ISIT 2014, Hawaii

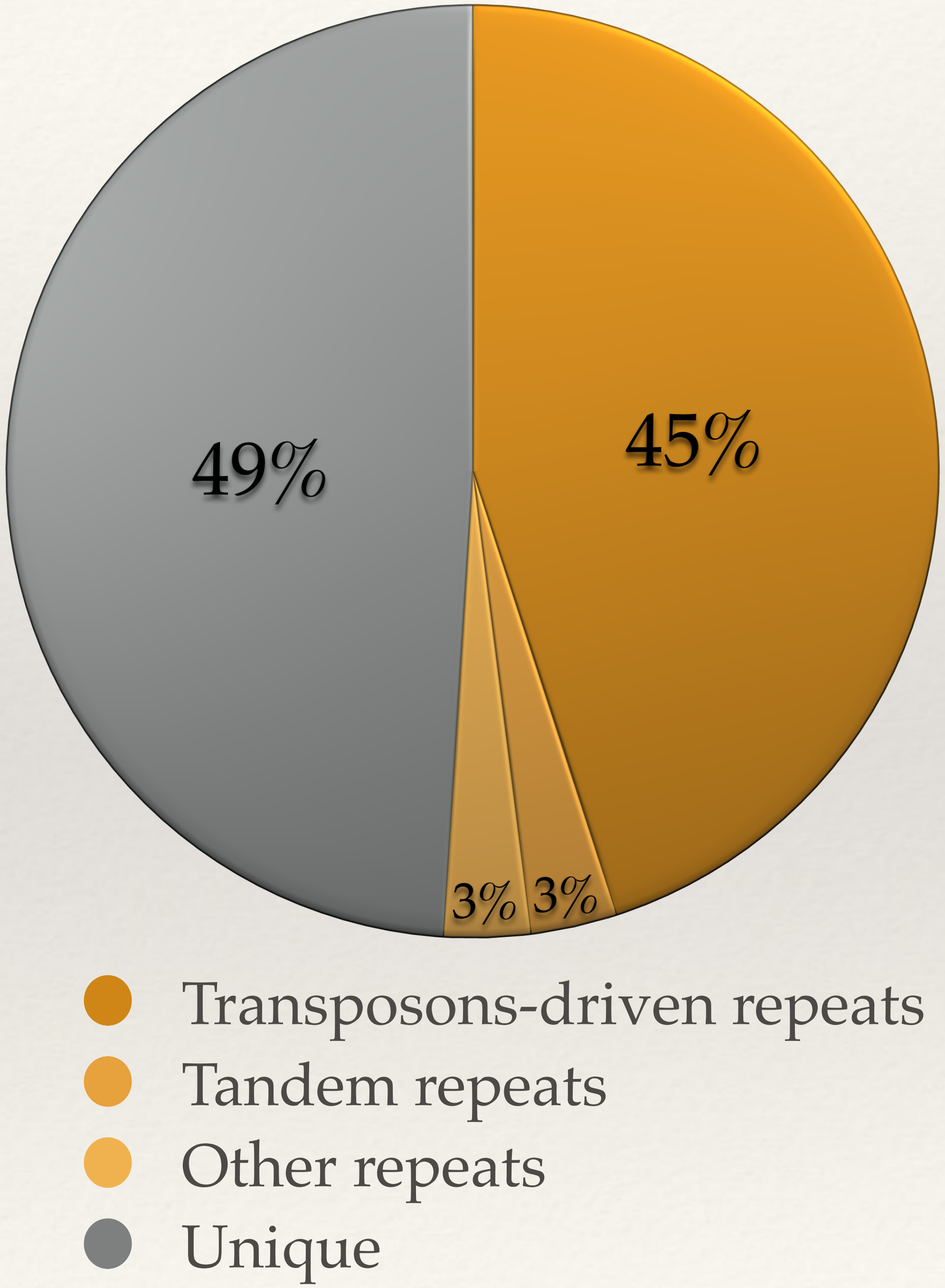
The capacity of String Duplication Systems

F. Farnoud🌴, M. Schwartz🏫, and J. Bruck🌴

🌴 California Institute of Technology
🏫 Ben-Gurion University of Negev

Repeated Sequences in Human Genome

- The majority of the human genome consists of repeated sequences.
  - Tandem repeats: TCATCATGCA
  - Transposon-driven repeats (interspersed repeats): TCATGCCATA
- Repeats provide a record of evolution and may cause chromosome fragility, expansion diseases, gene silencing, etc.



Expressive Power of Repetitions

- "Much of the remaining 'unique' DNA must also be derived from ancient transposable element copies that have diverged too far to be recognized as such." [Lander et al. Nature 2001]

- We investigate the possibility of generating a diverse family of sequences using repetitions.

- We take an information theoretic point of view by studying the capacity of *string duplication systems*.

String Duplication Systems

- A string system S is identified by the tuple (s,F) :
  - F: family of duplication rules
  - s: starting string on some alphabet A e.g., $A=\{0,1\}$ or $A=\{G,C,A,T\}$
    - $\delta(s)$ is the number of distinct symbols in s.
- The system S contains all sequences obtained by starting with s and applying functions $f \in F$
- The *capacity* of S is given by

$$\mathsf{cap}(S) = \limsup_{n\to\infty} \frac{\log_2 |S \cap A^n|}{n \log_2 \delta(s)}$$

- Note that $0 \leq \mathsf{cap}(S) \leq 1$.

Duplication Rules

- Four types of duplications
  - **End duplication**: TCATGC→ TCATGCCAT
  - **Tandem duplication**: TCATGC→ TCATCATGC
  - **Reversed tandem duplication**: TCATGC→ TCATTACGC
  - **Duplication with a gap**: TCATGC→ TCATGCATC
- Parameters: length of duplicate k, gap k'
- We find the capacity or bounds on the capacity of string systems with above duplication rules.

End Duplication Has Capacity 1

- Let $F_{k,\mathrm{end}}$ denote the set of functions that duplicate a k-substring and append it to the end
  - TCATGC→ TCATGCCAT (k=3)
- End duplication is relatively simple and easy to analyze.

Theorem: For any positive integer k, and $S=(s,F_{k,\mathrm{end}})$, we have

$$\mathrm{cap}(S)=1$$

End Duplication Has Capacity 1: Proof

Theorem: For any positive integer k, and $S=(s,F_{k,\mathrm{end}})$, we have $\mathrm{cap}(S)=1$

- Proof outline:
  - First, form a string with a substring of length $k\delta(s)^k$ that contains all possible k-substrings in a finite number of steps.
    - For $s=AGT$, and $k=2$: …AAAGATGAGGGTTATGTT…
  - Then, in each duplication step any k-substring can be duplicated.
    - For $s=AGT$, and $k=2$: …AAAGATGAGGGTTATGTT…AT

Tandem Duplication Has Capacity 0

- Let $F_{k,\mathrm{tan}}$ denote the set of functions that duplicate a k-substring and insert the duplicate immediately after the original copy.
  - T<u>CAT</u>GC→ T<u>CATCAT</u>GC (k=3)
- Capacity of tandem-duplication systems is in complete contrast to end-duplication systems:

Theorem: For any positive integer k, and $S=(s,F_{k,\mathrm{tan}})$, we have

$$\mathrm{cap}(S)=0.$$

Tandem Duplication Has Capacity 0: Proof

Theorem: For any positive integer k, and $S=(s,F_{k,\tan})$, $\mathrm{cap}(S)=0$.

- View a string of length n as a sequence of $n-k+1$ overlapping *circular* k-substrings (*super-symbols*).

GCATGC

- With this mapping, every duplication is equivalent to adding k identical super-symbols.

GCATCATGC

- The number of possible sequences can be shown to grow only polynomially.

Tandem Duplication with Non-uniform Length

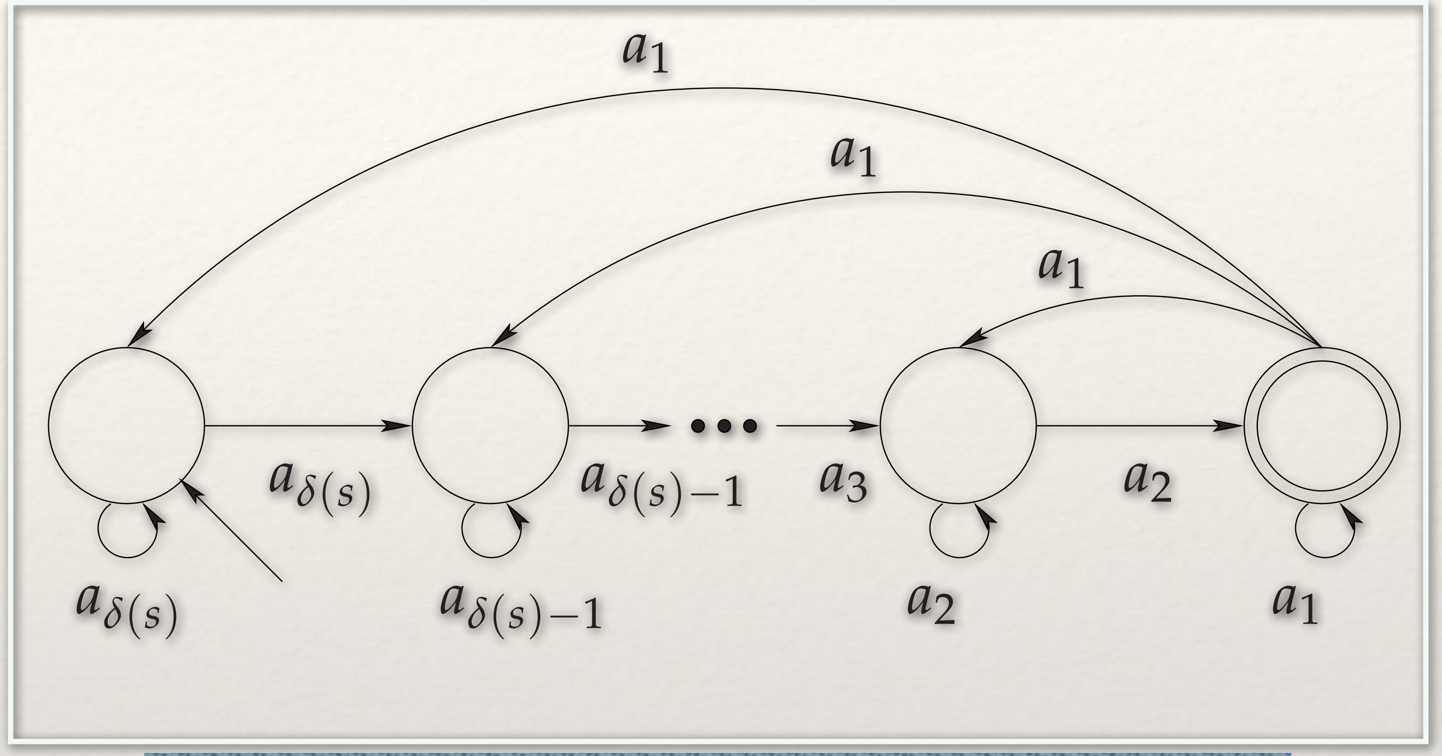
- Let $F_{\geq k,\tan}$ denote the set $\{F_{i,\tan} : i \geq k\}$.
  - T<u>CAT</u>GC→ T<u>CATCAT</u>GC→ TCA<u>TCTC</u>ATGC (k=2)
- Unlike tandem duplication with fixed length, the capacity is nonzero.

Theorem: For a nontrivial string s, let $S=(s,F_{\geq k,\tan})$ and $S'=(s,F_{\geq 1,\tan})$. We have $\mathrm{cap}(S)>0$ and $\mathrm{cap}(S')\geq \log_2(r+1)/\log_2\delta(s)$, where r is the largest (real) root of

$$x^{\delta(s)} - \sum_{i=0}^{\delta(s)-2}$$

Tandem Duplication with Non-uniform Length

- Outline of proof for $S'=(s,F_{\geq 1,\tan})$:
  - We show that S' has a regular language as a subset.
  - The capacity of this regular language is a lower bound for the capacity of S'



Finite-state automaton representing the regular language.

  - The number length n words in the regular language is the number of length n paths in the automaton.
  - The capacity of the regular language is the largest eigenvalue of the adjacency matrix of the finite-state automaton.

Reverse Tandem Duplication has Nonzero Capacity

- Let $F_{k,\mathrm{rt}}$ denote the set of functions that duplicate a k-substring and insert it immediately after the original copy in reverse.
  - T<u>CAT</u>GC→ T<u>CATTAC</u>GC (k=3)
- While allowing reversing the copy is seemingly a small change, unlike tandem duplication, reverse tandem duplication has nonzero capacity.

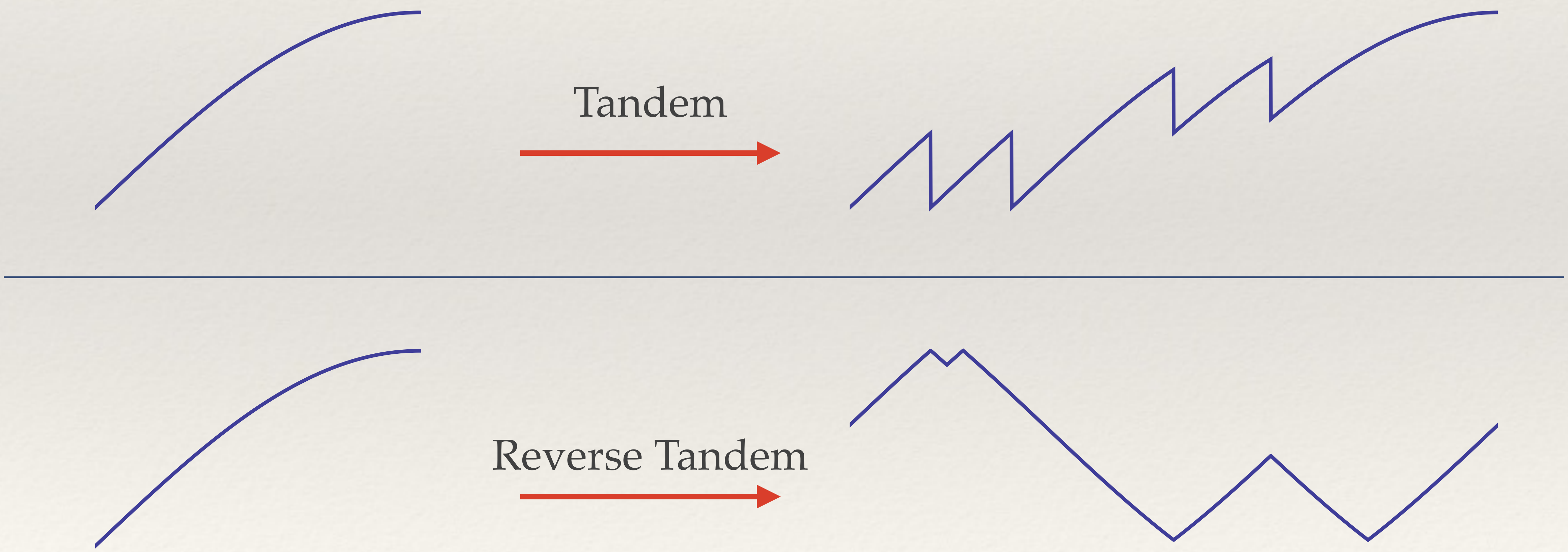
Theorem: For any positive integer k, and $S=(s,F_{k,\mathrm{rt}})$, we have

$$\mathrm{cap}(S)>0,$$

unless $\delta(s)=1$. Furthermore, capacity depends on s, only through $\delta(s)$.

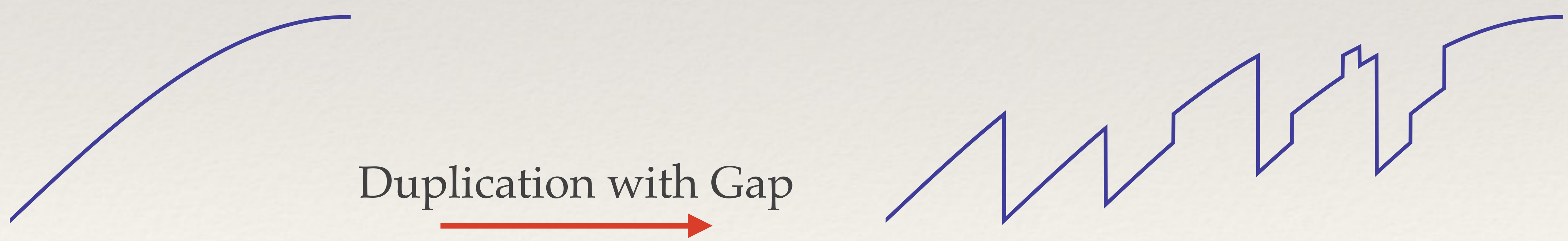
(Reverse) Tandem Duplication

- The main difference between tandem and reverse tandem duplication is that the former leads to near-periodic behavior with period *k*, but the latter does not.



Duplication with Gap

- Let $F_{k,k',\text{gap}}$ denote the set of functions that duplicate a k-substring and insert it k positions after the original copy.
  - T<u>CAT</u>GC→ T<u>CAT</u>G<u>CAT</u>C (k=3,k'=1)



Results on Duplication with Gap

Theorem: The capacity of S=(s, $F_{k,k',\mathrm{gap}}$) is zero if and only if s is periodic with period gcd(k,k').

Theorem: There are non-trivial strings s such that for S=(s, $F_{k,k',\mathrm{gap}}$) we have 0<cap(S)<1.

Theorem: If gcd(k,k')=1, then the capacity of S=(s, $F_{k,k',\mathrm{gap}}$) depends on s only through $\delta(s)$.

Conclusion

- We studied the expressive power of languages generated by different duplication rules from an information theoretic point of view.
- We showed rules that can produce nearly any sequence and rules that can produce a small set of sequences.
- These results *suggest* that it is plausible to have diverse genomic sequences solely using repetition.
- Rules that lead to interspersed repeats (end, gap) are generally more powerful than rules leading to tandem repeats (with fixed length/gap).
- Ongoing work: a probabilistic framework leading that takes into account the probabilities of different sequences and not only their count.