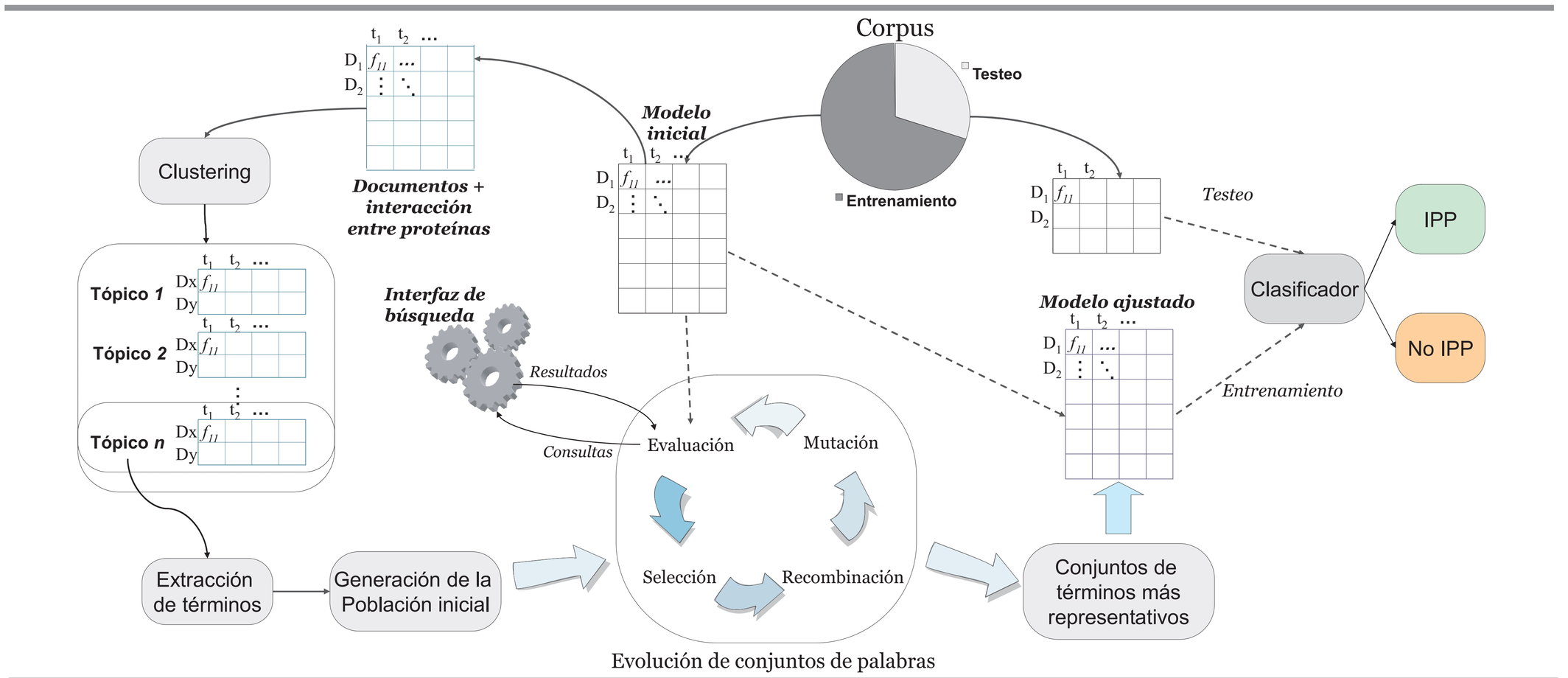
UNIVERSIDAD NACIONAL DEL SUR

DEPARTAMENTO DE CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN

Sistema híbrido para clasificación de documentos aplicado al problema bioinformático de inferencia de interacción proteína-proteína

El estudio de las interacciones existentes entre proteínas es una pieza clave en la compresión de los complejos mecanismos moleculares presentes en los procesos biológicos. La información acerca de interacciones proteína-proteína mejora el entendimiento sobre ciertas enfermedades, así como también puede ofrecer nuevas perspectivas para el desarrollo de tratamientos específicos. Sin embargo, gran parte del conocimiento sobre estas interacciones muchas veces permanece oculto en la inmensa cantidad de artículos científicos, reportes técnicos y demás trabajos científicos que día a día son publicados.

En esta línea de investigación proponemos diseñar, implementar y evaluar una infraestructura inteligente que sea capaz de inferir si un documento trata sobre interacciones entre proteínas.



Evolución de conjuntos de palabras

Arquitectura propuesta

El objetivo de la arquitectura es identificar las palabras más relevantes para documentos que tratan de interacciones entre proteínas. Este aprendizaje puede ser aprovechado para entrenar un clasificador para que sea capaz de inferir si un nuevo documento trata sobre interacción entre proteínas. Con el fin de alcanzar esta meta el sistema se vale de los siguientes componentes:

- el *Modelo inicial del corpus,* el cual se genera a partir de un corpus de documentos previamente identificados como pertenecientes al tópico de interacción entre proteínas,
- el *Módulo de clustering,* que es el encargado de identificar posibles subtópicos,
- el *Módulo evolutivo,* utilizado para inferir la relevancia de las palabras para cada tópico
- el *Modelo de ponderación ajustado,* utilizado para entrenar al módulo clasificador y
- el *Módulo clasificador,* utilizado para realizar la tarea de clasificación final.

Investigadores involucrados en esta línea

Ponzoni Ignacio.[1,3] ip@cs.uns.edu.ar
Maguitman Ana G.[2] agm@cs.uns.edu.ar

Cecchini Rocío L.[1] rlc@cs.uns.edu.ar
Lorenzetti Carlos M.[2] cml@cs.uns.edu.ar

Laboratorio de Investigación y Desarrollo en Computación Científica - LIDeCC

El LIDeCC es un grupo de científicos abocados a la tarea de diseño e implementación de soluciones computacionales efectivas para problemas científicos y tecnológicos de alta complejidad. Dentro del LIDeCC se encuentran en progreso varias líneas de investigación orientadas al estudio de herramientas para optimización y solución de sistemas no lineales. Estas herramientas están siendo utilizadas en áreas de gran interés como bioinformática e ingeniería logística.

Sub-áreas de interés en bioinformática:

- Inferencia de redes regulatorias de genes.
- Predicción de propiedades ADMET para el diseño de medicamentos.
- Inferencia de interacción proteína-proteína.

Grupo de Investigación en Recuperación de Información y Gestión de Conocimiento

Este grupo es parte del Laboratorio de Investigación y Desarrollo en Inteligencia Artificial y se dedica al estudio del tratamiento computacional de grandes volúmenes de información. Actualmente se está trabajando en el desarrollo de nuevos mecanismos de razonamiento en la Web que permitan integrar conceptos de inteligencia artificial con los algoritmos tradicionales usados por los buscadores.

[1] Laboratorio de Investigación y Desarrollo en Computación Científica
[2] Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
[3] Planta Piloto de Ingeniería Química, CONICET-UNS
Bahía Blanca - Buenos Aires - Argentina