PLANIFICACIÓN PROYECTO 4: EVALUACIÓN DE GRAFOS ALEATORIOS PARA TESTEO DE MODELOS NULOS EN REDES PPI

ELKIN NAVARRO (UNIVERSIDAD SIMÓN BOLIVAR) & CLAUDIO LÓPEZ-FERNÁNDEZ (UNIVERSIDAD DE CHILE)

1. Resumen del Problema.

Las redes de interacción proteína-proteína (PPI) son fundamentales para comprender la robustez biológica de microorganismos. Sin embargo, debido a las limitaciones experimentales para capturar todas las interacciones, estas redes se modelan frecuentemente mediante grafos aleatorios. Para distinguir las propiedades biológicas significativas del ruido estocástico, se utilizan modelos nulos: algoritmos que generan grafos análogos que preservan características de la red original.
No obstante, estos métodos suelen sufrir de una incapacidad de reproducir las topologías de la red original sin sacrificar variedad en las muestras o tiempos de cómputo. Para combatir este desbalance, Mornie et al (2025) construyen el algoritmo GRAIP, que mediante técnicas de *Simulated Annealing*, logra generar miles de muestras artificiales que preserven la proporción de subgrafos pequeños (3-5 nodos) de la red original en un tiempo razonable. No obstante, la validación de estas muestras se limita a comparaciones estadísticas mediante kernels Gaussianos sobre vectores de frecuencia, un método que puede ignorar matices topológicas globales.

La persona encargada de realizar este proyecto debe implementar un método de un método de comparación superior mediante la métrica de Wasserstein aplicada a kernels de grafos *Weisfeiler-Lehman* (WL). Esta técnica permite capturar diferencias estructurales con mayor precisión y, gracias a las recientes aproximaciones $\mathcal{O}(n)$ (Sando, 2025), es computacionalmente viable para grandes escalas. Además de la implementación, se busca demostrar analíticamente la pérdida de información (vía entropía) que ocurre al utilizar métricas de frecuencia simples frente al enfoque de transporte óptimo propuesto.

2. Entregables.

El practicante a cargo de este proyecto deberá entregar:

- Un repositorio de códigos comentados para la implementación del modelo. La idea es que pueda ser comprendido por personas ajenas al contexto matemático.
- Un pequeño informe en markdown que explique cómo usar el modelo, desde su instalación hasta su aplicación. Este sería parte del repositorio.
- Un breve borrador que incluya las demostraciones matemáticas pertinentes al proyecto, considerando posibles técnicas o ideas a desarrollar en un futuro para poder demostrar lo pedido.
- Una presentación al equipo de inmunología del CICV, mostrando resultados y explicando el desarrollo proyecto.

3. Desglose semanal

El trabajo se dividirá en 4 semanas con reuniones con el supervisor cada 1 o 2 días. Cada semana, en función del avance, con el supervisor se definirán entregables para esa semana. Las semanas debiesen seguir la siguiente línea:

- (**Semana 1: Análisis del estado del arte**): El practicante a cargo del proyecto deberá investigar sobre los algoritmos propuestos y cómo han sido implementados, notando las dificultades y/o problemas que puedan ocurrir.
- (**Semana 2: Implementación y pruebas**): El practicante a cargo del proyecto dedicará una semana a la implementación de los algoritmos y probará su desempeño con ejemplos pequeños.
- (**Semana 3: Montaje**): El practicante a cargo trabajará en la semana evaluando los algoritmos en los datasets biológicos de redes PPI utilizados por Mornie et al, haciendo uso de los computadores de alto rendimiento del CICV.
- (**Semana 4: Presentación**): El practicante a cargo del proyecto deberá comentar todos sus códigos para la entrega final y preparar un informe markdown que explique su instalación, funcionamiento y aplicación. También deberá preparar una presentación al resto del equipo de bioinformática del CICV, considerando las simplificaciones necesarias para un público ajeno al modelamiento matemático.

4. Referencias.

Para el desarrollo de este proyecto será útil consultar las siguientes referencias:

- Mornie, B., et al. (2025).*Generating random graphs with prescribed graphlet frequency bounds derived from probabilistic networks*. PLOS ONE.
- Sando, K., et al. (2025). *Tree Structure for the Categorical Wasserstein Weisfeiler-Lehman Graph Kernel*. Transactions on Machine Learning Research.