Duplication Correcting Codes for live DNA Storage

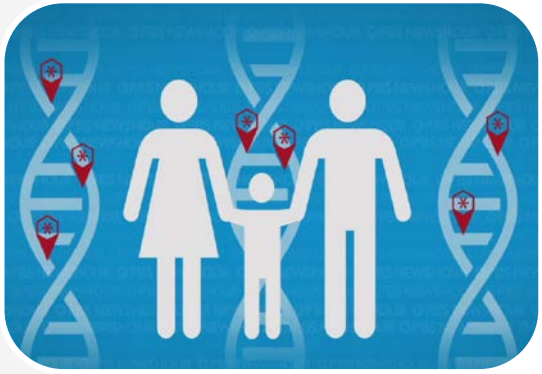
Farzad Farnoud

Siddharth Jain
Jehoshua Bruck

Moshe Schwartz

Allerton Conference, UIUC, 2016

Data Storage in DNA

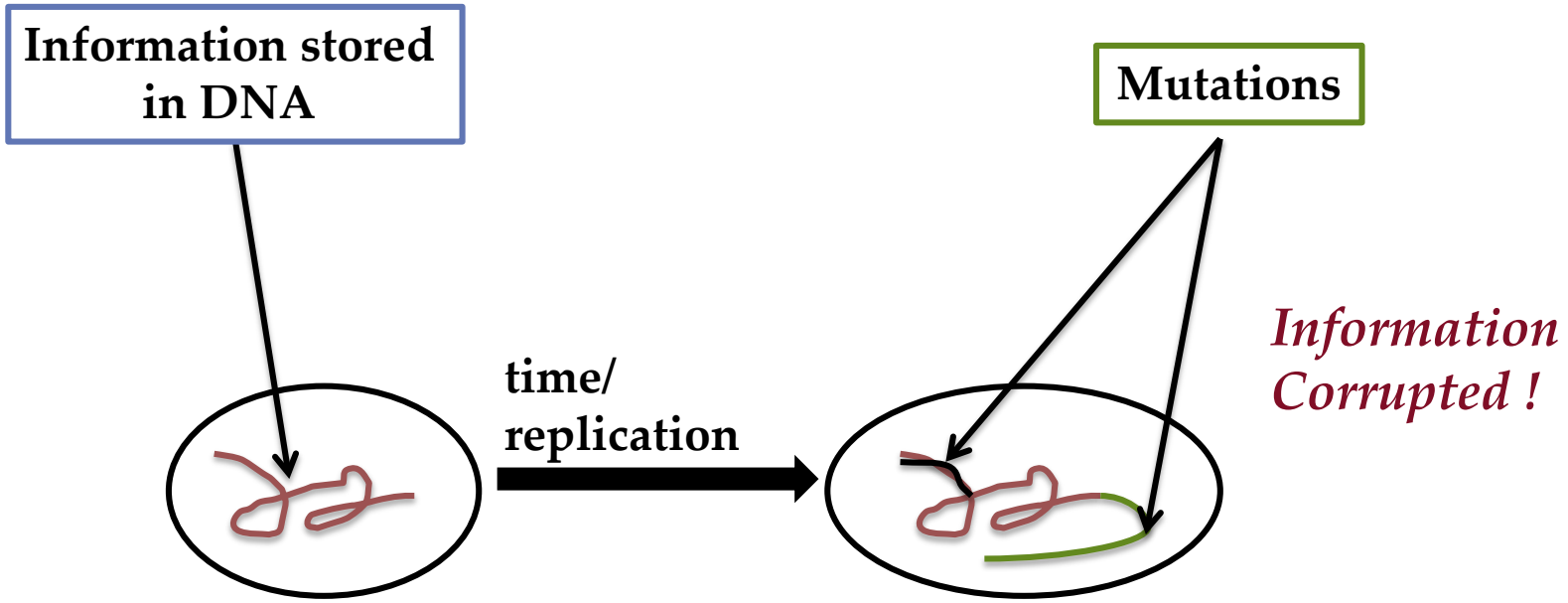


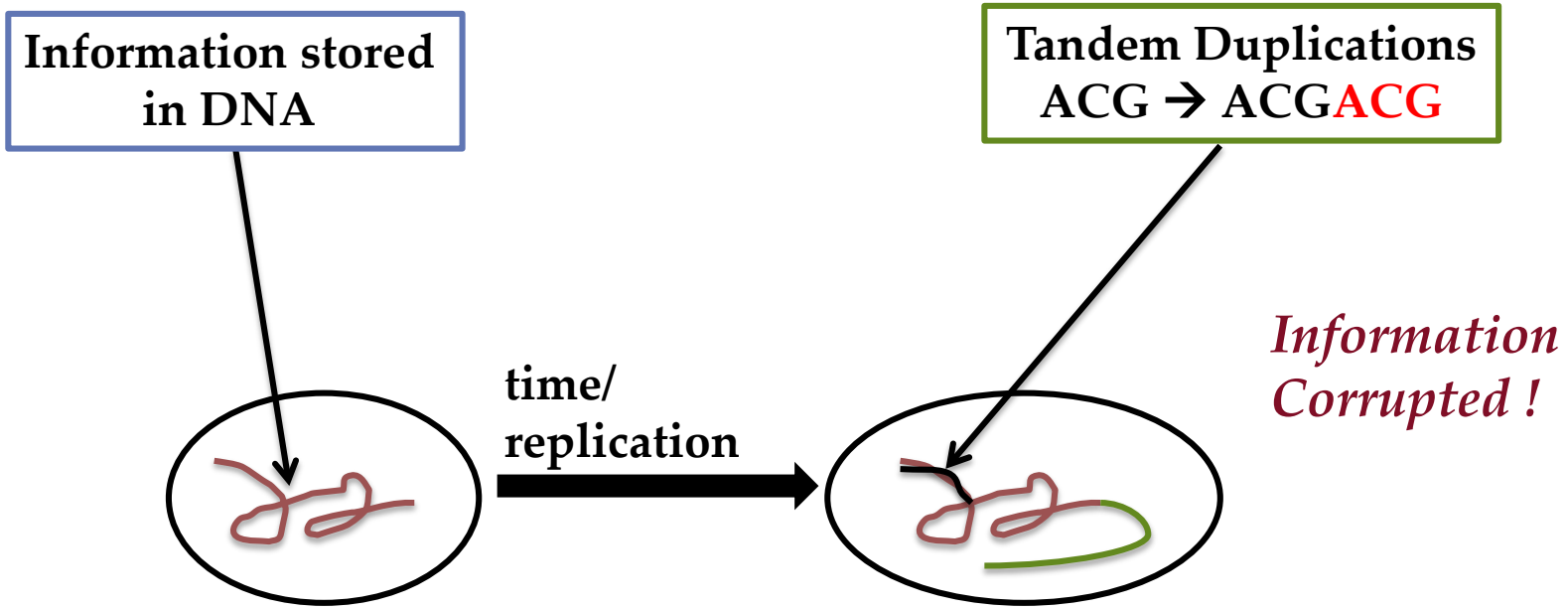
genetic information is stored in DNA

ex-vivo data storage

in-vivo data storage

Storage in live DNA


Information stored in DNA
Mutations
time/ replication
Information Corrupted !

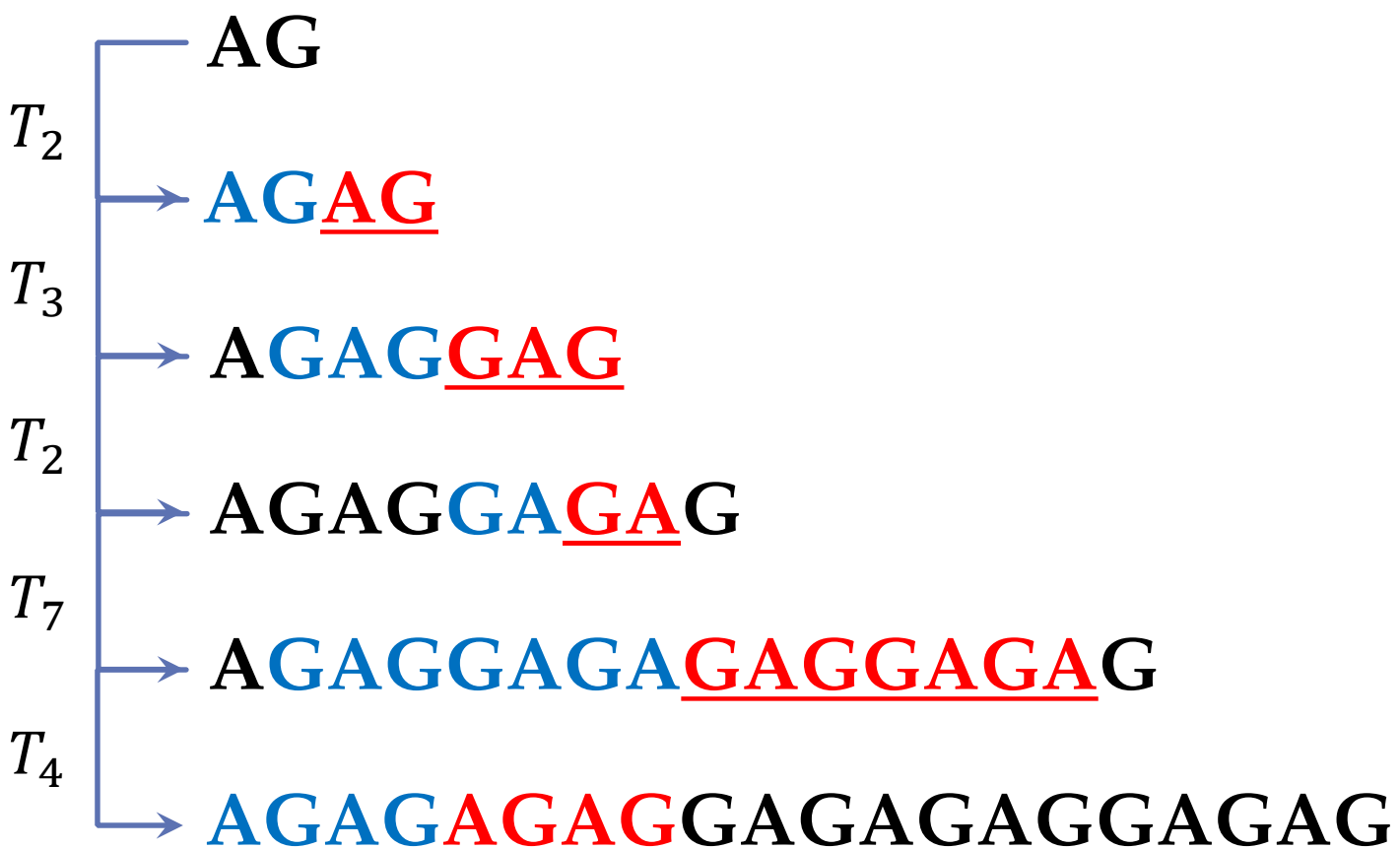

Information stored in DNA
Tandem Duplications
ACG → ACGACG
time/ replication
Information Corrupted !

Related Work

- Arita & Ohashi, 2004 – parity check bits
- Haider & Barenkow, 2007 - Hamming code or repetition code
- Yachie et. al, 2008 - copy data multiple times at different locations
- Haughton & Balado, 2013 - Coded for substitution

- Dolecek and Ananthram 2008 – Tandem duplication errors of length 1
- Mitzenmacher 2008 – Lower & upper bounds on sticky channel capacity

Tandem Duplications



Tandem Repeats in Genome

Channel Model

k-uniform Errors, T_k

Example : 2-uniform (T_2)

Input: x = ACGT

ACGT → ACG**CG**T → AC**AC**GCGT →
ACACGCGT**GT**

Output: y = ACACGCGTGT

k-bounded Errors, $T_{\leq k}$

Example : 4-bounded ($T_{\leq 4}$)

Input: x = ACGT

ACGT → ACG**CG**T → ACG**ACG**CGT →
A**A**CGACGCGT → AACGACGCGT**GCGT**

Output: y = AACGACGCGTGCGT

Decoding by deduplication

Encoding	• Repeat-free sequences
Decoding	• Remove all repeats

Decoding by Deduplication

Removing k-uniform errors

Example : 2-uniform (T_2)

Channel output: y = ACACGCGTGT

AC~~AC~~GCGTGT → ACG~~CG~~TGT → ACGT~~GT~~

Input estimate: $\hat{x}$ = ACGT

Decoding by Deduplication

Removing k-bounded errors

Example : 4-bounded ($T_{\leq 4}$)

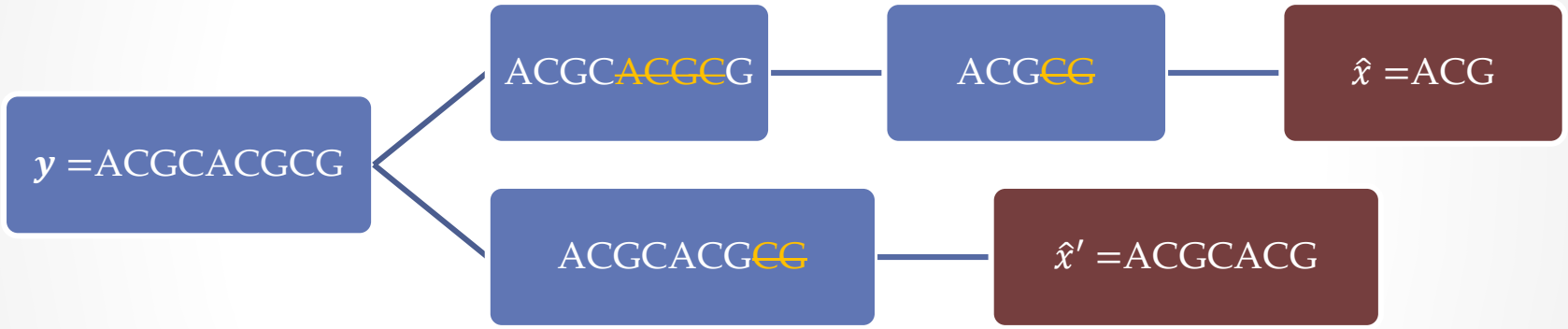
Channel output: y = AACGACGCGTGCGT

AAACGACGCGTGCGT → ACGACGCGTGCGT →
ACGCGTGCGT → ACGCGT

Input estimate: $\hat{x}$ = ACGT

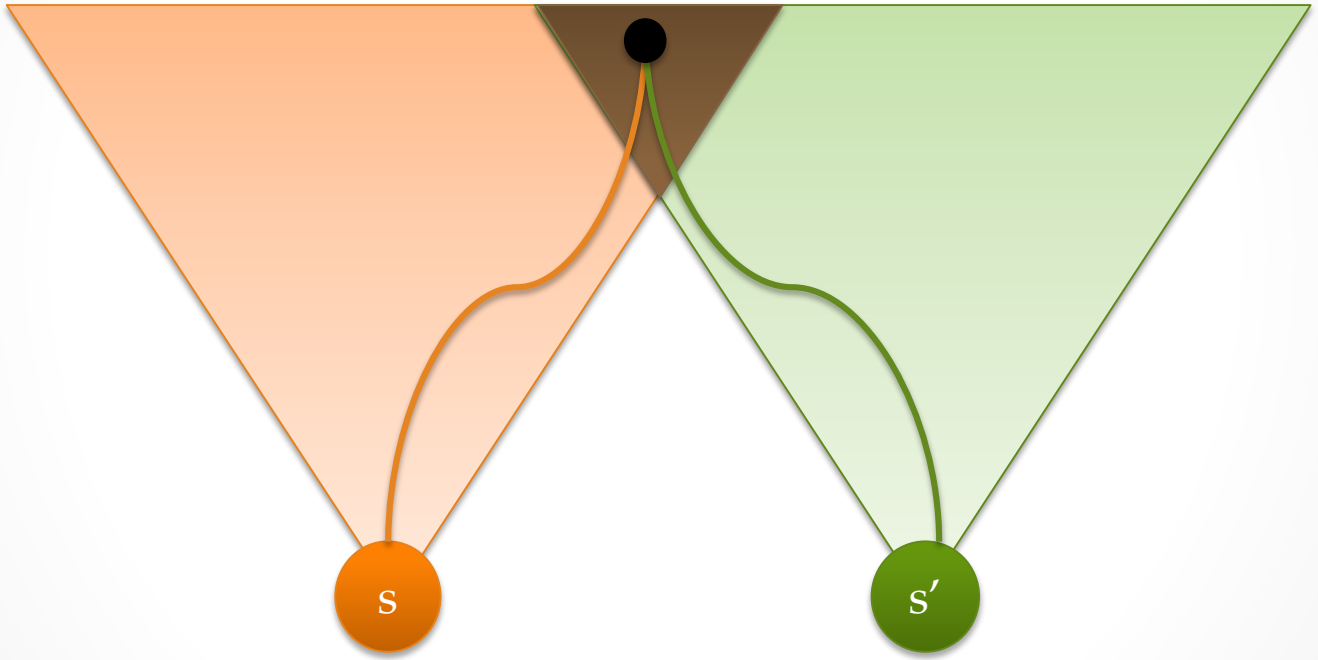
What Could Go Wrong?

Example: $T_{\leq 4}$



Root of s: repeat-free sequence that can be transformed to s via duplications

Duplication Cone



Uniqueness of Roots

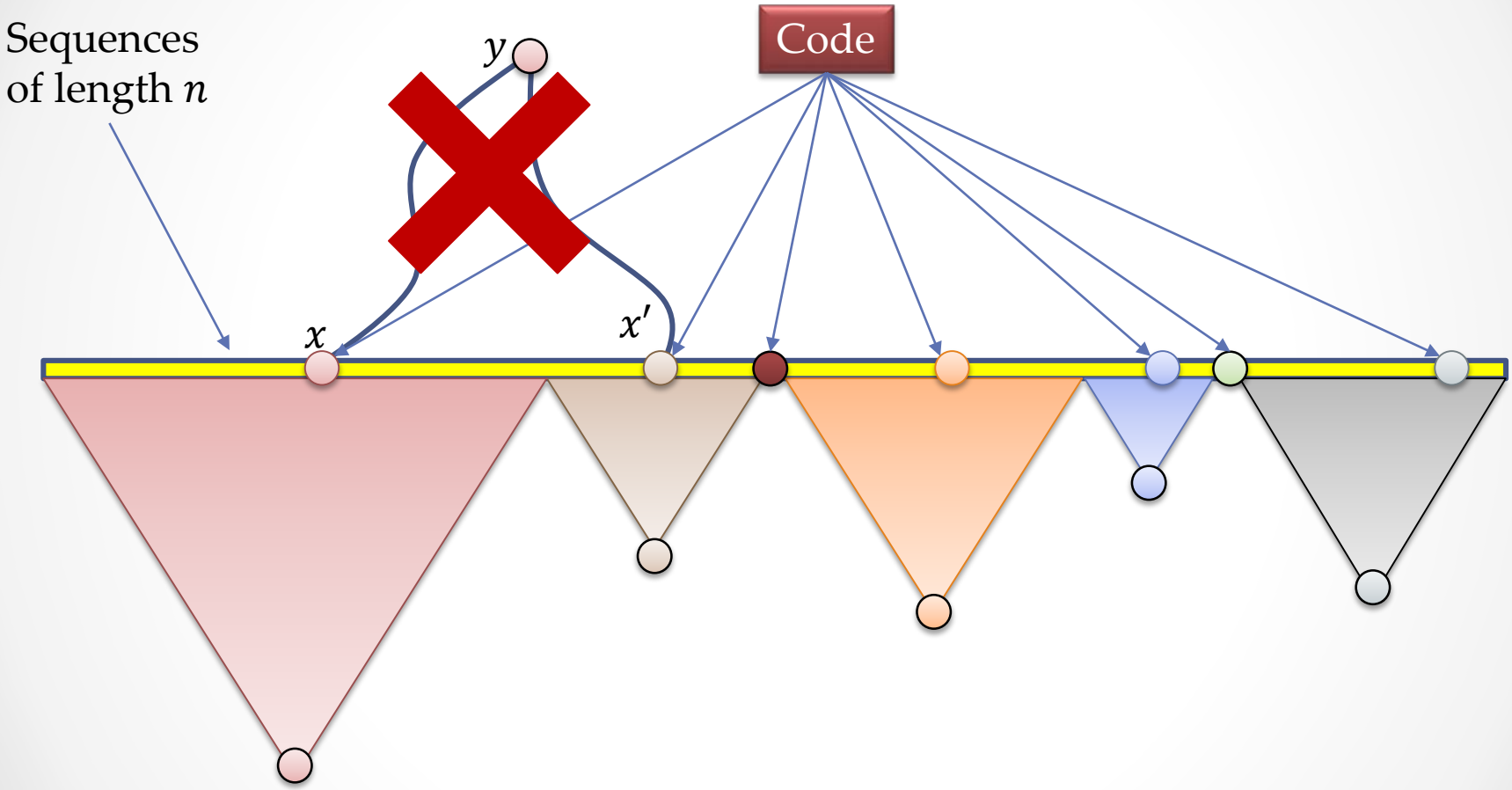
Theorem 1

For tandem duplication rule $\mathbf{T}_k$, *the root is unique for any* k.

Theorem 2

For tandem duplication rule $\mathbf{T}_{\leq k}$, *the root is unique for* $k \leq 3$.

$T_{k'}$ $T_{\leq 2'}$ $T_{\leq 3}$



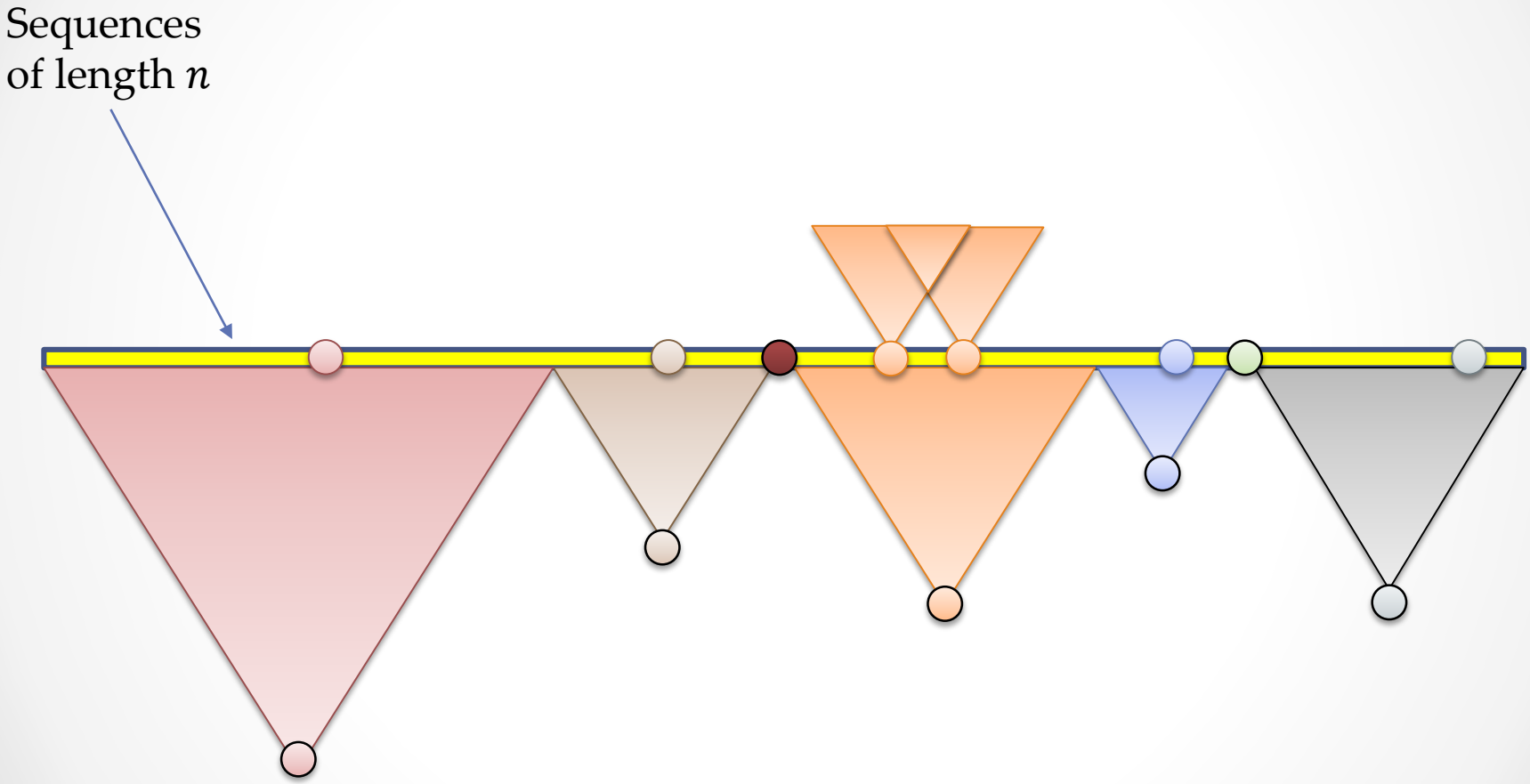
Codes for $T_k, T_{\leq 2}, T_{\leq 3}$ Channels

Extend each root to length n through T

Example: $T_2,\ n = 7, |\Sigma| = 4$:

ACTCTCT, AAAAAAA, CGGTATA, CATGCGA

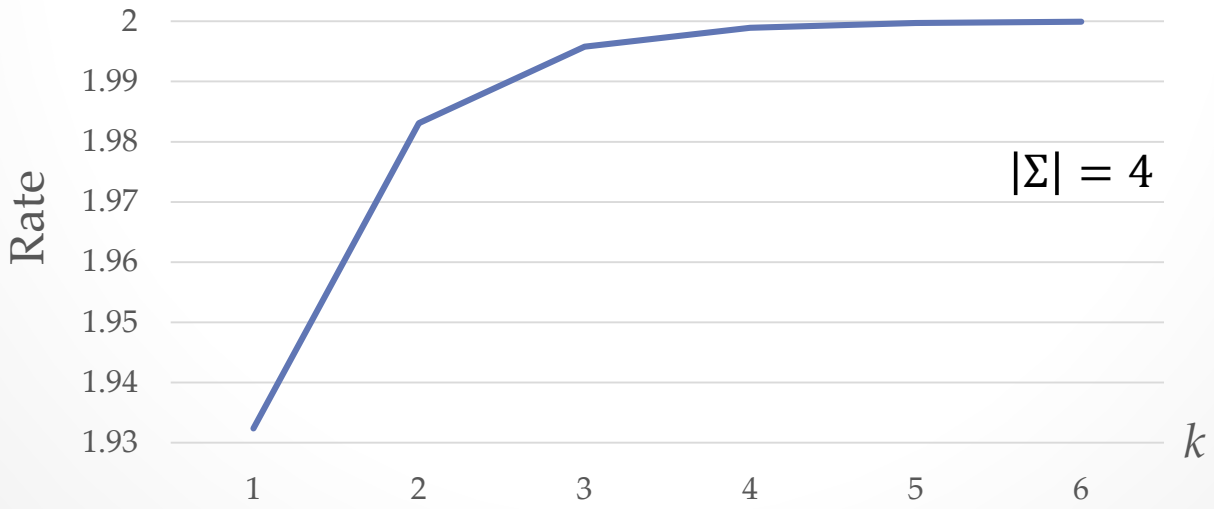
This code is optimal for T_k and $T_{\leq 2}$



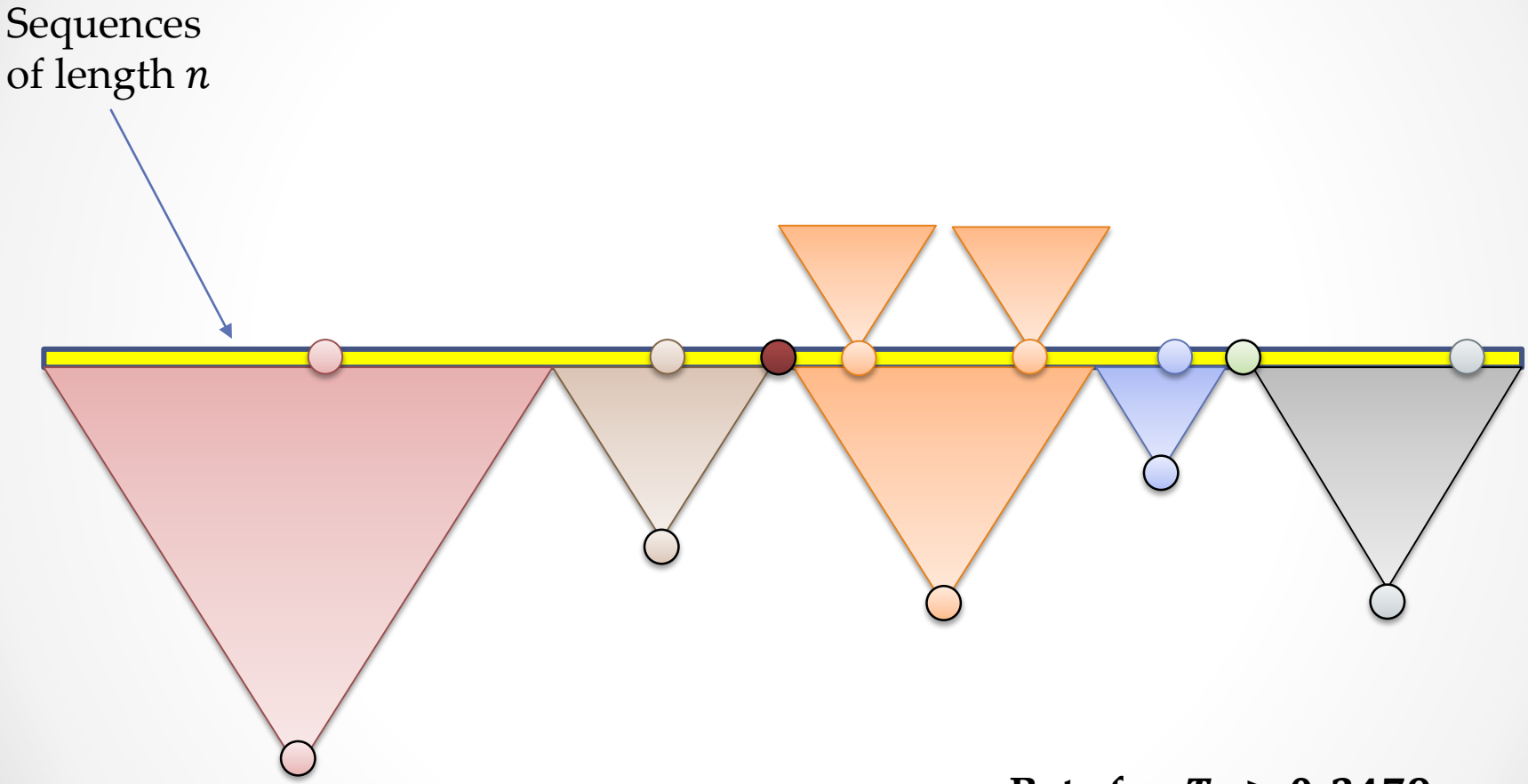
Codes for T_k Channel

Bijection: Roots ↔ RLL$(0, k-1)$

$$M = \sum_{i=0}^{\lfloor n/k \rfloor - 1} |\Sigma|^k M_{RLL(0,k-1)}(n-(i+1)k)$$



This code is not optimal for $T_{\leq 3}$



Rate for $T_3 \geq 0.3479$

Other Results

Construction: Optimal codes for t errors under T_k using codes in ℓ_1-metric

Theorem: Under T_U, the root is unique for all sequences if and only if

$\|\Sigma\| = 1$	$k \mid U$
$\|\Sigma\| = 2$	$U = \{k\}$ $U \supseteq \{1,2\}$
$\|\Sigma\| \geq 3$	$U = \{k\}$ $U \supseteq \{1,2\}$ $U \supseteq \{1,2,3\}$

Open Problems

- Optimal Code for ≤ 3 duplication error
- Codes for non-unique root regimes
- Codes for unbounded duplication error
- Code for duplication errors with point mutations