

Plan de Trabajo del Proyecto Final

Bases de datos no relacionales - 2025

1. Identificación del grupo

- **Número del grupo:** Grupo 17
- **Integrantes:** Jake Sheppard (5.803.137-4), Mathias Fuidio (4.695.065-3)

2. Familia de proyecto elegida

- **Familia 4:** Coast to coast (de la obtención al análisis, posible solución poliglota)

3. Objetivos del proyecto

Objetivo General: usar el Property Graph Model (PGM) y bases de datos de grafos para analizar diferencias en la expresión génica entre tejido tumoral y tejido sano utilizando datos transcriptómicos libremente disponibles en repositorios públicos.

- **Objetivo Específico 1**

Modelar datos ómicos de dos condiciones experimentales – tejido tumor y sano -- utilizando el PGM mediante un motor de bases de datos de grafos visto en el curso (Neo4j).

- **Objetivo Específico 2**

Estudiar las propiedades topológicas de los grafos generados: conectividad promedio, centralidad, diámetro de la red, detección de comunidades, camino más corto promedio entre dos nodos, etc. Analizar diferencias entre estas propiedades en los grafos creados a partir de los datos de cada condición.

- **Objetivo Específico 3**

Interrogar posible relevancia biológica de los patrones hallados en las etapas anteriores del análisis, a partir de revisión bibliográfica: identidad génica de los nodos *hub* y posibles evidencias de su rol en cáncer, vías metabólicas alteradas, etc.

4. Descripción general del proyecto

El cáncer es una patología de crítica importancia para la sociedad human y [una de las principales causas de muerte en el mundo](#), incluyendo en nuestro país. Los datos biológicos se prestan para ser modelados como grafos, y el análisis de redes es un campo subexplotado, pero de creciente interés en áreas como la genómica, la biología de sistemas, etc. Será interesante explorar las inferencias biológicas que se pueden extraer de bases de datos de grafos generadas a partir de datos ómicos, y enfrentarnos a los distintos problemas y desafíos que pueda conllevar.

Nuestro proyecto consiste en un análisis de redes biológicas a partir de datos de secuenciación masiva obtenidos de pacientes con cáncer. Estos datos se tratan de muestras (e.g., biopsias) obtenidas del tejido tumoral y tejido normal adyacente a partir de un mismo paciente, y que se encuentran disponibles

en repositorios públicos y bases de datos de uso libre, como son [Gene Expression Omnibus](#) (GEO) y [Sequence Read Archive](#) (SRA). Estos datos frecuentemente se depositan junto con su cuantificación en los estudios para los que fueron generados, en particular, en forma de matrices de conteo donde las filas designan los genes del genoma humano, y las columnas designan las muestras. Luego los valores en cada celda de la matriz representan un valor cuantitativo de la *expresión* de ese gen, donde expresión hace referencia al número de moléculas de ese gen detectado en la muestra.

Existen varios abordajes de redes en biología. En particular, para nuestro proyecto calcularemos matrices de correlación entre genes [usando librerías ya disponibles](#) para este tipo de análisis. Luego, generaremos grafos usando Neo4j donde las entidades, o nodos, representarán genes, que podrán ser poblados con atributos como: id en bases de datos conocidas, coordenadas genómicas, información acerca del biotipo, vías metabólicas en las que participa, si es de importancia conocida en cáncer, etc. Existirán relaciones entre aquellos genes que cumplan cierto criterio con respecto a la matriz de correlación, por ejemplo: existe una relación entre gen A y gen B si el coeficiente de correlación de Spearman entre ellos es mayor a 0.5 o menor a -0.5.

Una vez obtenidos, procesados, y transformados los datos en un formato adecuado para su importación con Neo4j, se podrá analizar las propiedades de los grafos resultantes. Existen diversos ejemplos de literatura científica para guiar nuestro análisis, y que sugieren algunos de los resultados que esperaríamos encontrar, por ejemplo, una disminución general de la conectividad en cáncer (*ver punto 6*).

5. Actividades previstas y cronograma tentativo

Completar la tabla con las actividades previstas para alcanzar los objetivos, estimando una duración tentativa para cada una (total aproximado: 60 horas).

Actividad	Breve descripción	Semana(s) prevista(s)
Recolección y validación de datos	Identificar y descargar datasets relevantes	Semana 1
Preprocesamiento de datos	Procesamiento y transformación de datos para importación con Neo4j	Semana 2
Prueba piloto	Intento inicial de modelado como grafo en Neo4j con pocos datos	Semanas 2-3
Creación de la base de datos	Generación de la base de datos en Neo4j	Semanas 3-4
Análisis de grafos	Análisis de las propiedades de los grafos generados como base de datos	Semanas 4-5
Informe y entrega final	Escritura de informe final y entrega	Semanas 5-6

6. Base teórica o artículos relacionados

1. Gene Co-expression Network Analysis (DOI: [10.1007/978-1-0716-2067-0_19](https://doi.org/10.1007/978-1-0716-2067-0_19))
2. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types (DOI: [10.1038/ncomms4231](https://doi.org/10.1038/ncomms4231))
3. Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer (DOI: [10.1038/s41598-017-05298-w](https://doi.org/10.1038/s41598-017-05298-w))
4. Applying graph database technology for analyzing perturbed co-expression networks in cancer (DOI: [10.1093/database/baaa110](https://doi.org/10.1093/database/baaa110))
5. A Guide to Conquer the Biological Network Era Using Graph Theory (DOI: [10.3389/fbioe.2020.00034](https://doi.org/10.3389/fbioe.2020.00034))
6. WGCNA: an R package for weighted correlation network analysis (DOI: [10.1186/1471-2105-9-559](https://doi.org/10.1186/1471-2105-9-559))

7. Datos y entornos

- **Juegos de datos a utilizar** (si ya se definieron):

- [GSE15471](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15471) / [GSE139038](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139038) / [GSE65485](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65485)

Descripción: Datos de expresión génica con muestras pareadas de tumores y tejido normal adyacente para cáncer de páncreas, mama y carcinoma hepatocelular (hígado) respectivamente.

(Son sets de ejemplo; se tendrá que analizar en profundidad luego cuales usamos exactamente)

Formato: matrices de conteos

- **Entornos de prueba y herramientas:**

- Python, R (paquete WGCNA), Neo4j, etc.

8. Otras observaciones

- Algunas dificultades podrían surgir en el análisis, por lo que prevemos la posibilidad de reducir la cantidad de muestras a analizar, ajustar criterios para reducir la cantidad de genes que pasan a la etapa de modelado para quedarnos con algún subconjunto reducido de ellos, etc. Anticipamos que manejar el volumen de datos puede ser uno de los mayores desafíos del proyecto.