**PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ**

**FACULTAD DE CIENCIAS E INGENIERÍA**

****

**Desarrollo de un modelo de aprendizaje profundo para identificar moléculas ARN no codificantes**

**Tesis para optar por el Título de Ingeniero Informático que presenta el bachiller:**

**José Luis Santillán Escudero**

**20030307**

**Asesor: Dr. Edwin Rafael Villanueva Talavera**

Lima, abril de 2019

# Tabla de Contenido

[Tabla de Contenido 2](#_Toc6609431)

[Índice de Figuras 5](#_Toc6609432)

[Índice de Tablas 6](#_Toc6609433)

[Capítulo 1. Generalidades 7](#_Toc6609434)

[1.1 Problemática 7](#_Toc6609435)

[1.2 Objetivos 9](#_Toc6609436)

[1.2.1 Objetivo general 9](#_Toc6609437)

[1.2.2 Objetivos específicos 9](#_Toc6609438)

[1.2.3 Resultados esperados 9](#_Toc6609439)

[1.2.4 Mapeo de objetivos, resultados y verificación 10](#_Toc6609440)

[1.3 Herramientas y Métodos 11](#_Toc6609441)

[1.3.1 Bases de datos de transcriptomas 12](#_Toc6609442)

[1.3.2 Lenguajes de programación y librerías 13](#_Toc6609443)

[1.3.3 Aprendizaje de Máquina 15](#_Toc6609444)

[1.3.4 Métodos de Validación 16](#_Toc6609445)

[Capítulo 2. Marco Conceptual 18](#_Toc6609446)

[2.1 Biología Molecular 18](#_Toc6609447)

[2.1.1 ADN 18](#_Toc6609448)

[2.1.2 ARN – proceso de transcripción 18](#_Toc6609449)

[2.1.3 Proteínas – síntesis mediante la traducción de ARN 19](#_Toc6609450)

[2.2 ARN no codificante 19](#_Toc6609451)

[2.2.1 Historia 19](#_Toc6609452)

[2.2.2 Clasificación 20](#_Toc6609453)

[2.2.3 Proyectos de secuenciación de transcriptomas 21](#_Toc6609454)

[2.3 Aprendizaje de máquina 21](#_Toc6609455)

[2.3.1 Definición 21](#_Toc6609456)

[2.3.2 Categorización 21](#_Toc6609457)

[2.3.3 Clasificadores con aprendizaje de máquina 22](#_Toc6609458)

[Capítulo 3. Estado del Arte 24](#_Toc6609459)

[3.1 Preguntas de investigación 24](#_Toc6609460)

[3.2 Estrategia de búsqueda 25](#_Toc6609461)

[3.2.1 Términos de búsqueda 25](#_Toc6609462)

[3.2.2 Proceso de búsqueda 25](#_Toc6609463)

[3.3 Criterios de inclusión y exclusión 25](#_Toc6609464)

[3.4 Extracción de la información 26](#_Toc6609465)

[3.5 Otras Tesis encontradas 28](#_Toc6609466)

[3.6 Resultados 28](#_Toc6609467)

[3.6.1 Métodos más usados 28](#_Toc6609468)

[3.6.2 Principales limitaciones 29](#_Toc6609469)

[3.7 Conclusiones 31](#_Toc6609470)

[Capítulo 4. Recopilación y preprocesamiento de datos 33](#_Toc6609471)

[4.1 Fuentes de información 33](#_Toc6609472)

[4.2 Selección de especies 33](#_Toc6609473)

[4.3 Método de obtención de las fuentes 37](#_Toc6609474)

[4.3.1 Ensembl 37](#_Toc6609475)

[4.3.2 CantataDB 38](#_Toc6609476)

[4.3.3 GreeNC 38](#_Toc6609477)

[4.4 Selección de secuencias a utilizar en el modelo 38](#_Toc6609478)

[4.5 Selección de características a utilizar para el modelo 38](#_Toc6609479)

[4.5.1 Características calculadas por scripts propios 39](#_Toc6609480)

[4.5.2 Características calculadas a través de CPAT 40](#_Toc6609481)

[4.5.3 Características calculadas a través de Diamond 43](#_Toc6609482)

[Referencias 45](#_Toc6609483)

[Anexos a](#_Toc6609484)

[Anexo A : Plan de Proyecto a](#_Toc6609485)

[Anexo B : Descarga de Ensembl a través de biomaRt i](#_Toc6609486)

[Anexo C : Procesamiento de descargas manuales de Ensembl k](#_Toc6609487)

[Anexo D : Procesamiento de archivos fasta de CantataDB m](#_Toc6609488)

[Anexo E : Procesamiento de archivos fasta de GreeNC o](#_Toc6609489)

[Anexo F : Selección aleatoria de transcriptomas para el modelo q](#_Toc6609490)

[Anexo G : Características – Longitud del transcriptoma r](#_Toc6609491)

[Anexo H : Características – Porcentaje de nucleótidos GC s](#_Toc6609492)

[Anexo I : Generación de archivos fasta para CDS y lncRNA t](#_Toc6609493)

[Anexo J : CPAT – Generación de tabla de frecuencias de hexámeros v](#_Toc6609494)

[Anexo K : Generación de archivos fasta para ARN codificantes w](#_Toc6609495)

[Anexo L : CPAT – Generación de regresión logística x](#_Toc6609496)

[Anexo M : CPAT – Ejecución de CPAT y](#_Toc6609497)

[Anexo N : CPAT – Importación de características a la base de datos aa](#_Toc6609498)

[Anexo O : Diamond – Generación de la base de datos cc](#_Toc6609499)

[Anexo P : Diamond – Búsqueda de alineamientos dd](#_Toc6609500)

[Anexo Q : Diamond – Importación de características a la base de datos ee](#_Toc6609501)

Índice de Figuras

Figura 1 - Síntesis de proteínas y ARN no codificante. Elaboración propia. 7

Figura 2 - Árbol de problema. Elaboración propia. 9

Figura 3 - k-fold cross validation. Elaboración propia. 16

Figura 4 - Tipos de resultados de un predictor. Elaboración propia. 17

Figura 5 - Esquema del ADN. Adaptado de (Risueño Pérez, 2012). 18

Figura 6 - Estructura del ARN. Adaptado de (Vieira, 2018). 19

Figura 7 - Diferencia entre aprendizaje tradicional y aprendizaje profundo. Adaptado de (Rouse, 2017). 22

Figura 8 - Número de investigaciones relacionadas a aprendizaje profundo en Bioinformática en Scopus. Elaboración propia. 23

Figura 9 - Diferencia visualizada en Biomart (Ensembl) entre transcriptoma y secuencia codificante 41

Figura 10 - Estructura de descomposición del trabajo. Elaboración propia. d

Figura 11 - Cronograma del proyecto por semanas. Elaboración propia. f

Índice de Tablas

Tabla 1 - Mapeo de objetivos, resultados y verificación. 10

Tabla 2 - Herramientas y métodos. 11

Tabla 3 - Clasificación del ARN. Adaptado de (Genomics, 2015). 20

Tabla 4 - Conceptos generales definidos con el método PICOC. 24

Tabla 5 - Componentes de la cadena de búsqueda. 25

Tabla 6 - Artículos encontrados por base de datos. 26

Tabla 7 - Artículos que discuten el estado del arte. 27

Tabla 8 - Métodos de aprendizaje de máquina más usados para la clasificación de lncRNA. 28

Tabla 9 - Número de especies en las bases de datos de transcriptomas 33

Tabla 10 - Especies con el mayor conteo de clases positivas, y sus fuentes 34

Tabla 11 - Número total de transcriptomas a utilizar según el modelo seleccionado 35

Tabla 12 - Ejemplo de tabla de frecuencia de hexámeros para Amborella trichopoda generada por CPAT 42

Tabla 13 - Riesgos del proyecto de investigación. c

Tabla 14 - Lista de tareas del proyecto. e

Tabla 15 - Costos del proyecto. h

# Generalidades

## Problemática

El ADN (ácido desoxirribonucleico) es una molécula compuesta principalmente por nucleótidos, que se encuentra en el núcleo de la célula y es responsable de generar proteínas y ARN (ácido ribonucleico) para que cumplan todas las funciones celulares (NHGRI, 2015).

Esta generación de proteínas en las células es posible a través del proceso de síntesis de proteínas. Para ello, se genera una copia de un segmento del ADN (gen) que contiene la información de la proteína a sintetizar a través del proceso de transcripción, produciéndose una molécula intermediaria de ARN con la información del gen, la cual mediante el proceso de traslación se puede expresar en proteína (NHGRI, 2015). Sin embargo, solo una pequeña proporción del ARN producido por el ADN se expresa en proteínas, y a estos se les conoce como ARN codificantes (o en inglés, como *protein coding transcripts* – *PCT*) (Kung, Colognori, & Lee, 2013). El ARN que no se expresa en proteínas es conocido como ARN no codificante (ver Figura 1).

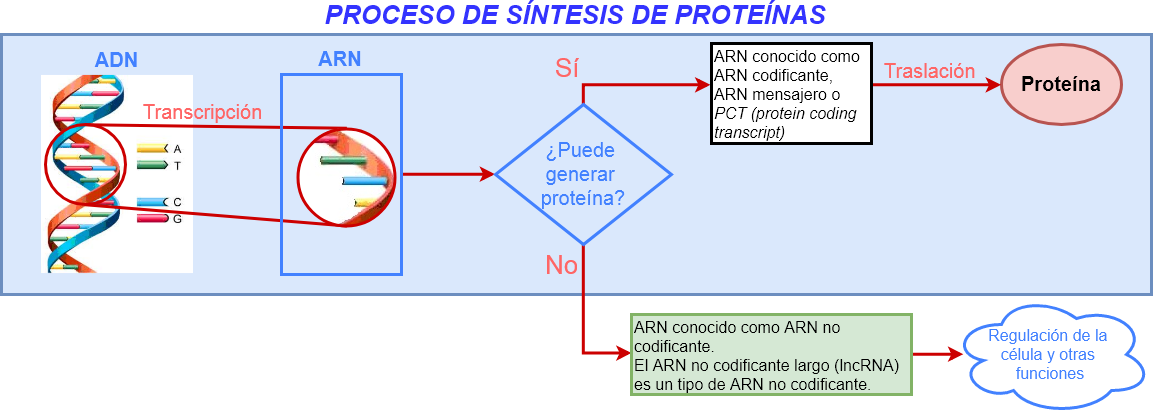


Figura 1 - Síntesis de proteínas y ARN no codificante. Elaboración propia.

Históricamente se pensó que las secciones del ADN que se transcriben a ARN no codificantes eran simplemente rezagos evolutivos de material genético, que perdió utilidad, y por lo tanto no tenían ninguna función biológica (Ponjavic, Ponting, & Lunter, 2007). Pero con el pasar del tiempo se fue descubriendo que estas secciones sí cumplían funciones, y en particular hoy en día se sabe que los ARN no codificantes largos (del inglés *long non-coding* *RNA*, o *lncRNA*) cumplen funciones reguladoras en la célula y que tienen implicancias en el desarrollo de enfermedades como el cáncer (Kung et al., 2013). Es así que estas regiones no codificantes toman importancia para el entendimiento del genoma, sobre todo al ser un componente mayoritario del mismo (Uszczynska-Ratajczak, Lagarde, Frankish, Guigo, & Johnson, 2018).

La definición clásica que se tiene sobre los *lncRNAs* es que son secciones no codificantes conformadas por más de 200 nucleótidos (Yao et al., 2018). Sin embargo, diferenciar *lncRNAs* de *PCTs* es una tarea desafiante, dada su similitud estructural (Quinn & Chang, 2016). La correcta clasificación de estas moléculas es importante, ya que pueden tener un impacto profundo en proyectos que se apoyan en estas anotaciones (Uszczynska-Ratajczak et al., 2018).

En el presente trabajo se pretende aplicar técnicas computacionales de aprendizaje de máquina (AM) para la clasificación de *lncRNAs* en especies desconocidas para el modelo. En la literatura existen técnicas basadas en AM para discriminar *lncRNAs* de *PCTs*. Sin embargo, el objetivo de tales estudios ha sido generar predictores para especies específicas, con base a muestras experimentalmente anotadas en esas mismas especies. El problema se presenta cuando se secuencian transcriptomas en especies no modeladas (Schneider, Raiol, Brigido, Walter, & Stadler, 2017). En tales casos, no se dispone de un conjunto de muestras anotadas para realizar el aprendizaje de modelos predictivos, y justamente es en esos casos donde más se necesita contar con tales modelos para hacer una discriminación inicial de *lncRNAs* y *PCTs* y ayudar así a focalizar la validación experimental (Simopoulos, Weretilnyk, & Golding, 2018).

Dentro de este contexto, se propone como proyecto de fin de carrera el desarrollo de un modelo computacional de aprendizaje profundo, para investigar la viabilidad de crear modelos de predicción de *lncRNAs* para especies desconocidas, entrenándolos con información de otras especies ya anotadas.

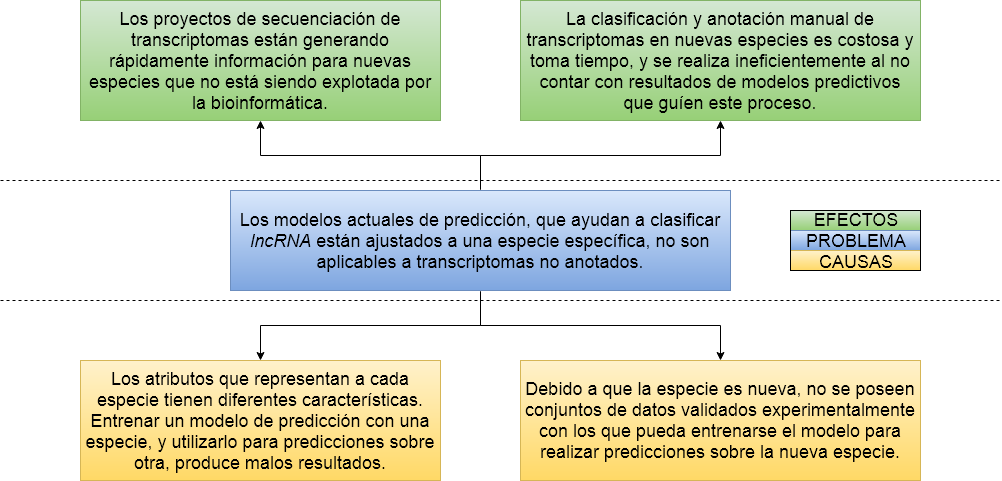


Figura 2 - Árbol de problema. Elaboración propia.

## Objetivos

### Objetivo general

Elaborar y evaluar un modelo computacional, basado en aprendizaje profundo, para la identificación de moléculas ARN no codificantes en transcriptomas de especies desconocidas al modelo.

### Objetivos específicos

1. Recopilar y preprocesar un conjunto de datos de transcriptomas para el entrenamiento y validación del modelo algorítmico.
2. Implementar la arquitectura del proceso de aprendizaje de máquina y que la misma produzca predictores aceptables.
3. Realizar un análisis comparativo de la precisión de los predictores generados contra otro modelo del estado del arte.

### Resultados esperados

1. Conjuntos de datos estructurados recolectados y preprocesados de transcriptomas (O1).
2. Arquitectura del proceso de Aprendizaje de Máquina (O2).
3. Análisis comparativo de la precisión de los predictores generados y otro modelo similar del estado del arte (O3).

### Mapeo de objetivos, resultados y verificación

En la Tabla 1 se presentan los objetivos específicos planteados, juntos a sus resultados esperados, mapeados con los entregables y su respectiva verificación.

Tabla 1 - Mapeo de objetivos, resultados y verificación.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Objetivo:** Recopilar y preprocesar un conjunto de datos de transcriptomas para el entrenamiento y validación del modelo algorítmico. | | | | |
| **Resultado** | | **Meta física** | **Medio de verificación** | |
| Conjuntos de datos estructurados recolectados y preprocesados de transcriptomas. | | Conjunto de datos listo para su uso.  Cuadros y gráficos estadísticos. | * Documento del proceso de recolección y preprocesamiento de transcriptomas. * Estadísticas descriptivas de los datos. | |
| **Objetivo:** Implementar la arquitectura del proceso de aprendizaje de máquina y que la misma produzca predictores aceptables. | | | | |
| **Resultado** | **Meta física** | | | **Medio de verificación** |
| Arquitectura del proceso de Aprendizaje de Máquina. | Arquitectura implementada.  Documento de reporte de resultados.  Cuadros y gráficos estadísticos. | | | * Documento detallando la arquitectura implementada. * Precisión de los predictores generados para clasificar especies para las cuales no fue entrenado. |
| **Objetivo:** Realizar un análisis comparativo de la precisión de los predictores generados contra otro modelo del estado del arte. | | | | |
| **Resultado** | | **Meta física** | **Medio de verificación** | |
| Análisis comparativo de la precisión de los predictores generados y otro modelo similar del estado del arte. | | Documento de reporte comparativo.  Cuadros y gráficos estadísticos. | * Comparación de la precisión de los predictores generados y otro modelo del estado del arte para clasificar especies desconocidas al modelo, al ser entrenados en otras. | |

## Herramientas y Métodos

A continuación, en Tabla 2, se detallan las herramientas y métodos que ayudarán con la obtención y/o validación de cada uno de los resultados esperados planteados.

Tabla 2 - Herramientas y métodos.

|  |  |
| --- | --- |
| **Objetivo:** Recopilar y preprocesar un conjunto de datos de transcriptomas para el entrenamiento y validación del modelo algorítmico. | |
| **Resultado** | **Herramientas o métodos** |
| Conjuntos de datos estructurados recolectados y preprocesados de transcriptomas. | * Ensembl * CantataDB * GreeNC * BLAST * Anaconda * Python * R * BiomaRt * Jupyter Notebook * Matplotlib y Seaborn |
| **Objetivo:** Implementar la arquitectura del proceso de aprendizaje de máquina y que la misma produzca predictores aceptables. | |
| **Resultado** | **Herramientas o métodos** |
| Arquitectura del proceso de Aprendizaje de Máquina. | * Redes neuronales profundas * Anaconda * Tensorflow * Keras * Python * Jupyter Notebook * Matplotlib y Seaborn * *k-Fold cross validation* * Medidas de precisión: exactitud, precisión y exhaustividad |
| **Objetivo:** Realizar un análisis comparativo de la precisión de los predictores generados contra otro modelo similar del estado del arte. | |
| **Resultado** | **Herramientas o métodos** |
| Análisis comparativo de la precisión de los predictores generados y otro modelo del estado del arte. | * Anaconda * Python * R * Jupyter Notebook * Matplotlib y Seaborn * *k-Fold cross validation* * Medidas de precisión: exactitud, precisión y exhaustividad |

### Bases de datos de transcriptomas

#### Ensembl

Los proyectos de secuenciación de transcriptomas buscan generar la secuencia completa de ARN de un organismo (Vieira, 2018). Debido a la aparición de nuevos métodos de secuenciación, los proyectos de secuenciación de transcriptomas ahora son más accesibles (Uszczynska-Ratajczak et al., 2018), y existe una multitud de fuentes disponibles para transcriptomas generados por estos proyectos.

Dentro de estas fuentes, Ensembl (http://www.ensembl.org) posee una gran cantidad de información de transcriptomas de diferentes especies. Adicionalmente, activamente desarrollan APIs públicas para acceder de forma más directa a sus servicios (Zerbino et al., 2018).

#### CantataDB

CantataDB es una base de datos de *lncRNAs* en línea ([http://cantata.amu.edu.pl](http://cantata.amu.edu.pl/)), que nació para suplir la necesidad de bases de datos más integrales de transcriptomas, específicamente de *lncRNAs*, para plantas (Szcześniak, Rosikiewicz, & Makałowska, 2016). Actualmente posee información para 36 especies de plantas, y permite realizar búsquedas en línea y descarga de transcriptomas.

#### GreeNC

GreeNC es otra una base de datos de *lncRNAs* en línea ([http://greenc.sciencedesigners.com](http://greenc.sciencedesigners.com/)), que también nació de la necesidad de tener anotaciones de ARN no codificantes largos en plantas, para lo que desarrollaron un flujo para la anotación de estas moléculas y lo aplicaron a 37 especies (Paytuví Gallart, Hermoso Pulido, Anzar Martínez de Lagrán, Sanseverino, & Aiese Cigliano, 2016). Actualmente posee información para 45 especies de plantas, y permite realizar búsquedas en línea y descarga de transcriptomas.

#### BLAST

BLAST son las siglas de *basic local alignment search tool* (herramienta de búsqueda de alineamiento local básico). Es un algoritmo que permite realizar búsquedas en base a la similitud que presenten secuencias de aminoácidos, o secuencias de nucleótidos de ADN y/o ARN (Mount, 2007).

### Lenguajes de programación y librerías

#### Anaconda

Anaconda (<https://www.anaconda.com>) es una plataforma de código abierto que puede gestionar la instalación completa de todos los elementos necesarios (incluyendo dependencias) para la elaboración de trabajos científicos. Debido a que centraliza la gestión de toda la plataforma necesaria para la investigación, se optó por su uso.

Todos los elementos mencionados en esta sección son instalables y administrables a través de Anaconda.

#### Python

Python es un lenguaje de programación que en los últimos años ha ganado popularidad. Entre sus principales ventajas está su facilidad de uso, el soporte que se puede encontrar para el lenguaje en la comunidad y la gran cantidad de librerías con integraciones a otros lenguajes o plataformas, como con el lenguaje R y con Tensorflow (Willems, 2015).

#### R

R es un lenguaje de programación está enfocado al análisis de data, estadísticas y gráficos, que también está gozando de una creciente popularidad en los últimos años (Willems, 2015). Su facilidad de uso para el análisis estadístico fue un factor clave para su elección de cara al análisis que se ha de realizar de los distintos modelos predictivos en la presente investigación.

#### BiomaRt

Este paquete para R forma parte del proyecto de código abierto Bioconductor, que desarrolla una multitud de herramientas estadísticas y de gráficos orientadas al análisis de información genómica, para el lenguaje R (Durinck et al., 2005). En particular el paquete biomaRt ofrece una interfaz de programación que permite consultar bases de datos a través de internet, como por ejemplo Ensembl (Durinck, Spellman, Birney, & Huber, 2009).

#### Tensorflow

Tensorflow es una librería de código abierto para aprendizaje de máquina, desarrollada por Google. Se ha optado por su uso debido a que se especializa en modelos con aprendizaje profundo, permitiendo la manipulación de estos a un bajo nivel y un gran nivel de flexibilidad en la implementación de distintos modelos (Abadi et al., 2016).

#### Keras

Keras es una API de alto nivel, orientada a facilitar la construcción y el manejo de modelos de redes neuronales profundas. Tensorflow posee una implementación específica de esta API, que facilita en gran medida el uso y ajustes que se necesiten realizar al modelo implementado en Tensorflow (Chollet, 2017).

#### Jupyter Notebook

Jupyter Notebook (<http://jupyter.org>) es una aplicación web que tiene varias funciones, como editar código en línea, compartir documentos y otras tareas especializadas para trabajos científicos. Se utilizará porque es ideal para el trabajo colaborativo científico, ya que expone códigos de programas con textos e imágenes descriptivos de los mismos, que pueden ser editados y ejecutados en línea (Shen, 2014).

#### Matplotlib y Seaborn

Matplotlib es una librería de código abierto para Python que permite generar gráficos y visualizaciones de data. Tiene una gran base de usuarios y soporte, además de varias otras APIs construidas sobre esta librería (VanderPlas, 2016).

Una de estas APIs es Seaborn, que ayuda a extender la funcionalidad de Matplotlib, que se estaba quedando desfasada, sobre todo en términos de la calidad de los gráficos que genera, además de ofrecer un uso más simple de Matplotlib (VanderPlas, 2016).

### Aprendizaje de Máquina

#### Redes neuronales profundas

Las redes neuronales son algoritmos de aprendizaje de máquina que fueron modelados en base al cerebro humano. La unidad básica de estas redes es la neurona, que recibe entradas numéricas y devuelve una salida entre 0 y 1. Estas neuronas se organizan en capas, y cuando la red neuronal posee más de dos capas, se le conoce como red neuronal profunda (Nielsen, 2018).

El uso de las redes neuronales profundas, y sus variantes, es cada vez mayor en el campo de la biología, apoyándose en su poder de abstracción y en el beneficio que obtienen de la cantidad de información disponible gracias a los proyectos de secuenciación de transcriptomas (Jabeen, Ahmad, & Raza, 2018).

Dentro de los distintos tipos de redes neuronales profundas, en el presente proyecto se hará uso de la red neuronal prealimentada (del inglés *feed-forward network*). Son el tipo más básico de red neuronal, en la que las conexiones no forman ciclos y viajan hacia adelante.

### Métodos de Validación

#### k-Fold cross validation

Es un método estadístico ampliamente usado en modelos de aprendizaje de máquina, que consiste, en su versión más simple, en la división de un conjunto de datos en un número k de grupos, de los cuales uno será tomado para la validación del modelo, y el resto para el entrenamiento (Brownlee, 2018).

El modelo será evaluado k veces, hasta que cada grupo haya sido usado para la validación del modelo (ver Figura 3). Al final del proceso obtendremos k mediciones de la precisión del modelo predictivo, de los cuales podemos extraer medidas representativas a través, por ejemplo, de un promedio simple de las mediciones (Brownlee, 2018).

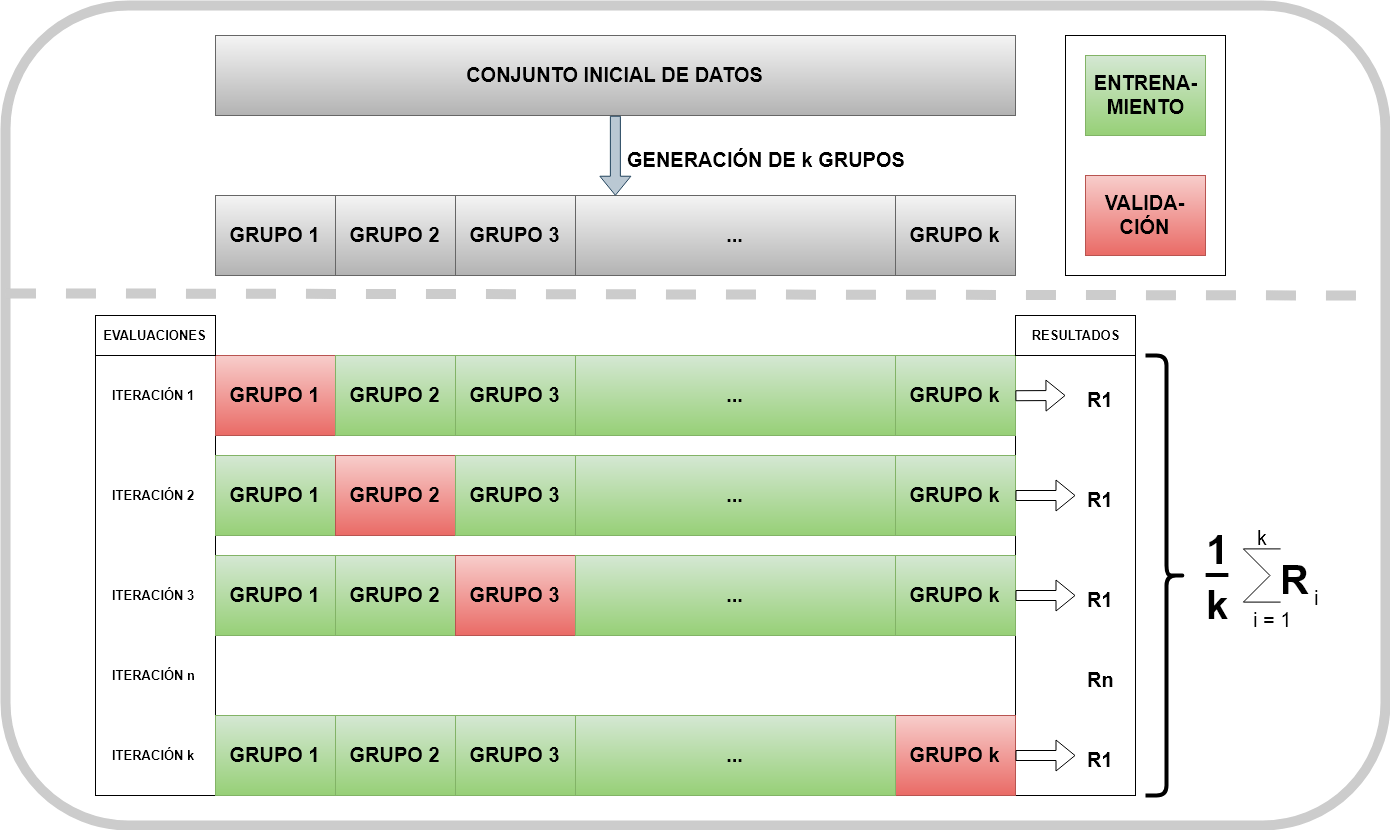


Figura 3 - k-fold cross validation. Elaboración propia.

#### Medidas de precisión: exactitud, precisión y exhaustividad

Para medir la precisión del modelo predictivo, se utilizarán tres medidas: la exactitud, la precisión y la exhaustividad. Estas medidas son clásicamente usadas para evaluar el desempeño de modelos predictivos (Wucher et al., 2017).

Cuando se genera un predictor binario (se llama así porque solamente predice dos clases, en este caso: *lncRNA* o *PCT*), los resultados pueden dividirse en cuatro secciones: verdaderos positivos (VP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (VN). La Figura 4 detalla el significado de cada uno de ellos en el contexto del predictor de *lncRNA*.

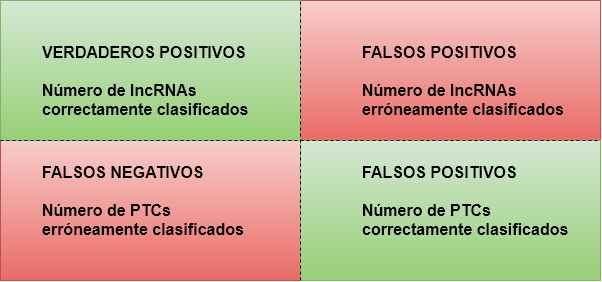


Figura 4 - Tipos de resultados de un predictor. Elaboración propia.

Entonces, para realizar la validación del modelo, las tres medidas se definen de la siguiente manera:

* Exactitud = (VP + VN) / (VP + FP + FN + VN)
* Precisión = (VP) / (VP + FP)
* Exhaustividad = (VP) / (VP + FN)

# Marco Conceptual

## Biología Molecular

La biología molecular estudia el funcionamiento de las moléculas que posibilitan los procesos biológicos. Entre las principales moléculas que estudia están el ADN (ácido desoxirribonucleico) y las proteínas (Setubal & Meidanis, 1997).

### ADN

El ADN está formado por un polímero de nucleótidos unidos en una doble hélice, y contiene la información genética del organismo. Un nucleótido está formado por una base nitrogenada, un grupo fosfato y un azúcar (ver Figura 5). Las bases nitrogenadas que puede contener un nucleótido en el ADN son la adenina (A), la guanina (G), la timina (T) y la citosina (C). El orden en el que se encuentran finalmente determina la secuencia genómica (Risueño Pérez, 2012).

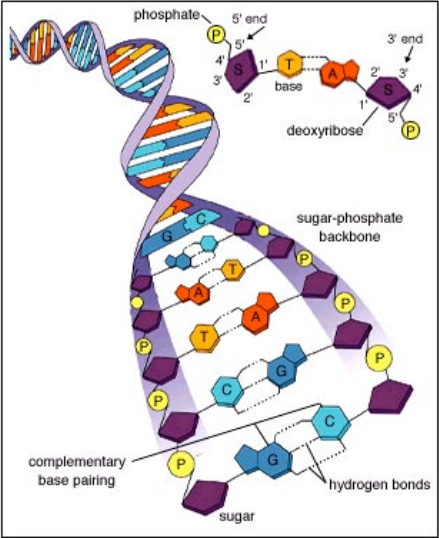


Figura 5 - Esquema del ADN. Adaptado de (Risueño Pérez, 2012).

La doble hélice está unida por puentes de hidrógeno formados por las bases nitrogenadas, y unen estos nucleótidos siempre en pares que se limitan a A-T (adenina con timina) y G-C (guanina con citosina).

El ADN no se usa de forma directa para la generación de proteínas, sino que su información se copia a una molécula de ARN intermediaria mediante el proceso de transcripción.

### ARN – proceso de transcripción

En el proceso de transcripción, secuencias de ADN son duplicadas en lo que se conoce como ARN (ácido ribonucleico). El ARN, al igual que el ADN, se compone de nucleótidos (ver Figura 6). Sin embargo, el ARN forma una sola hebra, y en el proceso de transcripción se reemplaza la timina, por el uracilo (U).

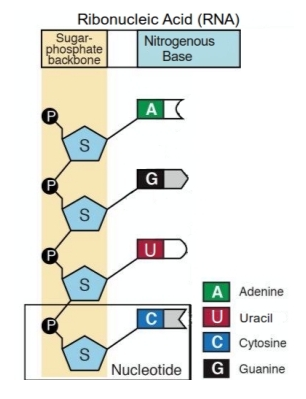


Figura 6 - Estructura del ARN. Adaptado de (Vieira, 2018).

En este punto, el ARN puede llegar o no a expresarse en proteínas. Los ARN que generarán proteínas mediante la traducción se conocen como ARNs mensajeros (*mRNA*), y los que no, se conocen como ARNs no codificantes.

### Proteínas – síntesis mediante la traducción de ARN

El ARN transcrito, viaja al ribosoma (sección de la célula dónde se lleva a cabo la síntesis de proteínas), ubicado en el citoplasma de la célula. El ribosoma se encarga entonces de la traducción del ARN a cadenas de aminoácidos, que finalmente son los que forman la proteína.

## ARN no codificante

### Historia

Como se mencionó, existe ARN transcrito que no se expresa en proteínas y se conoce como ARN no codificante. Al descubrirse inicialmente que gran parte del genoma no se expresaba en proteínas, y al no encontrarse ninguna función para estas secciones del ADN que no se codificaban (la biología molecular estaba en sus etapas iniciales), se asumió que estas secciones sin función existían debido a que en el pasado mutaron y hoy en día ya no cumplían ninguna función. También se teorizó que eran secciones que podían evolucionar y volverse codificantes, generando mutaciones en el organismo. En todo caso, esto originó que se acuñe el término “ADN basura” para estas secciones del ADN (Kung et al., 2013).

Entre 1990 y 2000, con la llegada de los proyectos de secuenciación del genoma, se haría evidente que una gran cantidad de este ADN basura era en realidad transcrito en algún punto (Kung et al., 2013), lo cual generó la pregunta de que si estas transcripciones, que no finalizaban en la síntesis de proteínas, servían algún propósito.

Así fue como el estudio del ARN no codificante se volvió relevante en la biología molecular. Sobre todo, a la luz de que se identificó que mientras un organismo es más complejo, posee mayor cantidad de ARN no codificante (Sebastián, 2017).

### Clasificación

Se presenta una tabla resumen con la clasificación del ARN:

Tabla 3 - Clasificación del ARN. Adaptado de (Genomics, 2015).

|  |  |
| --- | --- |
| Clasificación | Función |
| ARN mensajero (*mRNA*) | Lleva información genética transcrita del ADN. |
| ARN de la transferencia (*tRNA*) | Participa en el proceso de traducción del ARN |
| ARN ribosomal (*rRNA*) | Se encuentran en el ribosoma y forman el 80% del ARN de la célula. Participan en la síntesis de la proteína. |
| Micro ARN (*miRNA*) | Tiene funciones regulatorias. |
| ARN pequeño (*small* *RNA*) | Participa en el proceso de traducción del ARN. |
| ARN de telomerasa (*telomeraseRNA*) | Se relaciona con el proceso de replicación de cromosomas. |
| ARN anti-sentido (*antisense* *RNA*) | Directamente involucrado en la regulación de la expresión genética. |
| ARN largo no codificante (*lncRNA*) | ARN no codificante, que participa en la regulación de la célula. Suelen tener una longitud no menor a 200 nucleótidos y son difícil de distinguir de los codificantes. |

### Proyectos de secuenciación de transcriptomas

Los métodos más usados actualmente son los microarreglos y el secuencionamiento de ARN (*RNA-Seq*). En especial, el secuencionamiento de ARN permite obtener millones de secuencias con gran precisión estadística, y se pueden utilizar en nuevas especies para las cuáles no existe información previa sobre su secuenciación.

En los últimos años han habido muchas iniciativas de proyectos de secuenciación de transcriptomas, que han generado una enorme cantidad de información de ARN (Schneider et al., 2017). Estos proyectos son de gran importancia para el estudio del ARN, por ejemplo, son útiles para:

* El diagnóstico de enfermedades (Tolosa, 2017).
* La comprensión de cómo cada sección del ADN afecta el funcionamiento de distintos tipos de células para un mismo organismo.

## Aprendizaje de máquina

### Definición

El campo del aprendizaje de máquina se encarga del desarrollo de técnicas computacionales que aprenden a realizar tareas a partir del entrenamiento de conjuntos de datos preexistentes (Jabeen et al., 2018).

### Categorización

Se pueden dividir los métodos de aprendizaje de máquina en dos grandes grupos: el aprendizaje tradicional y el aprendizaje profundo.

El aprendizaje tradicional implica la extracción de atributos de las entidades que se quieren modelar, el entrenamiento del modelo en base a los atributos extraídos y la generación de resultados. La generación de atributos no es un proceso automático, y suele requerir el juicio de un experto.

El aprendizaje profundo integra todos estos pasos en el proceso de aprendizaje del modelo, aprovechando la gran cantidad de información disponible para inferir atributos y representarlos en un bajo nivel, y generar el modelo de predicción en base a ellas (ver Figura 7).



Figura 7 - Diferencia entre aprendizaje tradicional y aprendizaje profundo. Adaptado de (Rouse, 2017).

### Clasificadores con aprendizaje de máquina

Los métodos de clasificación de aprendizaje de máquina se favorecen de grandes cantidades de información, y dado que los proyectos de secuenciación de transcriptomas están generando una gran cantidad de información de transcriptomas, se prestan para su análisis (Min, Lee, & Yoon, 2017).

En particular, la generación de investigaciones relacionadas a algoritmos de aprendizaje profundo ha ido en aumento en el área de la bioinformática, debido a que apoyándose en la existencia de grandes cantidades de información, puede simplificar procesos al extraer por su propia cuenta atributos relevantes de las clases que estudia (Min et al., 2017).

Para visualizar este incremento en las investigaciones, se realizó un gráfico (ver Figura 8) de las investigaciones encontradas en la base de datos Scopus con dos cadenas de búsqueda:

* “Deep learning” (aprendizaje profundo).
* “Deep learning” AND “bio” (aprendizaje profundo en bioinformática).

Figura 8 - Número de investigaciones relacionadas a aprendizaje profundo en Bioinformática en Scopus. Elaboración propia.

La distinción de moléculas no codificantes, entonces, sería una tarea ideal de aprendizaje de máquina profundo, desde el punto de vista computacional (Schneider et al., 2017).

# Estado del Arte

En este capítulo se presentan estudios sobre la clasificación de ARNs no codificantes largos a través de métodos de aprendizaje de máquina en artículos científicos, tesis u otros trabajos de investigación relacionados. Se realiza esta revisión sistemática del estado del arte mediante los lineamientos definidos en el reporte “*Guidelines for performing systematic literature reviews in software engineering*” (Kitchenham & Charters, 2007).

## Preguntas de investigación

Primero se definirán con ayuda del método PICOC (*population, intervention, comparison, outcomes, context*) los conceptos generales. Estos servirán de apoyo para formular las peguntas de investigación.

Tabla 4 - Conceptos generales definidos con el método PICOC.

|  |  |
| --- | --- |
| Criterio | Descripción |
| Población | ARNs no codificantes largos. |
| Intervención | Clasificación a través de métodos de aprendizaje de máquina / aprendizaje profundo. |
| Comparación | *(\*) No aplica. Más que hacer una comparación de los métodos actuales, se desea saber cuáles son y qué necesidades hay en el campo de estudio.* |
| Resultados | Aplicables a múltiples especies. |
| Contexto | Contexto académico, bioinformática. |

En base a lo anterior, se plantearon las siguientes preguntas de investigación:

* ¿Cuáles son los métodos de aprendizaje de máquina más representativos que se han usado para la clasificación de *lncRNAs*?
* ¿Cuáles son las principales limitaciones de los métodos utilizados actualmente para la clasificación de *lncRNAs*?

## Estrategia de búsqueda

### Términos de búsqueda

A partir de los conceptos extraídos con el método PICOC, de las preguntas de investigación planteadas y tomando en cuenta que la literatura buscada se encuentra principalmente en inglés, se plantearon los siguientes componentes para la cadena de búsqueda (Tabla 5):

Tabla 5 - Componentes de la cadena de búsqueda.

|  |  |
| --- | --- |
| # | Componentes de la cadena de búsqueda |
| C1 | “lncRNA” OR “long non-coding RNA” OR “long noncoding RNA” OR “long non coding RNA” |
| C2 | “classification” OR “prediction” |
| C3 | “machine learning” OR “deep learning” |

Se usará la cadena de búsqueda:

C1 AND C2 AND C3

### Proceso de búsqueda

Se realizará la búsqueda en las siguientes bases de datos:

* Web of Science
* Scopus
* Springer Link
* BMC Bioinformatics

## Criterios de inclusión y exclusión

Los criterios de inclusión utilizados fueron los siguientes:

* Los artículos que en sí mismos sean una revisión del estado del arte son automáticamente incluidos.
* Los artículos que sean sobre la implementación de algún algoritmo de aprendizaje de máquina específico son revisados, y su inclusión se basa en:
  + Fecha de publicación, favoreciendo los más recientes.
  + Número de veces citados por otros artículos. Tomando en cuenta que el peso de este criterio es menor para los artículos más recientes, pero mayor para los más antiguos.
  + Se favorecen aquellos que tengan una implementación pública, con la intención de que más adelante se puedan realizar comparaciones. Pero si no la tienen no implica que serán excluidos.
* El artículo está en inglés o español (sin embargo, esto no modifica la decisión de que la búsqueda se realizará en inglés).

Los criterios de exclusión que se utilizaron fueron los siguientes:

* Artículos previos al 2008.
* Idioma diferente al inglés o español.
* Método de clasificación diferente a los de aprendizaje de máquina.

## Extracción de la información

Se presenta a manera de resumen, la cantidad de artículos encontrados en cada una de las bases de datos buscadas:

Tabla 6 - Artículos encontrados por base de datos.

|  |  |  |  |
| --- | --- | --- | --- |
| Base de datos | Artículos Encontrados | Artículos Duplicados | Artículos Seleccionados |
| Web of Science | 20 | 0 | 8 |
| Scopus | 40 | 14 | 2 |
| Springer Link | 154 | 14 | 4 |
| BMC Bioinformatics | 56 | 56 | 0 |
| Total | 270 | 84 | 14 |

Podemos dividir los artículos encontrados en dos clases:

* Artículos con implementaciones específicas de clasificadores de lncRNAs. De esta clase de artículos se encontraron 9.
* Artículos que discuten el estado del arte. De esta clase de artículos se encontraron 5.

De los artículos que discuten el estado del arte se extraerán otras implementaciones adicionales para expandir la lista de artículos con implementaciones específicas.

El objetivo final será, además de revisar la literatura para obtener el estado del arte con respecto a la clasificación de *lncRNAs* mediante métodos de aprendizaje de máquina, el generar una lista con los métodos más recientes que dispongan de implementaciones públicas que nos permita realizar comparaciones con lo desarrollado en la presente tesis para evaluar la utilidad del modelo predictivo que se quiere realizar.

Tomando en cuenta lo anterior, se presentan los artículos que discuten el estado del arte:

Tabla 7 - Artículos que discuten el estado del arte.

|  |  |  |
| --- | --- | --- |
| Año | Autor | Título |
| Yao, Yuhua Li, Xianhong Geng, Lili Nan, Xuying Qi, Zhaohui Liao, Bo | 2018 | Recent Progress in Long Noncoding RNAs Prediction (Yao et al., 2018) |
| Han, Siyu Liang, Yanchun Li, Ying Du, Wei | 2016 | Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination (Han, Liang, Li, & Du, 2016) |
| Jia-Pei, Yuan Hao-Wen, Zhang Zhi, Lu | 2013 | Progress on Bioinformatic Research of lncRNA (Jia-Pei, Hao-Wen, & Zhi, 2013) |
| Jabeen, Almas Ahmad, Nadeem Raza, Khalid | 2018 | Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data (Jabeen et al., 2018) |
| Yip, Kevin Y. Cheng, Chao Gerstein, Mark | 2013 | Machine learning and genome annotation: a match meant to be? (Yip, Cheng, & Gerstein, 2013) |

## Otras Tesis encontradas

Se realizó la búsqueda de tesis de bioinformática similares en el repositorio de tesis PUCP, pero no se encontraron resultados.

Por otro lado, se ha tenido acceso a una tesis de Brasil, “Métodos baseados em aprendizagem de máquina para distinguir RNAs longos não-codificadores intergênicos de transcritos codificadores de proteínas” (Vieira, 2018), donde se desarrolla un predictor nombrado PlantSniffer.

## Resultados

### Métodos más usados

A continuación, se discuten los métodos más usados para la clasificación de *lncRNAs* a través de predictores de aprendizaje de máquina. Para ello en la Tabla 8 se presentan los métodos agrupados según el tipo de clasificador que utilizan:

Tabla 8 - Métodos de aprendizaje de máquina más usados para la clasificación de lncRNA.

|  |  |  |
| --- | --- | --- |
| Método | Año | Clasificadores |
| Máquina de vectores de soporte (*support vector machine - SVM*) | 2007 | CPC (Kong et al., 2007) |
| **Ensamble de bosques aleatorios (*random forest ensemble*)** | **2015** | **LncRNA-ID (Achawanantakun, Chen, Sun, & Zhang, 2015)** |
| *Deep* *stacking* *networks* (*DSN*) | 2015 | lncRNA-MFDL (Fan & Zhang, 2015) |
| Redes neuronales profundas (*deep* *neural* *networks* - *DNN*) | 2016 | DeepLNC (Tripathi, Patel, Kumari, Chakraborty, & Varadwaj, 2016) |
| Máquina de vectores de soporte | 2017 | Clasificador de máquina de vectores de soporte de *lncRNA* (Schneider et al., 2017) |
| **Ensamble de bosques aleatorios** | **2017** | **PLncPRO (Singh, Khemka, Rajkumar, Garg, & Jain, 2017)** |
| **Ensamble de bosques aleatorios** | **2017** | **FEELnc** (Wucher et al., 2017) |
| **Ensamble de bosques aleatorios** | **2018** | **Predictor de *lncRNA* para plantas (Simopoulos et al., 2018)** |
| Máquina de vectores de soporte | 2018 | PlantSniffer (Vieira, 2018) |

En la Tabla 8 se resaltaron los métodos más usados (método por ensamble). Notar además que también son los más recientes. Los métodos por ensamble funcionan combinando múltiples clasificadores en un solo modelo, lo cual suele dar mayor generalización al clasificador que se obtiene como resultado y mejores tasas de predicción (Simopoulos et al., 2018).

Se encontraron también clasificadores que utilizan el modelo de máquina de vectores de soporte (*SVM*). Son métodos de aprendizaje supervisado, que requieren de la extracción de atributos de las entidades a clasificar por parte de expertos (Schneider et al., 2017). Esto difiere del método de aprendizaje profundo propuesto en la presente investigación, que incorpora de manera automática esta etapa dentro del proceso de aprendizaje.

Dentro de los métodos de aprendizaje profundo, encontramos el clasificador lncRNA-MFDL (Fan & Zhang, 2015) , que hace uso de *deep* *stacking* *networks* (*DSN*). Un segundo método de aprendizaje profundo encontrado es el clasificador DeepLNC, que hace uso de redes neuronales profundas (*DNN*), ideales para el reconocimiento de patrones (Tripathi et al., 2016).

### Principales limitaciones

Dentro de los modelos de predicción encontrados, analizaremos caso por caso sus limitaciones, prestando atención a la capacidad de los predictores de realizar la clasificación en especies desconocidas al modelo, es decir, especie bajo las cuales no han sido entrenados.

CPC (Kong et al., 2007) es uno de los primeros clasificadores popularizados. Utiliza un modelo máquina de vectores de soporte (*SVM*) para determinar la capacidad de un transcriptoma de ser codificado. Sin embargo, no es bueno para clasificar *lncRNAs*, ya que el tener baja capacidad para codificarse no es sinónimo de ser no codificante.

LncRNA-ID (Achawanantakun et al., 2015): A través de un ensamble de bosques aleatorios se buscó construir un modelo, que se entrenó en humanos y ratones por separado (dos modelos diferentes), que pueda manejar el problema de desbalance de información (la cantidad del conjunto de datos de una clase es mucho mayor a la de otra). Es posible entrenar este modelo para otras especies, pero se desconoce su utilidad para clasificar especies desconocidas al modelo.

lncRNA-MFDL (Fan & Zhang, 2015): Este predictor utiliza *deep stacking networks* (*DSN*) con un modelo de aprendizaje profundo. Posee un link público con su implementación, sin embargo, no está disponible. El modelo se usó para predicciones en humanos y no ha sido probado en otras especies.

DeepLNC (Tripathi et al., 2016): Este clasificador utiliza redes neuronales profundas (*DNN*), y su objetivo principal es el poder armar un clasificador para una especie sin tener que hacer una selección de atributos. Esto permite entrenarlo fácilmente para otras especies, pero no se probó su utilidad al clasificar especies desconocidas al modelo. Existe un link público para acceder al clasificador, pero no está disponible.

Clasificador de máquina de vectores de soporte de *lncRNA* (Schneider et al., 2017): A través de un clasificador de máquina de vectores de soporte (*SVM*) se buscó construir un modelo que pueda generalizar sus predicciones a otras especies con buenos resultados. El modelo se entrenó con información de humanos, ratones y danios rerios (pez cebra); obteniendo buenos resultados con especies cercanas.

PLncPRO (Singh et al., 2017): A través de un ensamble de bosques aleatorios se entrenó el modelo específicamente para predecir *lncRNAs* en plantas. También se realizaron pruebas en humanos (mediante una nueva extracción de atributos). No se ha comprobado su utilidad para especies desconocidas al modelo.

FEELnc (Wucher et al., 2017): A través de un ensamble de bosques aleatorios, este clasificador propone un método alternativo para el entrenamiento del modelo cuando se poseen *PCTs*, pero no se dispone de *lncRNAs* para la especie deseada: generar artificialmente *lncRNAs* para el entrenamiento del modelo a partir de las secuenciaciones iniciales.

Predictor de *lncRNA* para plantas (Simopoulos et al., 2018): A través de un ensamble de bosques aleatorios y potenciación del gradiente (*gradient boosting*), entrenado en base a información empíricamente confirmada de animales, plantas y virus, se buscó construir un clasificador para plantas. Este predictor muestra si un transcriptor es codificante o no, con un nivel de confianza. De esta forma se puede priorizar qué transcriptores pasarán por futuras validaciones empíricas, y retroalimentar el modelo con esta nueva información para que mejore continuamente.

PlantSniffer (Vieira, 2018): A través de un ensamble de máquina de vectores de soporte, cuyos resultados se cruzan contra BLAST, se busca predecir *lincRNAs* en plantas para las que no existen bases de datos de genomas y/o transcriptomas (caña de azúcar y maíz), utilizando en su lugar información de especies cercanas. Siendo este modelo predictivo que tiene mayor similitud con el nuestro, y para el cual se posee acceso público al código fuente, será utilizado para compararlo contra el modelo propuesto en el presente proyecto.

## Conclusiones

En los últimos años los proyectos de secuenciación de transcriptomas están generando una enorme cantidad de información que necesita ser explotada. La anotación y clasificación manual de los transcriptomas de estos proyectos es lenta y costosa. En este contexto los métodos computacionales de inteligencia artificial y aprendizaje de máquina son ventajosos, ya que se apoyan en la gran cantidad de información producida y generan resultados rápidos y sin mucho costo (Jabeen et al., 2018).

Dentro de los métodos de clasificación existentes, nos interesa saber su utilidad para la clasificación de especies desconocidas al modelo, y si se tiene acceso al clasificador con la intención de poder realizar comparaciones.

Los métodos más usados en los últimos años son los ensambles, debido a que combinan varios clasificadores y permiten evitar problemas de sobreajuste (ser muy específicos, del inglés *overfitting*). Aunque aún no son tan usados, los métodos de aprendizaje profundo también están ganando popularidad, ya que la extracción de atributos (en la que normalmente participa un experto) es un paso que realiza el algoritmo durante su entrenamiento, aprendiendo los atributos por su cuenta a partir de la información de la que se alimenta.

Sólo se encontraron dos modelos orientados a generalizar sus resultados a otras especies – el clasificador de máquina de vectores de soporte de *lncRNA* (Schneider et al., 2017) y PlantSniffer (Vieira, 2018). Al poseer acceso a las fuentes de PlantSniffer, será utilizado para realizar comparaciones de las precisiones contra el modelo propuesto en el presente proyecto de fin de carrera.

Los clasificadores suelen construirse con el objetivo de tener las mejores predicciones para las especies bajo las cuales son entrenadas. La carencia de estudios de modelos que intenten clasificar especies en base a información de otras especies, motiva la presente investigación de un modelo predictivo que pueda usarse en especies diferentes a la que se desea clasificar, pudiendo estas ser nuevas especies.

# Recopilación y preprocesamiento de datos

## Fuentes de información

La data está compuesta de transcriptomas identificados como ARN codificantes y no codificantes largos. Para la extracción de ARN codificante (PCT) se utilizó como fuente de información Ensembl (Zerbino et al., 2018), mientras que para los ARN no codificantes largos (lncRNA) se utilizaron las bases de datos especializadas en plantas CantataDB (Szcześniak et al., 2016) y GreeNC (Paytuví Gallart et al., 2016).

Se construirá un modelo de clasificación binaria en base a la información de varias especies ya conocidas, con el objetivo de realizar predicciones sobre especies no modeladas. Identificaremos como la clase positiva los ARN no codificantes largos, y como clase negativa los ARN codificantes.

Cada fuente de información posee un conjunto de datos para diferentes especies (ver Tabla 9). En este contexto, se desea que la data sea balanceada, por lo que el número de transcriptomas de cada especie a utilizar para la construcción del modelo deberá ser el mismo. Inclusive, dentro de cada especie, el número de casos positivos y negativos tendrá que ser igual.

Tabla - Número de especies en las bases de datos de transcriptomas

|  |  |  |
| --- | --- | --- |
| Fuente | Tipo | Conteo de especie |
| Ensembl v43 | PCT (clase negativa) | 59 especies |
| CantataDB v2.0 | lncRNA (clase positiva) | 43 especies |
| GreeNC v1.12 | lncRNA (clase positiva) | 45 especies |

## Selección de especies

Con el objetivo de construir un modelo que generalice adecuadamente a especies no modeladas, interesa cumplir con cuatro restricciones y/u objetivos:

* Elegir el mismo número de transcriptomas de cada especie.
* Dentro de cada especie, utilizar el mismo número de casos positivos y negativos.
* Hacer uso de la mayor cantidad de especies.
* Hacer uso del mayor número de transcriptomas.

Es común que el número de ARN codificantes disponibles siempre sea mucho mayor al número de ARN no codificantes largos disponibles, por lo que centraremos nuestra atención en la contabilización de la clase positiva, lncRNA.

Se debe tomar en cuenta que, para una misma especie, puede existir información de ARN no codificante tanto en CantataDB como en GreeNC. En estas casuísticas, se ha optado por utilizar solamente los transcriptomas de la fuente que posea el mayor número de transcriptomas, descartando los transcriptomas de la otra fuente.

En base a lo anterior, en la Tabla 10 podemos visualizar las 10 especies con la mayor cantidad de ARN no codificantes disponibles, junto con la fuente de dónde provienen estos transcriptomas.

Tabla - Especies con el mayor conteo de clases positivas, y sus fuentes

|  |  |  |  |
| --- | --- | --- | --- |
| Especie | Conteo PCT (clase negativa) | Conteo lncRNA (clase positiva) | Fuente de los lncRNA |
| Triticum aestivum | 107,545 | 38,820 | GreeNC |
| Brassica napus | 101,040 | 12,010 | CantataDB |
| Oryza rufipogon | 37,071 | 10,261 | CantataDB |
| Trifolium pratense | 39,917 | 10,179 | CantataDB |
| Physcomitrella patens | 32,447 | 9,690 | GreeNC |
| Medicago truncatula | 50,444 | 9,676 | GreeNC |
| Manihot esculenta | 33,044 | 9,504 | CantataDB |
| Oryza nivara | 36,313 | 8,955 | CantataDB |
| Brassica rapa | 41,018 | 8,501 | CantataDB |
| Hordeum vulgare | 37,705 | 7,970 | CantataDB |

En base a la tabla anterior, si elegimos trabajar con 3 especies, seleccionaríamos las primeras tres especies de la tabla, hasta Oryza rufipogon. El límite en la cantidad de clases negativas y positivas a utilizar por especie sería establecido por la cantidad de lncRNA disponibles en esta última especie (10,261). Entonces el número total de transcriptomas a utilizar sería 10,261 x 2 (clase positiva y negativa) x 3 (número de especies), en total 61,566 transcriptomas.

En cambio, si optamos por utilizar las 10 especies listadas, esta vez el límite viene dado por la última especie, Hordeum vulgare, y el total de transcriptomas a utilizar en el modelo sería 7,970 (transcriptomas en Hordeum vulgare) x 2 (clase positiva y negativa) x 10 (número de especies), en total 159,400 transcriptomas.

De esta forma podemos construir la fórmula del total de transcriptomas que utilizaríamos en el modelo al seleccionar cada una de las especies, en base a su posición o ranking, multiplicado por el número de transcriptomas ARN no codificantes que poseen:

Transcriptomas lncRNA de la especie x 2 x Ranking de la especie

Si realizamos esta operación para todas las especies disponibles, y las ordenamos de mayor a menor de acuerdo al resultado de la fórmula anterior, obtenemos lo mostrado en la Tabla 11 (se muestran los cinco primeros resultados).

Tabla - Número total de transcriptomas a utilizar según el modelo seleccionado

|  |  |  |  |
| --- | --- | --- | --- |
| Especie | Ranking | Conteo lncRNA | Número de transcriptomas total para el modelo |
| Oryza sativa Japonica Group | 25 | 5,237 | 261,850 |
| Setaria italica | 30 | 4,208 | 252,480 |
| Musa acuminata | 31 | 4,071 | 252,402 |
| Theobroma cacao | 24 | 5,256 | 252,288 |
| Vitis vinifera | 27 | 4,542 | 245,268 |

El número de transcriptomas a utilizar, viene maximizado al seleccionar las primeras 25 especies hasta Oryza sativa Japonica Group, especie para la cual se poseen 5,237 lncRNA, con lo cual nuestro modelo utilizaría un total de 261,850 transcriptomas.

Sin embargo, si seleccionamos el siguiente candidato, Setaria itálica, tendríamos como total de transcriptomas disponibles para la construcción del modelo 252,480. Serían 9,370 transcriptomas menos, pero con la ventaja de aumentar la variedad de especies utilizadas de 25 a 30. Además de utilizar el mayor número de transcriptomas disponibles, otro de los objetivos en la selección de especies era maximizar el número de especies utilizadas.

Debido a lo anterior, finalmente, se elige utilizar 30 especies para el modelo, teniendo cada especie un total de 4,208 clases positivas y 4,208 clases negativas. Se presenta el resumen de las especies seleccionadas y la fuente de donde se obtendrán los ARN no codificantes para las mismas:

* Amborella trichopoda (GreeNC)
* Arabidopsis lyrata (CantataDB)
* Arabidopsis thaliana (CantataDB)
* Brachypodium distachyon (GreeNC)
* Brassica napus (CantataDB)
* Brassica oleracea (CantataDB)
* Brassica rapa (CantataDB)
* Corchorus capsularis (CantataDB)
* Cucumis sativus (CantataDB)
* Glycine max (GreeNC)
* Gossypium raimondii (GreeNC)
* Hordeum vulgare (CantataDB)
* Leersia perrieri (CantataDB)
* Manihot esculenta (CantataDB)
* Medicago truncatula (GreeNC)
* Oryza barthii (CantataDB)
* Oryza brachyantha (CantataDB)
* Oryza nivara (CantataDB)
* Oryza rufipogon (CantataDB)
* Oryza sativa Japonica Group (GreeNC)
* Physcomitrella patens (GreeNC)
* Populus trichocarpa (GreeNC)
* Setaria italica (CantataDB)
* Solanum lycopersicum (CantataDB)
* Solanum tuberosum (GreeNC)
* Sorghum bicolor (GreeNC)
* Theobroma cacao (CantataDB)
* Trifolium pratense (CantataDB)
* Triticum aestivum (GreeNC)
* Vitis vinífera (CantataDB)

## Método de obtención de las fuentes

### Ensembl

Para el acceso a Ensembl Plants ([http://plants.ensembl.org](http://plants.ensembl.org/)) se utilizó el paquete en R biomaRt (Durinck et al., 2005) para migrar los transcriptomas de las especies a una base de datos. Los parámetros utilizados para la extracción de las clases negativas fueron:

* Como atributos a extraer se especificó ensembl\_transcript\_id y transcript\_exon\_intron. Esto quiere decir que se indicó como campos devueltos en la llamada a biomaRt el identificador del transcriptoma, y la secuencia de ARN completa.
* Como filtros utilizados, se usó transcript\_biotype, con el valor protein\_coding. Esto quiere decir que se filtraron los transcriptomas que codifican en ARN mensajero.
* Se extrajo la información de las bases de datos de cada una de las especies seleccionadas.

Se presenta en el Anexo B el script utilizado para la descarga de la información de Ensembl, a través del paquete biomaRt.

Sin embargo, debido a la cantidad de información contenida en la consulta, la extracción de la información no se pudo realizar automáticamente, por lo cual se tuvo que descargar manualmente la información de la web <https://plants.ensembl.org/biomart/martview/>, en archivos fasta, y de allí procesarlas a través de un script, disponible en el Anexo C.

Se limitó la longitud de las secuencias importadas a un máximo de 65,000 nucleótidos. Aquellas secuencias con una longitud mayor fueron ignoradas.

### CantataDB

Los transcriptomas de CantataDB fueron descargados en formato fasta de la web <http://cantata.amu.edu.pl/> e importados a la base de datos, ignorando secuencias con una longitud mayor a 65,000 nucleótidos. En el Anexo D se encuentra el script utilizado para procesar estos transcriptomas descargados.

### GreeNC

Los transcriptomas de GreeNC fueron descargados en formato fasta de la web [http://greenc.sciencedesigners.com](http://greenc.sciencedesigners.com/) e importados a la base de datos, ignorando secuencias con una longitud mayor a 65,000 nucleótidos. En el Anexo E se encuentra el script utilizado para procesar estos transcriptomas descargados.

## Selección de secuencias a utilizar en el modelo

Se realizó la descarga de información de todas las especies y transcriptomas disponibles en cada una de las fuentes listadas. El proceso que se realizó después de esto fue la selección de los transcriptomas que se utilizarán para la construcción del modelo.

El criterio de selección de los transcriptomas fue completamente aleatorio. Se anexa el script utilizado para este proceso (ver Anexo F).

## Selección de características a utilizar para el modelo

Para la representación de los transcriptomas a través de características (en inglés *features*), se hizo uso de una de las fuentes referenciadas durante la revisión del estado del arte, el predictor de lncRNA para plantas (Simopoulos et al., 2018), que tiene código fuente disponible en <https://github.com/gbgolding/crema>.

En el trabajo referenciado, las características utilizadas fueron:

1. Longitud del transcriptoma
2. Longitud del marco abierto de lectura (del inglés *open reading frame*, u *ORF*) y porcentaje del marco abierto de lectura con respecto a la longitud del transcriptoma
3. Porcentaje de nucleótidos GC presentes
4. Puntaje Fickett
5. Puntaje de hexámeros
6. Puntaje de identidad contra la base de datos SwissProt
7. Longitud del alineamiento contra la base de datos SwissProt
8. Proporción de la longitud del alineamiento con respecto a la longitud del transcriptoma
9. Proporción de la longitud del alineamiento con respecto a la longitud del marco abierto de lectura
10. Presencia de elementos transponibles (descartado)
11. Divergencia porcentual de la secuencia con respecto a los elementos transponibles (descartado)

El cálculo de características como la longitud del transcriptoma y el porcentaje de nucleótidos GC se puede realizar a través de scripts relativamente simples.

Por otro lado, para el cálculo del marco abierto de lectura, su porcentaje con respecto a la longitud del transcriptoma, el puntaje Fickett y el puntaje de hexámeros, se utilizó el programa CPAT (Wang et al., 2013).

Características relacionadas al alineamiento contra la base de datos SwissProt se realizaron con la aplicación del algoritmo Diamond (Buchfink, Xie, & Huson, 2015).

Las características relacionadas a elementos transponibles fueron extraídas con RepeatMasker (<http://www.repeatmasker.org/>), pero en el trabajo referenciado fueron descartadas debido a que de hecho afectaban negativamente el desempeño de los clasificadores (Simopoulos et al., 2018).

### Características calculadas por scripts propios

Se calcularon dos características a través de scripts propios: la longitud del transcriptoma y el porcentaje de nucleótidos GC.

Se anexa el script utilizado para el cálculo de la longitud del transcriptoma (ver Anexo G). A pesar de no ser una característica especialmente discriminante, es frecuentemente utilizado en la literatura.

El porcentaje de nucleótidos GC (se adjunta script en el Anexo H) es otra característica comúnmente utilizada en la literatura (Simopoulos et al., 2018).

### Características calculadas a través de CPAT

#### Definición de CPAT

CPAT es otra herramienta orientada a predecir el potencial codificante de transcriptomas (Wang et al., 2013). Al utilizar como entrada un archivo fasta, CPAT genera como salida sus predicciones con respecto al potencial codificante de cada transcriptoma que contiene, pero junto a esta información entrega características calculadas de estos transcriptomas, que son de interés para el modelo del presente trabajo de investigación: longitud del marco abierto de lectura, porcentaje de la longitud del marco abierto de lectura con respecto a la longitud total del transcriptoma, puntaje Fickett y el puntaje de hexámeros.

Entones, por cada modelo de que generemos, será necesario generar también un modelo CPAT específico al nuestro modelo, que nos permita construir las características de los transcriptomas a entrenar, validar y predecir. Será importante tomar nota de esto para la arquitectura del modelo.

#### Características extraídas

El marco abierto de lectura son secciones del transcriptoma con posibilidad de ser codificados, que suelen iniciar con codón de inicio (AUG) y se encuentran limitadas por un codón de parada. El marco abierto de lectura es una característica comúnmente utilizada en modelos predictivos de ARN no codificantes y ARN codificantes (Wang et al., 2013). Al poder poseer un transcriptoma múltiples marcos abiertos de lectura, CPAT utilizará la longitud del más largo.

El puntaje Fickett (en inglés *Fickett Testcode score*) se basa en la composición de cada nucleótido dentro de ARN codificante y no codificante, combinado con la preferencia del uso de cada nucleótido en una de las tres posiciones que puede ocupar en un codón (Fickett, 1982).

El puntaje de hexámeros (en inglés *hexamer usage bias*) es calculado en base a la probabilidad de la aparición de cada secuencia de seis nucleótidos en un ARN codificante/no codificante (Wang et al., 2013).

#### Requerimientos para CPAT

Instrucciones para la instalación y uso de CPAT se encuentran disponibles en <http://rna-cpat.sourceforge.net/>. Para hacer uso se CPAT se requieren las siguientes entradas:

* Archivo fasta con los transcriptomas a predecir.
* Tabla de frecuencias de hexámeros
* Modelo de regresión logística

La tabla de frecuencia de hexámeros y el modelo de regresión logística deben ser generados por modelo utilizado. CPAT ya tiene estas entradas preconstruidas para algunas especies mamíferas, pero para la presente investigación tendrán que ser generados.

#### Tabla de frecuencia de hexámeros

Para la generación de la tabla de frecuencias de hexámeros, se requiere como entrada las secuencias codificantes de los ARN codificantes (diferente de los transcriptomas de los ARN codificantes ya descargados de Ensembl, ver Figura 9), y los transcriptomas de ANR no codificantes. En el Anexo I se muestra el script utilizado para generar los archivos fasta de estas dos entradas.

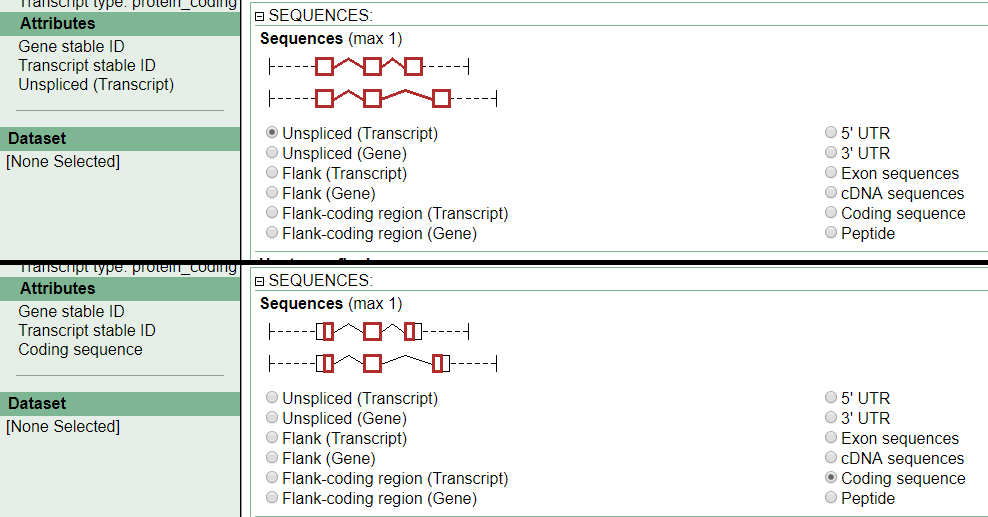


Figura - Diferencia visualizada en Biomart (Ensembl) entre transcriptoma y secuencia codificante

Con los archivos de entradas preparados, se debe ejecutar el script de CPAT “make\_hexamer\_tab.py” para generar la tabla de frecuencias de hexámeros. En el Anexo J se tiene el script utilizado para la generación de la tabla de frecuencia de hexámeros.

La salida de este script es un archivo con la tabla de hexámeros calculada, que es la todas las combinaciones posibles de secuencias de 6 nucleótidos, y la probabilidad de que formen parte de una secuencia codificante o un ARN no codificante. Se presenta un ejemplo en la Tabla 12. Todas las entradas en la columna de secuencias codificantes sumarán en total 1. De forma similar, todas las entradas en la columna de ARN no codificantes suman 1.

Tabla - Ejemplo de tabla de frecuencia de hexámeros para Amborella trichopoda generada por CPAT

|  |  |  |
| --- | --- | --- |
| Hexamer | coding | noncoding |
| AAAAAA | 0.0004820376938412906 | 0.00044355770099859063 |
| AAAAAC | 0.0003494773280349357 | 0.00030714278540845805 |
| AAAAAG | 0.0006740995874811799 | 0.0005264109319398369 |
| … | … | … |
| TTTTTC | 0.00030277992644406066 | 0.00027868814043873716 |
| TTTTTG | 0.000454923073562718 | 0.0003657258779931776 |
| TTTTTT | 0.0003095585815137038 | 0.00024270138356526658 |

#### Modelo de regresión logística

La generación del modelo de regresión logística requiere como entrada la tabla de hexámeros anteriormente calculada, los ARN codificantes (en esta ocasión el transcriptoma completo) y los ARN no codificantes.

El archivo fasta para ARN no codificantes ya se había generado con anterioridad, para la construcción de la tabla de hexámeros. En el Anexo K se puede visualizar el script utilizado para la generación del archivo fasta de ARN codificantes faltante.

Para generar el modelo de regresión logística, debe llamarse al script del CPAT “make\_logitModel.py”. Se anexa el script utilizado para la generación de la regresión logística en el Anexo L.

#### Ejecución de CPAT

Teniendo todos los archivos de entrada preparados (archivos fasta a predecir, archivo de tabla de frecuencias de hexámeros y modelo de regresión logística), procedemos con la ejecución de CPAT para obtener como salida los transcriptomas con las características de interés. Se anexa el script utilizado para ello en el Anexo M.

#### Importación de características a la base de datos

Se anexa el script utilizado para la importación de los archivos de salida de CPAT a la base de datos en el Anexo N.

### Características calculadas a través de Diamond

#### Definición de Diamond

Diamond (Buchfink et al., 2015) es un algoritmo que realiza búsquedas de alineamientos en base de datos de proteínas de manera rápida, en comparación con otras herramientas disponibles (Simopoulos et al., 2018).

#### Características Extraídas

Las características extraídas mediante el algoritmo Diamond serán aquellas relacionadas con alineamientos contra la base de datos de proteínas SwissProt, específicamente: el puntaje de identidad, la longitud del alineamiento, la proporción de la longitud del alineamiento con respecto a la longitud del transcriptoma, y la proporción de la longitud del alineamiento con respecto a la longitud del marco abierto de lectura antes calculado.

El puntaje de identidad se refiere al porcentaje del transcriptoma encontrado en la búsqueda en la base de datos SwissProt. Por otro lado, la longitud del alineamiento se refiere a la longitud de la coincidencia, y las otras dos medidas (proporción con respecto a la longitud del transcriptoma, y proporción con respecto a la longitud del marco abierto de lectura) se obtienen de características previamente extraídas.

#### Requerimientos para Diamond

Para ejecutar Diamond se requieren dos pasos. El primero será tener una base de datos de proteínas en un archivo fasta, que será utilizado como entrada para la generación de una base de datos en un formato propio utilizado por el algoritmo. Con esta base de datos construida, como segundo paso, podemos tomar como entrada los transcriptomas para los que se realizará la búsqueda de alineamientos, ejecutar Diamond, y tomar como salida los puntajes de alineamientos.

#### Construcción de la base de datos de proteínas de SwissProt

Para la construcción de la base de datos de proteínas, se descargó esta información desde UniProt (<https://www.uniprot.org/>). Dado que nos enfocamos en predicciones en plantas, se filtró la información para incluir solamente plantas (taxonomía Viridid Plantae), con data experimentalmente validada. Se puede visualizar el resultado de la búsqueda en [https://www.uniprot.org/uniprot/?query=taxonomy:viridiplantae&fil= reviewed%3Ayes&sort=score](https://www.uniprot.org/uniprot/?query=taxonomy:viridiplantae&fil=%20reviewed%3Ayes&sort=score).

La descarga del resultado de la búsqueda genera un archivo en formato fasta, que es utilizado para generar la base de datos de proteínas a través del comando de Diamond “makedb”. Se muestra el script utilizado para esta tarea en el Anexo O.

#### Ejecución de Diamond

Con el objetivo de obtener las características deseadas, se realizará la búsqueda de alineamientos de los ARN codificantes y ARN no codificantes, que ya se descargaron en archivos fasta en archivos previos, en la base de datos generada con el comando “makedb”. Se anexa el script utilizado para esta sección en el Anexo P.

Los parámetros utilizados en la ejecución de Diamond serán similares a los utilizados en el predictor lncRNA para plantas (Simopoulos et al., 2018) referenciado. El resultado del comando será un archivo tabulado todas las coincidencias encontradas, por lo que cada transcriptoma puede aparecer en ninguna, una o varias filas, según el número de coincidencias encontradas. Entonces, para pasar esta información a la base de datos de características, será necesario un script adicional.

#### Importación de características a la base de datos

Se adjunta en el Anexo Q el script utilizado para pasar el resultado de la búsqueda de alineamientos a través de Diamond a la base de datos. Para un transcriptoma, el script buscará el alineamiento con el mayor puntaje asignado por Diamond, y serán las características de identidad y longitud del alineamiento de este alineamiento los asignados al transcriptoma.

Referencias

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2016). *TensorFlow: A system for large-scale machine learning*. 21.

Achawanantakun, R., Chen, J., Sun, Y., & Zhang, Y. (2015). LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics*, *31*(24), 3897–3905. https://doi.org/10.1093/bioinformatics/btv480

Brownlee, J. (2018, mayo 22). A Gentle Introduction to k-fold Cross-Validation. Recuperado el 4 de noviembre de 2018, de Machine Learning Mastery website: https://machinelearningmastery.com/k-fold-cross-validation/

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Chollet, F. (2017). *Deep Learning with Python* (1st ed.). Greenwich, CT, USA: Manning Publications Co.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Fan, X.-N., & Zhang, S.-W. (2015). lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Molecular BioSystems*, *11*(3), 892–897. https://doi.org/10.1039/C4MB00650J

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, *10*(17), 5303–5318.

Genomics. (2015, septiembre 1). RNA Class:The Classification of RNA | CD Genomics Blog. Recuperado el 30 de septiembre de 2018, de https://www.cd-genomics.com/blog/index.php/rna-classthe-classification-of-rna/

Han, S., Liang, Y., Li, Y., & Du, W. (2016). Long Noncoding RNA Identification: Comparing Machine Learning Based Tools for Long Noncoding Transcripts Discrimination. *Biomed Research International*, 8496165. https://doi.org/10.1155/2016/8496165

Jabeen, A., Ahmad, N., & Raza, K. (2018). Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data. En *Lecture Notes in Computational Vision and Biomechanics*. *Classification in BioApps: Automation of Decision Making* (pp. 133–172). https://doi.org/10.1007/978-3-319-65981-7\_6

Jia-Pei, Y., Hao-Wen, Z., & Zhi, L. (2013). Progress on Bioinformatic Research of lncRNA. *Progress in Biochemistry and Biophysics*, *40*(7), 634–640. https://doi.org/10.3724/SP.J.1206.2013.00266

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, *35*(Web Server issue), W345–W349. https://doi.org/10.1093/nar/gkm391

Kung, J. T. Y., Colognori, D., & Lee, J. T. (2013). Long Noncoding RNAs: Past, Present, and Future. *Genetics*, *193*(3), 651–669. https://doi.org/10.1534/genetics.112.146704

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics*, *18*(5), 851–869. https://doi.org/10.1093/bib/bbw068

Mount, D. W. (2007). Using the Basic Local Alignment Search Tool (BLAST). *Cold Spring Harbor Protocols*, *2007*(7), pdb.top17. https://doi.org/10.1101/pdb.top17

NHGRI. (2015, octubre 21). Ácido desoxirribonucleico (ADN). Recuperado el 20 de octubre de 2018, de National Human Genome Research Institute (NHGRI) website: https://www.genome.gov/27562614/cido-desoxirribonucleico-adn/

Nielsen, M. A. (2018). *Neural Networks and Deep Learning*. Recuperado de http://neuralnetworksanddeeplearning.com

Paytuví Gallart, A., Hermoso Pulido, A., Anzar Martínez de Lagrán, I., Sanseverino, W., & Aiese Cigliano, R. (2016). GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Research*, *44*(D1), D1161–D1166. https://doi.org/10.1093/nar/gkv1215

Ponjavic, J., Ponting, C. P., & Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Research*, *17*(5), 556–565. https://doi.org/10.1101/gr.6036807

Quinn, J. J., & Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, *17*(1), 47–62. https://doi.org/10.1038/nrg.2015.10

Risueño Pérez, A. (2012). *Bioinformática aplicada a estudios del transcriptoma humano: análisis de expresión de genes, isoformas génicas y ncRNAs en muestras sanas y en cáncer*. Recuperado de https://gredos.usal.es/jspui/handle/10366/121408

Rouse, M. (2017, abril). ¿Qué es Aprendizaje profundo (deep learning)? - Definición en WhatIs.com. Recuperado el 30 de septiembre de 2018, de SearchDataCenter&nbsp;en&nbsp;Español website: https://searchdatacenter.techtarget.com/es/definicion/Aprendizaje-profundo-deep-learning

Schneider, H. W., Raiol, T., Brigido, M. M., Walter, M. E. M. T., & Stadler, P. F. (2017). A Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts. *Bmc Genomics*, *18*, 804. https://doi.org/10.1186/s12864-017-4178-4

Sebastián, I. (2017, diciembre 15). ADN codificante Archives. Recuperado el 5 de octubre de 2018, de Genética Médica Blog website: https://revistageneticamedica.com/blog/tag/adn-codificante/

Setubal, J. C., & Meidanis, J. (1997). *Introduction to computational molecular biology*. Boston: PWS Pub.

Shen, H. (2014). Interactive notebooks: Sharing the code. *Nature News*, *515*(7525), 151. https://doi.org/10.1038/515151a

Simopoulos, C. M. A., Weretilnyk, E. A., & Golding, G. B. (2018). Prediction of plant lncRNA by ensemble machine learning classifiers. *Bmc Genomics*, *19*, 316. https://doi.org/10.1186/s12864-018-4665-2

Singh, U., Khemka, N., Rajkumar, M. S., Garg, R., & Jain, M. (2017). PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Research*, *45*(22), e183–e183. https://doi.org/10.1093/nar/gkx866

Szcześniak, M. W., Rosikiewicz, W., & Makałowska, I. (2016). CANTATAdb: A Collection of Plant Long Non-Coding RNAs. *Plant and Cell Physiology*, *57*(1), e8–e8. https://doi.org/10.1093/pcp/pcv201

Tolosa, A. (2017, mayo 19). Utilidad diagnóstica de la secuenciación del transcriptoma. Recuperado el 30 de septiembre de 2018, de Genética Médica website: https://revistageneticamedica.com/2017/05/19/secuenciacion-transcriptoma/

Tripathi, R., Patel, S., Kumari, V., Chakraborty, P., & Varadwaj, P. K. (2016). DeepLNC, a long non-coding RNA prediction tool using deep neural network. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), 21. https://doi.org/10.1007/s13721-016-0129-2

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigo, R., & Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics*, *19*(9), 535–548. https://doi.org/10.1038/s41576-018-0017-y

VanderPlas, J. (2016). *Python Data Science Handbook*. Recuperado de http://shop.oreilly.com/product/0636920034919.do

Vieira, L. M. (2018). *Métodos baseados em aprendizagem de máquina para distinguir RNAs longos não-codificadores intergênicos de transcritos codificadores de proteínas*. Recuperado de http://repositorio.unb.br/handle/10482/32463

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, *41*(6), e74. https://doi.org/10.1093/nar/gkt006

Willems, K. (2015, mayo 12). Choosing R or Python for data analysis? An infographic. Recuperado el 3 de noviembre de 2018, de DataCamp Community website: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., … Derrien, T. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, *45*(8), e57. https://doi.org/10.1093/nar/gkw1306

Yao, Y., Li, X., Geng, L., Nan, X., Qi, Z., & Liao, B. (2018). Recent Progress in Long Noncoding RNAs Prediction. *Current Bioinformatics*, *13*(4), 344–351. https://doi.org/10.2174/1574893612666170905153933

Yip, K. Y., Cheng, C., & Gerstein, M. (2013). Machine learning and genome annotation: a match meant to be? *Genome Biology*, *14*(5), 205. https://doi.org/10.1186/gb-2013-14-5-205

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

# Anexos

## : Plan de Proyecto

* **Justificación**

El presente proyecto proveerá un modelo predictivo para clasificar *lncRNAs* y *PCTs*, diferenciándose de otros clasificadores existentes, en que para realizar predicciones sobre una especie específica puede ser entrenado en base a especies diferentes, y producir resultados aceptables en el proceso.

También, dentro del contexto de los proyectos de secuenciación de transcriptomas, beneficiará a otros proyectos de anotación experimental sobre nuevas especies, que no poseen anotaciones previas. Los proyectos de secuenciación del transcriptoma generan una enorme cantidad de información, y las predicciones iniciales del modelo pueden guiar la validación y anotación experimental.

* **Viabilidad**

El proyecto es viable desde el aspecto técnico, ya que en el campo de la bioinformática el aprendizaje profundo está siendo empleado con éxito. Adicionalmente, el aprendizaje profundo es ideal para problemas en los cuales se tiene gran cantidad de información (bases de datos de transcriptomas), y para el análisis de información secuencial (secuencias de nucleótidos); ambos factores están presentes en este proyecto.

Otro aspecto de viabilidad técnica, es el auge del aprendizaje de máquina en la actualidad. Se tienen herramientas, como la suite Anaconda, que centralizan la instalación, configuración, control de dependencias, edición de código y generación de gráficos y anotaciones para proyectos de investigación científica.

Adicionalmente, se tiene el apoyo del asesor de proyecto, especialista en bioinformática, para absolver posibles dudas que puedan surgir durante la elaboración del proyecto de fin de carrera.

Con respecto a la viabilidad temporal, se posee de medio ciclo para todas las labores de investigación previas a la presentación de resultados (planteamiento del problema y objetivos, marco teórico, estado del arte y planificación de actividades para llegar a los resultados esperados). También durante medio semestre universitario, que se traduce a dos meses y medios, se realizará en base al trabajo previo la implementación y análisis de los resultados. Se estima que estos tiempos serán suficientes para la realización de este proyecto de fin de carrera.

Con respecto a la viabilidad económica, la PUCP ofrece acceso gratuito a bases de datos de investigación. Adicionalmente, el acceso a las bases de datos transcriptómicas es gratuito, y se trabajará con software libre para la elaboración de la solución de aprendizaje de máquina. Finalmente, se requerirá equipo especializado para el entrenamiento y validación del modelo predictivo, el cuál será brindado por el grupo de inteligencia artificial de la PUCP (IA-PUCP), a través de servidores con GPUs especialmente orientados a estas labores.

* **Alcance**

El presente proyecto de investigación se encuentra enmarcado dentro de las Ciencias de la Computación, y es de carácter exploratorio, buscando una alternativa nueva para la predicción de moléculas ARN no codificantes en especies bajo las cuales el modelo no ha sido entrenado. Para ello se planteará una arquitectura para el proceso de aprendizaje de máquina, mediante el cual se defina el entrenamiento y validación del modelo y se generen predictores. Se probará el desempeño del modelo y sus predictores en tres instancias:

* La primera será la validación base del modelo predictivo bajo la técnica del aprendizaje profundo.
* La segunda será la validación del modelo predictivo, al utilizar para el entrenamiento del mismo, atributos utilizados en el modelo predictivo contra el cual será comparado. El motivo es el verificar si un modelo de aprendizaje profundo en este contexto genera mejores resultados que un modelo tradicional que realiza una extracción manual de atributos.
* Finalmente, se realizará una comparación de la precisión de los predictores generados y el modelo del estado del arte para clasificar especies desconocidas al modelo, al ser entrenado en otras.
* **Limitaciones**

Las limitaciones identificadas durante la revisión del estado del arte y la elaboración del proyecto fueron las siguientes:

* Debido a límites de tiempo la validación contra otros modelos no será exhaustiva. Es decir, a pesar de existir en la literatura una gran cantidad de modelos ya elaborados, se ha seleccionado un modelo específico para la comparación: PlantSniffer, de la tesis “Métodos baseados em aprendizagem de máquina para distinguir RNAs longos não-codificadores intergênicos de transcritos codificadores de proteínas” (Vieira, 2018).
* Debido al punto anterior, para la evaluación del modelo, se utilizará como data de entrenamiento transcriptomas (ARN no codificante, *lncRNA*) de plantas, y en específico, se simulará la predicción en maíz, como la especie desconocida al modelo.
* **Identificación de los riesgos del proyecto**

Los riesgos identificados en el presente proyecto y sus respectivas contingencias se presentan en la Tabla 13.

Tabla 13 - Riesgos del proyecto de investigación.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Descripción | Probabilidad | Impacto | Severidad | Síntomas, Mitigación y Contingencia |
| Cantidad insuficiente de transcriptomas para el entrenamiento y validación de los predictores. | Baja  10% | Alto  90% | Baja  9% | Síntomas: Bajo desempeño del modelo predictivo.  Mitigación: Evaluación del estado del arte, analizando si se dio el problema antes.  Contingencia: Uso de otras bases de datos de secuenciación de transcriptomas. |
| Pérdida de información del proyecto por malfunciona-miento o pérdida del equipo donde se trabaja el proyecto. | Media  50% | Alto  90% | Media  45% | Síntomas: Malfuncionamiento de Laptop de trabajo. Hurto.  Mitigación: Uso de sistemas de control de versiones para el almacenamiento en la nube de la información.  Contingencia: Se tiene una PC adicional de trabajo. Recuperación de fuentes de repositorios en la nube. |

* **Estructura de descomposición del trabajo (EDT)**

Se presenta en la Figura 10 la estructura de descomposición del trabajo para el proyecto de fin de carrera.

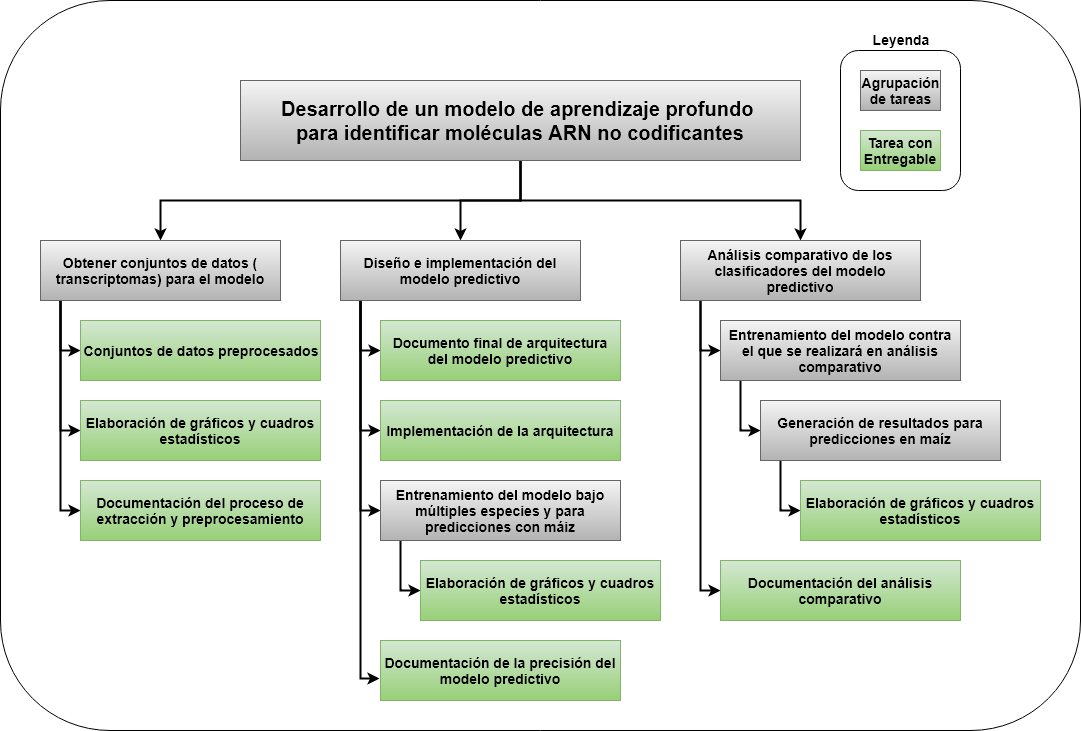


Figura 10 - Estructura de descomposición del trabajo. Elaboración propia.

* **Lista de tareas**

En la Tabla 14 se presenta la lista de tareas.

Se tomó en cuenta un solo recurso, con un costo mensual de mil soles, que se traslada en un costo por hora de 5.68 soles.

Tabla 14 - Lista de tareas del proyecto.



* **Cronograma del proyecto**

En la Figura 11 se presenta el cronograma del proyecto.

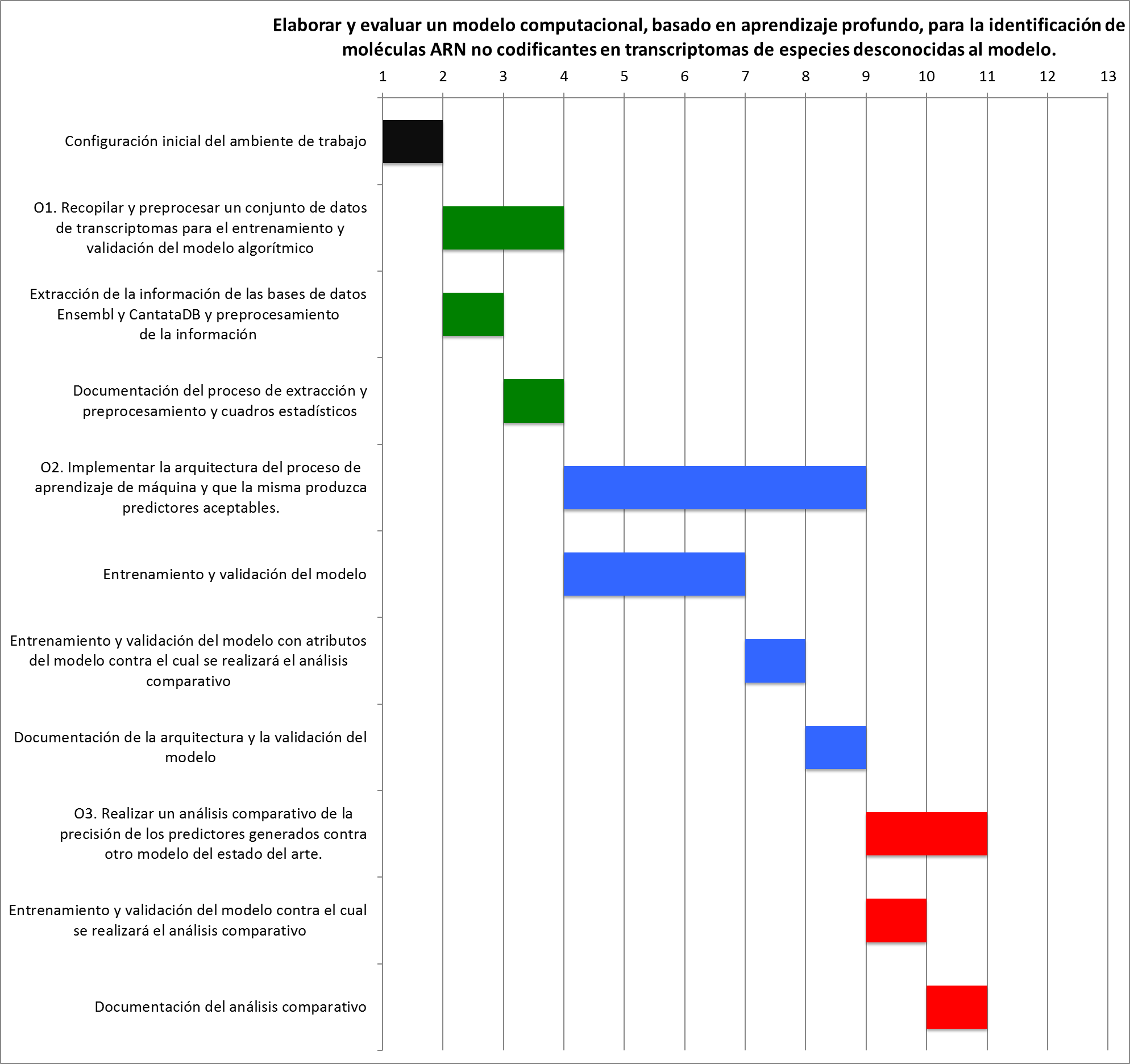


Figura 11 - Cronograma del proyecto por semanas. Elaboración propia.

* **Lista de recursos**
  + **Personas involucradas y necesidades de capacitación**
* Estudiante José Luis Santillán Escudero, quien elabora el proyecto de tesis. Debido a que el proyecto cubre el campo de la bioinformática, se requiere capacitación en temas relacionados a la biología, que será brindada por el asesor del presente proyecto.
* Dr. Edwin Rafael Villanueva Talavera, asesor del presente proyecto, brindará guía con respecto a temas de biología y el uso de aprendizaje de máquina para modelos predictivos.
  + **Materiales requeridos para el proyecto**
* Acceso a bases de datos de investigación, ofrecidos gratuitamente por la PUCP.
* Acceso a bases de datos de transcriptomas (Ensembl y CantataDB).
  + **Estándares utilizados en el proyecto**

No aplica.

* + **Equipamiento requerido**
* Equipo capaz de procesar gran cantidad de información para el entrenamiento y validación de algoritmos de aprendizaje de máquina. Se tendrá acceso al mismo a través del grupo IA-PUCP, que brindará servidores con GPUs especialmente diseñados para la labor.
* Laptop o PC para la elaboración del proyecto de tesis, y la interacción con el servidor durante la ejecución del proyecto de tesis. Se poseen ambos.
  + **Herramientas requeridas**
* Las herramientas requeridas, será un ambiente de trabajo especialmente diseñado para la investigación científica, en específico, para el trabajo con algoritmos de aprendizaje de máquina. Anaconda ofrece gratuitamente una suite, a través de la cual se pueden instalar los lenguajes de programación, librerías, dependencias, herramientas de análisis y visualización de datos y soporte a través de una comunidad activa.
* **Costeo del Proyecto**

En la Tabla 15 se presentan los costos que se identificaron para el presente proyecto.

Tabla - Costos del proyecto.

| **Ítem** | **Descripción** | | | **Unidad** | **Cantidad** | **Valor Unidad (S/.)** | **Monto Parcial (S/.)** | **Monto**  **Total (S/.)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | **Costo total del proyecto** | | | **---** | **---** | **---** | **---** | **5,209** |
| **1.** | **Estudiantes o tesistas** | | | **---** | **---** | **---** | **---** | **1,909** |
| 1.1 | José Luis Santillán Escudero | | | Horas | 336 | 5.68 | 1,909 |  |
| **2.** | **Otros participantes** | | | **---** | **---** | **---** | **---** | **800** |
| 2.1 | Asesor: Dr. Edwin Rafael Villanueva Talavera | | | Horas | 40 | 20 | 800 |  |
| **3.** | **Servicios y consultoría** | | | **---** | **---** | **---** | **---** | **0** |
|  | No aplica. | | |  |  |  |  |  |
| **4.** | **Materiales e insumos** | | | **---** | **---** | **---** | **---** | **0** |
| 4.1 | Acceso a base de datos PUCP | | | Costo desconocido | | | |  |
| 4.2 | Acceso a bases de datos de transcriptomas | | | Costo desconocido | | | |  |
| **5.** | **Bienes y equipos** | **Unid1** | **Cant1-** | **Unid2** | **Cant2-** | **-** | **-** | **2,500** |
| 5.1 | Computadora | Equipo | 1 | Horas | 100 | 5 | 500 |  |
| 5.2 | Servidor  GPU grupo IA-PUCP | Equipo | 1 | Horas | 200 | 10 | 2,000 |  |
| **6.** | **Pasajes y viáticos** | **Unid1** | **Cant1-** | **Unid2** | **Cant2-** | **-** | **-** | **0** |
|  | No aplica |  |  |  |  |  |  |  |

## : Descarga de Ensembl a través de biomaRt

%%time

%%R

library(RMySQL)

library(curl)

library(biomaRt)

library(stringr)

cat("Iniciando proceso...**\n**")

dsn\_database = "tesis2"

dsn\_hostname = "localhost"

dsn\_port = 3306

dsn\_uid = "tesis2"

dsn\_pwd = "tesis2"

conn = dbConnect(MySQL(), user=dsn\_uid, password=dsn\_pwd, dbname=dsn\_database, host=dsn\_hostname)

rs = dbSendQuery(conn, "SELECT id\_especie, especie FROM maestra\_especies m WHERE especie IN (SELECT c1.especie FROM conteo\_lncrna\_especie c1 LEFT JOIN conteo\_lncrna\_especie c2 ON c1.especie = REPLACE(c2.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c2.fuente = 'CantataDB' LEFT JOIN conteo\_lncrna\_especie c3 ON c1.especie = REPLACE(c3.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c3.fuente = 'GreeNC' WHERE c1.fuente = 'Ensembl' AND (c2.fuente IS NOT NULL OR c3.fuente IS NOT NULL)) AND NOT EXISTS (SELECT 1 FROM secuencias s WHERE s.id\_especie = m.id\_especie AND s.flg\_pct = 1)")

especies = dbFetch(rs)

dbClearResult(rs)

ensembl\_mart=useMart("plants\_mart", host="plants.ensembl.org")

lista\_datasets = listDatasets(ensembl\_mart)

**if** (nrow(especies) > 0) {

**for**(i **in** 1:nrow(especies)) {

tryCatch(

expr = {

cat("**\r**Procesando especie #", i, " ", especies[i,2], " ", sep = "")

la\_especie = especies[i,2]

especie = paste("^", la\_especie, sep="")

dataset = lista\_datasets[which(str\_detect(lista\_datasets$description, especie)), 1]

ensembl\_ds = useDataset(dataset,mart=ensembl\_mart)

genes = getBM(attributes=c('ensembl\_transcript\_id', 'transcript\_exon\_intron'), filters = 'transcript\_biotype', values = 'protein\_coding', mart = ensembl\_ds)

**for**(j **in** 1:nrow(genes)) {

**if** (j %% 1000 == 0) {

cat("**\r**Procesando especie #", i, " ", especies[i,2], " - procesadas ", j, " filas ", sep = "")

}

**if** (!(genes[j,2] == "Sequence unavailable" | str\_length(genes[j,2]) > 65000)) {

query = sprintf("INSERT INTO secuencias (id\_especie, cod\_secuencia, secuencia, flg\_pct, flg\_seleccionado) VALUES (**%d**, '**%s**', '**%s**', 1, NULL)", especies[i,1], genes[j,1], genes[j,2])

dbSendQuery(conn, query)

}

}

cat("**\r**Procesando especie #", i, " ", especies[i,2], " - procesadas ", nrow(genes), " filas ", sep = "")

},

error = function(e) {

cat("**\r**Error al procesar la especie ", especies[i,2], " **\n**", sep = "")

}

)

}

}

dbDisconnect(conn)

cat("**\r**Proceso finalizado... ")

## : Procesamiento de descargas manuales de Ensembl

%%time

**import** **mysql.connector**

**import** **os**

*#método para leer los transcriptomas de un archivo fasta de ensembl*

**def** leer\_transcriptomas\_ensembl(archivo, cur, id\_especie):

cabecera = ""

transcriptoma = ""

f = open(archivo, "r")

**for** linea **in** f:

**if** linea.startswith(">"):

**if** transcriptoma != "":

**if** transcriptoma != "Sequence unavailable" **and** len(transcriptoma) <= 65000:

query = "INSERT INTO secuencias (id\_especie, cod\_secuencia, secuencia, flg\_pct, flg\_seleccionado) VALUES (**%s**, **%s**, **%s**, 1, NULL)"

cur.execute(query, (id\_especie, cabecera, transcriptoma))

transcriptoma = ""

cabecera = linea.rstrip("**\n**").lstrip(">")

**else**:

transcriptoma += linea.rstrip("**\n**")

**if** transcriptoma != "" **and** len(transcriptoma) <= 65000:

**if** transcriptoma != "Sequence unavailable":

query = "INSERT INTO secuencias (id\_especie, cod\_secuencia, secuencia, flg\_pct, flg\_seleccionado) VALUES (**%s**, **%s**, **%s**, 1, NULL)"

cur.execute(query, (id\_especie, cabecera, transcriptoma))

transcriptoma = ""

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cur = conn.cursor()

directory = '../Semana 05/descarga\_manual\_ensembl'

**for** filename **in** os.listdir(directory):

**if** filename.endswith(".fasta"):

especie = os.path.splitext(filename)[0]

query = "SELECT COUNT(\*) FROM maestra\_especies m WHERE especie IN (SELECT c1.especie FROM conteo\_lncrna\_especie c1 LEFT JOIN conteo\_lncrna\_especie c2 ON c1.especie = REPLACE(c2.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c2.fuente = 'CantataDB' LEFT JOIN conteo\_lncrna\_especie c3 ON c1.especie = REPLACE(c3.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c3.fuente = 'GreeNC' WHERE c1.fuente = 'Ensembl' AND (c2.fuente IS NOT NULL OR c3.fuente IS NOT NULL)) AND m.especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

result=cur.fetchone()

**if** result[0] > 0:

print("**\r**Procesando especie ", especie, " ", end=" ")

query = "SELECT e.id\_especie FROM maestra\_especies e WHERE e.especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

id\_especie=cur.fetchone()

query = "delete from secuencias where id\_especie = **%s**"

cur.execute(query, [id\_especie[0]])

leer\_transcriptomas\_ensembl(directory + "/" + filename, cur, id\_especie[0])

conn.commit()

conn.close()

print("**\r**Proceso finalizado... ")

## : Procesamiento de archivos fasta de CantataDB

%%time

**import** **mysql.connector**

**import** **os**

*#método para leer los transcriptomas de un archivo fasta en cantatadb*

**def** leer\_transcriptomas\_cantatadb(archivo, cur, id\_especie):

cabecera = ""

transcriptoma = ""

f = open(archivo, "r")

**for** linea **in** f:

**if** linea.startswith(">"):

**if** transcriptoma != "":

**if** len(transcriptoma) <= 65000:

query = "INSERT INTO secuencias (id\_especie, cod\_secuencia, secuencia, flg\_pct, flg\_seleccionado) VALUES (**%s**, **%s**, **%s**, 0, NULL)"

cur.execute(query, (id\_especie, cabecera, transcriptoma))

transcriptoma = ""

cabecera = linea.rstrip("**\n**").lstrip(">")

**else**:

transcriptoma += linea.rstrip("**\n**")

**if** transcriptoma != "":

**if** len(transcriptoma) <= 65000:

query = "INSERT INTO secuencias (id\_especie, cod\_secuencia, secuencia, flg\_pct, flg\_seleccionado) VALUES (**%s**, **%s**, **%s**, 0, NULL)"

cur.execute(query, (id\_especie, cabecera, transcriptoma))

transcriptoma = ""

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cur = conn.cursor()

directory = '../Semana 02/httpcantata.amu.edu.pldownload.php/Otros'

**for** filename **in** os.listdir(directory):

**if** filename.endswith(".fasta"):

especie = os.path.splitext(filename)[0].split("\_lncrnas")[:-1][0]

query = "SELECT COUNT(\*) FROM maestra\_especies m WHERE especie IN (SELECT c1.especie FROM conteo\_lncrna\_especie c1 LEFT JOIN conteo\_lncrna\_especie c2 ON c1.especie = REPLACE(c2.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c2.fuente = 'CantataDB' LEFT JOIN conteo\_lncrna\_especie c3 ON c1.especie = REPLACE(c3.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c3.fuente = 'GreeNC' WHERE c1.fuente = 'Ensembl' AND (c2.fuente IS NOT NULL OR c3.fuente IS NOT NULL)) AND NOT EXISTS (SELECT 1 FROM secuencias s WHERE s.id\_especie = m.id\_especie AND s.flg\_pct = 0) AND m.especie = REPLACE(**%s**, '\_', ' ') AND EXISTS (SELECT 1 FROM especies\_seleccionadas es WHERE es.especie = m.especie AND fuente = **%s**)"

cur.execute(query, (especie, "CantataDB"))

result=cur.fetchone()

**if** result[0] > 0:

print("**\r**Procesando especie ", especie, " ", end=" ")

query = "SELECT e.id\_especie FROM maestra\_especies e WHERE e.especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

id\_especie=cur.fetchone()

leer\_transcriptomas\_cantatadb(directory + "/" + filename, cur, id\_especie[0])

conn.commit()

conn.close()

print("**\r**Proceso finalizado... ")

## : Procesamiento de archivos fasta de GreeNC

%%time

**import** **mysql.connector**

**import** **os**

*#método para leer los transcriptomas de un archivo fasta en cantatadb*

**def** leer\_transcriptomas\_cantatadb(archivo, cur, id\_especie):

cabecera = ""

transcriptoma = ""

f = open(archivo, "r")

**for** linea **in** f:

**if** linea.startswith(">"):

**if** transcriptoma != "":

**if** len(transcriptoma) <= 65000:

query = "INSERT INTO secuencias (id\_especie, cod\_secuencia, secuencia, flg\_pct, flg\_seleccionado) VALUES (**%s**, **%s**, **%s**, 0, NULL)"

cur.execute(query, (id\_especie, cabecera, transcriptoma))

transcriptoma = ""

cabecera = linea.rstrip("**\n**").lstrip(">")

**else**:

transcriptoma += linea.rstrip("**\n**")

**if** transcriptoma != "":

**if** len(transcriptoma) <= 65000:

query = "INSERT INTO secuencias (id\_especie, cod\_secuencia, secuencia, flg\_pct, flg\_seleccionado) VALUES (**%s**, **%s**, **%s**, 0, NULL)"

cur.execute(query, (id\_especie, cabecera, transcriptoma))

transcriptoma = ""

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cur = conn.cursor()

directory = '../Semana 03/httpgreenc.sciencedesigners.com'

**for** filename **in** os.listdir(directory):

**if** filename.endswith(".fasta"):

especie = os.path.splitext(filename)[0]

query = "SELECT COUNT(\*) FROM maestra\_especies m WHERE especie IN (SELECT c1.especie FROM conteo\_lncrna\_especie c1 LEFT JOIN conteo\_lncrna\_especie c2 ON c1.especie = REPLACE(c2.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c2.fuente = 'CantataDB' LEFT JOIN conteo\_lncrna\_especie c3 ON c1.especie = REPLACE(c3.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c3.fuente = 'GreeNC' WHERE c1.fuente = 'Ensembl' AND (c2.fuente IS NOT NULL OR c3.fuente IS NOT NULL)) AND NOT EXISTS (SELECT 1 FROM secuencias s WHERE s.id\_especie = m.id\_especie AND s.flg\_pct = 0) AND m.especie = REPLACE(**%s**, '\_', ' ') AND EXISTS (SELECT 1 FROM especies\_seleccionadas es WHERE es.especie = m.especie AND fuente = **%s**)"

cur.execute(query, (especie, "GreeNC"))

result=cur.fetchone()

**if** result[0] > 0:

print("**\r**Procesando especie ", especie, " ", end=" ")

query = "SELECT id\_especie FROM maestra\_especies WHERE especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

id\_especie=cur.fetchone()

leer\_transcriptomas\_cantatadb(directory + "/" + filename, cur, id\_especie[0])

conn.commit()

conn.close()

print("**\r**Proceso finalizado... ")

## : Selección aleatoria de transcriptomas para el modelo

%%time

**import** **mysql.connector**

**import** **pandas** **as** **pd**

print("Iniciando proceso...")

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

cursor = conn.cursor()

query = "UPDATE secuencias SET flg\_seleccionado = NULL"

cursor.execute(query)

query = "SELECT id\_especie FROM maestra\_especies m JOIN especies\_seleccionadas s ON m.especie = s.especie"

cursor.execute(query)

ids = cursor.fetchall()

**for** id\_especie **in** ids:

print("**\r**PCT for " + str(id\_especie[0]) + " ", end=" ")

query = "UPDATE secuencias SET flg\_seleccionado = 1 WHERE id\_especie = **%s** AND flg\_pct = 1 ORDER BY RAND() LIMIT 4208"

cursor.execute(query, [id\_especie[0]])

print("**\r**lncRNA for " + str(id\_especie[0]) + " ", end=" ")

query = "UPDATE secuencias SET flg\_seleccionado = 1 WHERE id\_especie = **%s** AND flg\_pct = 0 ORDER BY RAND() LIMIT 4208"

cursor.execute(query, [id\_especie[0]])

print("**\r**Proceso terminado... **\n**")

conn.close()

## : Características – Longitud del transcriptoma

%%time

**import** **mysql.connector**

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cursor = conn.cursor()

query = "SELECT id\_especie FROM maestra\_especies"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Features para id-especie " + str(especie[0]), " ", end="")

query = "INSERT INTO secuencias\_features (id\_especie, cod\_secuencia, flg\_pct, flg\_seleccionado, longitud) SELECT id\_especie, cod\_secuencia, flg\_pct, flg\_seleccionado, LENGTH(secuencia) FROM secuencias WHERE id\_especie = **%s**"

cursor.execute(query, [especie[0]])

print("**\r**Proceso finalizado. ")

conn.close()

## : Características – Porcentaje de nucleótidos GC

%%time

**import** **mysql.connector**

**import** **os**

**from** **Bio.SeqUtils** **import** GC

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cursor = conn.cursor()

query = "SELECT m.id\_especie, REPLACE(m.especie, ' ', '\_') FROM maestra\_especies m JOIN especies\_seleccionadas es ON m.especie = es.especie"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Generando GC content " + str(especie[0]), " ", end="")

query = "SELECT cod\_secuencia, secuencia FROM secuencias WHERE id\_especie = **%s** AND flg\_seleccionado = 1"

cursor.execute(query, [especie[0]])

transcriptomas = cursor.fetchall()

**for** transcriptoma **in** transcriptomas:

gc\_content = GC(transcriptoma[1])

query = "UPDATE secuencias\_features SET gc\_content = **%s** WHERE id\_especie = **%s** and cod\_secuencia = **%s**"

cur.execute(query, (gc\_content, especie[0], transcriptoma[0]))

conn.commit()

conn.close()

print("**\r**Proceso finalizado... ")

## : Generación de archivos fasta para CDS y lncRNA

%%time

**import** **mysql.connector**

**import** **os**

**def** generar\_fasta(secuencias, especie, tipo, directory):

t\_tamanio = 80

f = open(directory + especie + "." + tipo + ".fasta" ,"w+")

**for** transcriptoma **in** secuencias:

f.write(">**%s\n**" % (transcriptoma[0]))

seq = transcriptoma[1]

t\_partes = [seq[i:i+t\_tamanio] **for** i **in** range(0, len(seq), t\_tamanio)]

**for** t\_parte **in** t\_partes:

f.write("**%s\n**" % (t\_parte))

f.close()

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cursor = conn.cursor()

directory = '../Semana 05/fasta\_para\_hexamer/'

query = "SELECT m.id\_especie, REPLACE(m.especie, ' ', '\_') FROM maestra\_especies m JOIN especies\_seleccionadas es ON m.especie = es.especie"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Generando fasta CDS para " + str(especie[0]), " ", end="")

query = "select c.cod\_secuencia, c.coding from secuencias\_CDS c JOIN secuencias\_features f ON c.id\_especie = f.id\_especie AND c.cod\_secuencia = f.cod\_secuencia WHERE c.id\_especie = **%s** AND f.flg\_pct = 1 AND f.flg\_seleccionado = 1"

cursor.execute(query, [especie[0]])

secuencias = cursor.fetchall()

generar\_fasta(secuencias, especie[1], "CDS", directory)

print("**\r**Generando fasta lncRNA para " + str(especie[0]), " ", end="")

query = "SELECT cod\_secuencia, secuencia FROM secuencias WHERE id\_especie = **%s** AND flg\_pct = 0 AND flg\_seleccionado = 1"

cursor.execute(query, [especie[0]])

secuencias = cursor.fetchall()

generar\_fasta(secuencias, especie[1], "lncRNA", directory)

conn.close()

print("**\r**Proceso finalizado... ")

## : CPAT – Generación de tabla de frecuencias de hexámeros

%%time

**import** **mysql.connector**

**import** **os**

**from** **IPython.utils** **import** io

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cursor = conn.cursor()

input\_directory = '../Semana 05/fasta\_para\_hexamer/'

output\_directory = '../Semana 05/tablas\_hexamer/'

query = "SELECT m.id\_especie, REPLACE(m.especie, ' ', '\_') FROM maestra\_especies m JOIN especies\_seleccionadas es ON m.especie = es.especie"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Generando tabla hexamer para " + str(especie[0]), " ", end="")

script = "~/anaconda3/bin/make\_hexamer\_tab.py"

fasta\_cds = "'" + input\_directory + especie[1] + ".CDS.fasta" + "'"

lncRNA\_cds = "'" + input\_directory + especie[1] + ".lncRNA.fasta" + "'"

salida = output\_directory + especie[1] + ".tsv"

**with** io.capture\_output() **as** captured:

%run $script -c $fasta\_cds -n $lncRNA\_cds

f = open(salida ,"w+")

f.write(captured.stdout)

f.close()

conn.close()

print("**\r**Proceso finalizado... ")

## : Generación de archivos fasta para ARN codificantes

%%time

**import** **mysql.connector**

**import** **os**

**def** generar\_fasta(secuencias, especie, tipo, directory):

t\_tamanio = 80

f = open(directory + especie + "." + tipo + ".fasta" ,"w+")

**for** transcriptoma **in** secuencias:

f.write(">**%s\n**" % (transcriptoma[0]))

seq = transcriptoma[1]

t\_partes = [seq[i:i+t\_tamanio] **for** i **in** range(0, len(seq), t\_tamanio)]

**for** t\_parte **in** t\_partes:

f.write("**%s\n**" % (t\_parte))

f.close()

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cursor = conn.cursor()

directory = '../Semana 05/fasta\_pct\_para\_logit/'

query = "SELECT m.id\_especie, REPLACE(m.especie, ' ', '\_') FROM maestra\_especies m JOIN especies\_seleccionadas es ON m.especie = es.especie"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Generando fasta CDS para " + str(especie[0]), " ", end="")

query = "SELECT cod\_secuencia, secuencia FROM secuencias WHERE id\_especie = **%s** AND flg\_pct = 1 AND flg\_seleccionado = 1"

cursor.execute(query, [especie[0]])

secuencias = cursor.fetchall()

generar\_fasta(secuencias, especie[1], "PCT", directory)

conn.close()

print("**\r**Proceso finalizado... ")

## : CPAT – Generación de regresión logística

%%time

**import** **mysql.connector**

**import** **os**

**from** **IPython.utils** **import** io

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cursor = conn.cursor()

input\_directory = '../Semana 05/tablas\_hexamer/'

input\_directory\_pct = '../Semana 05/fasta\_pct\_para\_logit/'

input\_directory\_lncRNA = '../Semana 05/fasta\_para\_hexamer/'

output\_directory\_logit = '../Semana 05/modelo\_regresion\_logistica/'

query = "SELECT m.id\_especie, REPLACE(m.especie, ' ', '\_') FROM maestra\_especies m JOIN especies\_seleccionadas es ON m.especie = es.especie"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Generando modelo de regresión para " + str(especie[0]), " ", end="")

script = "~/anaconda3/bin/make\_logitModel.py"

archivo\_hexamer = "'" + input\_directory + especie[1] + ".tsv" + "'"

fasta\_pct = "'" + input\_directory\_pct + especie[1] + ".PCT.fasta" + "'"

lncRNA\_cds = "'" + input\_directory\_lncRNA + especie[1] + ".lncRNA.fasta" + "'"

salida = "'" + output\_directory\_logit + especie[1] + "'"

**with** io.capture\_output() **as** captured:

%run $script -x $archivo\_hexamer -c $fasta\_pct -n $lncRNA\_cds -o $salida

conn.close()

print("**\r**Proceso finalizado... ")

## : CPAT – Ejecución de CPAT

%%time

**import** **mysql.connector**

**import** **os**

**import** **csv**

**def** procesar\_features\_cpat(archivo, cur, id\_especie):

**with** open(archivo, "r") **as** tabular:

cpat\_reader = csv.reader(tabular, delimiter=("**\t**"))

**for** row **in** cpat\_reader:

cod\_secuencia = row[0]

orf\_length = float(row[2])

fickett = float(row[3])

hexamer = float(row[4])

query = "UPDATE secuencias\_features SET orf\_length = **%s**, orf\_coverage = **%s**/longitud, ficket\_score = **%s**, hexamer\_score = **%s** WHERE id\_especie = **%s** and cod\_secuencia = **%s**"

cur.execute(query, (orf\_length, orf\_length, fickett, hexamer, id\_especie, cod\_secuencia))

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cur = conn.cursor()

directory = '../Semana 05/cpat\_features'

**for** filename **in** os.listdir(directory):

**if** filename.endswith(".dat"):

especie = filename.split(".")[0]

query = "SELECT COUNT(\*) FROM maestra\_especies m WHERE especie IN (SELECT c1.especie FROM conteo\_lncrna\_especie c1 LEFT JOIN conteo\_lncrna\_especie c2 ON c1.especie = REPLACE(c2.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c2.fuente = 'CantataDB' LEFT JOIN conteo\_lncrna\_especie c3 ON c1.especie = REPLACE(c3.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c3.fuente = 'GreeNC' WHERE c1.fuente = 'Ensembl' AND (c2.fuente IS NOT NULL OR c3.fuente IS NOT NULL)) AND m.especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

result=cur.fetchone()

**if** result[0] > 0:

print("**\r**Procesando especie", especie, filename.split(".")[1]," ", end=" ")

query = "SELECT e.id\_especie FROM maestra\_especies e WHERE e.especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

id\_especie=cur.fetchone()

procesar\_features\_cpat(directory + "/" + filename, cur, id\_especie[0])

conn.commit()

conn.close()

print("**\r**Proceso finalizado... ")

## : CPAT – Importación de características a la base de datos

%%time

**import** **mysql.connector**

**import** **os**

**import** **csv**

**def** procesar\_features\_cpat(archivo, cur, id\_especie):

**with** open(archivo, "r") **as** tabular:

cpat\_reader = csv.reader(tabular, delimiter=("**\t**"))

**for** row **in** cpat\_reader:

cod\_secuencia = row[0]

orf\_length = float(row[2])

fickett = float(row[3])

hexamer = float(row[4])

query = "UPDATE secuencias\_features SET orf\_length = **%s**, orf\_coverage = **%s**/longitud, ficket\_score = **%s**, hexamer\_score = **%s** WHERE id\_especie = **%s** and cod\_secuencia = **%s**"

cur.execute(query, (orf\_length, orf\_length, fickett, hexamer, id\_especie, cod\_secuencia))

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

print("Iniciando proceso...")

cur = conn.cursor()

directory = '../Semana 05/cpat\_features'

**for** filename **in** os.listdir(directory):

**if** filename.endswith(".dat"):

especie = filename.split(".")[0]

query = "SELECT COUNT(\*) FROM maestra\_especies m WHERE especie IN (SELECT c1.especie FROM conteo\_lncrna\_especie c1 LEFT JOIN conteo\_lncrna\_especie c2 ON c1.especie = REPLACE(c2.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c2.fuente = 'CantataDB' LEFT JOIN conteo\_lncrna\_especie c3 ON c1.especie = REPLACE(c3.especie, '\_', ' ') AND c1.fuente = 'Ensembl' AND c3.fuente = 'GreeNC' WHERE c1.fuente = 'Ensembl' AND (c2.fuente IS NOT NULL OR c3.fuente IS NOT NULL)) AND m.especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

result=cur.fetchone()

**if** result[0] > 0:

print("**\r**Procesando especie", especie, filename.split(".")[1]," ", end=" ")

query = "SELECT e.id\_especie FROM maestra\_especies e WHERE e.especie = REPLACE(**%s**, '\_', ' ')"

cur.execute(query, [especie])

id\_especie=cur.fetchone()

procesar\_features\_cpat(directory + "/" + filename, cur, id\_especie[0])

conn.commit()

conn.close()

print("**\r**Proceso finalizado... ")

## : Diamond – Generación de la base de datos

%%time

**import** **os**

!~/anaconda3/bin/diamond makedb --**in** '../Semana 05/diamond/uniprot-viridiplantae-reviewed.fasta' -d '../Semana 05/diamond/uniprot-viridiplantae-reviewed'

## : Diamond – Búsqueda de alineamientos

%%time

**import** **mysql.connector**

**import** **os**

**from** **IPython.utils** **import** io

print("Iniciando proceso...")

cursor = conn.cursor()

input\_directory\_pct = '../Semana 05/fasta\_pct\_para\_logit/'

input\_directory\_lncRNA = '../Semana 05/fasta\_para\_hexamer/'

output\_directory = '../Semana 05/diamond\_features/'

query = "SELECT m.id\_especie, REPLACE(m.especie, ' ', '\_') FROM maestra\_especies m JOIN especies\_seleccionadas es ON m.especie = es.especie"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Realizando búsqueda para " + str(especie[0]), " PCT ", end="")

diamond\_db = "'" + '../Semana 05/diamond/uniprot-viridiplantae-reviewed.dmnd' + "'"

fasta\_pct = "'" + input\_directory\_pct + especie[1] + ".PCT.fasta" + "'"

lncRNA\_cds = "'" + input\_directory\_lncRNA + especie[1] + ".lncRNA.fasta" + "'"

salida = "'" + output\_directory + especie[1] + ".PCT.tsv" + "'"

**with** io.capture\_output() **as** captured:

!~/anaconda3/bin/diamond blastx -d $diamond\_db -q $fasta\_pct -o $salida -k 5 --gapopen 11 --gapextend 1 --more-sensitive -f 6 qseqid pident length qframe qstart qend sstart send evalue bitscore

print("**\r**Realizando búsqueda para " + str(especie[0]), " lncRNA ", end="")

lncRNA\_cds = "'" + input\_directory\_lncRNA + especie[1] + ".lncRNA.fasta" + "'"

salida = "'" + output\_directory + especie[1] + ".lncRNA.tsv" + "'"

**with** io.capture\_output() **as** captured:

!~/anaconda3/bin/diamond blastx -d $diamond\_db -q $lncRNA\_cds -o $salida -k 5 --gapopen 11 --gapextend 1 --more-sensitive -f 6 qseqid pident length qframe qstart qend sstart send evalue bitscore

conn.close()

print("**\r**Proceso finalizado... ")

## : Diamond – Importación de características a la base de datos

%%time

**import** **mysql.connector**

**import** **csv**

**import** **os**

conn = mysql.connector.connect(

host="localhost",

database="tesis2",

user="tesis2",

passwd="tesis2"

)

*#adaptado de https://github.com/gbgolding/crema/blob/master/bin/featuresetup\_module.py*

**def** transcript\_info\_dict(id\_especie, blast\_file, cursor, flg\_pct):

transcript\_dict = {}

**with** open(blast\_file, "r") **as** f:

tab\_reader = csv.reader(f, delimiter=("**\t**"))

line\_1 = next(tab\_reader)

first = line\_1[0]

score = [float(line\_1[9])]

with\_len = [[first, float(line\_1[1]), float(line\_1[2]), float(line\_1[3]), float(line\_1[9])]] *# name identity length frame score*

**for** row **in** tab\_reader:

**if** row[0] == first:

score.append(float(row[9]))

with\_len.append([row[0], float(row[1]), float(row[2]), float(row[3]), float(row[9])])

**else**:

transcript\_dict[first] = {}

transcript\_dict[first]["identity"] = float(0)

transcript\_dict[first]["align\_length"] = float(0)

max\_value = max(score)

max\_index = score.index(max\_value)

max\_len\_ident = with\_len[max\_index]

**if** max\_len\_ident[3] > 0:

transcript\_dict[first]["identity"] = float(max\_len\_ident[1])

transcript\_dict[first]["align\_length"] = float(max\_len\_ident[2])

score = [float(row[9])]

first = row[0]

with\_len = [[first, float(row[1]), float(row[2]), float(row[3]), float(row[9])]]

transcript\_dict[first] = {}

transcript\_dict[first]["identity"] = float(0)

transcript\_dict[first]["align\_length"] = float(0)

max\_value = max(score)

max\_index = score.index(max\_value)

max\_len\_ident = with\_len[max\_index]

**if** max\_len\_ident[3] > 0:

transcript\_dict[first]["identity"] = float(max\_len\_ident[1])

transcript\_dict[first]["align\_length"] = float(max\_len\_ident[2])

*#fin de código adaptado de https://github.com/gbgolding/crema/blob/master/bin/featuresetup\_module.py*

query = "SELECT cod\_secuencia, IFNULL(longitud, 0), IFNULL(orf\_length, 0) FROM secuencias\_features WHERE flg\_seleccionado = 1 AND id\_especie = **%s** AND flg\_pct = **%s**"

cursor.execute(query, (id\_especie, flg\_pct))

transcriptomas = cursor.fetchall()

**for** transcriptoma **in** transcriptomas:

cod\_secuencia = transcriptoma[0]

longitud = transcriptoma[1]

orf\_length = transcriptoma[2]

**if** transcriptoma[0] **in** transcript\_dict:

transcript\_dict[cod\_secuencia]["align\_perc\_len"] = float(transcript\_dict[cod\_secuencia]["align\_length"]/longitud)

**if** orf\_length == 0:

transcript\_dict[cod\_secuencia]["align\_perc\_ORF"] = float(0)

**else**:

transcript\_dict[cod\_secuencia]["align\_perc\_ORF"] = float(transcript\_dict[cod\_secuencia]["align\_length"]/orf\_length)

**else**:

transcript\_dict[cod\_secuencia] = {}

transcript\_dict[cod\_secuencia]["identity"] = float(0)

transcript\_dict[cod\_secuencia]["align\_length"] = float(0)

transcript\_dict[cod\_secuencia]["align\_perc\_len"] = float(0)

transcript\_dict[cod\_secuencia]["align\_perc\_ORF"] = float(0)

query = "UPDATE secuencias\_features SET alignment = **%s**, alignment\_length = **%s**, alignment\_proportion = **%s**, alignment\_proportion\_orf = **%s** WHERE id\_especie = **%s** AND cod\_secuencia = **%s**"

cursor.execute(query, (transcript\_dict[cod\_secuencia]["identity"], transcript\_dict[cod\_secuencia]["align\_length"], transcript\_dict[cod\_secuencia]["align\_perc\_len"], transcript\_dict[cod\_secuencia]["align\_perc\_ORF"], id\_especie, cod\_secuencia))

print("Iniciando proceso...")

cursor = conn.cursor()

output\_directory = '../Semana 05/diamond\_features/'

query = "SELECT m.id\_especie, REPLACE(m.especie, ' ', '\_') FROM maestra\_especies m JOIN especies\_seleccionadas es ON m.especie = es.especie"

cursor.execute(query)

especies = cursor.fetchall()

**for** especie **in** especies:

print("**\r**Exportando features para " + str(especie[0]), " PCT ", end="")

salida = output\_directory + especie[1] + ".PCT.tsv"

transcript\_info\_dict(especie[0], salida, cursor, 1)

conn.commit()

print("**\r**Exportando features para " + str(especie[0]), " lncRNA ", end="")

salida = output\_directory + especie[1] + ".lncRNA.tsv"

transcript\_info\_dict(especie[0], salida, cursor, 0)

conn.commit()

conn.close()

print("**\r**Proceso finalizado... ")