**Sistema de identificación y vinculación gen–función a partir de literatura científica** Octubre de 2025

Marian Bastidas Cruz, Catalina Criollo Castelblanco, Jonathan Andres Jimenez Trujillo,

[mbastidasc@unal.edu.co](mailto:mbastidasc@unal.edu.co) , [ccriollo@unal.edu.co](mailto:ccriollo@unal.edu.co) , [jojimenezt@unal.edu.co](mailto:jojimenezt@unal.edu.co)

Universidad Nacional de Colombia

**Resumen** - Este

*Palabras claves:*

1. INTRODUCCIÓN

En la actualidad, el crecimiento exponencial de la literatura biomédica representa un desafío significativo para investigadores y profesionales del área de la salud. La cantidad de artículos publicados diariamente en bases de datos como PubMed o PMC dificulta la lectura manual y la extracción eficiente del conocimiento científico disponible. En este contexto, surge la necesidad de desarrollar herramientas automáticas capaces de procesar grandes volúmenes de texto y transformar la información no estructurada en datos organizados y consultables.

El presente proyecto propone el diseño e implementación de un sistema basado en Procesamiento de Lenguaje Natural (PLN) orientado a la identificación automática de genes y sus funciones biológicas en artículos científicos. Mediante el uso de modelos preentrenados especializados en biomedicina, el sistema busca reconocer entidades relevantes (genes y funciones) y extraer las relaciones entre ellas, con el fin de generar una base de datos estructurada que consolide la información dispersa en la literatura científica.

1. ENTENDIMIENTO DEL NEGOCIO
2. Objetivos

El presente proyecto tiene como objetivo desarrollar una herramienta de Procesamiento de Lenguaje Natural (PLN) aplicada al análisis de literatura biomédica, con el fin de identificar automáticamente genes y sus funciones biológicas en artículos científicos. Esta herramienta permitirá estructurar el conocimiento disperso en grandes volúmenes de textos y almacenarlo en una base de datos consultable que relacione genes, funciones y evidencia textual.

Desde la perspectiva del negocio, este proyecto representa el punto de partida para una solución tecnológica innovadora orientada a optimizar los procesos de revisión, análisis y extracción de información biomédica. Su implementación puede beneficiar a laboratorios farmacéuticos, grupos de investigación, universidades y empresas del sector biotecnológico, al facilitar la integración y sistematización del conocimiento científico.

1. Descripción del conjunto de textos seleccionado

El conjunto de textos seleccionado está conformado por artículos científicos especializados en biología molecular y genética, centrados en la descripción de genes y sus funciones. Los documentos pertenecen principalmente a estudios relacionados con familias génicas como RAS y PLA2G2A, involucradas en diversos procesos celulares y patologías.

Entre los títulos que componen el corpus se encuentran, por ejemplo:

* *“The Secretory Phospholipase A2 Gene Is a Candidate”*
* *“RAS Is Regulated by the let-7 MicroRNA Family”*
* *“Multiple Forms of Phospholipase A2 in Arthritic Synovial Fluid”*
* *“NRAS Mutation Causes a Human Autoimmunelymphoproliferative Disorder”*

De cada artículo se extrajeron los siguientes elementos:

* Autor(es): responsables de la publicación científica.
* Título: nombre del artículo, indicador principal del tema tratado.
* Abstract: resumen del contenido, empleado para la detección automática de entidades (genes y funciones).
* Referencias: conjunto de fuentes bibliográficas citadas, que permiten establecer relaciones entre trabajos y autores.

Estos textos conforman el corpus de entrenamiento y análisis sobre el cual se aplican las técnicas de PLN para la identificación de entidades y la extracción de relaciones gen-función.

1. Proceso de carga u obtención de textos

El proceso de obtención de los textos se realizó a partir de la plataforma OMIM (Online Mendelian Inheritance in Man), específicamente desde la sección Gene Map, que recopila información genética y publicaciones científicas asociadas a genes humanos y sus mutaciones.

Para cada gen de interés, se consultó su entrada en la Gene Map, desde la cual se accedió a los artículos vinculados. Los documentos fueron descargados en formato PDF, organizados en carpetas correspondientes a cada gen (por ejemplo, *RAS* y *PLA2G2A*), y posteriormente procesados para la extracción automática de metadatos relevantes: autor, título, abstract y referencias.

1. ENTENDIMIENTO Y PREPARACIÓN DEL CONJUNTO DE TEXTO

- Análisis exploratorio del conjunto de textos.  
- Análisis descriptivo  
- Visualización de los textos  
- Preprocesamiento en NLP

1. TRANSFORMACIÓN DEL TEXTO PARA EXTRACCIÓN DE CARACTERÍSTICAS
2. MODELO DE APRENDIZAJE SUPERVISADO

REFERENCES

1. Preparación de los documentos IEEE Transaction and Journals (diciembre, 2020).