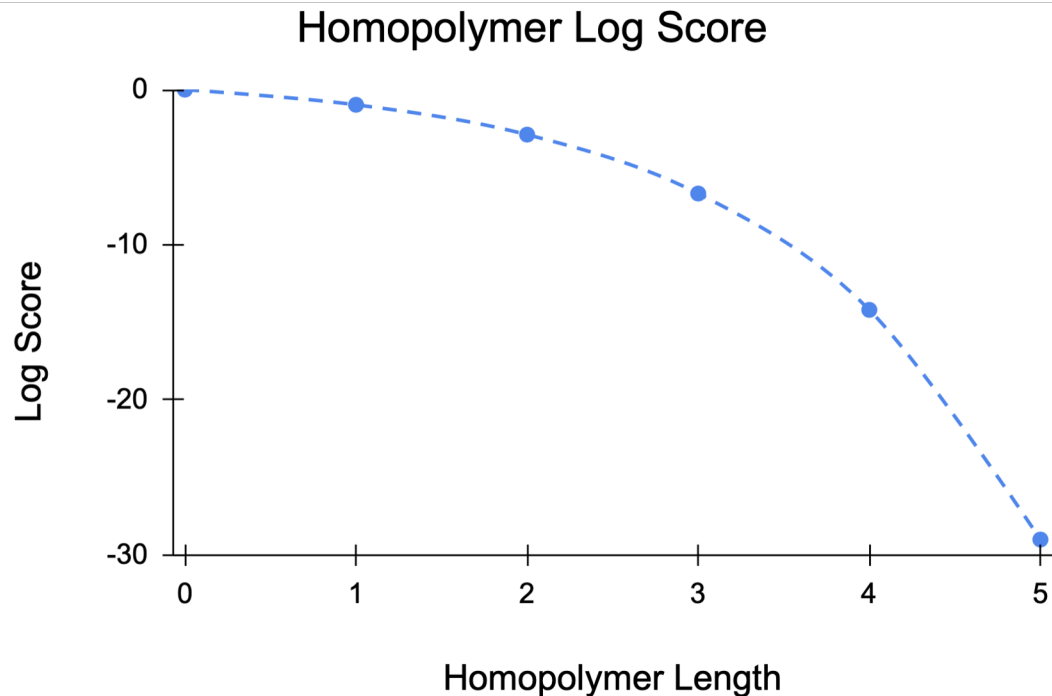
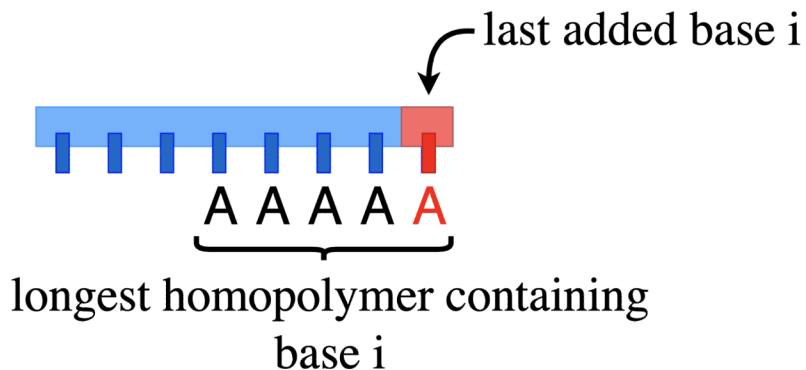
$$score(i) = e^{\logScore(i)}$$

$$p_i = \frac{score(i)}{\sum_{j \in \{A, T, C, G\}} score(j)}$$

Homopolymer Log Score

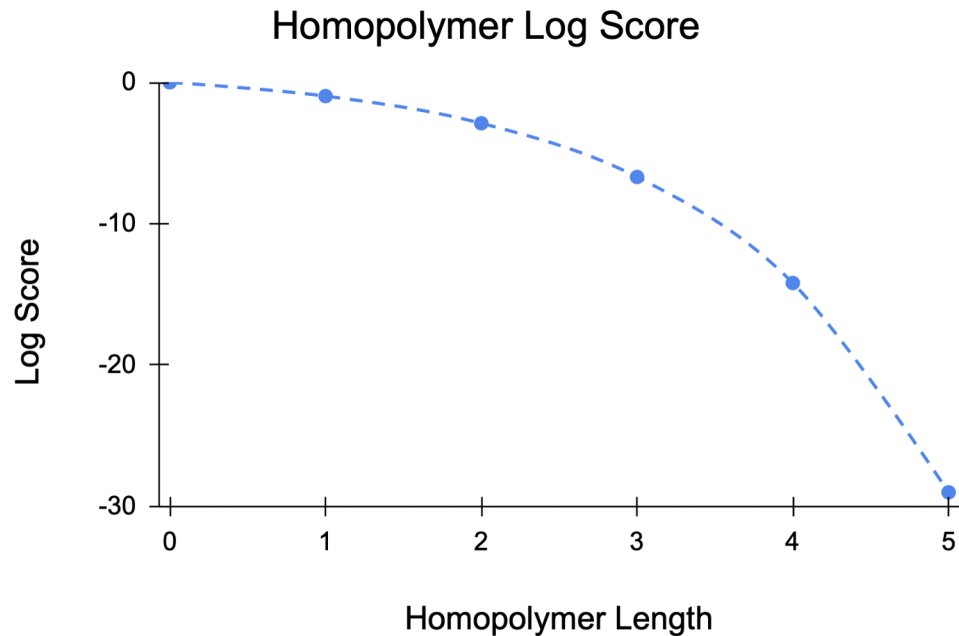


Homopolymer Log Score



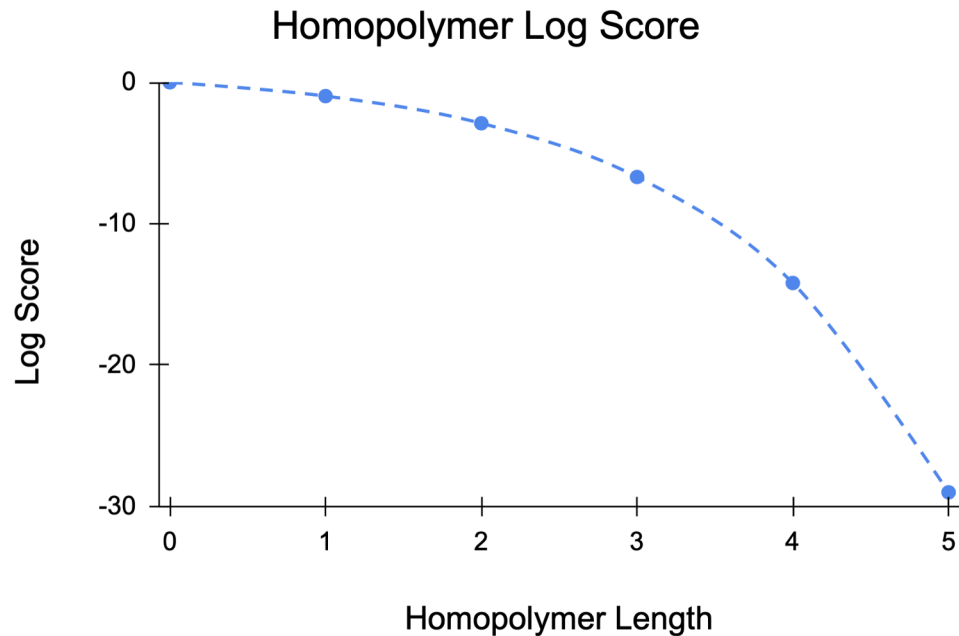
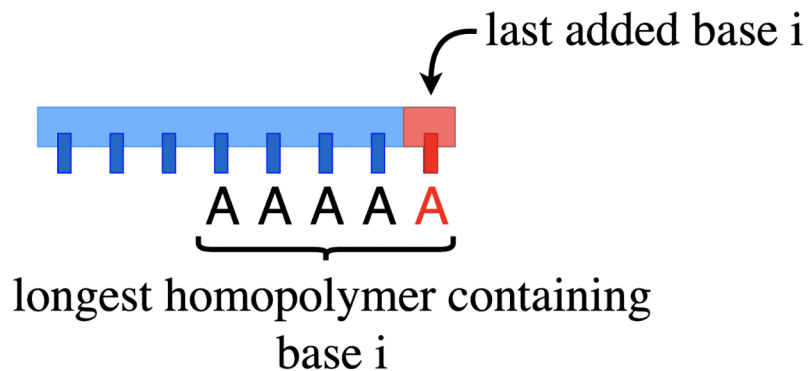
Homopolymer Log Score

$$ls_{hom}(i) = -(\mathbf{h}_{hom})^{homLen/\maxHom} + 1$$



Homopolymer Log Score

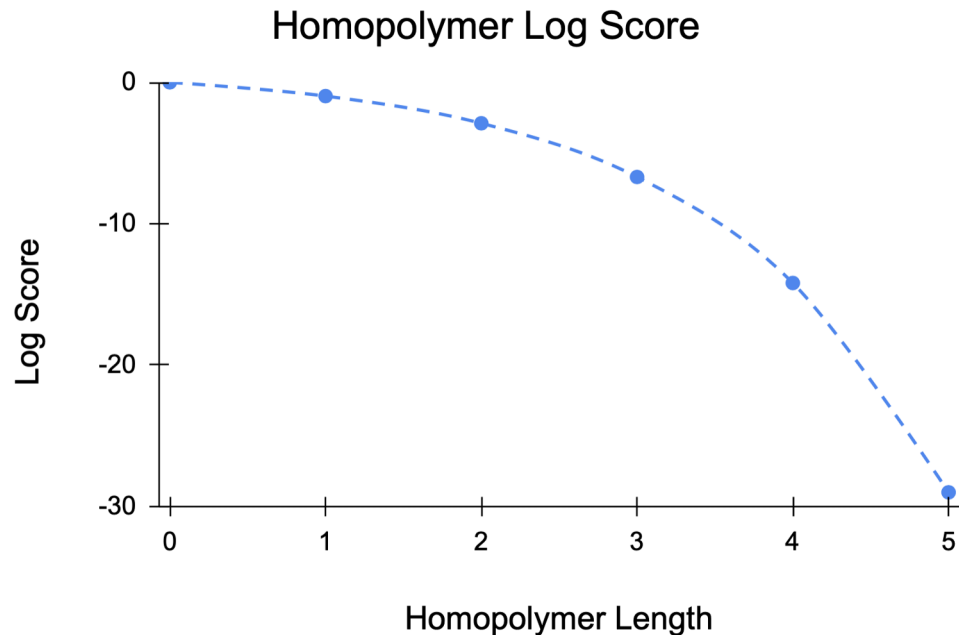
$$ls_{hom}(i) = -(\mathbf{h}_{hom})^{\text{homLen}/\text{maxHom}} + 1$$



Homopolymer Log Score

$$ls_{hom}(i) = -(\mathbf{h}_{hom})^{homLen/\mathbf{maxHom}} + 1$$

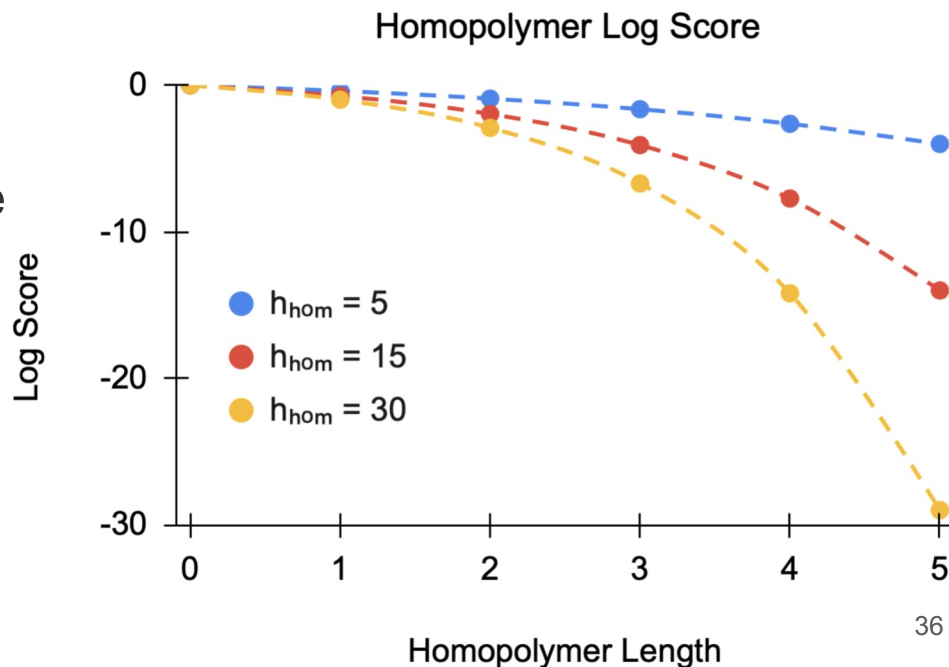
● **maxHom** : Constraint



Homopolymer Log Score

$$ls_{hom}(i) = -(\mathbf{h_{hom}})^{homLen/\maxHom} + 1$$

- $\mathbf{h_{hom}}$: controls shape of log score



Evaluation

Evaluation Setup

Other tools for comparison:

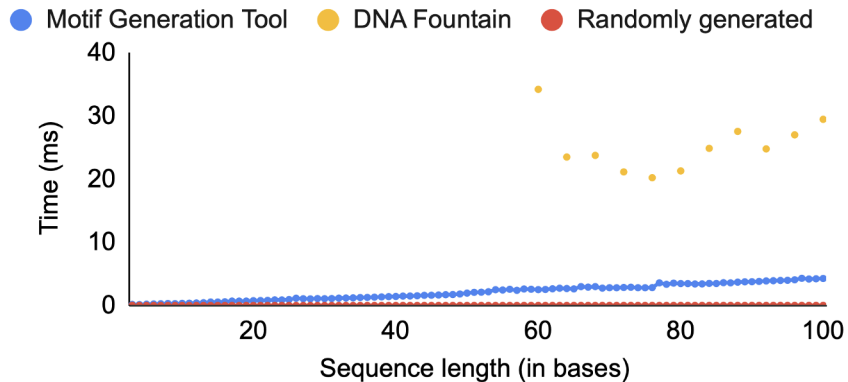
- DNA Fountain, Erlich et al.
- Finite State Machine Encoding, Sella et al.
- Shortmer combinatorial encoding scheme, Preuss et al.
- Randomly generated DNA sequences

Evaluation:

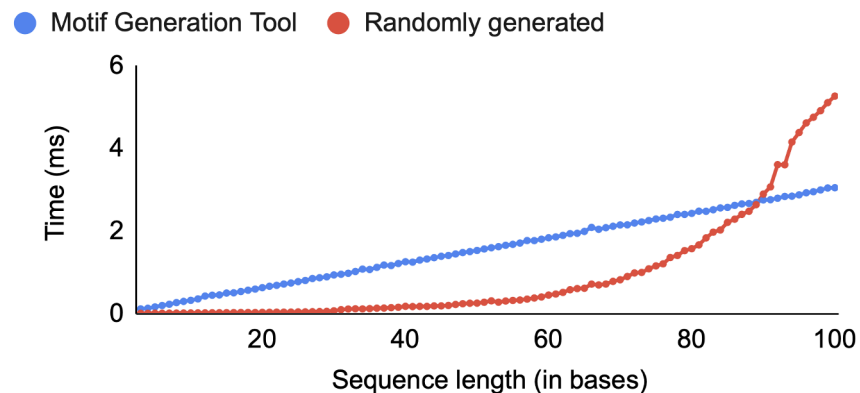
- For each constraint separately
- Case Study

Single Constraint

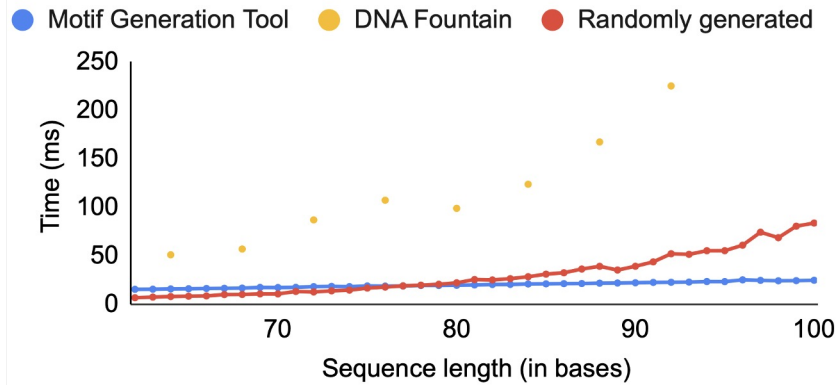
GC-Content between 25% to 65%



No homopolymers with more than 2 bases

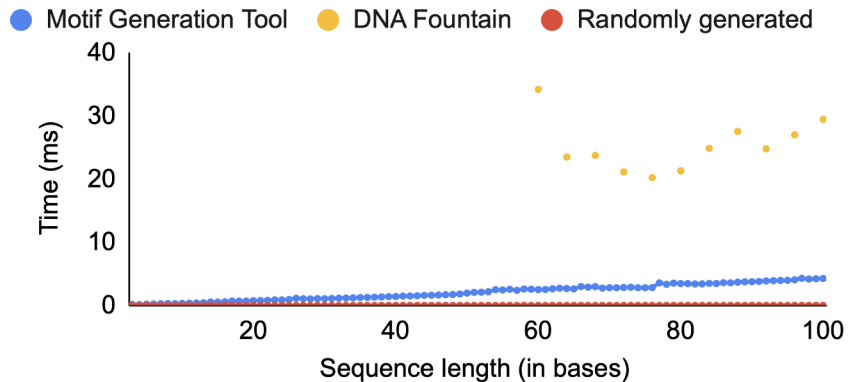


No hairpin with Stem ≥ 2 bases and Loop Size 1 base

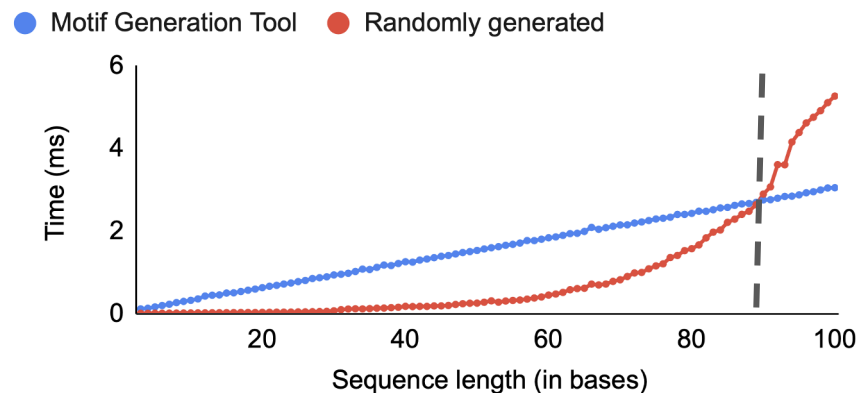


Single Constraint

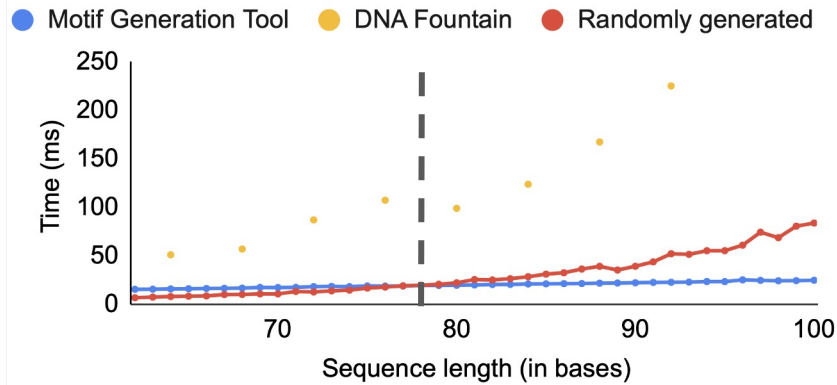
GC-Content between 25% to 65%



No homopolymers with more than 2 bases



No hairpin with Stem ≥ 2 bases and Loop Size 1 base



Case Study

- Currently applicable constraints

Parameters		Values
Motifs	<i>keySize</i>	20
	<i>keyNum</i>	8
	<i>payloadSize</i>	60
	<i>payloadNum</i>	15
Homopolymers	<i>maxHom</i>	5
Hairpins	<i>maxHairpin</i>	1
	<i>minLoopSize</i>	6
	<i>maxLoopSize</i>	7
GC-Content	<i>minGC</i>	25
	<i>maxGC</i>	65

Case Study

	Motif Generation Tool	DNA Fountain [11]	Euclid [21]	Preuss et al. [17]	Randomly generated
Time taken to generate a set of motifs conforming to the constraints in Table 1 (seconds)	2.54 sec	>5min	>5min	>5min	>5min

Summary

- Motif-based DNA data storage - cost efficient
- **Motif Generation Tool** for automated reliable motif generation
- Outperforms other tools based on currently industry standard biological constraints

Summary

- Motif-based DNA data storage - cost efficient
- **Motif Generation Tool** for automated reliable motif generation
- Outperforms other tools based on currently industry standard biological constraints

Thank you for your attention!

Q&A

Sponsors: DNAMIC and NEO