



## **GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM**

**CÓDIGO: GIN-GUI-001**

**VERSIÓN: 01**

**FECHA ELABORACIÓN  
DEL DOCUMENTO:  
23/ENE/2015**



**UNIVERSIDAD AUTÓNOMA DE MANIZALES**

**VICERRECTORÍA ACADÉMICA**

**UNIDAD DE INVESTIGACIÓN**

**UNIDAD DE POSGRADOS**



## GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM

CÓDIGO: GIN-GUI-001

VERSIÓN: 01

FECHA ELABORACIÓN  
DEL DOCUMENTO:  
23/ENE/2015

### TÓPICOS PARA LA PRESENTACIÓN DE INFORMES FINALES<sup>1</sup>

#### UNIVERSIDAD AUTÓNOMA DE MANIZALES

**PROYECTO:** Diseño de un identificador automático de LTR retrotransposones en plantas mediante el uso de técnicas de Machine Learning

**GRUPO DE INVESTIGACIÓN:** Ingeniería de Software y Automática

**ESTUDIANTE:** Mariana Sofía Candamil Cortés

**TUTOR DE TESIS:** Simón Orozco Arias


**CO TUTOR DE TESIS:** Reinel Tabares Soto

**DATOS DE IDENTIFICACIÓN:** C.C. 1010012738 – Correo:

mariana.candamilc@autonoma.edu.co

**AÑO: 2020**

<sup>1</sup>Formato adaptado de COLCIENCIAS. 2006

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


## RESUMEN

Los elementos transponibles (ET), son estructuras específicas del genoma de las especies, los cuales, tienen la capacidad de trasladarse de una ubicación a otra. Debido a esta característica, son causantes de mutaciones o cambios, que pueden ser negativos como la aparición de enfermedades o benéficas como participar en roles fundamentales en la evolución de los genomas y en la diversidad genética. Estos elementos pueden dividirse, dependiendo de su estructura y modo de propagación, en clases, órdenes, superfamilias, linajes y familias; en especial, uno de los órdenes más relevantes en plantas son los *Long Terminal Repeat* (LTR) retrotransposones, pertenecientes a la Clase I, pues son los más abundantes en estas especies, de ahí la importancia de estudiar en particular estas estructuras; es por esto que en el presente proyecto, se busca desarrollar un identificador, que permita detectar los LTR retrotransposones mediante técnicas de aprendizaje de máquina, para el cual, se realiza la construcción de conjuntos de datos, y se establece una comparación entre las técnicas de preprocesamiento, esquemas de codificación y modelos de machine Learning, en la cual, se observó que las técnicas de preprocesamiento que mejor rendimiento muestran son PCA (Análisis de Componentes Principales) y escalamiento en conjunto con el esquema de codificación k-mers. Usando estas técnicas se obtuvo una eficiencia superior al 98% en F1-score para los modelos de aprendizaje de máquina analizados; en particular con el modelo de MLP (*multi-layer perceptron*) se obtuvo el mayor porcentaje, siendo de 99,7%, demostrando el funcionamiento y optimización de los resultados obtenidos.

**PALABRAS CLAVES:** Bioinformática, Machine Learning, Elementos transponibles, LTR retrotransposones, genómica.

## ABSTRACT

Transposable elements (TE) are specific structures in the genome of the species, which have the ability to move from one location to another. Due to this characteristic, they are the cause of mutations or changes, which can be negative, such as the appearance of diseases, or beneficial, such as participating in fundamental roles in the evolution of genomes and genetic diversity. These elements can be divided, depending on their structure and mode of propagation, into classes, orders,

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

superfamilies, families and lineages; in particular, one of the most relevant orders in plants are the *Long Terminal Repeat* (LTR) retrotransposons, belonging to Class I, because they are the most abundant in these species, hence the importance of studying these structures in particular; this is why in this project, we develop an identifier that allows detecting the LTR retrotransposons using machine learning techniques, for which the construction of data sets is carried out, and a comparison is established between the techniques of preprocessing, coding schemes and machine learning models, in which it was observed that the appropriate preprocessing techniques should be PCA (*Principal Component Analysis*) and scaling and the appropriate coding scheme should be kmers. With which an efficiency greater than 98% was obtained, demonstrating the operation and optimization of the results obtained.

**KEY WORDS:** Bioinformatics, Machine Learning, Transposable elements, LTR retrotransposons, genomics.



## GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM


CÓDIGO: GIN-GUI-001

VERSIÓN: 01

FECHA ELABORACIÓN  
DEL DOCUMENTO:  
23/ENE/2015


### TABLA DE CONTENIDO

1. PRESENTACIÓN.....	7
2. INTRODUCCIÓN .....	8
3. ÁREA PROBLEMÁTICA Y JUSTIFICACIÓN.....	9
4. REFERENTE TEÓRICO .....	10
5. LOS OBJETIVOS.....	14
6. METODOLOGÍA .....	14
7. RESULTADOS.....	15
8. DISCUSIÓN DE RESULTADOS.....	21
9. CONCLUSIONES .....	22
10. RECOMENDACIONES.....	23
11. EVIDENCIA DE RESULTADOS .....	24
11.1. Formación de recurso humano.....	24
11.2. Apropiación social del conocimiento .....	24
11.3. Generación de Nuevo Conocimiento:.....	24
12. IMPACTOS LOGRADOS.....	25
13. BIBLIOGRAFÍA.....	26
14. ANEXOS.....	31

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

## Índice de figuras

Figura 1. Resultados del proceso de llenado con "N" y sin preprocesamiento .....	16
Figura 2. Resultados del proceso de llenado con "N" y con escalamiento .....	16
Figura 3. Resultados del proceso de llenado con "N" y PCA .....	17
Figura 4. Resultados del proceso de llenado con "N" y PCA con escalamiento .....	17
Figura 5. Resultados del proceso de llenado con self replication y sin preprocesamiento .....	18
Figura 6. Resultados del proceso de llenado con self replication y escalamiento .....	18
Figura 7. Resultados del proceso de llenado con self replication y PCA .....	19
Figura 8. Resultados del proceso de llenado con self replication y PCA con escalamiento ...	19
Figura 9. Resultados de identificación y clasificación con KNN .....	20
Figura 10. Resultados de identificación y clasificación con LDA .....	20
Figura 11. Resultados de identificación y clasificación con LR .....	21


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

## 1. PRESENTACIÓN

El proyecto que se expone a continuación se realizó en el grupo de investigación de ingeniería de software y automática de la Universidad Autónoma de Manizales – UAM, en la línea de Bioinformática e Inteligencia Artificial, consiste en la ejecución de un identificador de LTR retrotransposones, un orden de ET que contribuye de manera significativa en el genoma de las especies de plantas, mediante técnicas de Machine Learning, generando un impacto para el conocimiento del genoma completo de las especies de plantas.

El equipo de investigación consta de dos ramas Bioinformática e Inteligencia Artificial, el presente proyecto trabaja bajo la línea de la bioinformática, la cual está conformada por:

- Simón Orozco Arias, tutor principal del semillero. Es candidato a doctorado en Ingeniería en la Universidad de Caldas, ha desarrollado diversos proyectos, entre los que se encuentran: aplicación de técnicas de HPC para acelerar procesos bioinformáticos, análisis de las dinámicas y estructuras de elementos transponibles en diferentes especies de plantas (café robusta, café arábica, caña de azúcar, entre otros), en la roya del café y en la mosca de la fruta y últimamente está interesado en la aplicación de técnicas de aprendizaje de máquina para anotar elementos transponibles.
- Romain Guyot, doctor en biología de la Universidad de Zürich, Suiza, investigador senior de Minciencias y director de investigación en genómica y evolución en bioinformática en Institut De Recherche Pour Le Développement (IRD) en Montpellier, Francia. Ha desarrollado múltiples proyectos internacionales de secuenciación y anotación de genomas como el café robusta, café arábigo, piña, lupín blanco, entre otros. Ha desarrollado múltiples estudios de las dinámicas de elementos transponibles y tiene más de 15 años de experiencia en genética, genómica, biotecnología y bioinformática.
- Reinel Tabares Soto, coordinador de Ingeniería Electrónica de la Universidad Autónoma de Manizales, magister en Ingeniería – Automatización Industrial, ha participado en proyectos de supercomputación, minería de datos, bioinformática y aprendizaje profundo. Actualmente está haciendo su doctorado en la aplicación de redes neuronales convolucionales en estaganálisis.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


El Semillero de Bioinformática e Inteligencia Artificial inicia a finales del año 2018 y cuenta con cerca de 15 estudiantes de Ingeniería Biomédica, Ingeniería Electrónica e Ingeniería de Sistemas. Se ha divulgado los resultados obtenidos en diversos eventos como: Congreso Internacional de Ingeniería Biomédica y Bioingeniería llevado a cabo en la ciudad de Manizales en noviembre de 2019, participación en el XII Encuentro Departamental de Semilleros de Investigación –RREDSI, llevado a cabo en Manizales en abril de 2020, participación en el primer Congreso Latinoamericano de Mujeres en Bioinformática y Ciencias de Datos (1st-WBDS), llevado a cabo en Buenos Aires, Argentina en septiembre de 2020.

## 2. INTRODUCCIÓN

El genoma de las especies es el conjunto de genes que se encuentran en los cromosomas, es decir, el material genético que corresponde a cada una de las especies. Este brinda información de cada organismo y está conformado por numerosas estructuras que interfieren en su funcionalidad. Específicamente, una de las estructuras de mayor relevancia actualmente son los elementos transponibles (ET), descubiertos por McClintock (1953), los cuales son entidades genéticas móviles que podrían cambiar la expresión genética y ser causantes de reordenamientos cromosómicos.

Desde el momento de su descubrimiento, se han realizado numerosas investigaciones con el fin de profundizar en dichos elementos, su estructura, funcionalidad e impacto en las especies (Cui y Cao, 2014). En estos estudios, se pudo observar que los ET tienen un papel fundamental en la evolución adaptativa, debido a la variación genética, de igual forma, estos elementos pueden afectar la regulación genética, pues existe una delección o una inserción de estas estructuras en el genoma, causando un cambio en su fenotipo, es decir, en las características que se pueden apreciar del organismo. Así mismo, debido a investigaciones como las realizadas por Lisch (2013), cabe destacar que el movimiento de un ET, puede ser afectado por las condiciones de estrés del medio, debido a esto, se puede activar o silenciar su propagación, cuando está expuesta a condiciones diferentes como calor extremo o radiación. Gracias a estos análisis, se pudo identificar que los ET son necesarios en las especies y pueden contribuir positiva o negativamente a su desarrollo.



	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


En particular, al avanzar en las investigaciones, se clasificaron los ET, según su estructura, mecanismo de transposición y proliferación, en diversos niveles (Wicker et al., 2007), (Neumann et al., 2019); de los cuales, uno de los más relevantes son los LTR retrotransposones, los cuales, contienen secuencias de ADN no codificantes al comienzo y al final del mismo y son los ET que presentan mayor contribución al tamaño del genoma de las plantas, reprogramando múltiples procesos en su funcionalidad, de ahí la importancia de identificar dichas estructuras en el genoma, con el fin de evitar cambios negativos en el mismo. Aunque existen diversas herramientas bioinformáticas para su detección, estas requieren del desarrollo de trabajos manuales, por esto, el propósito de este proyecto consiste en diseñar un identificador de LTR retrotransposones en especies de plantas basado en la implementación de técnicas de aprendizaje de máquina, lo cual automatiza el proceso aumentando el porcentaje de precisión al momento de detectar la estructura de interés.

### **3. ÁREA PROBLEMÁTICA Y JUSTIFICACIÓN**

Los elementos transponibles (ET) conforman una gran proporción del genoma de las especies y tienen la capacidad de moverse de una ubicación cromosómica a otra (Mita & Boeke, 2016). Esta movilización de elementos puede promover la inactivación de genes, modular la expresión de genes o la generación de la recombinación ilegítima o no homóloga (Munoz-Lopez y Garcia-Perez 2010), es por esto que son causantes de múltiples mutaciones en el genoma, que pueden ser tanto beneficiosas como negativas, teniendo un impacto en la evolución del genoma de la especie (Bourque et al., 2018).

Dependiendo del mecanismo de replicación que empleen estos elementos, en el dominio eucariota, se pueden clasificar en Clase I o Retrotransposones y Clase II o Transposones (Makalowski et al., 2019), siendo los de la Clase I, los más abundantes en las especies, específicamente, el orden, de los LTR retrotransposones, es el que mayor cantidad posee en plantas. Este presenta una contribución mayor al tamaño del genoma y contiene las dos superfamilias más relevantes del dominio eucariota: Gypsy y Copia (Esposito et al., 2019).

Debido a la gran importancia de los ET en las especies, se han desarrollado diversas técnicas y metodologías, con el fin de identificar estos elementos en el genoma, como la basada en estructura, basada en homología, de novo y genómica comparativa (Loureiro et al., 2012); sin embargo, se han detectado diversos inconvenientes con el uso de estas metodologías, como una difícil identificación


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

mediante la técnica basada en estructura, pues existe una constante variación o mutación de las secuencias, de igual forma, si se recurre al uso de la homología, se debe realizar un extenso trabajo manual, debido a que los elementos son específicos para cada especie, lo cual hace que se requieran diversas bases de datos para una identificación más completa. Así mismo, al usar la técnica de novo, se requiere una gran cantidad de repeticiones de los elementos y la técnica de genómica comparativa, necesita un genoma bien anotado, lo cual es bastante complejo, pues hay muchos componentes que se desconocen en el genoma. Estos enfoques ofrecen diferentes especificidades y todos poseen de una tasa relativamente alta de detecciones de falsos positivos (Orozco-Arias et al., 2019).

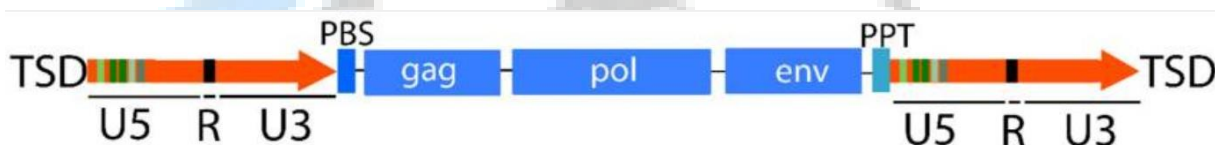
De igual forma, considerando el apogeo de las técnicas de Machine Learning enfocado a diversos aspectos de la genética y genómica (Libbrecht y Noble 2015), se propone una estrategia novedosa para la identificación de LTR retrotransposones, mediante el uso de técnicas de Machine Learning, logrando realizar una detección más eficiente de los mismos. Es así como se busca responder la siguiente pregunta de investigación, ¿Cuál técnica de preprocesamiento, forma de codificación y modelo de Machine Learning presenta una mejor eficiencia para la identificación de LTR retrotransposones en especies de plantas?

#### **4. REFERENTE TEÓRICO**

Los ET son componentes específicos del genoma que tienen la facilidad moverse de una ubicación a otra (Mita & Boeke, 2016). Estos elementos contribuyen de manera significativa en la diversidad del genoma de las especies y pueden promover la generación de mutaciones y cambios en el mismo, aumentando de igual forma, la cantidad de copias de ET presentes en este (Arango-López et al., 2017). Aquellos ET que utilizan la vía de la molécula de ARN para moverse son llamados Retrotransposones y se catalogan en la Clase I, mientras que aquellos que se mueven usando la molécula de ADN, se denominan transposones y se encuentran en la Clase II (Wicker et al., 2007). Así mismo, los Retrotransposones (Clase I), pueden clasificarse en cuatro órdenes: LTR (Long Terminal Repeat), no-LTR que pueden ser LINEs (Long Interspersed Nuclear Elements) o SINEs (Short Interspersed nuclear elements), PLEs (Penelope-like elements), y elementos DIRs (Schietgat et al., 2018).

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

Específicamente, los LTR retrotransposones son los más abundantes en los genomas de plantas (Gao et al., 2012), estos no son codificantes y están conformados por dos secuencias idénticas al inicio y al final, posee el dominio de U5, R y U3, proteínas gag, env y pol, las cuales son necesarias para el funcionamiento y estructura del retrotransposon. De igual forma, posee regiones que funcionan como primer e inician replicación, estos sitios son PBS (primer binding site) y PPT (Poly-Purine Tract); esta estructura se puede observar a detalle en la figura 1 (Orozco-Arias et al., 2019).




**Figura 1.** Estructura de los LTR retrotransposones. Recuperado de Orozco-Arias et al. 2019

De igual forma, los LTR retrotransposones se dividen en diversas superfamilias, entre las que se encuentran Gypsy y Copia, las cuales son las más importantes en el dominio eucariota (Esposito et al., 2019). Al ver la importancia de este orden de elementos transponibles, se han desarrollado múltiples herramientas y estrategias que permitan identificarlos y clasificarlos, basadas en bioinformática, la cual es la rama de la informática encargada de aplicar las herramientas computacionales en el ámbito de la biología (Guio y González 2019); es de esta manera, que existen diversas metodologías para la detección de elementos transponibles como: basadas en estructura, basadas en homología, de novo y genómica comparativa (Orozco-Arias et al., 2019).

La técnica basada en estructura, es la más usada para la identificación de LTR retrotransposones, esta hace uso de información de características que se conocen con anterioridad sobre la estructura del elemento, sin necesidad de tener una base de datos de los mismos (Loureiro et al., 2013); al existir una varianza de los dominios de los LTR retrotransposones, en muchas ocasiones, no se realiza correctamente la identificación de los mismos (Orozco-Arias et al., 2019). Existen diversos softwares basados en esta técnica, como: LTR\_FINDER (Xu & Wang, 2007), LTR\_retriever (Ou & Jiang, 2018), LTR\_STRUC (McCarthy & McDonald, 2003) y LTRharvest (Ellinghaus et al., 2008)

La técnica basada en homología es la más usada para la clasificación, esta requiere una referencia de elementos transponibles y mediante alineamientos realiza la comparación y detecta si la secuencia es

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>


un elemento transponible (Hermann et al., 2014), esta metodología necesita una base de datos completa de los elementos transponibles para que realice correctamente la identificación (Orozco-Arias et al., 2019). Algunos softwares basados en esta metodología son: Inpactor (Orozco-arias et al., 2018) y LTRclassifier (Monat et al., 2016).

La técnica de novo busca secuencias similares que se encuentren en múltiples posiciones del genoma, esta utiliza la característica de repetición de los elementos transponibles y no requiere información adicional al genoma inicial (Loureiro et al., 2013), sin embargo, cuando un elemento transponible no se encuentra repetido múltiples veces, puede que no se pueda identificar (Orozco-Arias et al., 2019).

En la estrategia de genómica comparativa, se realiza una comparación de las secuencias de un genoma completo para detectar inserciones o deleciones debidas a elementos transponibles; para esto, se debe contar con un genoma bien anotado, que permita identificar estos comportamientos (Orozco-Arias et al., 2019).

Por otro lado, al momento de realizar una identificación, se quiere distinguir los LTR retrotransposones, de otros elementos del genoma, como ARN, CDS, intrones, exones y otros elementos transponibles que no sean de este orden, es así como se pueden ver como instancias positivas, los LTR retrotransposones e instancias negativas, lo que no corresponde a dichos elementos (Ventola et al., 2017). La creación de estas bases de datos resultan ser altamente efectivas si se desea optimizar la identificación mediante estrategias bioinformáticas o haciendo uso del aprendizaje de máquina o Machine Learning (Orozco-arias et al., 2020).


En los reciente años, los métodos de Machine Learning han sido utilizados para resolver diversos problemas genómicos y de evolución de sistemas biológicos (Larrañaga et al., 2006), este se centra en el estudio de algoritmos computacionales basados en la experiencia previa, que permiten el aprendizaje de máquina y como resultado una inferencia estadística de los datos ingresados (Mjolsness & DeCoste, 2001). En investigaciones actuales, realizadas por Orozco-arias et al. (2019) y por Tabares-Soto et al. (2020) se destacan diversos algoritmos para realizar la clasificación de los TEs, entre los que se encuentran KNN (*K-nearest neighbors*), SVM (*support vector machine*), modelos lineales como, LR (*logistic regression*), LDA (*linear discriminant analysis*), NB (*naive Bayesian classifier*), MLP (*multi-layer*

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

*perceptron*) y modelos basados en árboles de decisión (DT) como RF (*random forest*), la principal diferencia con los lineales, es que no varían los resultados al tener características en diferentes escalas.

La principal función de los modelos de Machine Learning es lograr optimizar el rendimiento para que funcione, no únicamente con los datos de entrenamiento, sino, también para datos ingresados posteriormente (Zou et al., 2019). Para medir ese porcentaje de efectividad del modelo, se utilizan diversas métricas utilizadas para observar la efectividad de los modelos computacionales, se encuentran *Recall* (sensibilidad), F1-score, exactitud, precisión, especificidad, coeficiente de rendimiento, tasa de falsos positivos y curva ROC (Orozco-arias et al., 2020), para garantizar un buen entrenamiento del modelo y generalizando su comportamiento.

Para mejorar la efectividad del algoritmo, se utiliza igualmente, diversas estrategias de preprocesamiento de los datos, entre las que se encuentran, aplicación de esquemas de codificación y técnicas con escalamiento y PCA (Análisis de Componentes Principales), dependiendo de los datos. Para el estudio genómico, se han utilizado diversos esquemas de codificación de secuencias, entre los que se encuentran: DAX, EIIP, Complementary, Enthalpy, Galois4, Físico Químico (pc) y k-mers (Orozco-arias et al., 2020).

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

## 5. LOS OBJETIVOS

### General:

Diseñar un identificador automático de LTR retrotransposones en plantas usando técnicas de Machine Learning.

### Específicos:

- Construir una base de datos negativa de LTR retrotransposones en plantas.
- Implementar un pipeline de esquemas de codificación para la detección de los LTR retrotransposones.
- Diseñar y realizar experimentos usando diversos modelos de Machine Learning.
- Establecer los mejores modelos, la técnica de preprocesamiento y la forma de codificación de los datos más óptimos.


## 6. METODOLOGÍA

La presente investigación se define como aplicada con un enfoque mixto, pues se realizará una selección de fuentes de información y análisis cuantitativo de resultados, producto de la identificación de la eficiencia de los algoritmos de Machine Learning para la detección de LTR retrotransposones.

El tipo de investigación es experimental, debido a que se propone una nueva estrategia para la identificación de LTR retrotransposones, basada en técnicas de aprendizaje de máquina, teniendo en cuenta la comparación entre diferentes esquemas de codificación, técnicas de preprocesamiento y modelos de Machine Learning.

Para el desarrollo del proyecto, se llevará a cabo diferentes etapas. En la primera, se realizará una construcción de una base de datos negativa, la cual contenga, las instancias diferentes a los LTR retrotransposones, en esta se incluyen elementos como CDS, diferentes tipos de ARN y otros elementos transponibles que no corresponden a este orden (ET Clase II, PLEs, DIRs, y no-LTR), para esto, se usaron diversas bases de datos como: PGSB PlantsDB (Spannagl et al., 2016), Repbase (Bao



	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

et al., 2015), RepetDB (Amselem et al., 2019), Ensembl Plants (Bolser et al., 2017) y JGI (Joint Genome Institute) (Nordberg et al., 2014).

Después de esto, se extrajeron aleatoriamente 10 mil muestras de cada una de las instancias, para proceder a realizar las pruebas correspondientes. Se utilizan los esquemas de codificación DAX, EIIP, Complementary, Enthalpy, Galois4, Físico Químico (pc) y k-mers para los datos de las secuencias iniciales, cuyo algoritmo se encuentra en el Anexo 1, y se procede a realizar las pruebas con cada uno de los siguientes modelos: KNN, SVM, LR, LDA, NB, MLP, DT y RF, mediante el uso de preprocesamiento de escalamiento, PCA y PCA+escalamiento, como se realizó en Orozco-arias et al. (2020), para la ejecución de esto se usó el algoritmo que se observa en el Anexo 2; cabe destacar que como las secuencias deben tener igual tamaño, dentro de la etapa de preprocesamiento de datos, se encuentra llenar las secuencias con valores de “N” o mediante una técnica de *self replication*, en la que se llenen con repeticiones de la misma secuencia.

Finalmente, se realiza una comparación de los resultados obtenidos, teniendo en cuenta la métrica de F1-score, con el fin de identificar las técnicas de preprocesamiento, formas de codificación y modelo computacional más óptimo para la identificación de LTR retrotransposones y se realizan las primeras pruebas utilizando la base de datos para clasificación de LTR retrotransposones con el fin de unificar ambos problemas y ejecutar una herramienta con la capacidad de identificar y a su vez clasificar dichos elementos esta se puede apreciar en el Anexo 3.

## 7. RESULTADOS

Después de construir la base de datos negativa de LTR retrotransposones, se varía la técnica para el preprocesamiento de los datos, los esquemas de codificación y los modelos planteados anteriormente, obteniendo los siguientes resultados.

### 8.1. Preprocesamiento llenando las secuencias con “N”

Dentro de esta técnica, se encuentran las otras técnicas de preprocesamiento

### 8.1.1. Sin preprocesamiento (Ver figura 1):

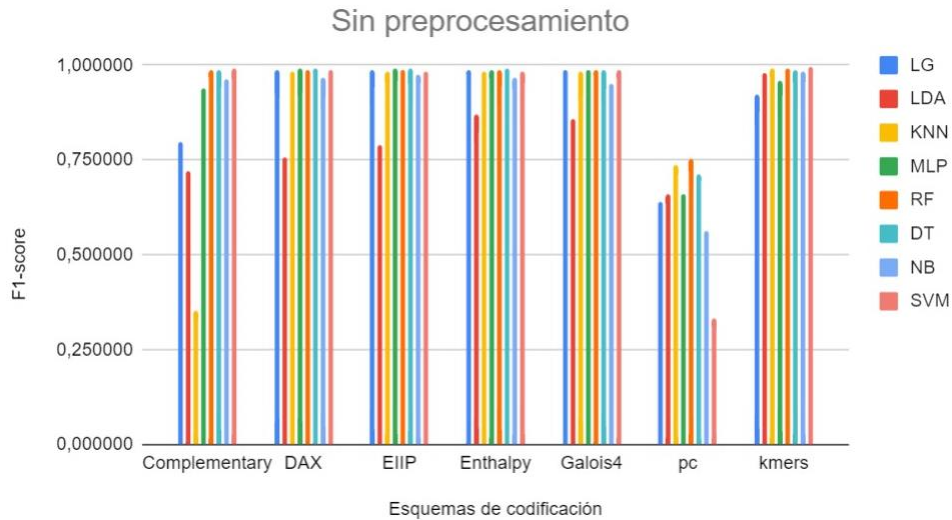


Figura 1. Resultados del proceso de llenado con "N" y sin preprocesamiento

### 8.1.2. Escalamiento (Ver figura 2):

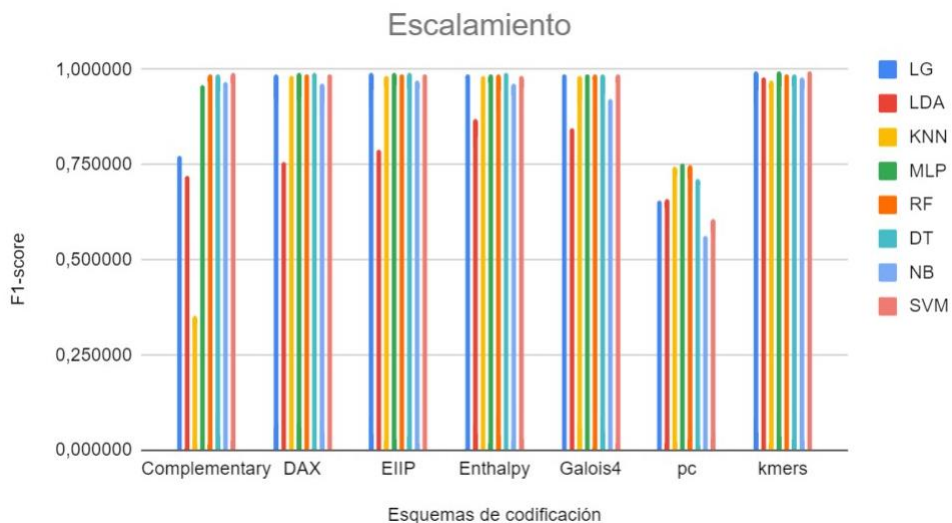


Figura 2. Resultados del proceso de llenado con "N" y con escalamiento



### 8.1.3. PCA (Ver figura 3):

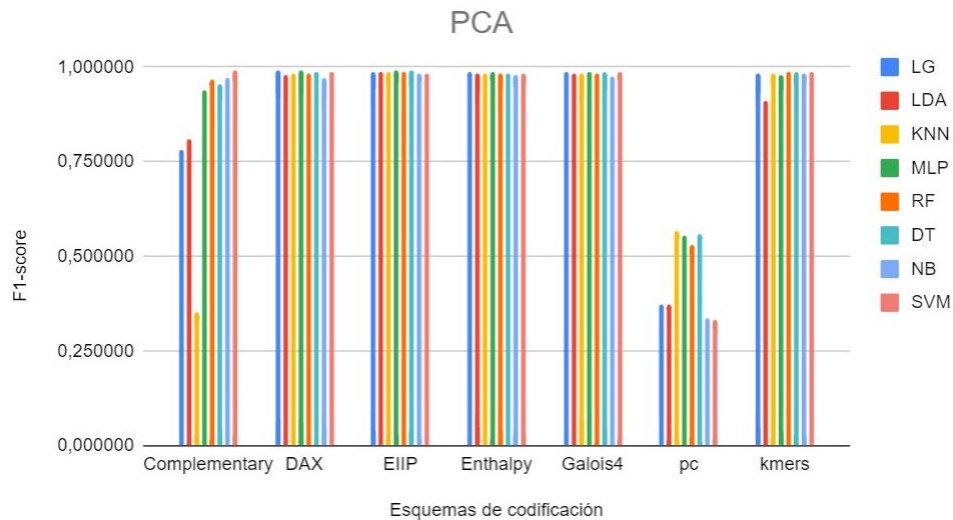


Figura 3. Resultados del proceso de llenado con "N" y PCA

### 8.1.4. PCA+escalamiento (Ver figura 4):

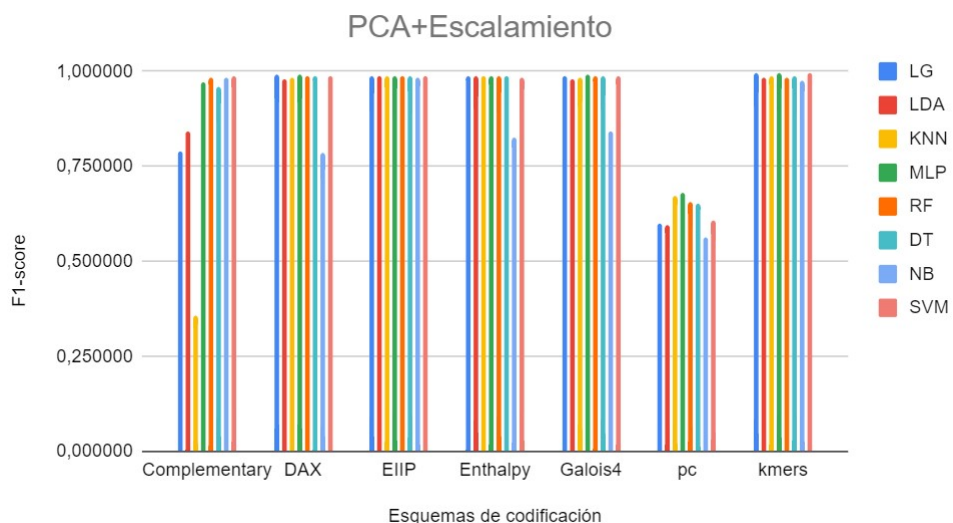
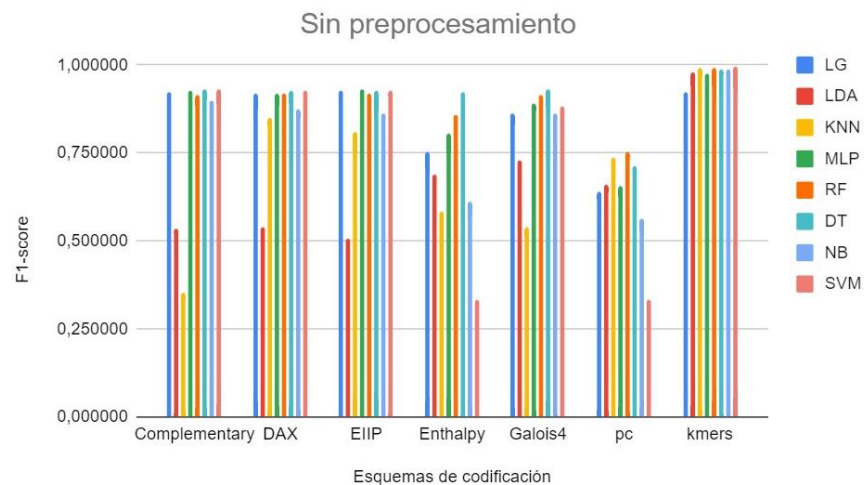


Figura 4. Resultados del proceso de llenado con "N" y PCA con escalamiento

## 8.2. Preprocesamiento llenando las secuencias con técnica de *self replication*

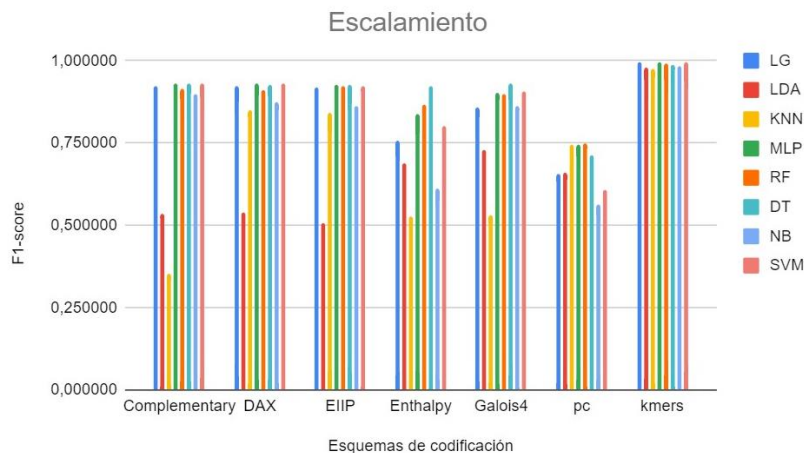
Dentro de esta técnica, se encuentran las otras técnicas de preprocesamiento

### 8.2.1. Sin preprocesamiento (Ver figura 5):



*Figura 5. Resultados del proceso de llenado con self replication y sin preprocesamiento*

### 8.2.2. Escalamiento (Ver figura 6):



*Figura 6. Resultados del proceso de llenado con self replication y escalamiento*

### 8.2.3. PCA (Ver figura 7):

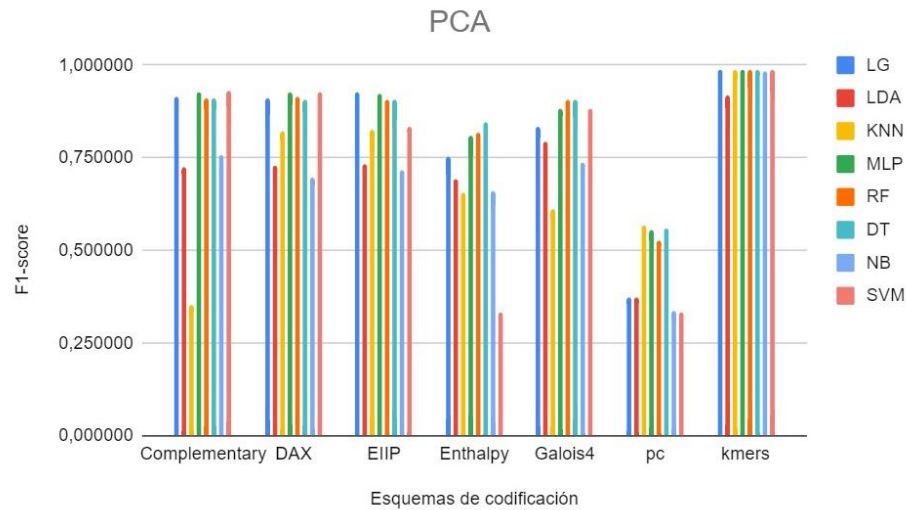


Figura 7. Resultados del proceso de llenado con self replication y PCA

### 8.2.4. PCA+escalamiento (Ver figura 8):

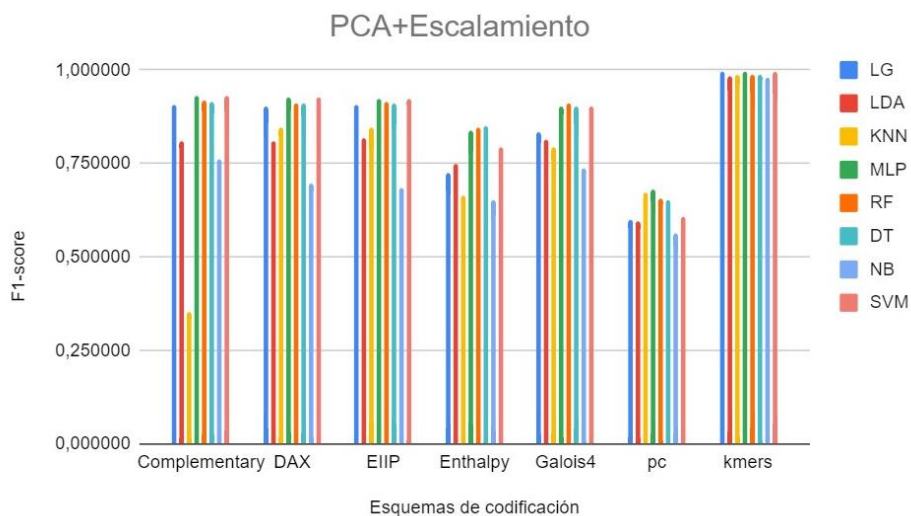


Figura 8. Resultados del proceso de llenado con self replication y PCA con escalamiento

De igual forma, se realizaron las primeras pruebas para unificar los problemas de identificación y clasificación de LTR retrotransposones, utilizando las técnicas de preprocesamiento y esquemas de codificación con los que se obtuvieron mejores resultados: PCA+escalamiento y kmers, utilizando en primera instancia los modelos de KNN, LDA y LR, y se aprecian los siguientes resultados:

- **KNN** (Ver figura 9):

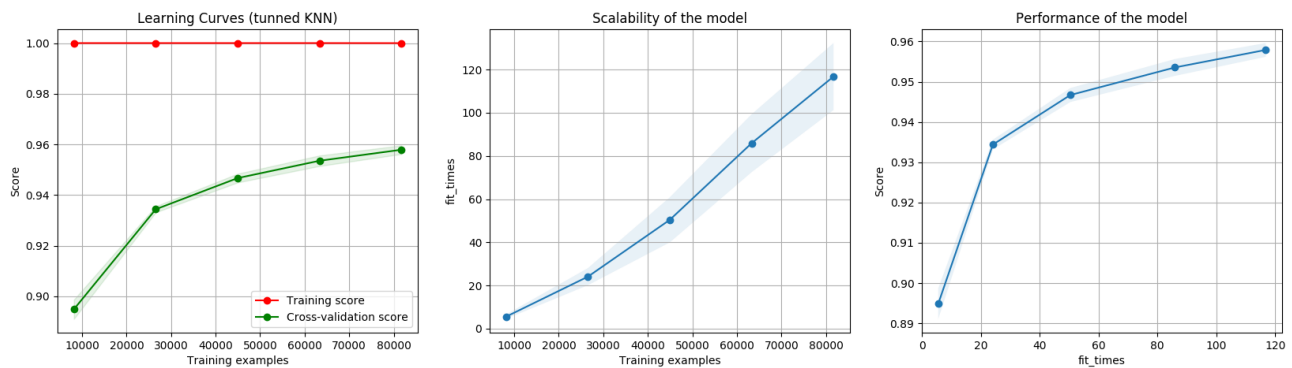


Figura 9. Resultados de identificación y clasificación con KNN

- **LDA** (Ver figura 10):

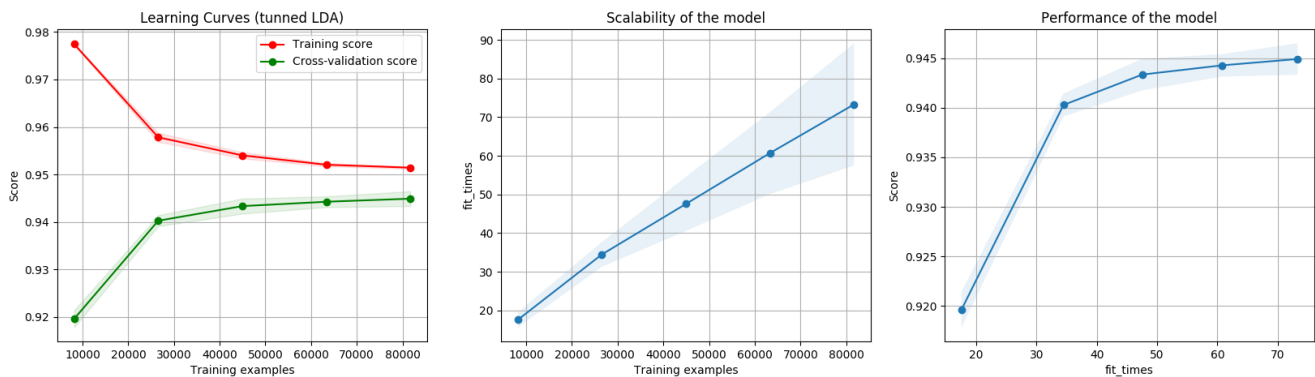


Figura 10. Resultados de identificación y clasificación con LDA

- LR (Ver figura 11):

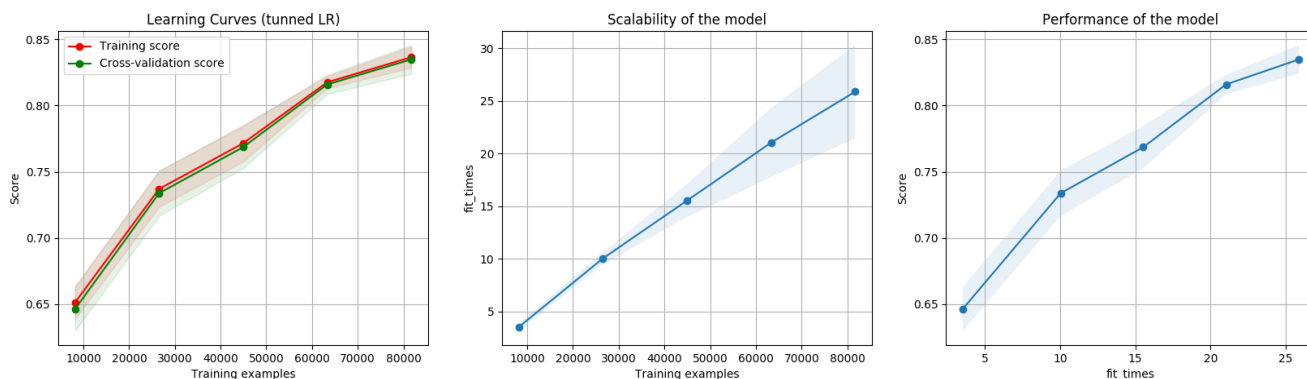



Figura 11. Resultados de identificación y clasificación con LR

## 8. DISCUSIÓN DE RESULTADOS

El proceso de identificación de LTR retrotransposones en plantas es de vital importancia para profundizar en el conocimiento del genoma y la evolución de las especies, considerando el impacto que poseen en la funcionalidad y la manera en la que afecta la capacidad de regulación genética. Dicho proceso, puede tener una alta dificultad, considerando que los LTR retrotransposones suelen poseer dinámicas muy complejas, dificultando su detección y aumentando el trabajo manual. De ahí, la importancia de usar técnicas como el aprendizaje de máquina para automatizar el proceso y optimizar el porcentaje de precisión obtenido.

Gracias a los resultados obtenidos en las gráficas, se puede deducir que el esquema de codificación k-mers, es el que presenta un mejor rendimiento para la mayoría de los modelos analizados, de igual forma, para las técnicas de preprocesamiento, se puede establecer que la técnica de PCA con escalamiento, junto con un llenado de *self replication*, es aquella que presenta un mejor resultado, considerando el porcentaje de precisión y los recursos computacionales que requiere, lo cual se puede confirmar, teniendo en cuenta la revisión realizada de la literatura, en la que se estableció que para clasificar los ET, se requiere primero la realización de preprocesamiento PCA con escalamiento. Al realizar el análisis para observar el modelo de Machine Learning con el que se obtuvo mejores resultados, se puede observar que, para todos los modelos, después de realizar las mejores técnicas

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

de preprocesamiento mencionadas anteriormente, se obtuvo un porcentaje superior al 98% en F1-score, sin embargo, cabe destacar que el modelo con el porcentaje de precisión más alto es MLP con el que se obtuvo 99,7%, considerando este modelo como el más apto para la realización del proceso de identificación de LTR retrotransposones.

Los resultados obtenidos anteriormente, se pueden extrapolar, teniendo en cuenta la literatura en la que se clasifican los ET, con el fin de diseñar una única herramienta automática con la capacidad de identificar y clasificar LTR retrotransposones con un alto porcentaje de precisión, debido a que en la prueba realizada con los tres modelos de KNN, LDA y LR, se obtuvo un porcentaje de f1-score de 95%, 94% y 84% respectivamente, con lo cual se demuestra una prueba satisfactoria, con un porcentaje de identificación y clasificación alto, lo que representa una base indispensable para continuar con las pruebas utilizando los demás modelos planteados en la metodología del presente informe y poder ejecutar una herramienta automática para la resolución de identificación y clasificación de LTR retrotransposones.


## **9. CONCLUSIONES**

La base de datos negativa de LTR retrotransposones en plantas, al ser realizada teniendo en cuenta otros elementos del genoma como CDS, ARN, y ET no LTR, puede ser usada para la ejecución de otros proyectos, en los que se involucre la identificación de otras estructuras en el genoma de las especies, encargadas de diversas funciones, que podrían ser de gran relevancia en otros estudios.

El pipeline de esquemas de codificación, permite codificar cada secuencia, teniendo en cuenta una equivalencia para cada nucleótido, lo cual es de vital importancia, a la hora de ejecutar cualquier proyecto relacionado con aprendizaje de máquina, que involucre secuencias de ADN.

Este trabajo, permite evidenciar la importancia de identificar los ET, en específico los LTR retrotransposones, debido al impacto que presentan para el desarrollo evolutivo de las especies, lo cual permite profundizar en el conocimiento del genoma de las mismas.

Cabe destacar que la herramienta computacional desarrollada permite identificar con un porcentaje de precisión superior al 98% los LTR retrotransposones en plantas, siendo una base para investigaciones


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

futuras, en las que sea necesario la detección de estas estructuras, garantizando un proceso automático, en el que se considere las variaciones existentes de dichos elementos.

Para finalizar, es posible afirmar que el problema de identificación y clasificación de LTR retrotransposones, puede ser unificado en una herramienta computacional, que tenga un porcentaje superior al 82%, mediante la utilización de un modelo automático para la resolución de un problema multiclase, en la cual, si se identifica un LTR retrotransposon, puede ser clasificado correctamente según la superfamilia a la que pertenece.

## 10. RECOMENDACIONES

Se recomienda al momento de crear la base de datos negativa, tener en cuenta la mayor cantidad de elementos que pertenecen al genoma de especies de plantas y que no sean LTR retrotransposones. De igual forma, al momento de correr el esquema de codificación de k-mers, utilizar la mayor cantidad de procesadores, pues este algoritmo corre en paralelo.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

## 11. EVIDENCIA DE RESULTADOS

Se presentan los resultados, teniendo en cuenta, la generación de conocimiento, fortalecimiento de la capacidad y apropiación social del conocimiento y formación de recurso humano.

### 11.1. Formación de recurso humano

<b>Resultado/Producto esperado</b>	<b>Indicador</b>	<b>Beneficiario</b>
Formación de pregrado	Vinculación de estudiantes de pregrado	Estudiantes de la Universidad Autónoma de Manizales


### 11.2. Apropiación social del conocimiento

<b>Resultado/Producto esperado</b>	<b>Indicador</b>	<b>Beneficiario</b>
Ponencia	Participación en Encuentro departamental de Semilleros de investigación	Estudiantes UAM
Divulgación	Presentación de resultados en el foro de investigación UAM	Comunidad UAM

### 11.3. Generación de Nuevo Conocimiento:

<b>Resultado/Producto esperado</b>	<b>Indicador</b>	<b>Beneficiario</b>
Artículo de Revisión y Comparación	Artículo en revista indexada	Comunidad Académica Nacional e Internacional




	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

## 12. IMPACTOS LOGRADOS


<b>Impacto esperado</b>	<b>Plazo (años) después de finalizado el proyecto: corto (1-4), mediano (5-9), largo (10 o más)</b>	<b>Indicador verificable</b>	<b>Supuestos<sup>2</sup></b>
Contribuir a la formación de estudiantes	Corto (1 - 4 años)	Número de estudiantes vinculados al semillero	Divulgación de resultados en revista indexada y foros de investigación
Contribuir a posteriores estudios acerca de la identificación de LTR retrotransposones	Corto (1 - 4 años)	Cantidad de proyectos relacionados con la identificación de LTR retrotransposones	Divulgación de resultados en revista indexada
Contribuir al desarrollo de anotación de genomas	Corto (1 - 4 años)	Mejora de la anotación de genomas, identificando con eficiencia los LTR retrotransposones	Divulgación de resultados en revista indexada

<sup>2</sup> Los supuestos indican los acontecimientos, las condiciones o las decisiones, necesarios para que se logre el impacto esperado.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

### 13. BIBLIOGRAFÍA

- Amselem, J., Cornut, G., Choise, N., Alaux, M., Alfama-Depauw, F., Jamilloux, V., Maumus, F., Letellier, T., Luyten, I., Pommier, C., Adam-Blondon, A. F., & Quesneville, H. (2019). RepetDB: A unified resource for transposable element references. *Mobile DNA*, 10(1), 4–11. <https://doi.org/10.1186/s13100-019-0150-y>
- Arango-López, J., Orozco-Arias, S., Salazar, J. A., Guyot, R., Arango-Lopez, J., Orozco-Arias, S., Salazar, J. A., & Guyot, R. (2017). Application of Data Mining Algorithms to Classify Biological Data: The Coffea canephora Genome Case. In *Communications in Computer and Information Science* (Vol. 735, pp. 156–170). [https://doi.org/10.1007/978-3-319-66562-7\\_12](https://doi.org/10.1007/978-3-319-66562-7_12)
- Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6(1), 4–9. <https://doi.org/10.1186/s13100-015-0041-9>
- Bolser, D. M., Staines, D. M., Perry, E., & Kersey, P. J. (2017). Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomic data. *Methods in Molecular Biology*, 1533, 1–31. [https://doi.org/10.1007/978-1-4939-6658-5\\_1](https://doi.org/10.1007/978-1-4939-6658-5_1)
- Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., & others. (2018). Ten things you should know about transposable elements. *Genome Biology*, 19(1), 199.
- Cui, X., & Cao, X. (2014). Epigenetic regulation and functional exaptation of transposable elements in higher plants. *Current Opinion in Plant Biology*, 21(Figure 1), 83–88. <https://doi.org/10.1016/j.pbi.2014.07.001>
- Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9. <https://doi.org/10.1186/1471-2105-9-18>
- Esposito, S., Barteri, F., Casacuberta, J., Mirouze, M., Carputo, D., & Aversano, R. (2019). LTR-TEs abundance, timing and mobility in Solanum commersonii and S. tuberosum genomes following cold-stress conditions. *Planta*, 250(5), 1781–1787.


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

<https://doi.org/10.1007/s00425-019-03283-3>

- Gao, D., Jimenez-Lopez, J. C., Iwata, A., Gill, N., & Jackson, S. A. (2012). Functional and Structural Divergence of an Unusual LTR Retrotransposon Family in Plants. *PLoS ONE*, 7(10), 1–12. <https://doi.org/10.1371/journal.pone.0048595>
- Guio, L., & González, J. (2019). *Evolutionary Genomics Statistical and Computational Methods Second Edition Methods* (M. Anisimova & Wadenswil Suiza. Humana Press (eds.)). <http://www.springer.com/series/7651>
- Hermann, D., Egue, F., Tastard, E., Nguyen, D. H., Casse, N., Caruso, A., Hiard, S., Marchand, J., Chénais, B., Morant-Manceau, A., & Rouault, J. D. (2014). An introduction to the vast world of transposable elements - What about the diatoms? *Diatom Research*, 29(1), 91–104. <https://doi.org/10.1080/0269249X.2013.877083>
- Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1), 86–112. <https://doi.org/10.1093/bib/bbk007>
- Libbrecht, M. W., & Noble, W. S. (2015). Machine learning in genetics and genomics. *Nature Review Genetics*, 16(6), 321–332. <https://doi.org/10.1038/nrg3920>
- Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, 14(1), 49–61. <https://doi.org/10.1038/nrg3374>
- Loureiro, T., Camacho, R., Vieira, J., & Fonseca, N. A. (2013). Improving the performance of Transposable Elements detection tools. *Journal of Integrative Bioinformatics*, 10(3), 231. <https://doi.org/10.2390/biecoll-jib-2013-231>
- Loureiro, T., Fonseca, N., & Camacho, R. (2012). *Application of Machine Learning techniques on the Discovery and annotation of Transposons in genomes*. 1, 1–3. <http://paginas.fe.up.pt/~ei07087/dokuwiki/files/Abstract.pdf>
- Makałowski, W., Gotea, V., Pande, A., & Makałowska, I. (2019). Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. In Maria Anisimova. *Evolutionary Genomics. Methods in Molecular Biology* (Ed.),


	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

- Evolutionary Genomics* Maria Anisimova Editor *Statistical and Computational Methods* (2nd ed., Vol. 1910). [https://doi.org/https://doi.org/10.1007/978-1-4939-9074-0\\_6](https://doi.org/https://doi.org/10.1007/978-1-4939-9074-0_6)
- McCarthy, E. M., & McDonald, J. F. (2003). LTR STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics*, 19(3), 362–367. <https://doi.org/10.1093/bioinformatics/btf878>
- McClintock, B. (1953). Induction of Instability at Selected Loci in Maize. *Genetics*, 38(6), 579–599. <http://www.ncbi.nlm.nih.gov/pubmed/17247459> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1209627>
- Mita, P., & Boeke, J. D. (2016). How retrotransposons shape genome regulation. *Current Opinion in Genetics & Development*, 37, 90–100. <https://doi.org/10.1016/j.gde.2016.01.001>.
- Mjolsness, E., & DeCoste, D. (2001). Machine learning for science: State of the art and future prospects. *Science*, 293(5537), 2051–2055. <https://doi.org/10.1126/science.293.5537.2051>
- Monat, C., Tando, N., Tranchant-Dubreuil, C., & Sabot, F. (2016). LTRclassifier: A website for fast structural LTR retrotransposons classification in plants. *Mobile Genetic Elements*, 6(6), e1241050. <https://doi.org/10.1080/2159256x.2016.1241050>
- Munoz-Lopez, M., & Garcia-Perez, J. (2010). DNA Transposons: Nature and Applications in Genomics. *Current Genomics*, 11(2), 115–128. <https://doi.org/10.2174/138920210790886871>
- Neumann, P., Novák, P., Hošťáková, N., & MacAs, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA*, 10(1), 1–17. <https://doi.org/10.1186/s13100-018-0144-1>
- Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, 42(D1), 26–31.

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

<https://doi.org/10.1093/nar/gkt1069>

- Orozco-Arias, S., Isaza, G., & Guyot, R. (2019). Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning. *International Journal of Molecular Sciences*, 20(15), 1–29. <https://doi.org/10.3390/ijms20153837>
- Orozco-arias, S., Isaza, G., Guyot, R., & Tabares-soto, R. (2019). A systematic review of the application of machine learning in the detection and classification of transposable elements. *Peerj*, 7, 18311. <https://doi.org/10.7717/peerj.8311>
- Orozco-arias, S., Liu, J., Id, R. T., Ceballos, D., Silva, D., Id, D., Ming, R., & Guyot, R. (2018). Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. *Biology*. <https://doi.org/10.3390/biology7020032>
- Orozco-arias, S., Piña, J. S., Tabares-soto, R., & Castillo-ossa, L. F. (2020). Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. *Processes*, 8(638), 1–20. <https://doi.org/10.3390/pr8060638>
- Ou, S., & Jiang, N. (2018). LTR\_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, 176(2), 1410–1422. <https://doi.org/10.1104/pp.17.01310>
- Schietgat, L., Vens, C., Cerri, R., Fischer, C. N., Costa, E., Ramon, J., Carareto, C. M. A., & Blockeel, H. (2018). A machine learning based framework to identify and classify long terminal repeat retrotransposons. *PLOS Computational Biology*, 14(4), e1006097. <https://doi.org/10.1371/journal.pcbi.1006097>
- Spannagl, M., Nussbaumer, T., Bader, K. C., Martis, M. M., Seidel, M., Kugler, K. G., Gundlach, H., & Mayer, K. F. X. (2016). PGSB plantsDB: Updates to the database framework for comparative plant genome research. *Nucleic Acids Research*, 44(D1), D1141–D1147. <https://doi.org/10.1093/nar/gkv1130>
- Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V. S., Rodríguez-Sotelo, J. L., & Jiménez-Varón, C. F. (2020). A comparative study of machine learning and deep

	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

learning algorithms to classify cancer types based on microarray gene expression data.

*PeerJ Computer Science*, 2020(4), 1–22. <https://doi.org/10.7717/peerj-cs.270>

Ventola, G. M. M., Noviello, T. M. R., D’Aniello, S., Spagnuolo, A., Ceccarelli, M., & Cerulo, L. (2017). Identification of long non-coding transcripts with feature selection: a comparative study. *BMC Bioinformatics*, 18(1), 187. <https://doi.org/10.1186/s12859-017-1594-z>

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12), 973–982. <https://doi.org/10.1038/nrg2165>

Xu, Z., & Wang, H. (2007). LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35(SUPPL.2), 265–268. <https://doi.org/10.1093/nar/gkm286>

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, 51(1), 12–18. <https://doi.org/10.1038/s41588-018-0295-5>



## GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM

CÓDIGO: GIN-GUI-001

VERSIÓN: 01

FECHA ELABORACIÓN  
DEL DOCUMENTO:  
23/ENE/2015

### 14. ANEXOS

#### Anexo 1. Algoritmo para implementar esquemas de codificación

Ver en carpeta





## GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM

CÓDIGO: GIN-GUI-001

VERSIÓN: 01


FECHA ELABORACIÓN  
DEL DOCUMENTO:  
23/ENE/2015

### Anexo 2. Algoritmo de los modelos de Machine Learning

Ver en carpeta





	<b>GUÍA PARA PRESENTACIÓN DE INFORMES FINALES UAM</b>	<b>CÓDIGO: GIN-GUI-001</b>
		<b>VERSIÓN: 01</b>
		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 23/ENE/2015</b>

### **Anexo 3. Algoritmo para integrar identificación y clasificación**

Ver en carpeta

