MPP, January 2015

Biological Diversity through Duplication

F. Farnoud 🌴
with
M. Schwartz🔥, J. Bruck🌴

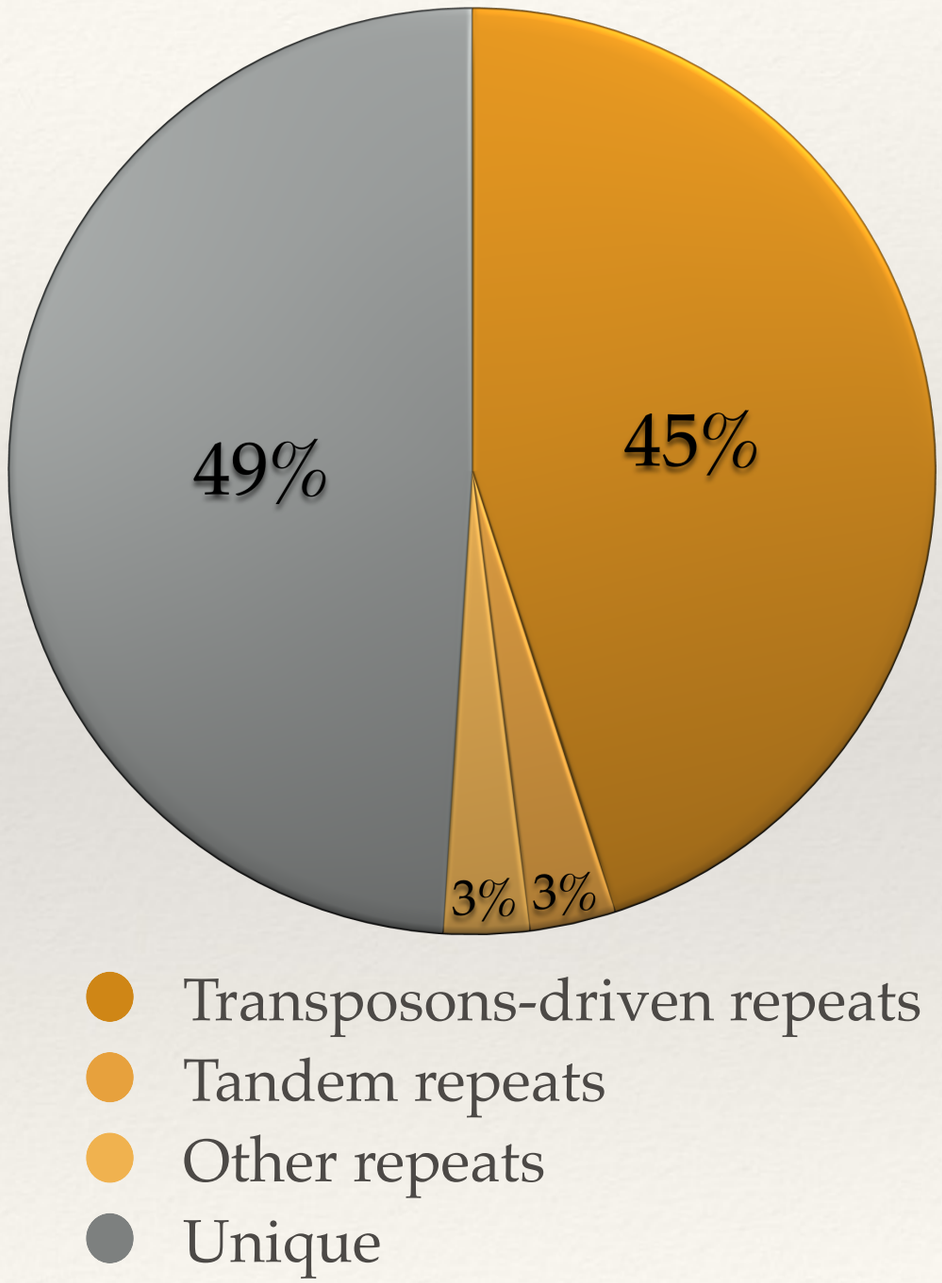
🌴 California Institute of Technology
🔥 Ben-Gurion University of Negev

Mutation and Diversity

- Mutation plays a crucial role in evolution by creating diversity.
- Types of mutation include:
  - Point mutation: TCATGCA → TGATGCA
  - Deletion/insertion: TCATGCA → TCATAGCA
  - **Duplication**:
    - Tandem: TCATGCA → TCATCATGCA
    - Transposon driven: TCATGCA → TCATGCATCA
- A mathematical study of duplication systems.

Repeated Sequences in Human Genome

- The majority of the human genome.
- *"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]*



[Lander et al. Nature 2001]

Repeated Sequences in Human Genome

- The majority of the human genome.
- *"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]*
- Cause chromosome fragility, expansion diseases, gene silencing, *rapid morphological changes*.

1931

1950

1976

[Myers et al., Nature 2007]

Role of Duplication in Generating Diversity

- Is it possible/probable to generate a diverse family of sequences by duplication?
- Two types of models:
  - **Combinatorial**: Study of what is *possible*.
    - Information theoretic view: capacity and expressiveness of *string duplication systems*.
  - **Stochastic**: Study of what is *probable*.
    - Asymptotic properties of likely outcomes.

A Tandem Duplication String System

- String system: *S*(seed, rule)
- Example:

  AGT

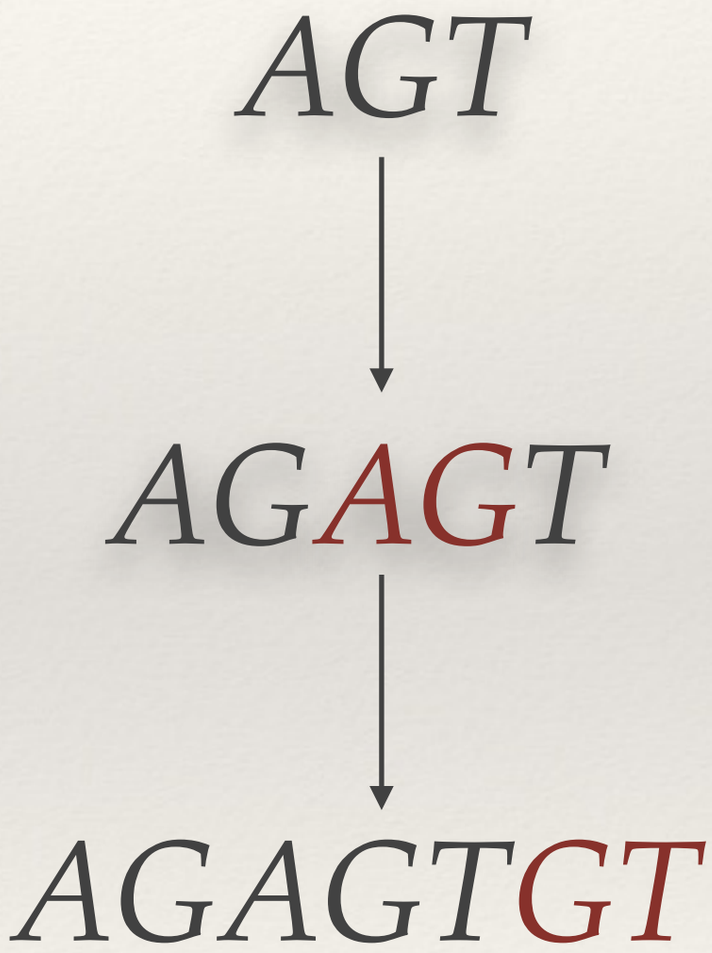
  - Seed = AGT
  - Rule: a substring of length 2 may be duplicated in tandem

A Tandem Duplication String System

- String system: *S*(seed, rule)
- Example:
  - Seed = AGT
  - Rule: a substring of length 2 may be duplicated in tandem

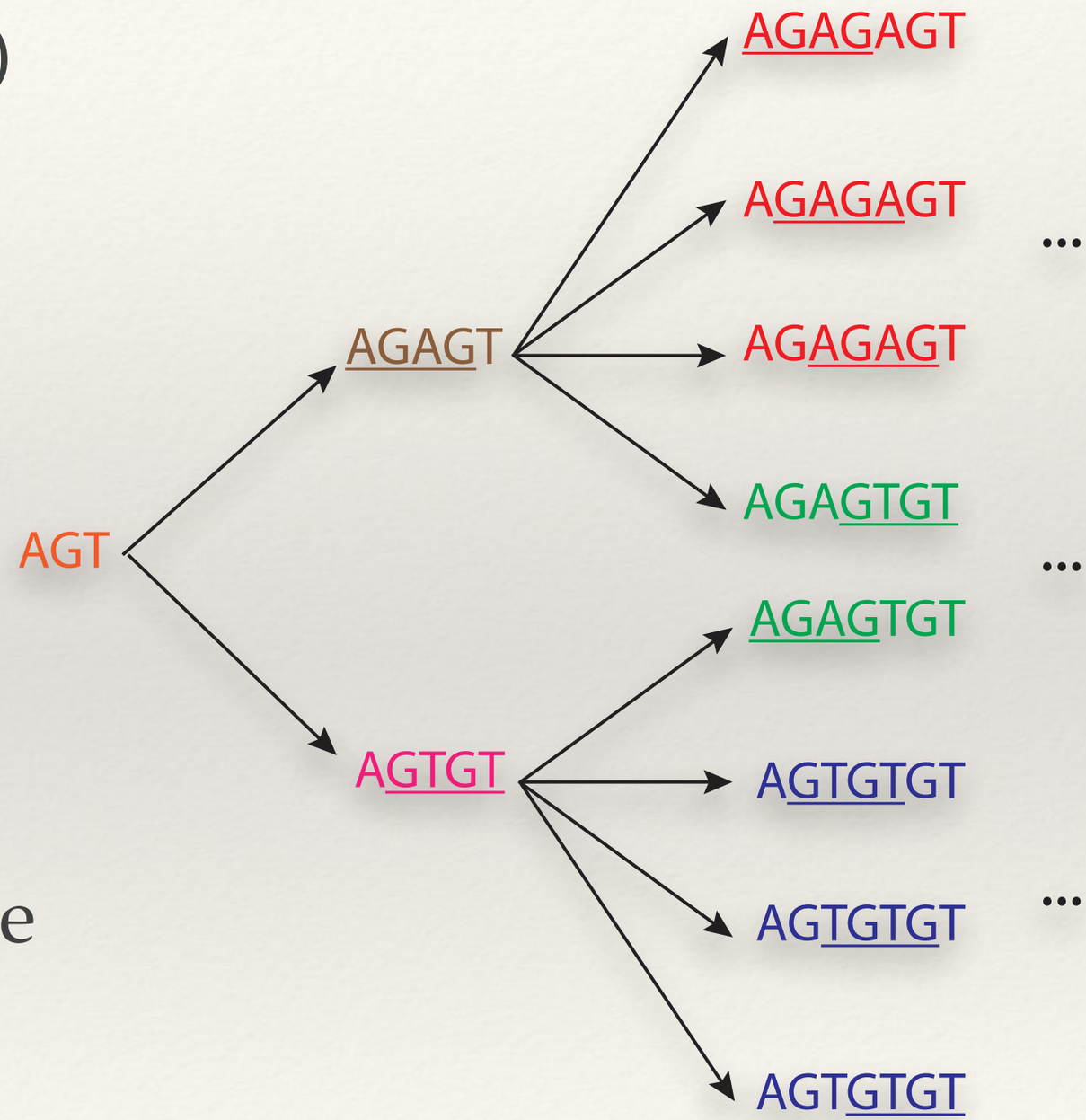
A Tandem Duplication String System

- String system: S(seed, rule)
- Example:
  - Seed = AGT
  - Rule: a substring of length 2 may be duplicated in tandem



A Tandem Duplication String System

- String system: S(seed, rule)
- Example:
  - Seed = AGT
  - Rule: a substring of length 2 may be duplicated in tandem
- System contains all possible strings.



Capacity and Expressiveness

- The *capacity* of S is given by

$$\mathsf{cap}(S) = \limsup_{n\to\infty} \frac{\log(\#\text{ strings of length } n \text{ in } S)}{n}$$

  - Base of log = #distinct symbols in *seed*
  - Used to measure the amount of information a constrained coding system can store.
- A string system is *fully expressive* if it can generate *every* string as a substring.
- Full capacity (capacity = 1) ⇒ Fully expressive

Duplication Rules

Duplication Rules

- Duplication rules:

Duplication Rules

- Duplication rules:
  - **Tandem duplication**: TCATGC→ TCATCATGC

Duplication Rules

- Duplication rules:
  - **Tandem duplication**: TCATGC→ TCATCATGC
  - **Reverse tandem duplication**: TCATGC→ TCATTACGC

Duplication Rules

- Duplication rules:
  - **Tandem duplication**: TCATGC→ TCATCATGC
  - **Reverse tandem duplication**: TCATGC→ TCATTACGC
  - **Displaced duplication**: TCATGC→ TCATGCATC

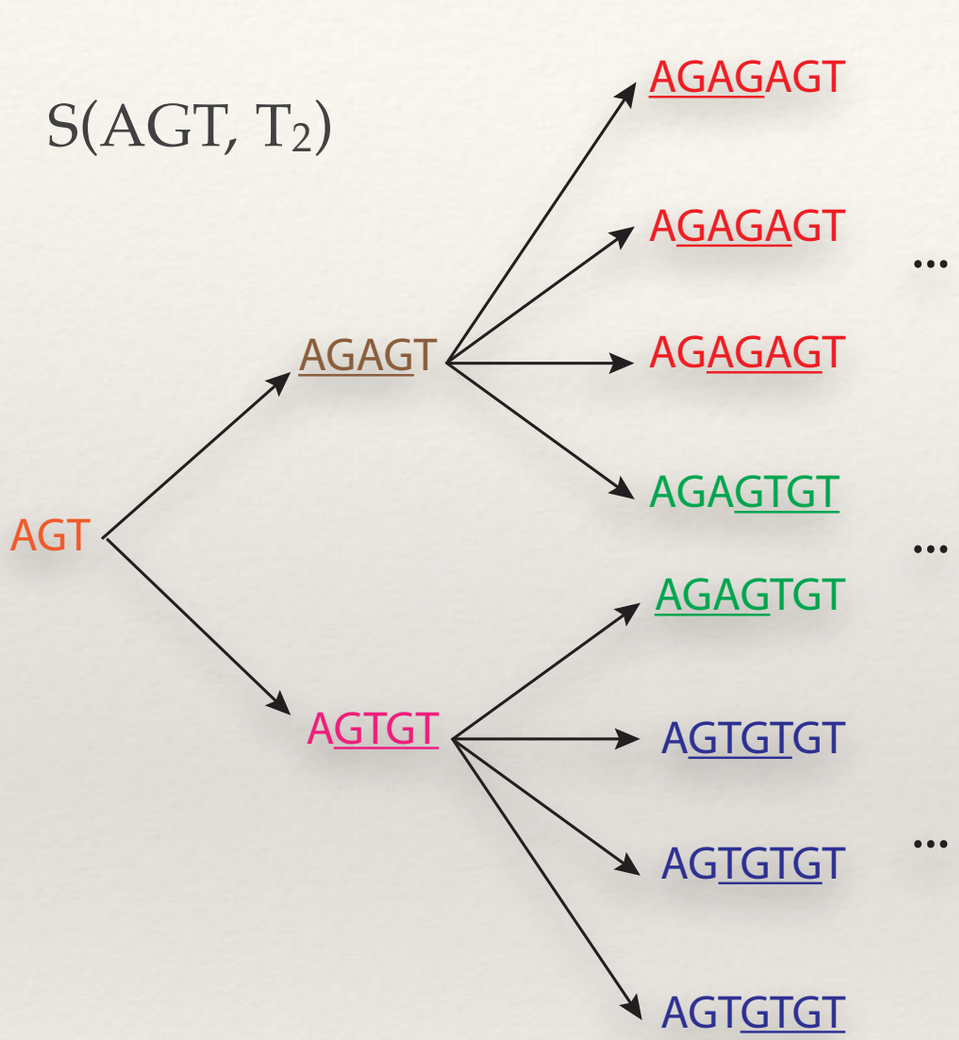
Duplication Rules

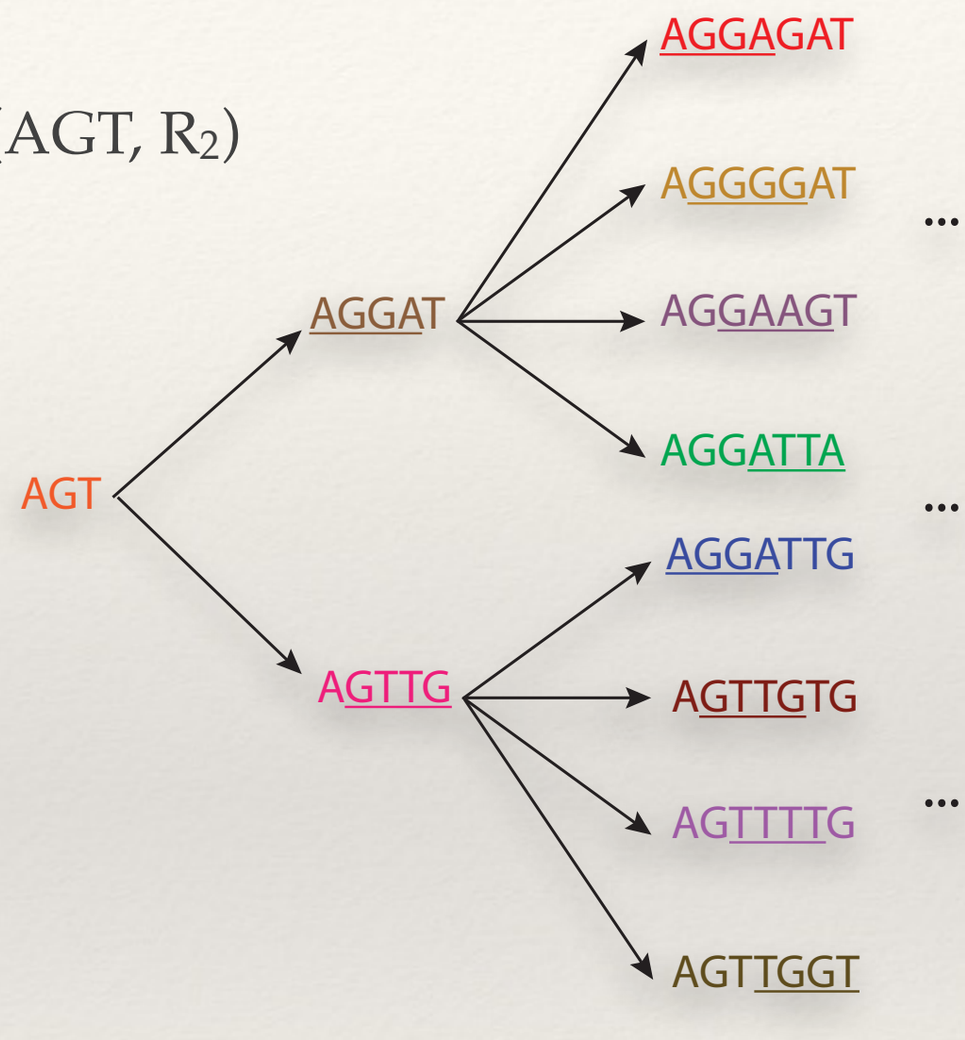
- Duplication rules:
  - **Tandem duplication**: TCATGC→ TCATCATGC
  - **Reverse tandem duplication**: TCATGC→ TCATTACGC
  - **Displaced duplication**: TCATGC→ TCATGCATC
- Parameters: length of duplicate *k*, gap *k'*

Duplication Rules

- Duplication rules:
  - **Tandem duplication**: TCATGC→ TCATCATGC
  - **Reverse tandem duplication**: TCATGC→ TCATTACGC
  - **Displaced duplication**: TCATGC→ TCATGCATC
- Parameters: length of duplicate *k*, gap *k'*
- Tandem duplication studied in literature: [Dassow'99,'02], [Leupold'04,'05]: Concerned with position in Chomsky hierarchy of formal languages.

A Tale of Two Systems

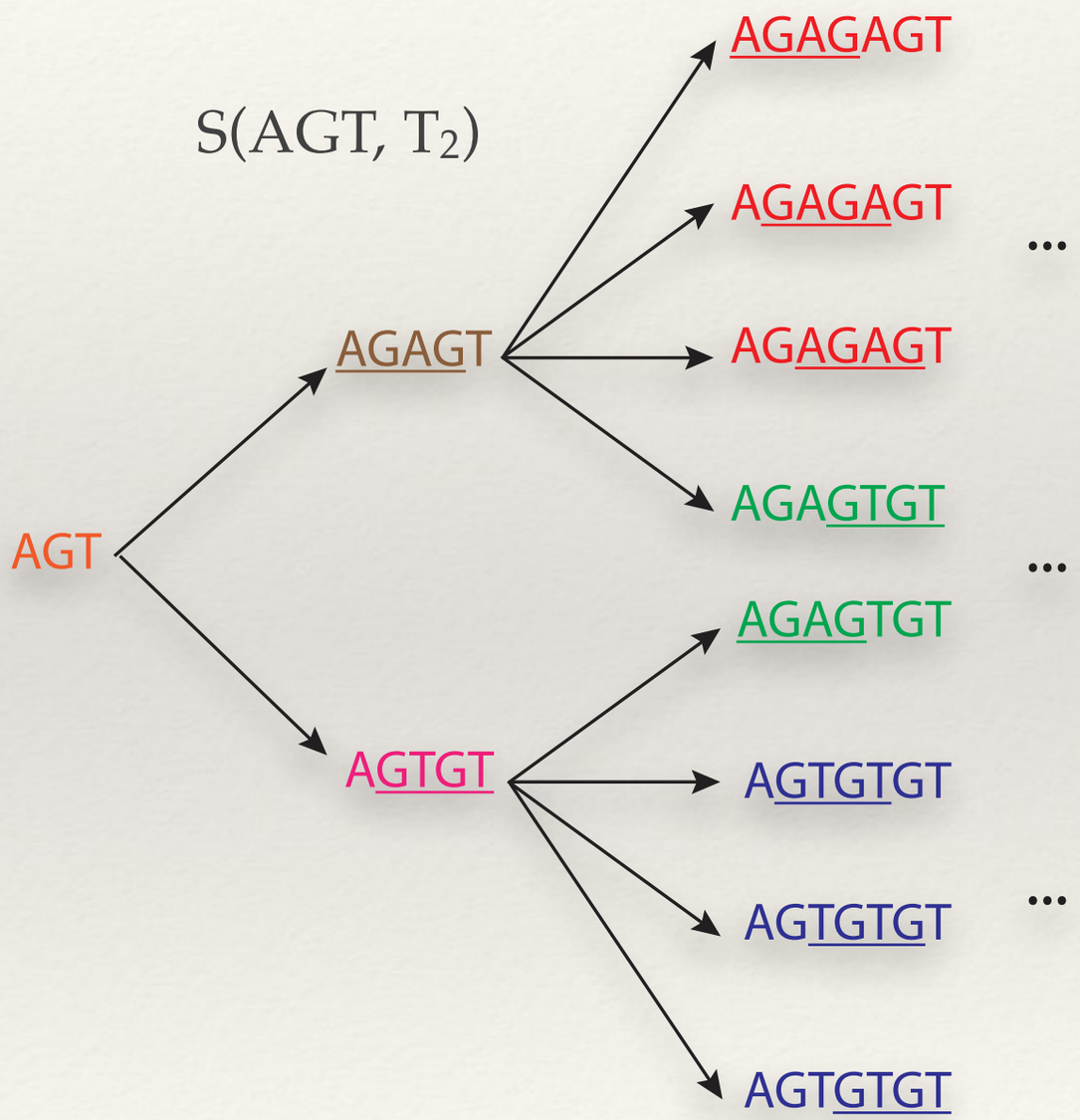




- **Capacity** measures the ability of systems to generate a large number of sequences.
- $S(AGT, T_2)$ is not fully **expressive** (e.g., cannot have GAT as a substring), but $S(AGT, R_2)$ is fully expressive (e.g., AGGATTG is in the system).

Tandem Duplication

- T_k: Tandem duplication of length k
  - TCATGC→ TCATCATGC (k=3)

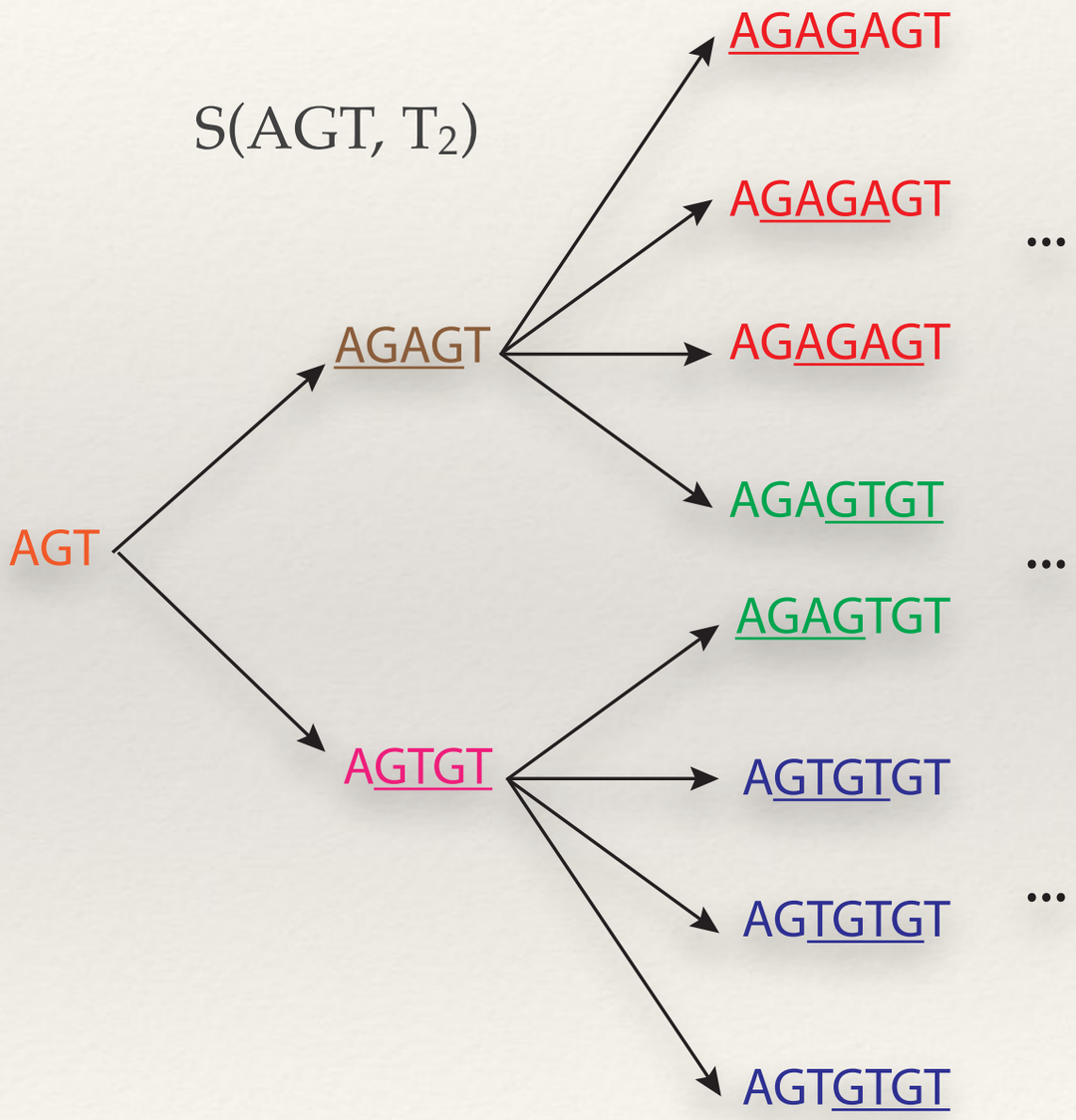


Tandem Duplication

- T_k: Tandem duplication of length k
  - TCATGC→ TCATCATGC (k=3)

For any s and positive integer k, $S(s,T_k)$ has capacity zero!

- Number of sequences grows polynomially.
- It is **never** fully expressive.



Tandem Duplication with Variable Length

- $T_{\geq k}$ = Tandem duplication of length $\geq k$
  - $T_{\geq 2}$: TCATGC→ TCATCATGC→ TCATCTCATGC

Tandem Duplication with Variable Length

- $T_{\geq k}$ = Tandem duplication of length $\geq k$
  - $T_{\geq 2}$: TCATGC→ TCATCATGC→ TCATCTCATGC

For a nontrivial string s, $S(s,T_{\geq k})$ is *fully expressive* and $\text{cap}(S(s,T_{\geq k}))>0$.

Tandem Duplication with Variable Length

- $T_{\geq k}$ = Tandem duplication of length $\geq k$
  - $T_{\geq 2}$: TCATGC→ TCATCATGC→ TCATCTCATGC

For a nontrivial string s, $S(s,T_{\geq k})$ is *fully expressive* and $\mathrm{cap}(S(s,T_{\geq k}))>0$.

Also, $\mathrm{cap}(S(s,T_{\geq 1}))\geq \log(r+1)$, where r is the largest (real) root of $x^{\delta}-(1+x+\cdots+x^{\delta-2})$, and δ=# distinct symbols in s.

$\delta(s)$	2	3	4	5
$\mathrm{cap}(S')\geq$	1	0.77	0.65	0.58

*See the poster by **Siddhartha Jain** for tandem duplications with short duplication lengths.*

Tandem Duplication with Variable Length Example

A substring from chromosome 1 of human genome

GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGG
GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGG
GGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGG
GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGG
GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGG
GGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

Tandem Duplication with Variable Length Example

Tandem Duplication with Variable Length Example

GTCTG → GTCTGCTG → GTGTGTGTCTGCGCTG → GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

Tandem Duplication with Variable Length Example

GTCTG → GTCTGCTG → GTGTGTGTCTGCGCTG → GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

Tandem Duplication with Variable Length Example

GTCTG → GTCTGCTG → GTGTGTCTGCGCTG → GGGGTGTGGTGGTCTGCGGGCCCTGGGGGGG

→GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

Tandem Duplication with Variable Length Example

GTCTG → GTCTGCTG → GTGTGTCTGCGCTG → GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGT
G TGGTGGGGTCTGCGGGGCCCTGGGGGGG

Tandem Duplication with Variable Length Example

GTCTG → GTCTGCTG → GTGTGTGTCTGCGCTG → GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGT
G TGGTGGGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGT
GTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTG
GGGTCTGCGGGGCCCTGGGGGGG

Tandem Duplication with Variable Length Example

GTCTG → GTCTGCTG → GTGTGTGTCTGCGCTG → GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTG TGGTGGGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

Tandem Duplication with Variable Length Example

GTCTG → GTCTGCTG → GTGTGTGTCTGCGCTG → GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGT
G TGGTGGGGTCTGCGGGGCCCTGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGT
GTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTG
GGGTCTGCGGGGCCCTGGGGGGG

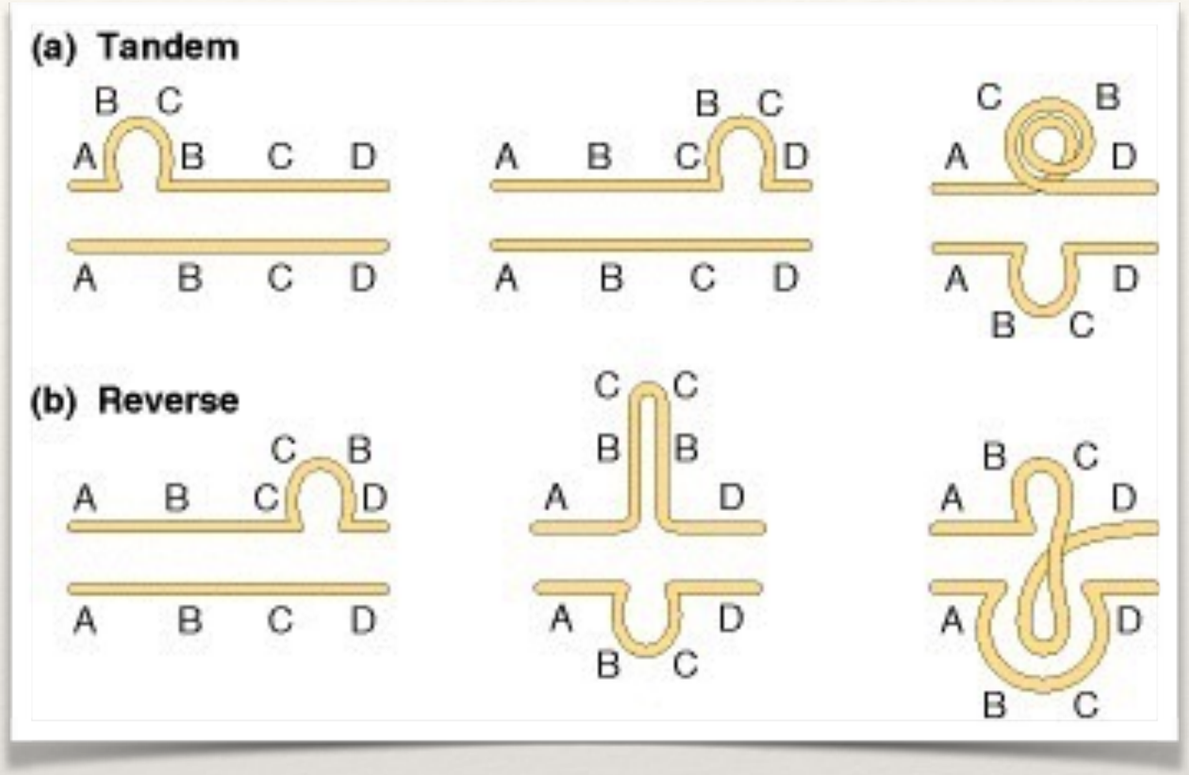
→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGT
GGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCC
CTGGGGGGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTG
GTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGG
GGCCCTGGGGGGGG

→GGGGTGTGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTG
TGGTGTGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGG
TCTGCGGGGCCCTGGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGGTGTGGTGTGGTCTGCGGG
GCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGTGGTCTGCGGGGCCCTGGG
GGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGGTGTGGTGGGGTCTGCGGGGCCCTGGGGGGG

Reverse Tandem Duplication

- R_k: the duplicate of length k is inserted in *reverse*.
  - R_3 : T<u>CAT</u>GC→

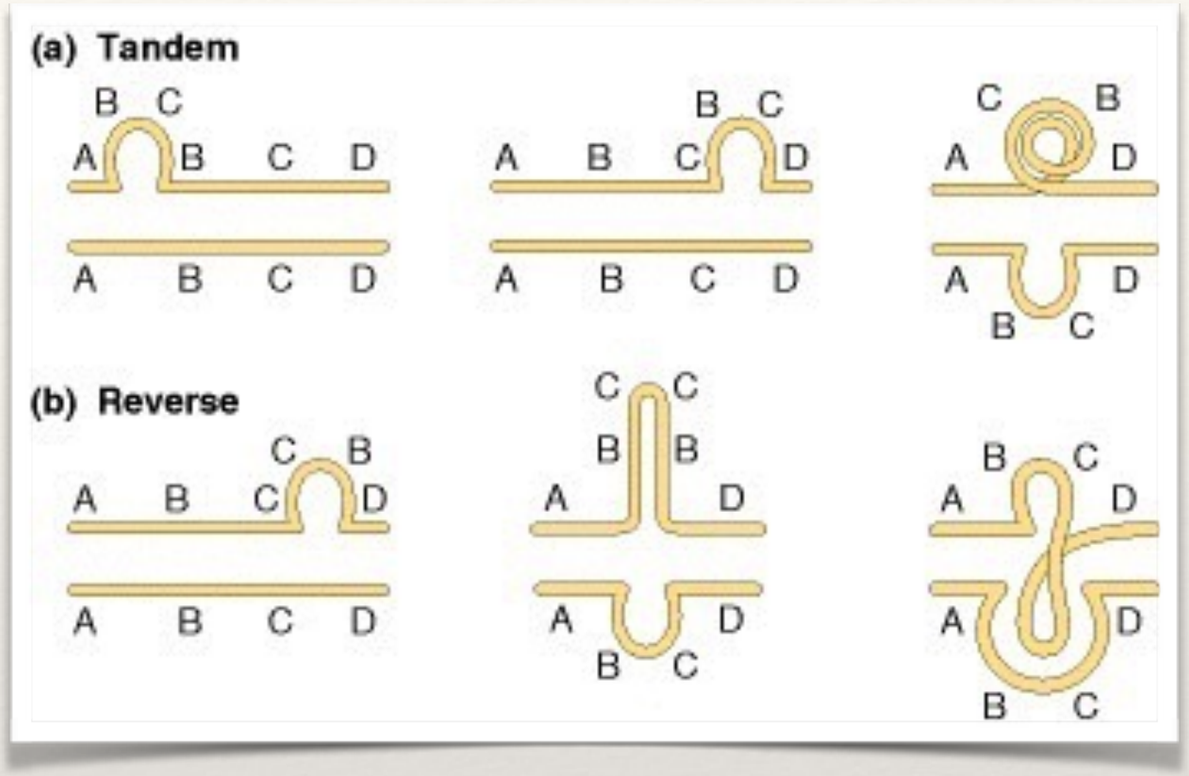
T<u>CAT</u><u>TAC</u>GC



[Griffiths et al, An Introduction to Genetic Analysis, 2000]

Reverse Tandem Duplication

- R_k: the duplicate of length k is inserted in *reverse*.
  - R_3 : T<u>CAT</u>GC→ T<u>CAT</u><u>TAC</u>GC

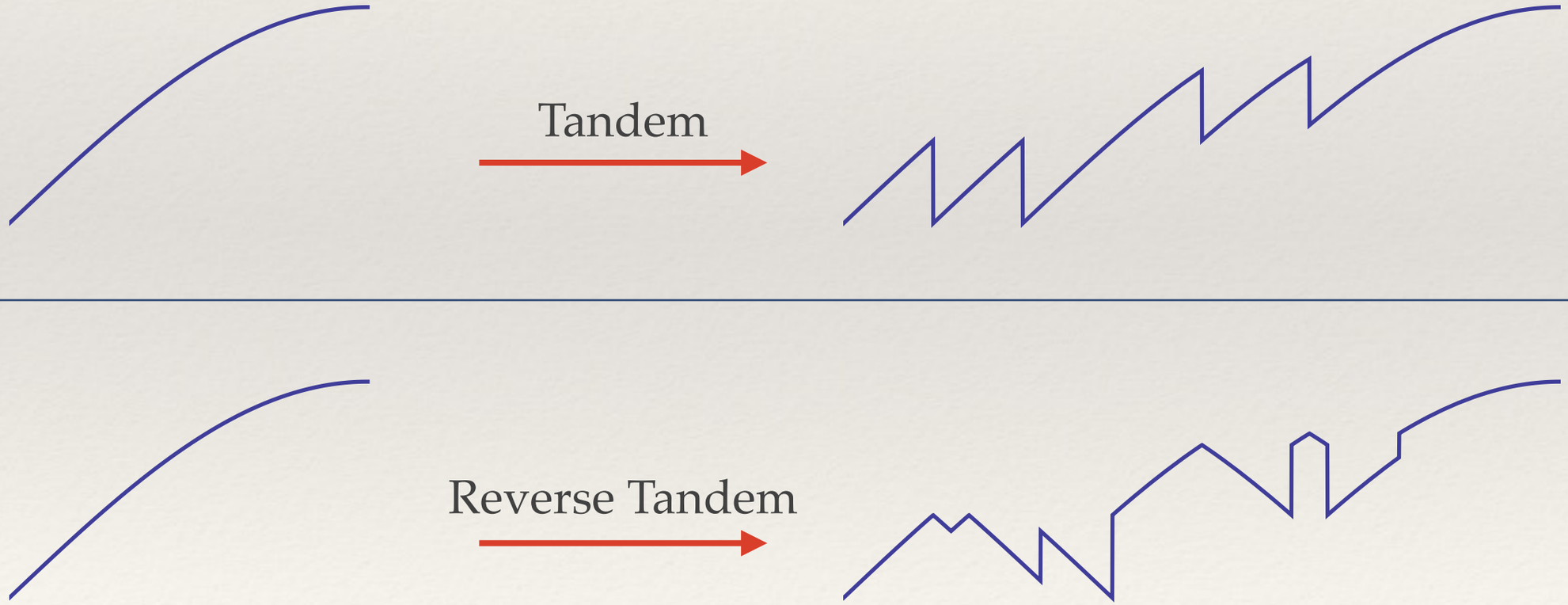


[Griffiths et al, An Introduction to Genetic Analysis, 2000]

For nontrivial s and positive integer $k>1$, and $S=(s,R_k)$, we have $cap(S)>0$ and S is fully expressive.

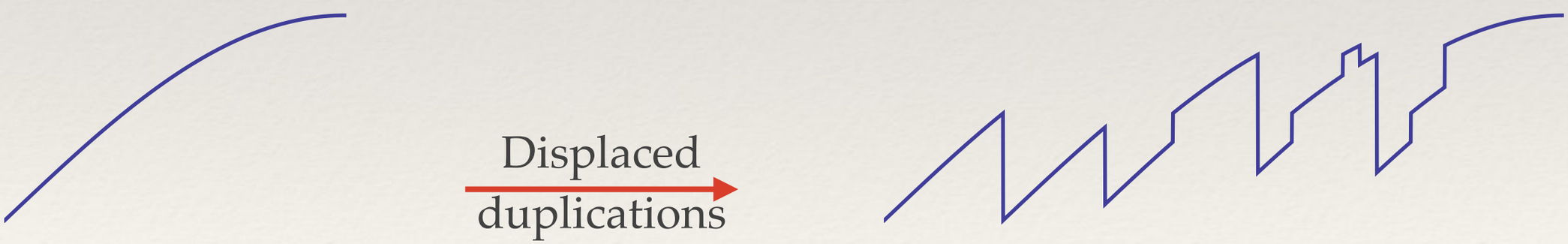
Tandem vs Reverse Tandem

- 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.



Displaced Duplication

- $D_{k,k'}$: Duplicates a k-substring and inserts it k' positions after original copy.
  - TCATGC→ TCATGCATC (k=3,k'=1)



Displaced Duplication

The capacity of $S=(s, D_{k,k'})$ is zero if and only if s is periodic with period gcd(k,k').

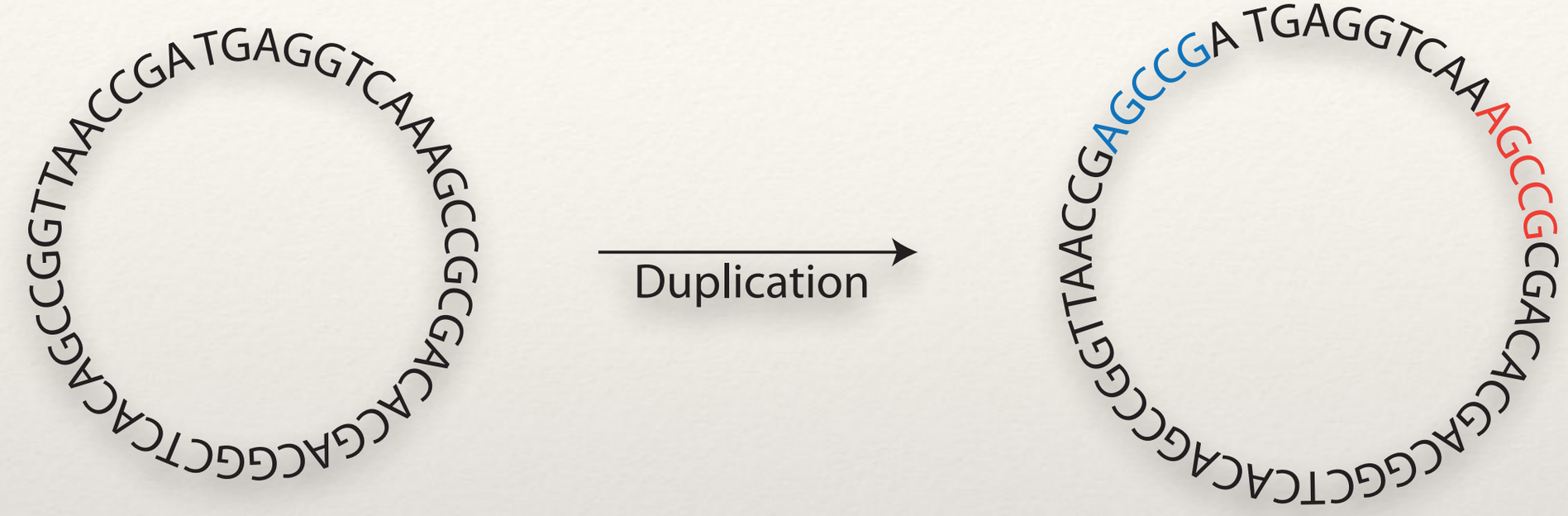
- "if" direction: if s is periodic with period gcd(k,k'), then so is every other string in S:
  - $k=2$, $k'=4$, $s=AGAGAGAG \Rightarrow S=\{(AG)^m : m \geq 4\}$

If gcd(k,k')=1, then S is fully expressive.

Summary of Results for Combinatorial Models

	0	$0 < \text{cap}(S) < 1$	1	Expressive
Tandem k	✓	✗	✗	✗
Tandem $\geq k$	✗	?	?	✓
Reverse Tandem	✗	?	?	✓
Displaced (k,k')	✓	✓	?	✓✗

Stochastic Duplication Systems



- s: evolving string
- At each time step, a substring of length l is duplicated.
- Length l of duplicated string is bounded, has distribution q_l.
- The position of the copy depends on type of duplication.
- After a long time, what does the outcome look like?

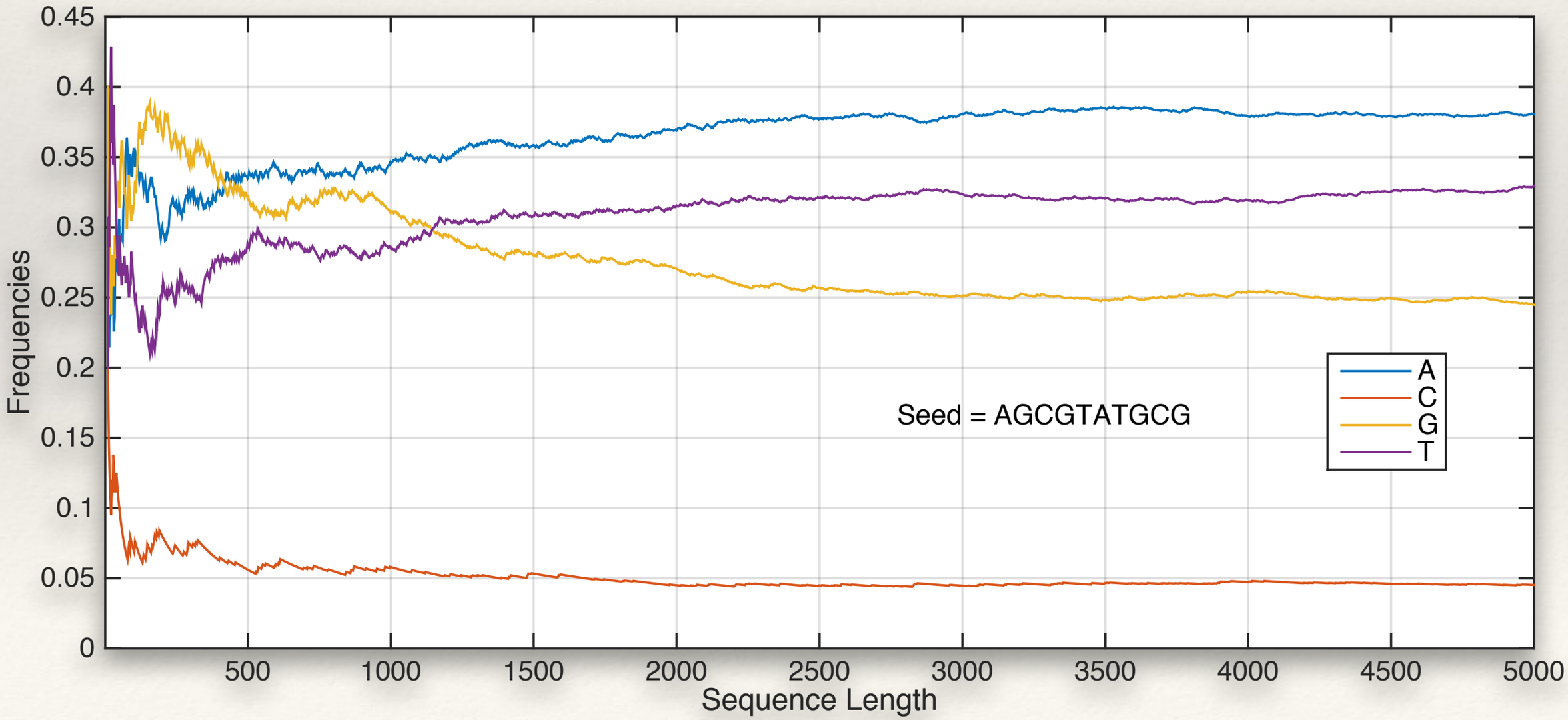
Symbol Frequencies

- $\mu_n(a)$: # occurrences of a in s at time n
- L_n: length of s at time n
- $x_n(a) = \mu_n(a)/L_n$: frequency of a in s at time n
- How do symbol frequencies change in duplication systems?
  - Convergence?
  - Dominated by one symbol?

Symbol Frequencies

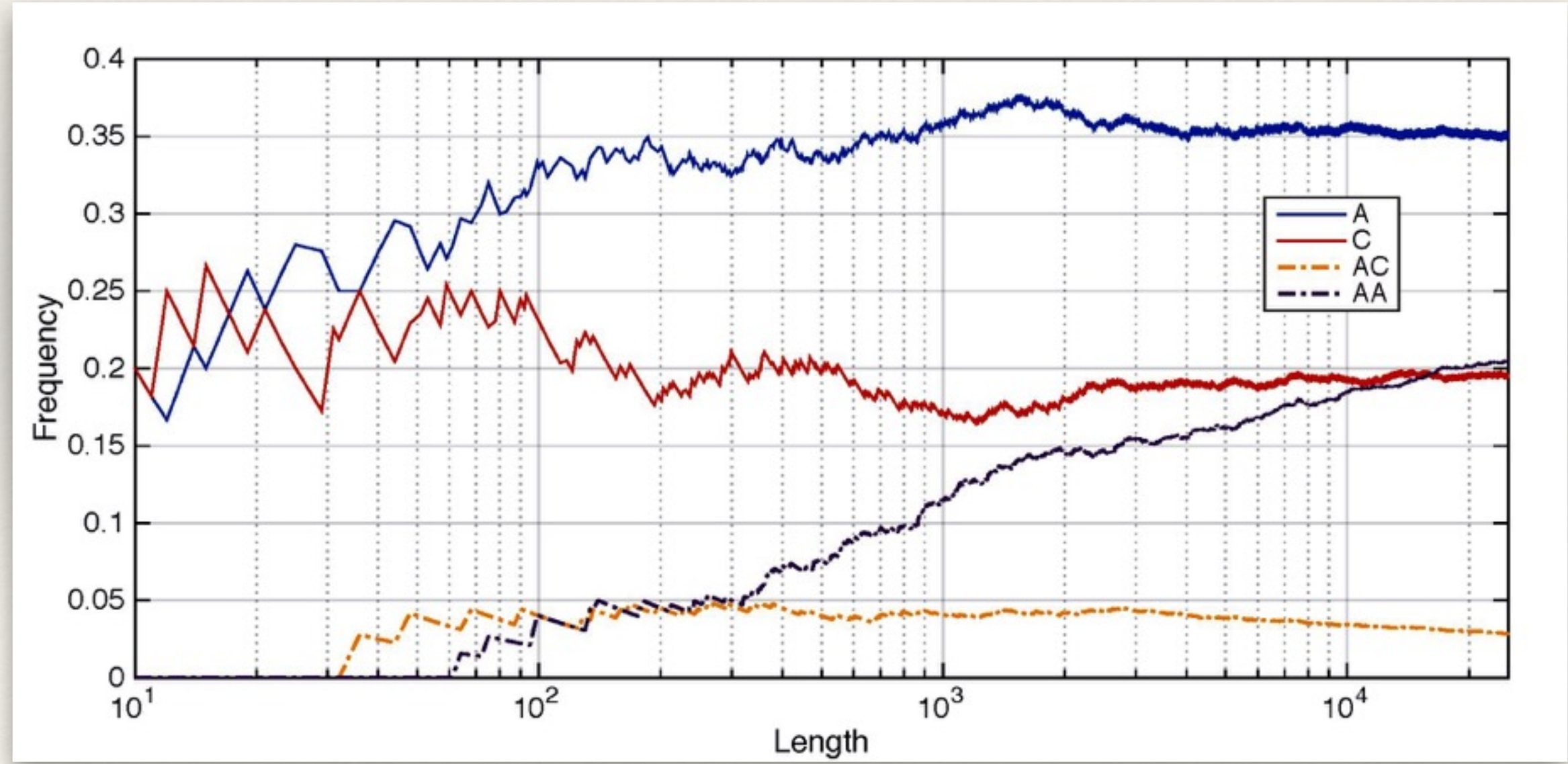
- Symbol frequencies are *martingales* and converge almost surely. If K is the bound on distribution q_l,

$$P(|x_n(u) - x_0(u)| \geq \lambda) \leq 2e^{-\lambda^2 L_0^2/(2K^4)}.$$



String Frequencies

- How do frequencies of strings (length>1), e.g. AA and AC, change? Not martingale!



Stochastic Approximation

Suppose a discrete random process x satisfies:

$$x_{n+1} - x_n = a_n(h(x_n) + M_{n+1})$$

for a Lipschitz function h, martingale difference M, and

$$\sum a_n = \infty, \quad \sum a_n^2 < \infty, \quad \text{e.g., } a_n = \frac{1}{n}.$$

Then x_n converges almost surely to a compact connected internally chain transitive invariant set of the ode

$$\dot{x}_t = h(x_t).$$

Tandem Duplication

- Start with a seed s. Repeat:
  - Choose a bounded length $l \sim q_l$.
  - Choose random l-substring of s and insert a copy after original.
- Example: Seed = 0110, $T_{1,2}$, for n=100:
  01110000100111111111000000000000000000001111111111111111111111111010100111
  11111110111111010100000010101011011001111101111010011000000011010100000000000

Tandem Duplication: Autocorrelation

- Autocorrelation of s: $R(r) = \sum_{i=1}^{|s|} \mathbb{I}(s_i = s_{i+r})$
- $\rho_n = \frac{1}{L_n}(R_n(0), \ldots, R_n(m-1))$.

- If we choose m such that q_i=0 for $i \geq m$ then

$$\dot{\rho} = A\rho,$$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & \\ q_1 & -1+q_2 & q_3 & q_4 & \\ q_2 & q_1+q_3 & -1+q_4 & q_5 & \\ q_3 & q_2+q_4 & q_1+q_5 & -1+q_6 & \cdots \\ & & \vdots & & \end{pmatrix}$$

Tandem Duplication

- **Eigenvalues**: either 0 or negative imaginary part → Stable system.
- **Null space**: depends only on the set of indexes i with q_i.
  - All vectors in null space are periodic with period $\gcd\{i{:}q_i{>}0\}$.

Tandem Duplication

- **Eigenvalues**: either 0 or negative imaginary part → Stable system.
- **Null space**: depends only on the set of indexes i with q_i.
  - All vectors in null space are periodic with period $\gcd\{i{:}q_i{>}0\}$.

$\rho_n(d)$ converges almost surely to 1. That is, with high probability, every two symbols at distance d are the same—the string becomes locally periodic.

Tandem Duplication

- **Eigenvalues**: either 0 or negative imaginary part → Stable system.
- **Null space**: depends only on the set of indexes i with q_i.
  - All vectors in null space are periodic with period $\gcd\{i{:}q_i{>}0\}$.

$\rho_n(d)$ converges almost surely to 1. That is, with high probability, every two symbols at distance d are the same—the string becomes locally periodic.

- Example: seed = ACGTCATG, with $q_4 = q_6 = ½ \rightarrow d{=}2$.
- At n=15000:

…TATGTGTATGTGTATGTGTATGTGTGTGTGTGTGTGTATGTGTATGTGTAT…

Conclusion

- We presented tandem duplication systems that have nonzero capacity and are fully expressive: capable of creating diversity.
- Stochastic tandem systems asymptotically show periodic behavior: many novel sequences not a likely outcome.
- Stochastic interspersed systems are indistinguishable from iid systems in certain respects.
- These results *suggest* that it is plausible to generate diverse genomic sequences using duplications.