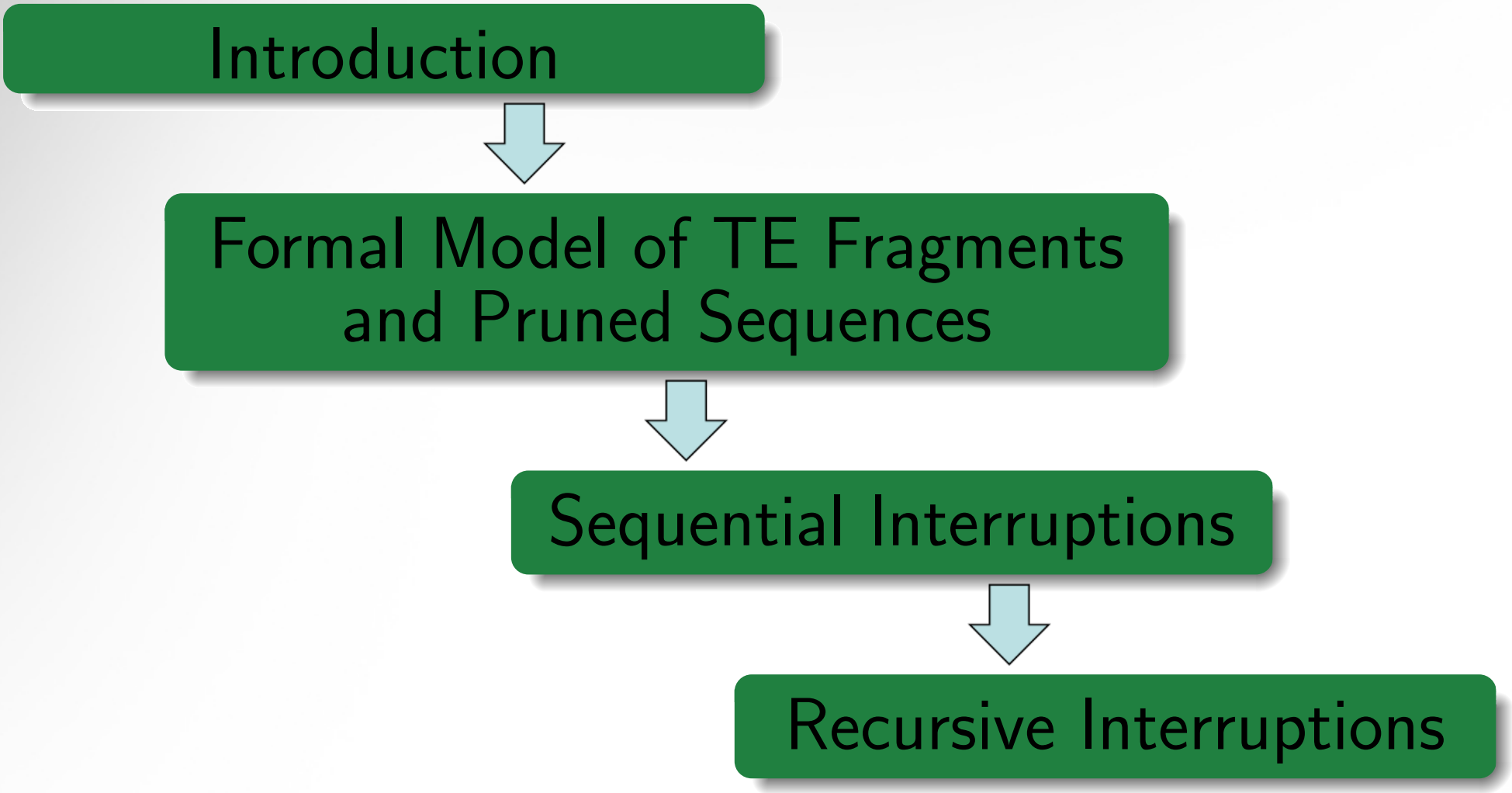
Computational Modelling of the Interruptional Activities between Transposable Elements

Lingling Jin and Ian McQuillan

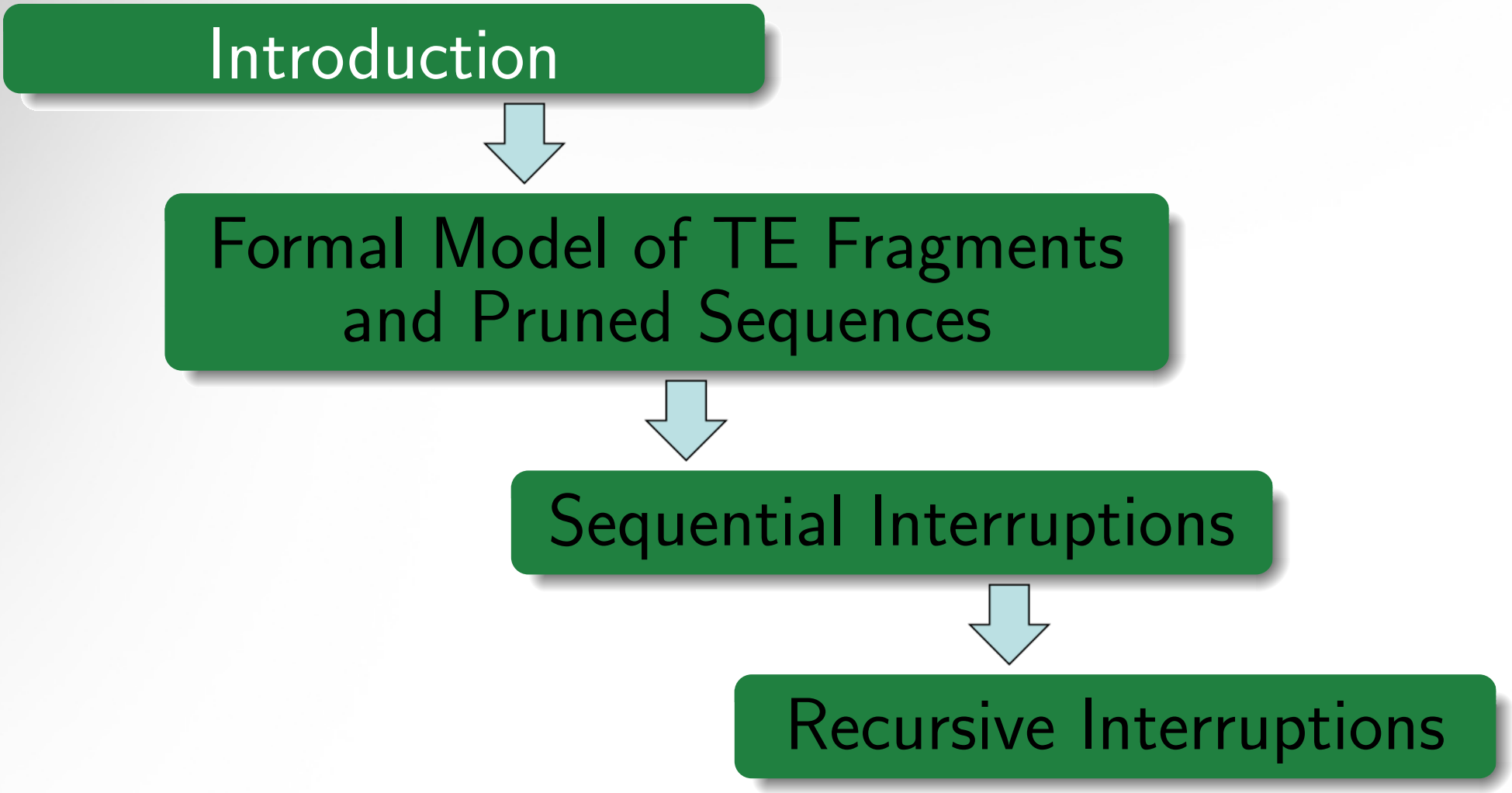
Department of Computer Science
University of Saskatchewan

December 5, 2013

Outline


Introduction
Formal Model of TE Fragments and Pruned Sequences
Sequential Interruptions
Recursive Interruptions

Part I

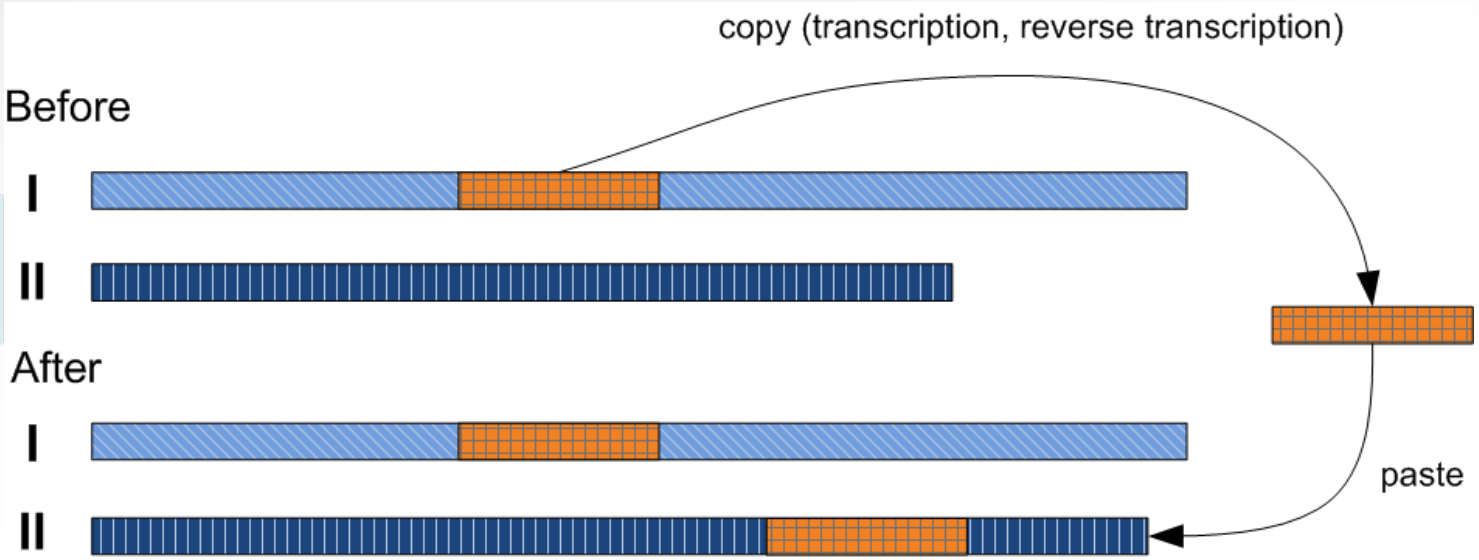

Introduction
Formal Model of TE Fragments and Pruned Sequences
Sequential Interruptions
Recursive Interruptions

Transposable Elements (TEs)

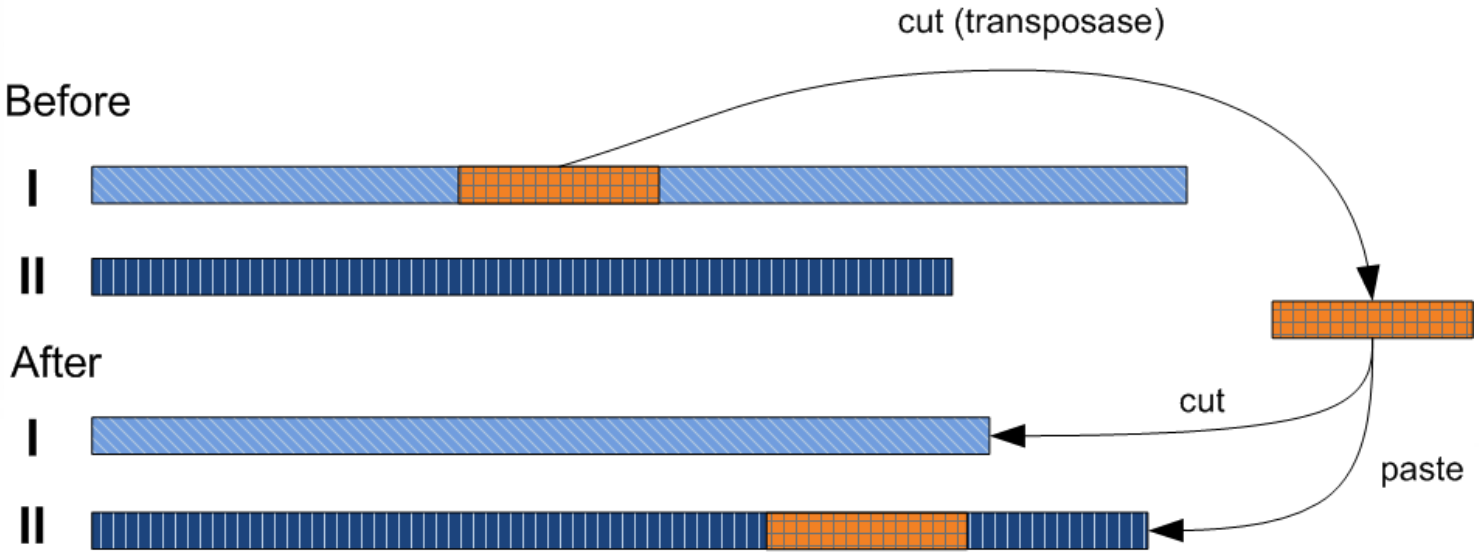
- Transposable Elements (TEs) account for **45%** of the Human Genome.
- TEs are interspersed DNA sequences that can **move or transpose** themselves to new positions within the genome.
- TEs have been reported to **cause** human diseases, including several types of cancer.

Classifications

Class I
Retrotransposons



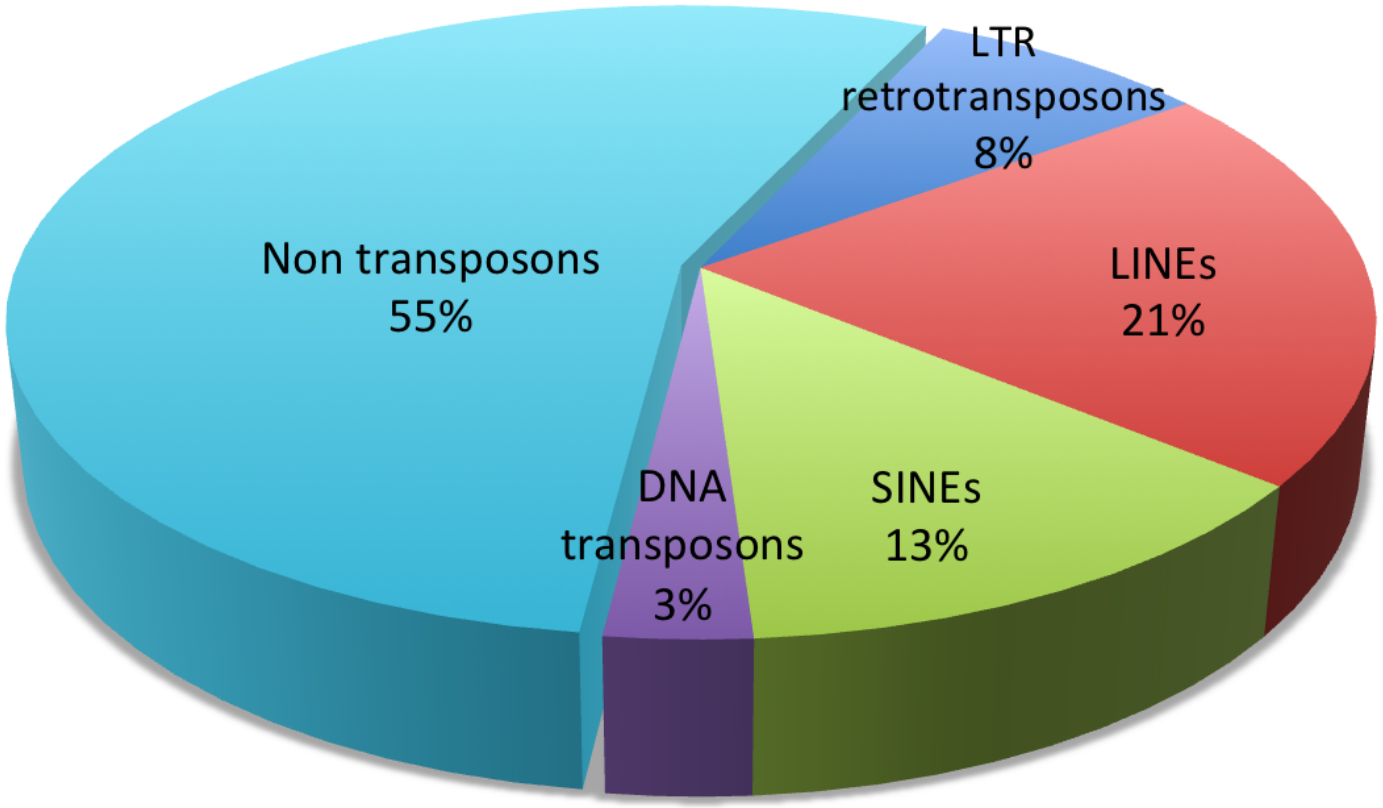
Class II
DNA transposons



Types of TEs based on Structural Features

TE type	TE Class	Mode of Transportation	Length	Copy Number
LINEs	Class I: Retrotransposons	Autonomous	6-8 kb	850,000
SINEs	Class I: Retrotransposons	Non-autonomous	100-300 bp	1,500,000
LTR retrotransposons	Class I: Retrotransposons	Autonomous	6-11 kb	450,000
		Non-autonomous	1.5-3 kb	
DNA transposons	Class II: DNA transposons	Autonomous	2-3 kb	300,000
		Non-autonomous	80-3,000 bp	

The Human Genome



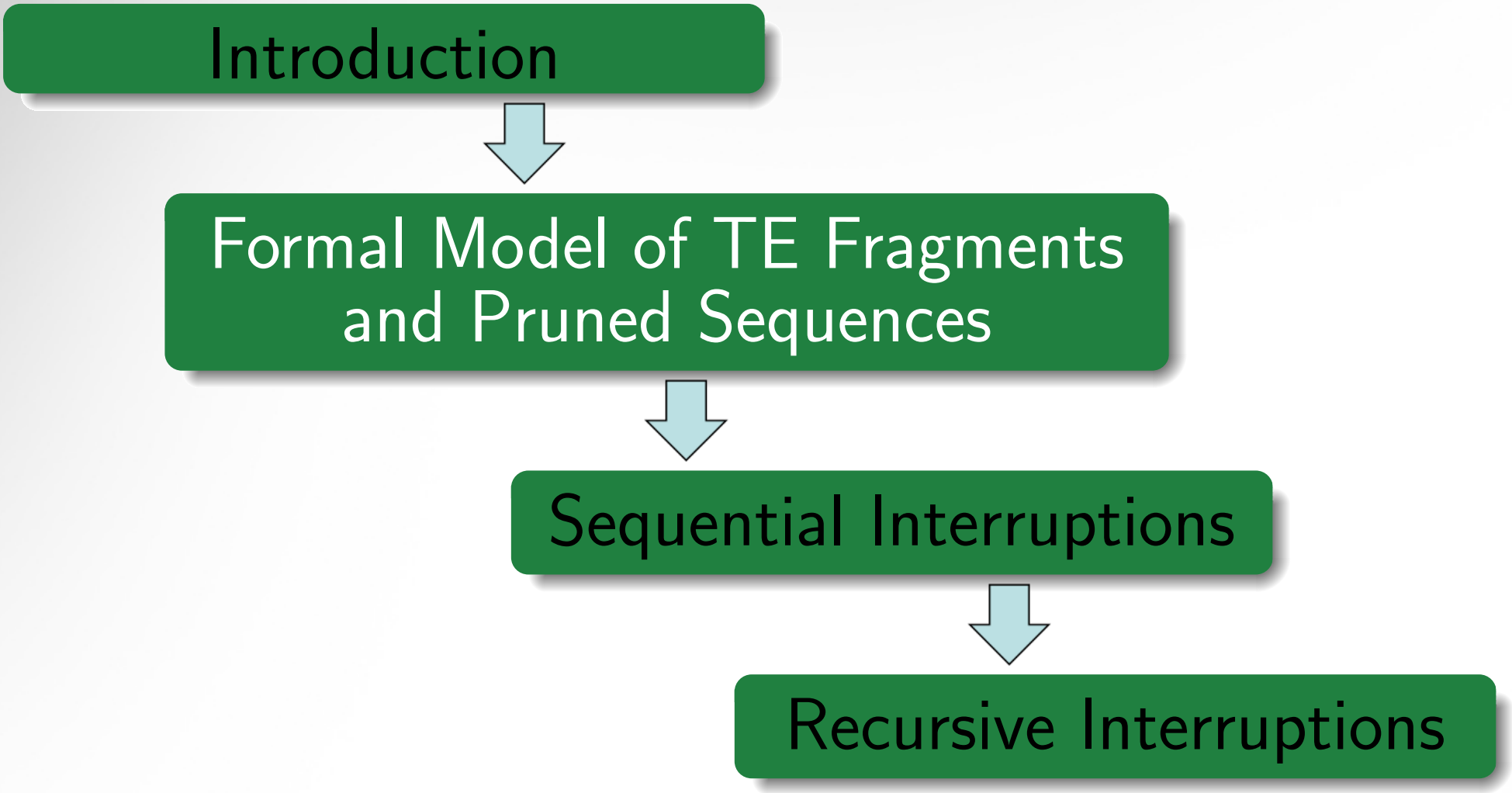
Transpositional Activites

- **Older** TEs are heavily interrupted by younger TEs, but have not inserted into younger elements;
- **Younger** TEs have inserted into older elements that were present in the genome, but are **not** interrupted by older elements;
- Elements of **intermediate age** have both inserted into older elements and been themselves fragmented by younger elements.
- Younger TEs interrupt not only older TEs, but also the **fragmented TEs** that had been previously interrupted.

Motivations

- Creating a standard formal model is necessary for further computational analysis.
- Knowing the evolution of TEs is helpful in understanding the activities of these elements and their impacts on genomes.

Part II


Introduction
Formal Model of TE Fragments and Pruned Sequences
Sequential Interruptions
Recursive Interruptions

Formal Definitions about TEs

Definition (Transposable Element)

A *transposable element (TE)* X is a finite set of strings with $X \subseteq \Sigma^*$.

Definition (Instance of a TE)

An *instance of a TE* X is an element $x \in X$.

Definition (Consensus TE)

A *consensus TE* is a consensus sequence of the elements of X.

Definition (Set of TEs)

A *set of TEs* χ is a finite set of TEs. That is, $\chi \subseteq 2^{\Sigma^*}$, χ is finite and each element of χ is finite.

Fragments of TEs

Definition (Set of Fragments)

Let $s \in \Sigma^*$, then $frag(s)$ is the *set of fragments* of s. We extend this to sets of strings $S \subseteq \Sigma^*$ by $frag(S) = \bigcup_{s \in S} frag(s)$.

Definition (TE Fragments Set)

Let χ be a finite set of TEs. Then we call $\bar{\chi} \subseteq 2^{\Sigma^*}$ a *TE fragments set* if each element $\bar{X} \in \bar{\chi}$ is a subset of $frag(X)$, for some $X \in \chi$.

RepeatMasker TE Fragments Set I

Although we defined TE fragments sets in a general way, we would also like to create a restriction to **the transposable elements that occur in present-day sequences**.

Definition (RepeatMasker TE fragments set)

Let s be a string representing some genomic sequence and χ_s be the set of TEs existing in s, then $\bar{\chi}(s \overset{\text{RM}}{\longleftrightarrow} \chi_s)$ is a *RepeatMasker*[a] *TE fragments set*, running the program with a set of consensus TEs, χ_s, against the genomic sequence s.

In other words, each element of $\bar{\chi}(s \overset{\text{RM}}{\longleftrightarrow} \chi_s)$ is a subset of some element of $\bar{\chi}$, where only TE fragments detected by the RepeatMasker program are selected.

[a] RepeatMasker is the predominant library-based tool used in repeat identification, which has become a standard tool for any search of repeats in genomes.

RepeatMasker TE Fragments Set II

Definition (RepeatMasker TE fragment)

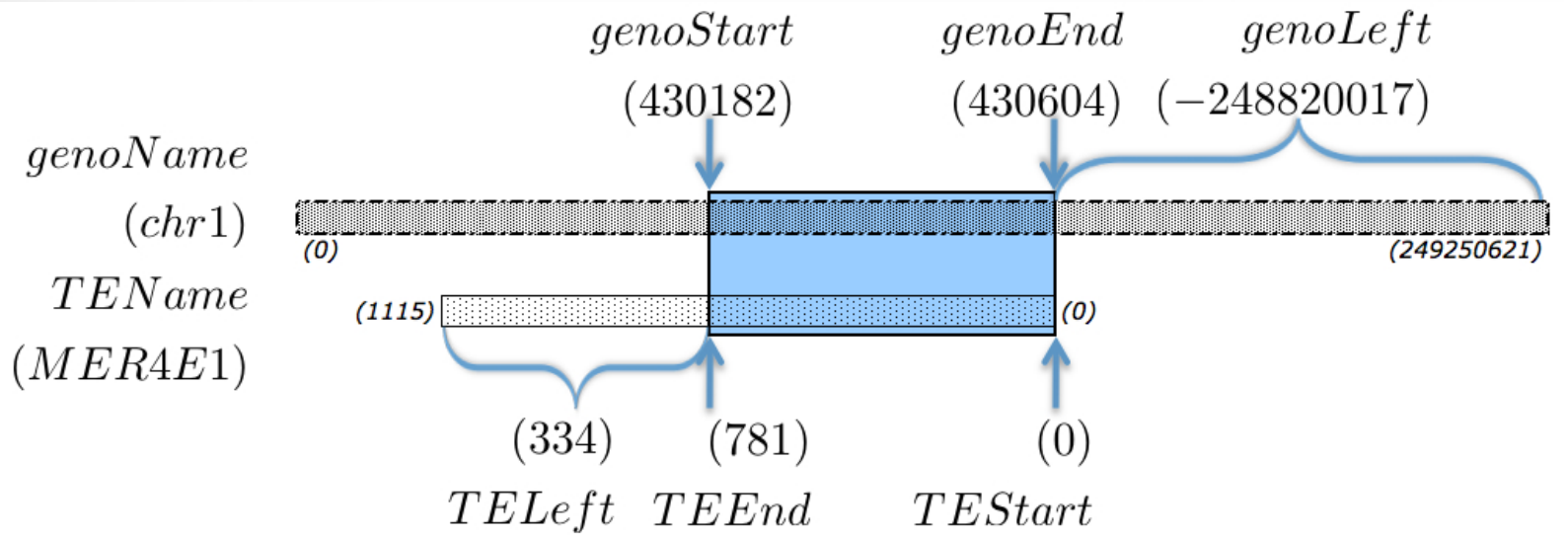
Given a genomic sequence s and a set of TEs χ_s, each *TE fragment* z in each $\bar{X} \in \bar{\chi}(s \overset{\text{RM}}{\longleftrightarrow} \chi_s)$ is a tuple:

$$\begin{aligned} info(z) =& (genoName, genoStart, genoEnd, genoLeft, strand, \\ & TEName, TEClass, TEStart, TEEnd, TELeft). \end{aligned} \tag{1}$$

We use the operator "." to access the attributes, for example, $z.TEname$ is the name of the TE to which fragment z belongs.

RepeatMasker TE Fragment: Example I

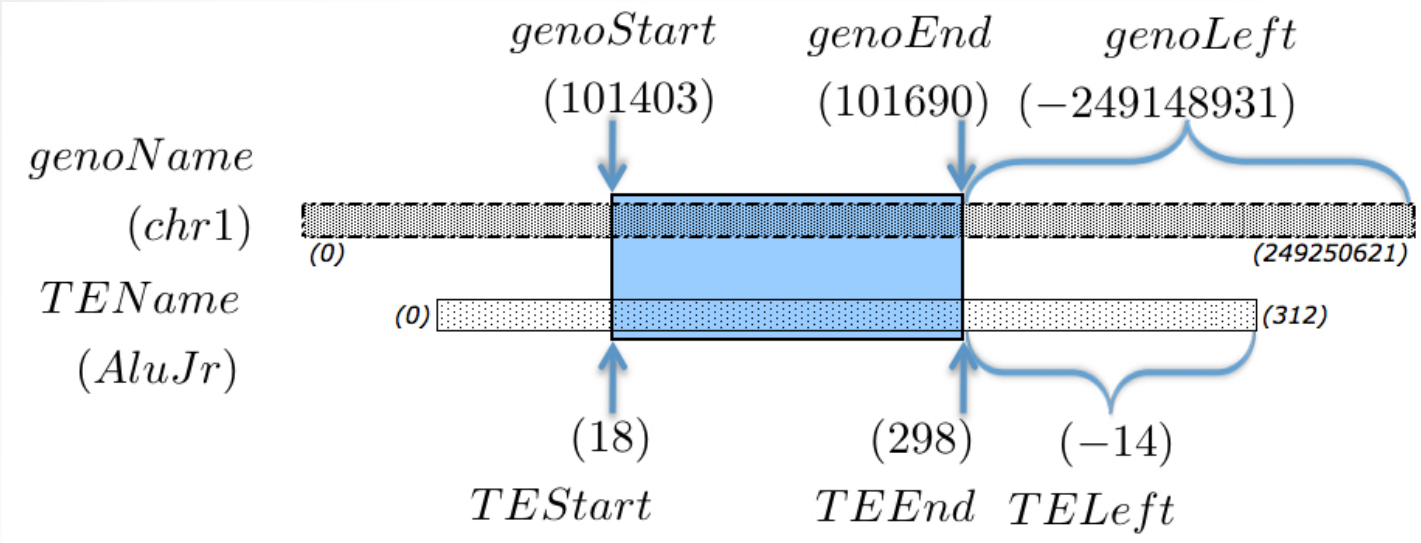
A fragment of MER4E1, which was detected in the "-" strand of Human chromosome 1.



genoName	genoStart	genoEnd	genoLeft	strand	TEName	TEClass	TEStart	TEEnd	TELeft
chr1	430182	430604	-248820017	-	MER4E1	LTR	0	781	334

RepeatMasker TE Fragment: Example II

A fragment of AluJr, which was detected in the "+" strand of Human chromosome 1.



genoName	genoStart	genoEnd	genoLeft	strand	TEName	TEClass	TEStart	TEEnd	TELeft
chr1	101403	101690	-249148931	+	AluJr	SINE	18	298	-14

Pruned Sequence and Order Pruned Sequence

A genomic sequence

$s = w_0 z_1 w_1 z_2 \ldots z_k w_k$, with $z_1, \ldots, z_k$ in sets in $\bar{\chi}_s$, and no fragment of $w_0, \ldots, w_k$ in sets in $\bar{\chi}_s$.

A pruned sequence

$\bar{s} = \beta_0 z_1 \beta_1 z_2 \beta_2 \ldots z_k \beta_k$, where $\beta_i = |w_i|$, $0 \leq i \leq k$.

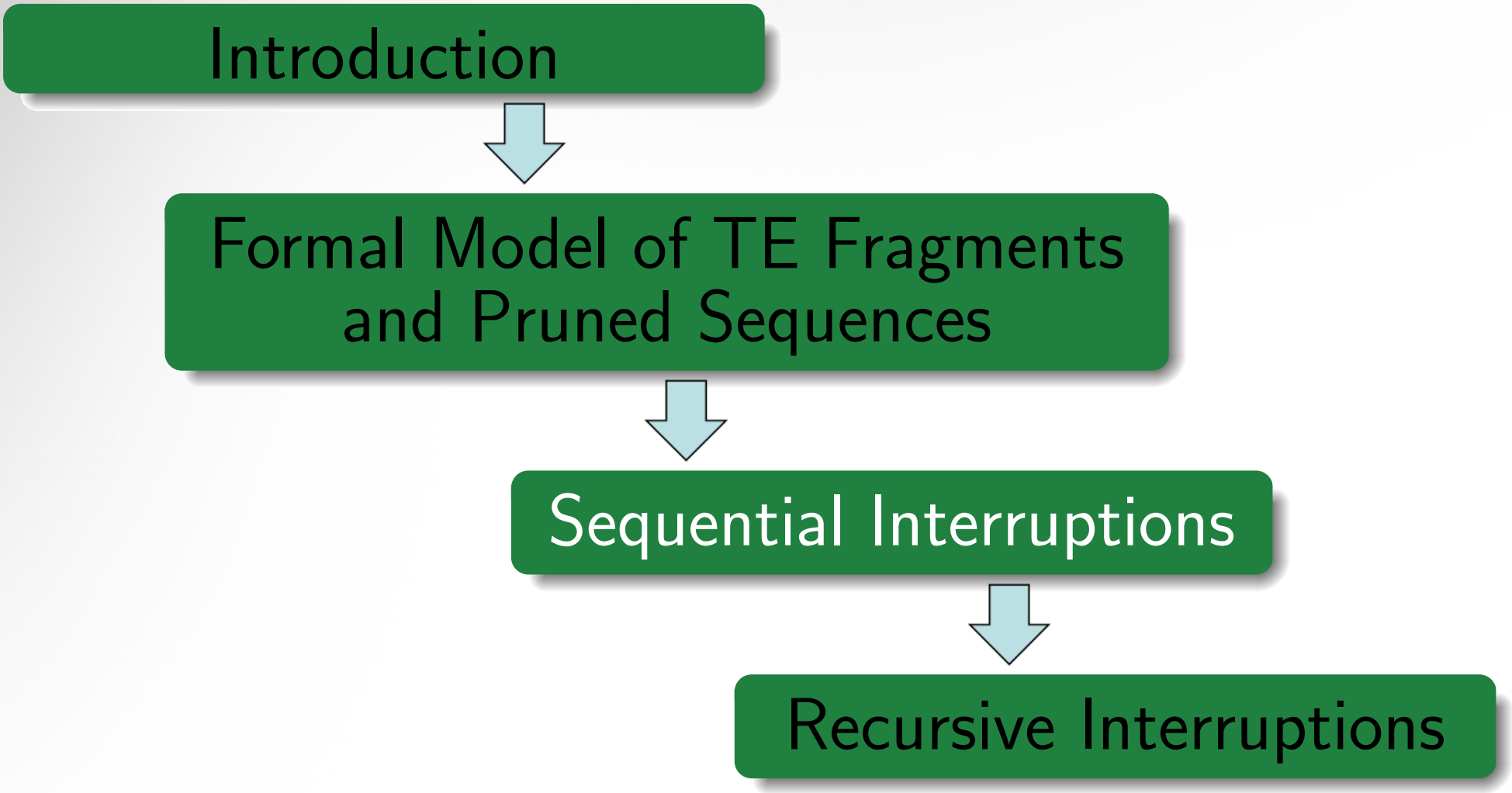
For Sequential Interruptions

An order pruned sequence

$\bar{s}_o = j_1 j_2 \ldots j_k$, where $z_i \in X_{j_i}$, for all i, $1 \leq i \leq k$.

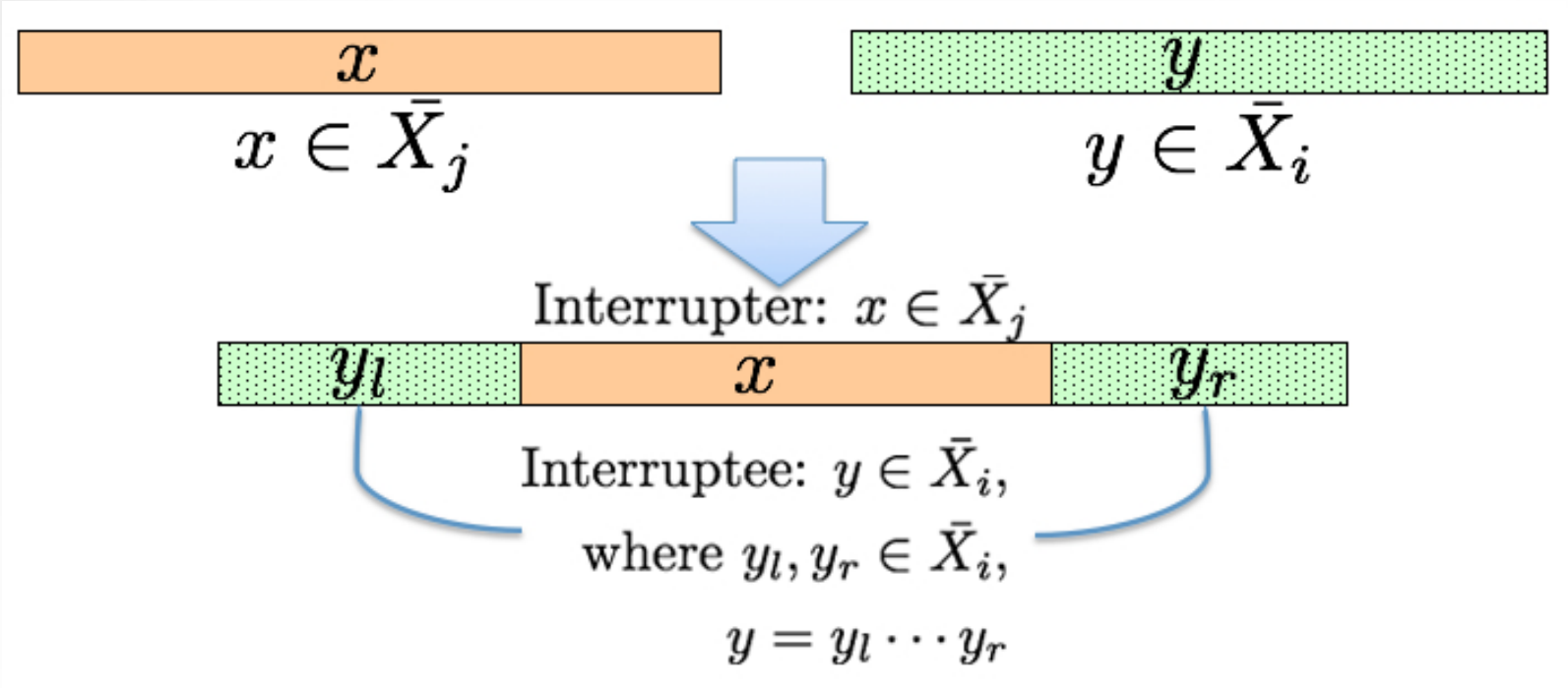
For Recursive Interruptions

Part III


Introduction
Formal Model of TE Fragments and Pruned Sequences
Sequential Interruptions
Recursive Interruptions

An Interruption

Newer TEs tend to interrupt older TEs, thereby fragmenting older TEs within the single linear sequence.



Continuous Fragments

Definition (Continuous TE Fragments)

Two TE fragments z_i and z_j $(i < j)$ are *continuous TE fragments*, $z_i \overset{\varepsilon,E}{\sim} z_j$, with threshold $\varepsilon \in \mathbb{N}$ and distance $E \in \mathbb{N}$, if they satisfy the following conditions:

1. they belong to the same transposable element;
2. they are detected in the same DNA strand;
3. they are either separated or overlap with less than or equal to a threshold, ε, with respect to the TE consensus sequence to which they belong;
4. They are in the genomic sequence within a distance, E, of non-transposon DNA sequence.

Sequential Interruptions

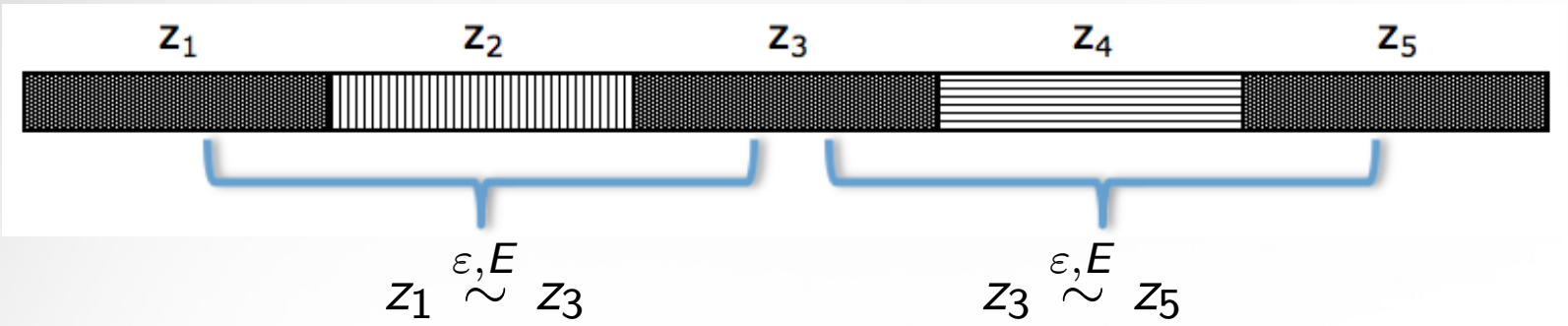
Definition (Sequential Interruptions)

Given a genomic sequence s, a set of TEs with a fixed ordering on its elements $\chi_s = \{X_1, X_2, \ldots, X_m\}$, a threshold $\varepsilon \in \mathbb{N}$ in TE consensus sequence, a distance $E \in \mathbb{N}$ in genomic sequence, and a pruned sequence $\bar{s} = \beta_0 z_1 \beta_1 z_2 \ldots z_k \beta_k$, we define *sequential interruptions* of X_j by X_i as

$$\Xi_s^{\varepsilon,E}(X_i,X_j)=\{k|z_k\in\bar{X}_i,z_{k-\eta_1},z_{k+\eta_2}\in\bar{X}_j,z_{k-\eta_1}\overset{\varepsilon,E}{\sim}z_{k+\eta_2},\text{and }\eta_1,\eta_2\in\mathbb{N}\}.$$

Thus X_i is called *interrupter*, and X_j is called *interruptee*.

An example of sequential interruptions in chromosome 1 I



genoName	*genoStart*	*genoEnd*	*genoLeft*	*strand*	*TEName*	*TEClass*	*TEStart*	*TEEnd*	*TELeft*
chr1	448062	448139	-248802482	+	L1MD3	LINE	6988	7068	-814
chr1	448150	448328	-248802293	+	AluYc	SINE	122	299	0
chr1	448332	448403	-248802218	+	L1MD3	LINE	7068	7148	-847
chr1	448403	448710	-248801911	+	AluSq	SINE	1	313	0
chr1	448710	449273	-248801348	+	L1MD3	LINE	7149	7753	-242

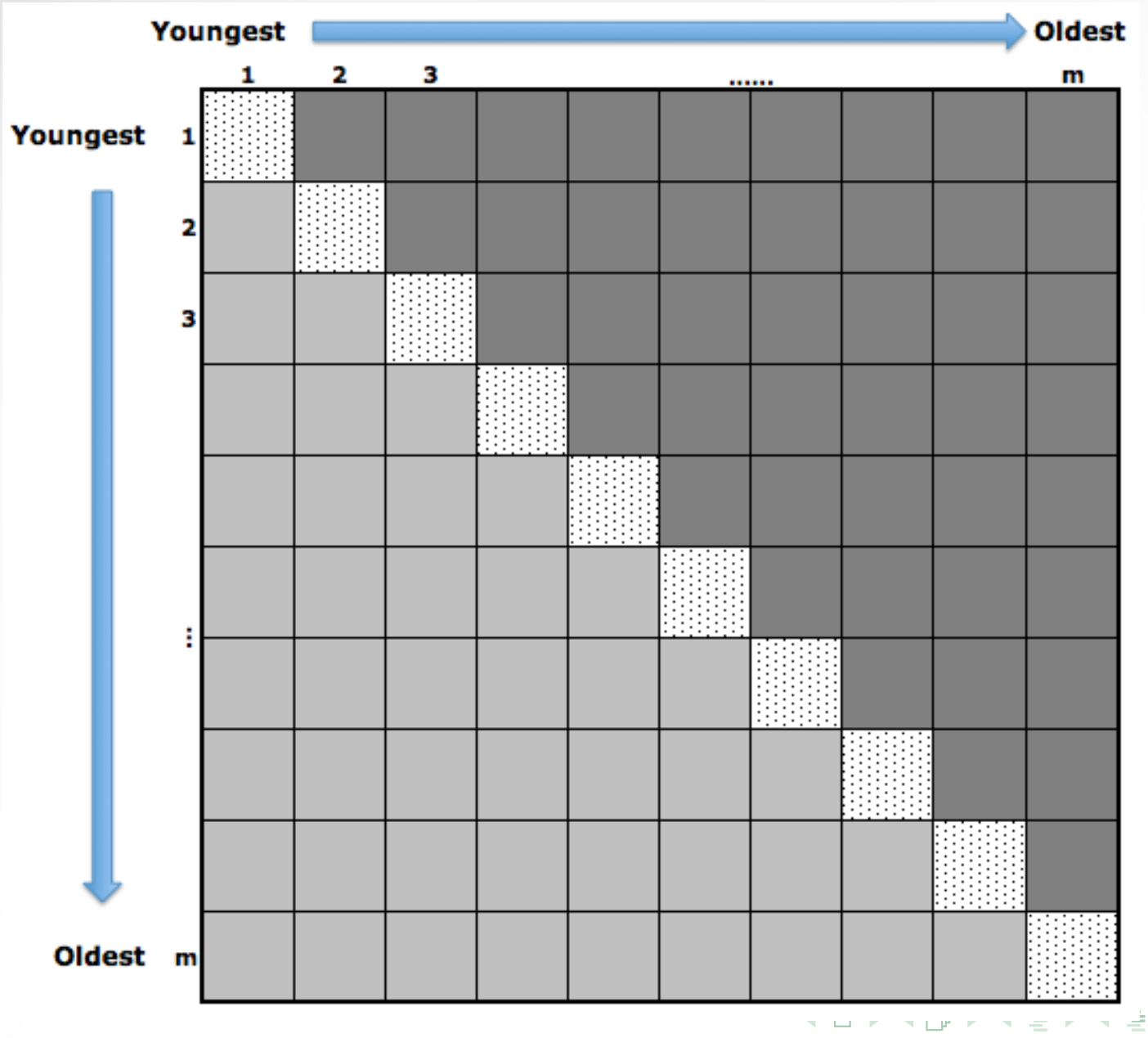
- The pruned sequence is $\bar{s} = \beta_0 z_1 \beta_1 z_2 \beta_2 z_3 \beta_3 z_4 \beta_4 z_5 \beta_5$, where $\beta_0, \ldots, \beta_5 \in \mathbb{N}$, and $z_1, z_3, z_5 \in \bar{X}_1$, $z_2 \in \bar{X}_2$, $z_4 \in \bar{X}_3$, $\bar{X}_1, \bar{X}_2, \bar{X}_3 \in \bar{\chi}(s \overset{\text{RM}}{\longleftrightarrow} \chi_s)$.

An example of sequential interruptions in chromosome 1 II

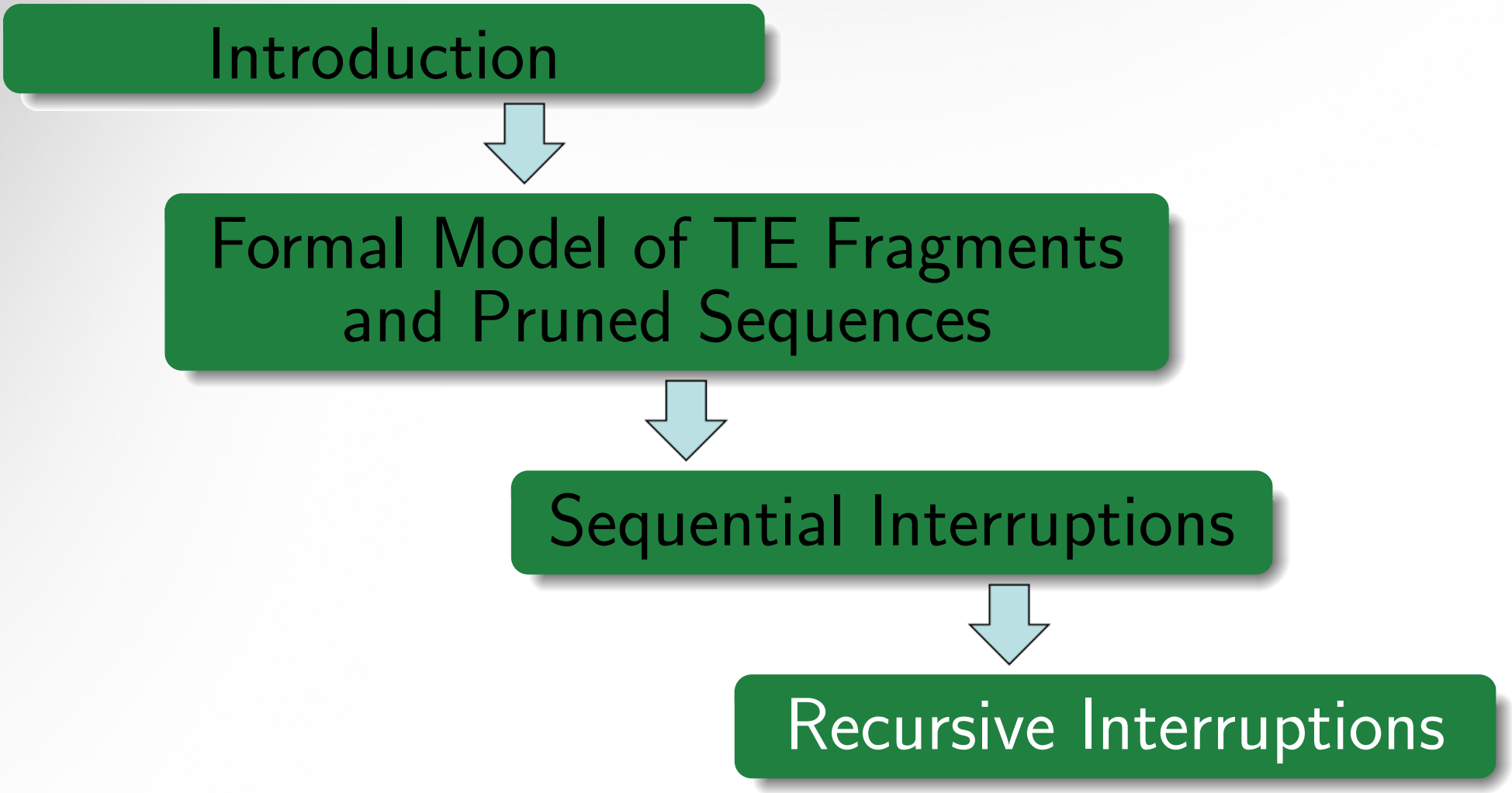
- There are two potential interruptions in s: an instance of X_1 is present in the sequence, then an instance of X_2 and an instance of X_3 potentially inserted themselves into the instance of X_1 to break it into three segments z_1, z_3 and z_5; that is, $|\Xi_s^{\varepsilon,E}(X_2, X_1)| = 1$ and $|\Xi_s^{\varepsilon,E}(X_3, X_1)| = 1$.
- Given a fixed order of the set of TEs as $\chi_s = \{\ldots, X_1, \ldots, X_2, \ldots, X_3 \ldots\}$, the **interruption matrix** showing only the rows and columns of these TEs is

$$M = \begin{bmatrix} & \vdots & & \vdots & & \vdots & \\ \ldots & 0 & \ldots & 0 & \ldots & 0 & \ldots \\ & \vdots & & \vdots & & \vdots & \\ \ldots & 1 & \ldots & 0 & \ldots & 0 & \ldots \\ & \vdots & & \vdots & & \vdots & \\ \ldots & 1 & \ldots & 0 & \ldots & 0 & \ldots \\ & \vdots & & \vdots & & \vdots & \end{bmatrix}.$$

Interruption Matrix



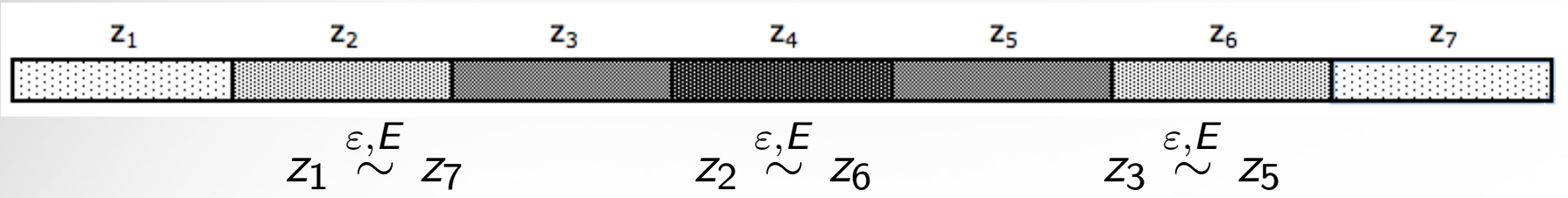
Part IV


Introduction
Formal Model of TE Fragments and Pruned Sequences
Sequential Interruptions
Recursive Interruptions

Recursive Interruptions

- When many insertions occurred throughout the evolution of a genomic sequence, the interruptions nest in a **recursive pattern**;
- Recursive interruptions cannot be represented accurately with the interruptional matrix that only counts the abundance without storing the **hierarchical relationships** of interruptions.

An Example of Recursive Interruptions in chromosome X



genoName	*genoStart*	*genoEnd*	*genoLeft*	*strand*	*TEName*	*TEClass*	*TEStart*	*TEEnd*	*TELeft*
chrX	53437061	53437143	-101833417	+	MIR	SINE	3	88	-174
chrX	53437143	53437277	-101833283	+	AluJb	SINE	1	132	-170
chrX	53437277	53437448	-101833112	+	AluSx	SINE	39	192	-120
chrX	53437448	53437761	-101832799	+	AluSq2	SINE	1	312	0
chrX	53437761	53437887	-101832673	+	AluSx	SINE	193	312	0
chrX	53437887	53438055	-101832505	+	AluJb	SINE	133	293	-9
chrX	53438055	53438226	-101832334	+	MIR	SINE	89	261	-1

The pruned sequence is $\bar{s} = \beta_0 z_1 \beta_1 z_2 \beta_2 z_3 \beta_3 z_4 \beta_4 z_5 \beta_5 z_6 \beta_6 z_7 \beta_7$, where

$$\beta_0, \ldots, \beta_7 \in \mathbb{N},\ z_1, z_7 \in \bar{X}_1,\ z_2, z_6 \in \bar{X}_2,\ z_3, z_5 \in \bar{X}_3,\ z_4 \in \bar{X}_4,$$

$$\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4 \in \bar{\chi}(s \overset{\text{RM}}{\longleftrightarrow} \chi_s)\ z_1 \overset{\varepsilon,E}{\sim} z_7,\ z_2 \overset{\varepsilon,E}{\sim} z_6,\ z_3 \overset{\varepsilon,E}{\sim} z_5.$$

Recursive Interruption Context-free Grammar

A **recursive interruption context-free grammar** can model the generation of recursive interruptions.

Definition (recursive interruption context-free grammar)

Given a set of TEs with a fixed order on its elements, $\chi = \{X_1, X_2, \ldots, X_m\}$, the *recursive interruption context-free grammar* is a grammar $G = (V, T, \delta, S)$, where $V = \{S, X_1, X_2, \ldots, X_m\}$, $T = \{1, 2, \ldots, m\}$, and δ contains the following productions:

$$S \rightarrow X_i S, \quad 1 \leq i \leq m, \tag{1}$$

$$S \rightarrow X_i, \quad 1 \leq i \leq m, \tag{2}$$

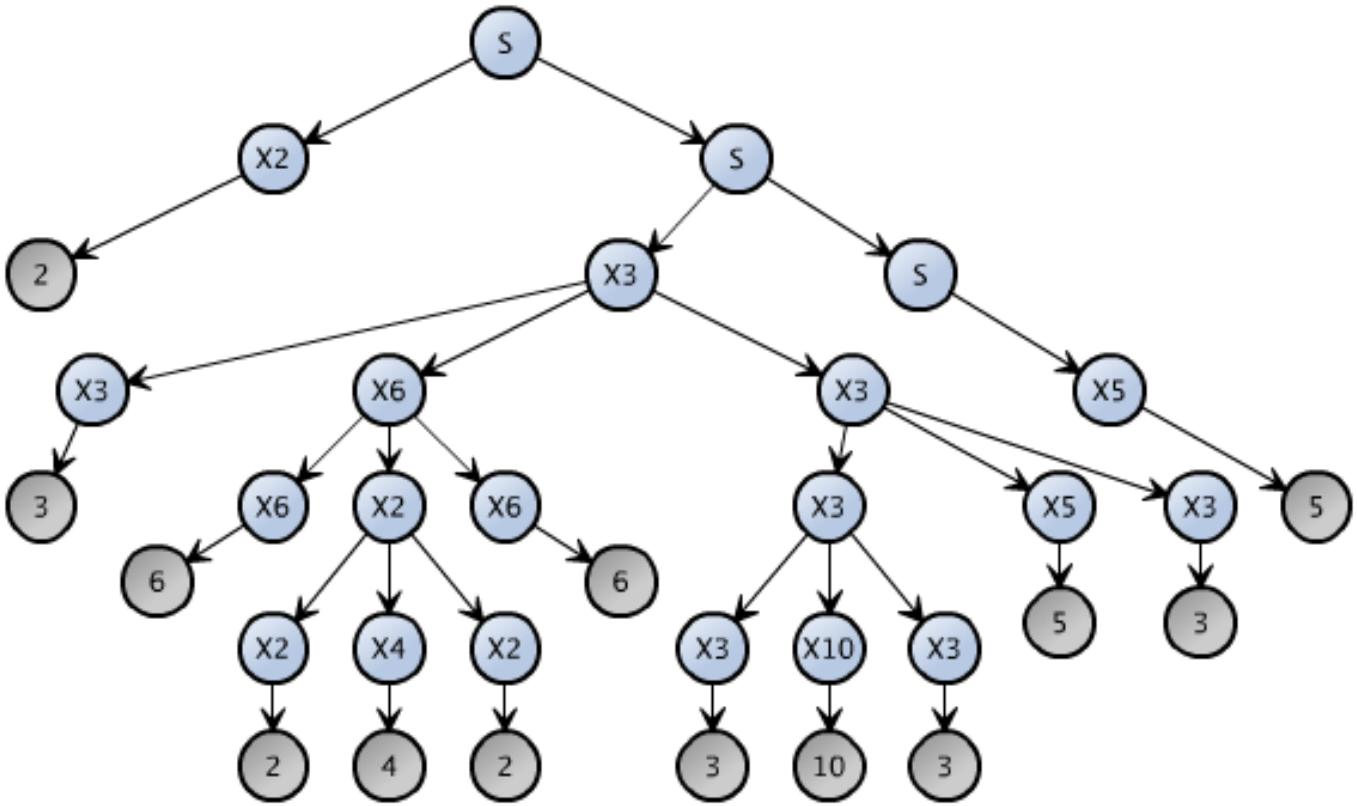
$$X_i \rightarrow X_i X_j X_i, \quad 1 \leq i \leq m, \ 1 \leq j \leq m, \tag{3}$$

$$X_i \rightarrow i, \quad 1 \leq i \leq m. \tag{4}$$

Example

Given a genomic sequence s and a set of TEs with a fixed order on its elements $\chi_s = \{X_1, X_2, \ldots, X_{10}\}$, and assume $s = w_0 z_1 w_1 \ldots z_{13} w_{13}$, with $z_1, z_4, z_6 \in \bar{X}_2$, $z_2, z_8, z_{10}, z_{12} \in \bar{X}_3$, $z_5 \in \bar{X}_4$, $z_{11}, z_{13} \in \bar{X}_5$, $z_3, z_7 \in \bar{X}_6$, $z_9 \in \bar{X}_{10}$. Then an order pruned sequence $\bar{s}_o = 2\,3\,6\,2\,4\,2\,6\,3\,10\,3\,5\,3\,5$ is the yield of the parse tree:

$$
\begin{aligned}
S &\rightarrow X_i S, && (1)\\
S &\rightarrow X_i, && (2)\\
X_i &\rightarrow X_i X_j X_i, && (3)\\
X_i &\rightarrow i. && (4)
\end{aligned}
$$



Summary

- We created a formal model capturing TE types and fragments within genomic sequences.
- The model can be used to capture and calculate interruptions and their frequencies in a general way.
- We provided a stochastic context-free grammar to capture recursive TE nesting.
- The grammar allows standard polynomial time parsing to be used to calculate a prediction of the nesting.

Thank You!