Universidad Autónoma Metropolitana
Unidad Iztapalapa
División de Ciencias Básicas e Ingeniería
Posgrado en Ciencias (Física)

PROPIEDADES ESPECTRALES DE REDES DE REGULACIÓN GENÉTICA EN CÁNCER

Tesis

para obtener el título de

Maestro en Ciencias (Física)

presenta

Alejandro Juárez Toribio

2182800793

alex.fis.uam@gmail.com

Directores de tesis:

Dr. Leonardo Dagdug Lima
Dr. Enrique Hernández Lemus

Jurado:

Presidente
Dr. David Philip Sanders
Secretario
Dr. Leonardo Dagdug Lima
Vocal
Dr. Ricardo Marcelin Jiménez

Iztapalapa, Ciudad de México a 28 de mayo de 2021

Índice general

Dedicado a todos los jóvenes científicos que como yo, tienen la esperanza de construir un mundo en el que prevalezca la cooperación científica en beneficio del bienestar social y económico de la humanidad.

Agradecimientos

Todo el bagaje científico que la humanidad ha adquirido a los largo de miles de años, no es más que el resultado de un esfuerzo colectivo de una ingente cantidad de personas que han entregado su vida a desentrañar los secretos de la naturaleza. Las ecuaciones de Maxwell habrían tardado muchos más años en formularse si Michael Faraday no hubiera llevado a cabo sus experimentos condensados en las *Experimental Researches in Electricity*; Adolf Fick habría tenido un camino más difícil al intentar predecir la forma en que la difusión causa que la concentración cambie con el tiempo, sin los experimentos previos de Thomas Graham; todas las modernas teorías de la mecánica cuántica no hubieran sido posibles sin los estudios previos de Max Planck y Niels Bohr y esto de la mano de un largo etcétera. Por ello es que cualquier trabajo de investigación como el que se presentará a continuación debería estar acompañado siempre de los agradecimientos a esas personas que directa o indirectamente hicieron posible la materialización de ideas que podrían contribuir al desarrollo de la ciencia en nuestro país y el mundo.

Por lo anterior, quisiera agradecer en primer lugar a los dos ejes más importantes en mi vida, las personas que me construyeron intelectual y moralmente y a quienes les debo todos los logros que he cosechado hasta ahora: mi madre y mi padre.

A mis hermanos; Antonio, quien a los 14 años me dio un libro que desviaría mi sendero hacia la Física, ese libro de portada llamativa con estrellas deslumbrantes en la oscuridad del espacio tenía por título *Historia del tiempo*; a Mariana, que a pesar de ser menor que yo, siempre me ha dado útiles consejos que han cambiado positivamente mi vida.

A mis asesores, Dr. Leonardo Dagdug Lima y Dr. Enrique Hernández Lemus, quienes además de tomarse el tiempo de revisar y corregir la tesis, han compartido todos su conocimientos y me han proporcionado las herramientas necesarias para llevar a cabo el presente trabajo.

A todos mis compañeros de la UAM como a los del INMEGEN que sin tener la obligación de ayudarme, me apoyaron incondicionalmente, en especial, quiero agradecer a Iván Pompa García y Diana García Cortés, por su valioso apoyo en

temas de programación.

Finalmente, esto no podría ser posible sin el pueblo trabajador de México, que a través de CONACYT, hace posible que miles de estudiantes concluyan sus estudios de posgrado en todo el país.

Índice de figuras

Resumen

Dentro de un contexto nacional, según datos de la Secretaría de Salud del gobierno de México, el cáncer del cuello uterino es la segunda causa de muerte por cáncer en la mujer. Anualmente se estima una ocurrencia de 13,960 casos en mujeres, con una incidencia de 23.3 casos por 100,000 mujeres. En el año 2013, en el grupo específico de mujeres de 25 años y más, se registraron 3,771 defunciones con una tasa de 11.3 defunciones por 100,000 mujeres. Las entidades con mayor mortalidad por cáncer de cuello uterino son Morelos (18.6), Chiapas (17.2) y Veracruz (16.4)[10]. Así podemos apreciar que es un problema urgente que debe atenderse no solo en el mundo sino también en México. De aquí surge la importancia de contribuir en la creación de métodos matemáticos y computacionales que puedan ser útiles para dilucidar el comportamiento genético de las células cancerígenas y específicamente, como se verá en este trabajo, del cáncer cérvico-uterino.

El cáncer es una enfermedad de desregulación genética, donde las células adquieren alteraciones genéticas que provocan una señalización aberrante[9]. Estas alteraciones afectan negativamente a los programas transcripcionales y causan cambios profundos en la expresión génica. Uno de los objetivos generales de este trabajo es identificar genes que sean impulsores esenciales de los procesos celulares en el cáncer.

Para ello se usarán fudamentalmente cuatro medidas usadas en el anásis espectral de grafos: la matriz de grado, la centralidad de vector propio, la conectividad algebraica y la matriz laplaciana sin signo.

En el primer capítulo se estudiarán los aspectos más relevantes sobre la teoría de la información, aquellos que nos permitirán entender, por lo menos de manera general, el mecanismo con el que ARACNe[1](Algoritmo para la reconstrucción de redes celulares precisas)[1] procesa datos de perfiles de expresión génica y nos entrega como resultado la matriz de adyacencia de información mutua, una medida de la teoría de la información, que a grandes rasgos, mide la dependencia mutua

[1]Algoritmo que utiliza perfiles de expresión génica, diseñado para hacer desconvolución de redes reguladoras de genes .

de dos variables aleatorias, en este caso de la co-expresión de los genes asociados a muestras de tejido sano o tumoral.

En el segundo capítulo se estudia de manera más profunda a ARACNe, cubriendo no sólo los fundamentos matemáticos, sino cómo usarlo al introducir los datos de entrada, además de la primera interpretación de los datos de salida.

En el tercer capítulo, se cubren varios aspectos sobre la teoría de grafos, piedra angular de este trabajo, pues es la que permitirá entender la motivación de todo el código escrito que se utilizó para hacer el análisis de datos de la matriz de adyacencia.

En el cuarto capítulo , se realiza el análisis de datos de las matrices de adyacencia para los datos de tejido sano y tumoral, correspondientes a casos de cáncer en el útero. En ambos casos se hacen los análisis de la matriz de grado, centralidad de vector propio, conectividad algebraica y de la matriz laplaciana sin signo. Todo el código relacionado con dichos análisis de encuentra en el repositorio https://github.com/Alejandro1848/SGT-in-cancer-GRN.

En el último capítulo se realiza una discusión sobre las conclusiones dadas por el análisis de datos y las perspectivas que podrían dar pie a futuros trabajos tomando como base al actual.

Capítulo 1

Introducción

1.1. La teoría de la información en la red reguladora de genes de inferencia

Un problema importante en la biología computacional contemporánea, es el de reconstruir el mejor conjunto posible de interacciones reguladoras entre genes (una llamada **red reguladora de genes GRN**, por sus siglas en inglés) a partir de un conocimiento parcial. Las redes reguladoras de genes revelan cómo los genes trabajan juntos para llevar a cabo sus funciones biológicas. En las GRN, cada variable del conjunto de datos está representada por un **nodo** (o vértice) en el grafo. Hay un enlace (**arista**) que une dos nodos de variables si estas variables exhiben una forma particular de dependencia (la forma particular de dependencia depende explícitamente del método de inferencia elegido). Algunos genes pueden producir una proteína (u otras biomoléculas, como un microARN) que puede activar o reprimir la producción de la proteína de otro gen. Las reconstrucciones de redes de genes a partir de datos de expresión génica facilitan enormemente nuestra comprensión de los mecanismos biológicos subyacentes y brindan nuevas oportunidades para el descubrimiento de biomarcadores[1] y fármacos. En las redes de genes, un gen que tiene muchas interacciones con otros genes se denomina gen concentrador, que suele desempeñar un papel esencial en la regulación de genes y los procesos biológicos [36].

Una variedad de algoritmos arraigados en la teoría de la información y los

[1] Molécula biológica que se encuentra en la sangre, otros líquidos o tejidos del cuerpo, y cuya presencia es un signo de un proceso normal o anormal, de una afección o de una enfermedad. Un marcador biológico se utiliza a veces para determinar la respuesta del cuerpo a un tratamiento para una enfermedad o afección. También se llama biomarcador, marcador molecular y molécula distintiva.

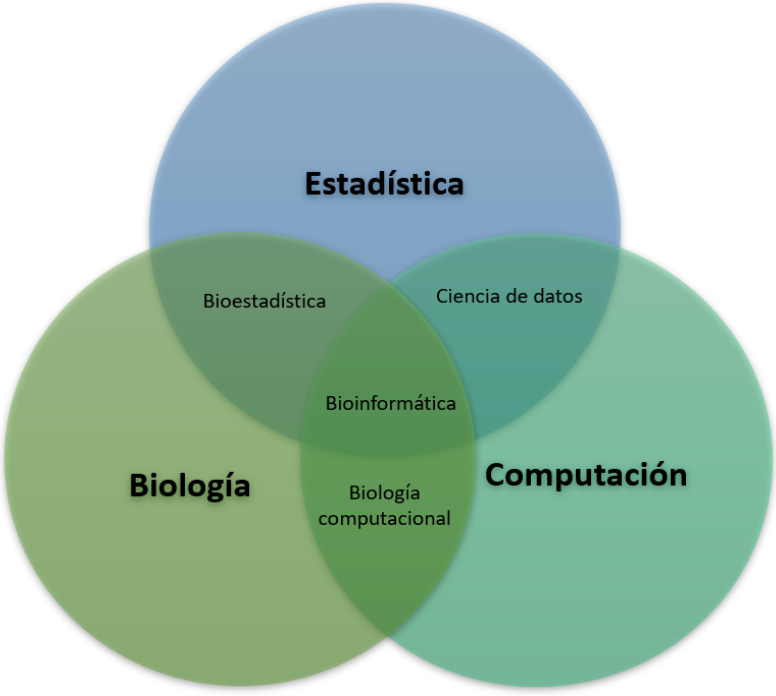

Figura 1.1: Un diagrama de Venn que muestra lo que se puede encontrar en las intersecciones de estadística, computación y biología.

métodos de máxima entropía se han desarrollado y han solucionado el problema con éxito con ciertas limitaciones[30]. La información mutua condicional [32], los campos aleatorios de Markov [31], el uso de la desigualdad en el procesamiento de datos [33], la longitud mínima de descripción [34] y la divergencia de Kullback-Liebler [35] son algunos de ellos.

La construcción de estas interacciones genéticas (GRN) se basa en la comprensión de la interacción entre miles de genes. De esta surgen problemas en el análisis de datos relacionados con la función de los genes: los procesos de medición generan señales altamente ruidosas; hay muchas más variables involucradas (número de genes e interacciones entre ellos) que las muestras experimentales. Otra fuente de complejidad es el carácter altamente no lineal de la dinámica bioquímica subyacente [4].

En el caso de la inferencia de red, consiste en representar el conjunto (en general no lineal) de dependencias estadísticas entre variables en un conjunto (que

puede ser todo el conjunto de datos de entrada o un subconjunto de características seleccionadas) por medio de un grafo. Cuando se aplica a los datos de expresión genómica (por ejemplo, de experimentos de micromatrices), la inferencia de red es capaz de realizar ingeniería inversa de la red reguladora del gen transcripcional (GRN) de la célula relacionada.

La **teoría de la información**[2] **(TI)** ha dado como resultado una poderosa base teórica para desarrollar algoritmos y técnicas computacionales para tratar tanto la selección de características como los problemas de inferencia de red aplicados a datos reales [37]. Sin embargo, existen objetivos y desafíos relacionados con la aplicación de TI al análisis genómico. Los algoritmos aplicados deben devolver modelos inteligibles, también deben confiar en un conocimiento a priori escaso, lidiar con miles de variables, detectar dependencias no lineales y todo esto a partir de decenas (o, como mucho, cientos) de muestras muy ruidosas.

Las GRN son construcciones teóricas gráficas que describen el estado integrado de una célula (o una pequeña población de células similares) bajo ciertas condiciones biológicas en un momento dado. Las GRN son medios para identificar interacciones genéticas a partir de datos experimentales mediante el uso de modelos teóricos y análisis computacional. La inferencia de dicha red de conectividad de interacción implica la solución de un **problema inverso**[3] (una deconvolución) que pretende descubrir las interacciones de las propiedades y la dinámica del comportamiento observable en forma de, por ejemplo, niveles de transcripción de ARN en un perfil característico de expresión génica. Su objetivo es proporcionar una representación bien definida de la topología de la red celular a partir de las interacciones transcripcionales, tal como lo revelan las mediciones de expresión génica que luego se tratan como muestras de una distribución de probabilidad conjunta.

Por ejemplo en las figuras 1.2-1.4 se puede observar que la diferencia en las redes reguladoras de genes entre el estado normal y el cáncer de mama maligno fue la más significativa, y la diferencia entre el cáncer de mama normal y benigno fue menos significativa que entre el cáncer de mama benigno y maligno, esto en términos de la alta concentración de grado[4] que se da en ciertas regiones del grafo en el de cáncer maligno siendo evidente en los *clusters* que se forman en él [22](figura 1.4).

En resumen, hay dos deficiencias importantes relacionadas con la selección de características y los procedimientos de inferencia de red: i) no linealidad y ii) gran número de variables. Los métodos de la TI son a menudo técnicas eficientes para

[2]Estudia los principios matemáticos que rigen la transmisión y el procesamiento de la información y se ocupa de la medición de la información y de la representación de la misma, así como también de la capacidad de los sistemas de comunicación para transmitir y procesar información.

[3]Esto quedará mejor explicado en la sección 1.3.1, cuando se estudie el papel del modelo de Ising en genética estadística.

[4]En teoría de grafos, el grado vértice es el número de aristas incidentes al mismo.

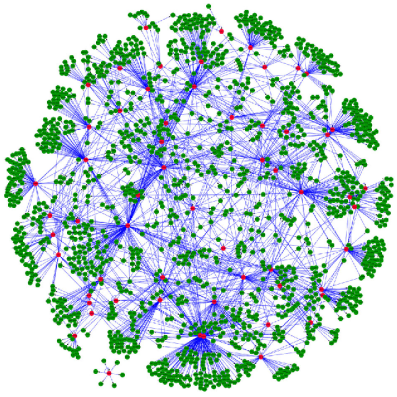

Figura 1.2: Red reguladora de genes en condiciones normales. Los nodos rojos son factores de transcripción, los nodos verdes son genes objetivo. Imagen tomada de [22].

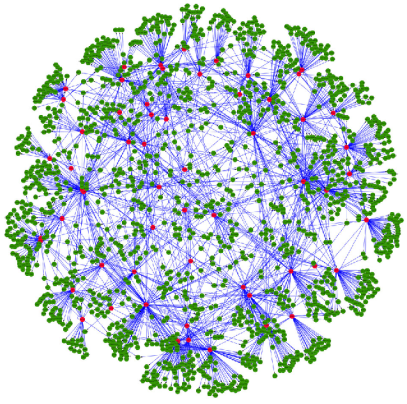

Figura 1.3: Red reguladora de genes en muestras de cáncer de mama benigno. Los nodos rojos son factores de transcripción, los nodos verdes son genes objetivo. Imagen tomada de [22].

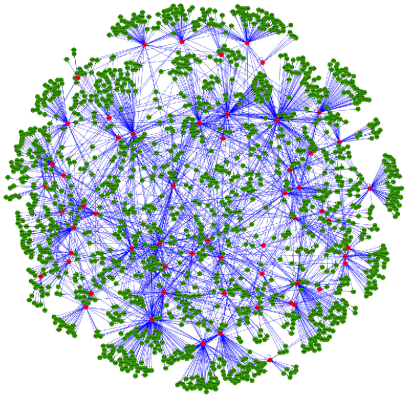

Figura 1.4: Red reguladora de genes en muestras de cáncer de mama maligno. Los nodos rojos son factores de transcripción, los nodos verdes son genes objetivo. Basándose en genes relacionados y factores de transcripción, en este estudio D.B. Chen y H.J. Yang encontraron que 8 genes desempeñan un papel importante en todo el proceso del cáncer de mama. Imagen tomada de [22].

tratar los problemas i) y ii). Se puede ver que la mayoría de estos métodos se basan en algún tipo de métrica de información mutua. **La información mutua (MI)** es una medida teórica de la dependencia de la información que es independiente del modelo y se ha utilizado para definir (y cuantificar) la relevancia, la redundancia y la interacción en conjuntos de datos tan ruidosos. La MI tiene la enorme ventaja de que captura dependencias no lineales . Finalmente, la MI es bastante rápida de computar, por lo que se puede calcular un gran número de veces en un tiempo razonable [4]. En las siguientes secciones se ahondara más al respecto de esta medida de la información.

1.2. Medidas teóricas de información y de probabilidad

Entropía de Shannon

Presentaremos aquí las nociones esenciales de TI que se utilizarán, como la entropía, la información mutua y otras medidas. Para hacerlo, denotaremos a X e Y como dos variables aleatorias discretas que tienen las siguientes características [B]:

- Alfabeto finito[5] $\mathcal{X}$ e $\mathcal{Y}$ respectivamente
- Distribución de probabilidad conjunta de masa $p(X, Y)$
- Distribución de probabilidad marginal de masa $p(X)$ y $p(Y)$

Sean además $\hat{X}$ y $\hat{Y}$ dos variables aleatorias discretas adicionales definidas sobre $\mathcal{X}$ y $\mathcal{Y}$ respectivamente, las distribuciones de probabilidad de masa asociadas serán $p(\hat{X})$ y $p(\hat{Y})$ y su distribución de probabilidad conjunta de masa será $p(\hat{X}, \hat{Y})$ y definido sobre $\mathcal{J}$, el espacio de muestreo de probabilidad conjunta, $\mathcal{J} = \mathcal{X} \times \mathcal{Y}$. Además sean $p(x) = P(X = x)$ y $p(y) = P(\hat{Y} = y)$. Ahora, para cada distribución de probabilidad discreta X es posible definir la entropía teórica de información H de dicha distribución de la siguiente manera:

$$H(X) = -\sum_{\nu} p_\nu(X) \log_2 p_\nu(X). \tag{1.1}$$

Aquí H se llama entropía de Shannon. Para esta primera parte usaremos logaritmos base 2; la entropía se medirá en bits y $p_\nu(X)$ es la densidad de probabilidad de masa para el estado ν de la variable aleatoria dada por $X = x$.

[5] Veáse la definición de alfabeto finito en el apéndice A.

Ejemplo 1: Considere una variable aleatoria que tiene una distribución uniforme sobre 32 resultados. Para identificar un resultado, necesitamos una etiqueta que adopte 32 valores diferentes. Por lo tanto, las cadenas de 5 bits[6] son suficientes como etiquetas ya que $2^5 = 32$; una de esas cadenas podría ser por ejemplo 00101. La entropía de esta variable aleatoria es:

$$H(X) = -\sum_{i=1}^{32} p(i)\log_2\ p(i) = -\sum_{i=1}^{32} \frac{1}{32}\log_2\frac{1}{32} = \log_2 32 = 5 \text{ bits}$$

que concuerda con la cantidad media de bits necesarios para describir X [12]. En este caso, todos los resultados tienen representaciones de la misma longitud. Ahora considere un ejemplo con una distribución no uniforme.

Ejemplo 2: Supongamos que tenemos una carrera con ocho caballos participando. Supongamos que las probabilidades de ganar para estos ocho son ($\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{64}$, $\frac{1}{64}$, $\frac{1}{64}$, $\frac{1}{64}$) para cada caballo, respectivamente. Podemos calcular la entropía de la carrera como:

$$H(X) = -\frac{1}{2}\log\ \frac{1}{2} - \frac{1}{4}\log\ \frac{1}{4} - \frac{1}{8}\log\ \frac{1}{8} - \frac{1}{16}\log\ \frac{1}{16} - 4\frac{1}{64}\log\ \frac{1}{64} = \ \ 2 \text{ bits}$$

intuitivamente, este resultado nos muestra la cantidad de información promedio que contienen el total de caballos . Los caballos con menor probabilidad de ganar, por ejemplo los 4 caballos con probabilidad $\frac{1}{64}$ de ganar son los que aportan mayor información y el caballo con probabilidad $\frac{1}{2}$ de ganar es el que aporta menor información respecto a la entropía. Así se necesita una media de 2 bits para recuperar cualquier valor de X.

La entropía se desarrolló originalmente para servir como una medida de la cantidad de incertidumbre asociada con el valor de X, por lo que se relaciona la previsibilidad de un resultado con la distribución de probabilidad.

Divergencia de Kullback-Leibler

La divergencia de Kullback-Leibler, $\mathcal{KL}(\cdot,:)$ es una medida no conmutativa de la diferencia entre dos distribuciones de probabilidad discretas.

$$\mathcal{KL}\left[p(Y);\tilde{p}(Y)\right] = \sum_{y\in\mathcal{Y}} p(y)\log\frac{p(y)}{\tilde{p}(y)} \tag{1.2}$$

[6] Veáse la definición de información en el apéndice A

Generalmente $p(y)$ representa la "verdadera" distribución de los datos, observaciones, o cualquier distribución teórica. La medida $\tilde{p}(y)$ generalmente representa una teoría, modelo, descripción o aproximación de $p(y)$.

La divergencia conjunta de Kullback-Leibler entre dos distribuciones de masa de probabilidad $p(X,Y)$ y $\tilde{p}(X,Y)$ está dada por:

$$\mathcal{KL}\left[p(X,Y);\tilde{p}(X,Y)\right] = \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y \mid x) \log \frac{p(x,y)}{\tilde{p}(x,y)} \tag{1.3}$$

Ejemplo 3:

Sea el alfabeto $\mathcal{H} = \{0,1\}$ y considere dos distribuciones p y q sobre $\mathcal{H}$. Sea $p(0) = 1-r$,$p(1) = r$, y sea $q(0) = 1-s$, $q(1) = s$. Entonces

$$\mathcal{KL}\left[p;q\right] = (1-r)\log\frac{1-r}{1-s} + r\ \log\frac{r}{s}$$

y

$$\mathcal{KL}\left[q;p\right] = (1-s)\log\frac{1-s}{1-r} + s\ \log\frac{s}{r}$$

si $r = s$, entonces $\mathcal{KL}\left[p;q\right] = \mathcal{KL}\left[q;p\right] = 0$. Si $r = 1/2$, $s = 1/4$ podemos calcular

$$\mathcal{KL}\left[p;q\right] = \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2}\ \log\frac{\frac{1}{2}}{\frac{1}{4}} = 1 - \frac{1}{2}\ \log 3 = 0.2075 \text{ bits}$$

mientras que

$$\mathcal{KL}\left[q;p\right] = \frac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4}\ \log\frac{\frac{1}{4}}{\frac{1}{2}} = \frac{3}{4}\log 3 - 1 = 0.1887 \text{ bits}$$

Note que $\mathcal{KL}\left[p;q\right] \neq \mathcal{KL}\left[q;p\right]$ en general. Además, como se ha descrito antes la divergencia de KL mide la distancia entre dos distrbuciones de probabilidad- una de las cuales actúa como referencia- definidas sobre la misma variable aleatoria Y. Supongamos que tenemos un canal de comunicación que nos proporciona información con símbolos de cierta probabilidad. En el caso del ejemplo anterior podemos representar nuestro canal mediante una sucesión de posibles eventos $\{0,1\}$ con sus respectivas probabilidades de ocurrencia $p(0)$, $p(1)$, $q(0)$, $q(1)$. Recuérdese además que debe cumplirse que $\sum_{i=1}^{n} p_i = 1$ y $p_i \geq 0$ (lo mismo se debe cumplir para q_i). Por otro lado observemos que la divergencia de KL se puede reescribir (usando propiedades de logaritmos) en términos de la entropía de Shanon como:

$$\mathcal{KL}\left[p;q\right] = -H(p) - \sum_i p_i \log q_i.$$

De forma inmediata podemos notar que esta llamada divergencia de KL se puede interpretar como la cantidad de información extra(en bits) [38] que uno necesita si observa un código basado en una distribución de probabilidad q cuando en realidad esperaba uno basado en una distribución de probabilidad p. En el ejemplo anterior se fijo a los valores $r = 1/2$ $s = 1/4$ por lo que las probabilidades quedaron como $p(0) = 1/2$, $p(1) = 1/2$, $q(0) = 3/4$, $q(1) = 1/4$. Para el caso de p_i, vemos que la probabilidad de que nos llegue información en forma de un 0 o 1 es la misma, mientras que para las q_i la probabilidad de que nos llegue información en forma de 0 es mayor que la que nos llegará en forma de 1 con una probabilidad 3 veces mayor de ser observada. De tal manera que se necesitarán 0.2075 bits extras para observar un código basado en la distribución q (con probabilidades de ocurrencia diferentes) cuando en realidad se esperaba la distribución p (con probabilidades de ocurrencia iguales). Por otro lado tiene sentido que para el caso contrario la cantidad extra de bits para observar un código ahora basado en p cuando en realidad se esperaba q sea menor e igual a 0.1887 bits, pues en este caso al ser la probabilidad de ocurrencia igual para ambos casos de p_i, no hay un código que sea más probable que otro de ser observado y por lo tanto la incertidumbre en la informnación se verá reducida y por consiguiente la cantidad de bits extras totales para observar el código basado en la distribución p tendría que ser menor.

De una forma similar, es posible definir la divergencia condicional de Kullback-Leibler entre $p(Y \mid X)$ y $\tilde{p}(Y \mid X)$ como sigue:

$$\mathcal{KL}\left[p(Y \mid X); \tilde{p}(Y \mid X)\right] = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log \frac{p(y \mid x)}{\tilde{p}(y \mid x)} \tag{1.4}$$

la ecuación anterior significa que una divergencia condicional de Kullback-Leibler también se puede definir como el valor esperado de la divergencia Kullback-Leibler de las funciones de masa de probabilidad condicional promediadas sobre las variables aleatorias condicionantes.

Por propiedades de los logaritmos es fácil ver que (1.2) se puede reescribir como:

$$\mathcal{KL}\left[p(Y); \tilde{p}(Y)\right] = \sum_{y \in \mathcal{Y}} p(y) \log\ p(y) - \sum_{y \in \mathcal{Y}} p(y) \log\ \tilde{p}(y) \tag{1.5}$$

Podríamos ver que el primer término en el lado derecho de la ecuación (1.5) es precisamente el negativo de la entropía $H(Y)$ como se indica en la ecuación (1.1). La entropía de Shannon depende de la distribución $p(Y)$ y, como se puede mostrar, es máxima para una distribución uniforme $u(Y)$. $H[u(Y)] = \log\ \mid \mathcal{Y} \mid$ [7]. Si reemplazamos $\tilde{p}(y)$ por $u(Y)$ en la ecuación (1.5) obtenemos:

[7] Veáse la demostración apéndice D

$$H[p(Y)] = \log \mid \mathcal{Y} \mid -\mathcal{KL}[p(Y); u(Y)]. \tag{1.6}$$

Como podemos ver, la ecuación (1.6) establece que la entropía de una variable aleatoria Y es el logaritmo del tamaño del conjunto de soporte menos la divergencia de Kullback-Leibler entre la distribución de probabilidad de Y y la distribución uniforme sobre el mismo dominio Y. Por lo tanto, cuanto más cerca esté la distribución de probabilidad de una distribución uniforme, mayor será la entropía. Por lo tanto, la entropía mide la aleatoriedad y la imprevisibilidad de una distribución.

Ahora, consideremos un par de variables aleatorias discretas (Y, X) con una Distribución de Probabilidad Conjunta (**JPD**, por sus siglas en inglés) $p(Y, X)$. Para estas variables aleatorias, la entropía conjunta $H(Y, X)$ se da en términos de JPD como:

$$H(Y, X) = -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(y, x) \log p(y, x) \tag{1.7}$$

La entropía conjunta máxima se alcanza en condiciones de independencia de las variables aleatorias Y y X, es decir, cuando el JPD está factorizado $p(Y, X) = p(Y)p(X)$; en este caso la entropía del JPD es solo la suma de sus respectivas entropías. Un teorema de desigualdad [12] podría establecerse como una cota superior para la entropía de unión:

$$H(Y, X) \leq H(Y) + H(X) \tag{1.8}$$

la igualdad solo se mantiene si X e Y son estadísticamente independientes.

Además, dada una Distribución de Probabilidad Condicional (**CPD**, por sus siglas en inglés), la entropía condicional correspondiente de Y dada X puede definirse como:

$$H(Y \mid X) = -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(y, x) \log p(y \mid x) \tag{1.9}$$

Las entropías condicionales son útiles para medir la incertidumbre de una variable aleatoria una vez que se conoce otra. Se puede probar que [12]:

$$H(Y, X) = H(X) + H(Y \mid X) \leq H(Y) + H(X), \tag{1.10}$$

justificándose la última desigualdad como consecuencia del resultado mostrado en (1.8) y (1.10) se reescribe de forma equivalente como:

$$H(Y \mid X) \leq H(Y). \tag{1.11}$$

La igualdad solo se cumple cuando X e Y son estadísticamente independientes. La expresión (1.11) es extremadamente útil en el escenario de inferencia / predicción: si Y es una variable objetivo y X es un predictor, la adición de variables solo puede disminuir la incertidumbre sobre el objetivo Y. Esto resultará casi esencial para los métodos de inferencia de GRN de teoría de la información. La reducción de la entropía por condicionamiento puede explicarse de manera bastante formal si consideramos una medida llamada información mutua, $I(Y, X)$ que es una medida simétrica (es decir, $I(Y, X) = I(X, Y)$, como puede verse de la propia definición de MI) que se escribe como:

$$I(Y, X) = H(Y) - H(Y \mid X) \quad \text{o} \quad I(X, Y) = H(X) - H(X \mid Y). \tag{1.12}$$

Si recurrimos a la definición de entropía de Shannon dada en (1.1) y la sustituimos en la ecuación (1.12) obtenemos:

$$I(Y, X) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{1.13}$$

La información mutua (**MI**) se puede escribir como el producto de la divergencia de Kullback-Leibler entre la JPD y la distribución del producto (usando la ecuación (4):

$$I(Y, X) = \mathcal{KL}\left[p(X, Y); p(X)p(Y)\right]. \tag{1.14}$$

La información mutua también viene dada por la divergencia de Kullback-Leibler entre la distribución marginal $p(X)$ y la distribución condicional $p(X|Y)$

$$I(Y, X) = \mathcal{KL}\left[p(X \mid Y); p(X)\right]. \tag{1.15}$$

Adicionalmente se puede extender la medida de la información mutua para tres variables aleatorias como se verá a continuación. Dadas tres variables aleatorias X, Y y Z, la información mutua condicional es una medida de la reducción en la incertidumbre de X debido al conocimiento de Y cuando se da Z. En otras palabras,

$$I(X; Y \mid Z) = -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} p(x, y, z) \log \frac{p(x, y \mid z)}{p(x \mid z)p(y \mid z)} \tag{1.16}$$

Ejemplo 4: Supongamos que se tiene la siguiente tabla de datos experimentales y si quisieramos calcular la información mutua condicional a este conjunto de datos, bastaría calcular las entropías de Shannon $H(X, Z)$, $H(Y, Z)$

X	0	1	1	1	1	1	1	0	0	0	0
Y	0	0	0	1	1	0	0	1	1	1	0
Z	1	1	0	0	0	1	1	0	0	0	1

Figura 1.5: Ejemplo de cálculo de Información mutua condicional [18]

,$H(Z)$, $H(X,Y,Z)$ ya que se pude mostar que la CMI puede expresarse en términos de la entropía de Shannon como [18]

$$I(X;Y \mid Z) = H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z)$$

así, usando la figura anterior podemos calcular

$$H(Z) = H(0) + H(1) = -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10}$$

$$0.9709$$

esto porque se observa que hay un total de 6 celdas con el número 1 y 4 celdas con el número 0, además el número total de datos para H(X) es de 10 por lo que al usar la definición de probabilidad clásica se obtiene el resultado antes mencionado. Análogamente se calcula:

$$H(X,Z) = H(0,0) + H(0,1) + H(1,1) + H(1,0) =$$

$$-\frac{3}{10}\log\frac{3}{10} - \frac{1}{10}\log\frac{1}{10} - \frac{3}{10}\log\frac{3}{10} - \frac{3}{10}\log\frac{3}{10}$$

$$= 1.8954.$$

En el anterior cálculo se ha usado la noción de probabilidad conjunta , pues se han calculado entropías conjuntas explicadas con anterioridad en la introducción. También se calcula:

$$H(Y,Z) = H(0,0) + H(0,1) + H(1,1) + H(1,0) =$$

$$-\frac{2}{10}\log\frac{2}{10} - \frac{3}{10}\log\frac{3}{10} - \frac{1}{10}\log\frac{1}{10} - \frac{4}{10}\log\frac{4}{10}$$

$$= 1.8464$$

y finalmente se calcula $H(X,Y,Z)$. Para este caso es necesario , usando teoría básica de combinatoria, calcular todas las posibles combinaciones de la triada (X,Y,Z) con $X,Y=0,1$, así vemos que las entropías de Shannon posibles serán 8 en total para $H(X,Y,Z)$, esto es :

$$H(X,Y,Z) = H(0,0,0) + H(0,1,0) + H(0,0,1) + H(0,1,1) + H(1,0,0)$$

$$+H(1,1,0) + H(1,0,1) + H(1,1,1)$$

$$= -\frac{1}{10}\log\frac{1}{10} - \frac{2}{10}\log\frac{2}{10} - \frac{1}{10}\log\frac{1}{10} - \frac{1}{10}\log\frac{1}{10} - \frac{1}{10}\log\frac{1}{10}$$

$$-\frac{2}{10}\log\frac{2}{10} - \frac{2}{10}\log\frac{2}{10} + 0$$

$$= 2.7219$$

observamos que para los datos de la tabla mostrada anteriormente no existe una entropía de Shannon asociada a la combinación H(1,1,1)

Así al sumar todas las contribuciones se obtiene :

$$I(X;Y \mid Z) = H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z)$$

$$= 1.8954 + 1.8464 - 0.9709 - 2.7219$$

$$= 0.049$$

1.3. Métodos de Inferencia de Red Reguladora

Uno de los campos más activos en biología cuantitativa es la inferencia de redes de interacción biológica (v.gr. proteínas o redes reguladoras de genes) a partir de datos de alto rendimiento, como los microarrays de expresión o en los tiempos más modernos con **RNA-seq** En estos problemas, uno mide los valores (simultáneos o en serie) de las expresiones de genes en diferentes condiciones y los trata como muestras de una **distribución de probabilidad conjunta** (**JPD** por sus siglas en inglés). El objetivo es inferir la red genética basada en dependencias estadísticas en esta JPD [13].

La deconvolución de un GRN podría basarse en una optimización de entropía máxima del JPD de las interacciones gen-gen tal como se indica en la expresión génica, los datos experimentales podrían implementarse de la siguiente manera . Para poder entender de mejor manera el modelo que se explicará a continuación es necesario recurrir al conocido Modelo de Ising usado en la física estadística que resultó interesante en su papel en el desarrollo histórico de la comprensión del ferromagnetismo y de las transiciones de fase, en cuyo proceso representó un papel fundamental. Hoy se sabe que el modelo de Ising y generalizaciones del mismo sirven para explicar una variedad de fenómenos, no solo físicos sino también de diversas áreas de la biología. Así, es importante primero ver la relación que existe entre el modelo de Ising y la genética estadística.

1.3.1. La relación entre el modelo de Ising en física y la genética estadística

El modelo de Ising en física

Según el modelo de Ising unidimensional, se considera una cadena de N partículas, las cuales están limitadas a interactuar solo con sus vecinos más cercanos. Un campo magnético puntual externo actúa sobre una partícula dada. En una posición particular (i) a lo largo de la cadena, cada partícula tiene un valor de espín de $+1$ o -1 (representando un alineamiento paralelo o antiparalelo respecto al campo). Denotamos este valor de espín como σ_i. La configuración de la cadena completa $\{\sigma_i\}$ se determina especificando σ_i en todas las posiciones i. Adicionalmente se imponen condiciones de frontera de tipo periódicas dada por $\sigma_{i+N} = \sigma_i$. Así, la energía(hamiltoniano) de la configuración de N partículas se puede expresar como:

$$H\left[\{\sigma_i\}\right] = -J\sum_{i=1}^{N}\sigma_i\sigma_{i+1} - B\sum_{i=1}^{N}\sigma_i$$

$$= -\sum_{i=1}^{N}\left[J\sigma_i\sigma_{i+1} + \frac{B}{2}(\sigma_i + \sigma_{i+1})\right] \tag{1.17}$$

donde donde J es la fuerza de acoplamiento entre los espines σ_{i+1} y σ_i y B es el campo magnético local

En esta última expresión ya se han usado las condiciones periódicas para reescribir la ecuación. De acuerdo a la mecánica estadística, la probabilidad de observar una configuración específica está dada por

$$P\left[\{\sigma_i\}\right] = \frac{\exp(-\frac{H[\{\sigma_i\}]}{\kappa T})}{Z} \tag{1.18}$$

donde κ es la contante de Boltzmann, T es la temperatura de la cadena, y Z es la función de partición que es igual a la suma de las energías de todas las posibles configuraciones:

$$Z_N = \sum_{\{\sigma_i\}} \exp(-H\left[\{\sigma_i\}\right]) \tag{1.19}$$

Un campo magnético externo aplicado hará que las partículas ferromagnéticas se alineen en la dirección del campo, con las partículas vecinas alineadas en la misma dirección. El efecto de la temperatura es introducir aleatoriedad adicional en el sistema. El grado de magnetización de la cadena está determinado por la fuerza del campo y la energía de acoplamiento, en relación con la energía térmica. [14]

El modelo de Ising adaptado a datos genéticos

Para las parejas de hermanos afectados, la estadística principal de interés es el número de alelos compartidos **IBD**. Para un marcador locus determinado, cada padre de un **ASP** transfiere ($x = 1$) o no ($x = -1$) el mismo alelo a dos descendientes.

Por lo tanto, los datos pueden representarse en forma de una matriz simple de $n.m$ con filas correspondientes a n padres y columnas correspondientes a m marcadores loci [8]. Para un padre dado i, el estado de compartir IBD para cada marcador tipado en un cromosoma corresponde a la fila i-ésima, y es análogo a una configuración de partículas m en el modelo de Ising. Se supone que los datos para cada padre representan un sorteo independiente de una distribución de probabilidad subyacente.

Ahora nos concentraremos en las ecuaciones (1.16) y (1.17). El primer término en la ecuación (1.16) J, representa el hecho de que si se comparte un marcador de IBD, los marcadores vecinos tienen una mayor probabilidad de ser también compartidos debido a la vinculación genética.

El segundo término, B, es el parámetro real de interés para los estudios genéticos, ya que un campo magnético local B es análogo a un efecto genético que causa un aumento en el intercambio de IBD en el locus i-ésimo. En los hermanos afectados, el aumento en el intercambio de IBD es el resultado de la proximidad de un gen causante de enfermedades. Para una enfermedad mendeliana simple, habrá un aparente 'campo' fuerte cerca del gen de la enfermedad. Para una enfermedad compleja, puede haber múltiples genes que influyen en la enfermedad con fuerzas variables

[8] Consultar definición en el apéndice A

Así, de manera análoga a la ecuación (1.17) la distribución de probabilidad conjunta para la expresión estacionaria de todos los genes, $P(\{g_i\}), i = 1, ..., N$ puede escribirse como sigue :

$$P(\{g_i\}) = \frac{1}{Z}\exp^{H_{gen}} \tag{1.20}$$

$$H_{gen} = \left[-\sum_{i}^{N} \Phi_i(g_i) - \sum_{i,j}^{N} \Phi_{i,j}(g_i, g_j) - \sum_{i,j,k}^{N} \Phi_{i,j,k}(g_i, g_j, g_k) - ... \right] \tag{1.21}$$

aquí N es el número de genes, Z es un factor de normalización (la función de partición), las Φ son los potenciales de interacción. Un procedimiento de truncamiento en la ecuación (1.20) se utiliza para definir una H_p hamiltoniana aproximada que tiene como objetivo describir las propiedades estadísticas del sistema. Un conjunto de variables (genes)Ω , interactúa entre sí, si y solo si el potencial entre dicho conjunto de variables es distinto de cero. La contribución relativa de Φ_Ω se toma como proporcional a la fuerza de la interacción entre este conjunto. La ecuación (1.20) no define los potenciales de manera única, por lo tanto, se deben proporcionar restricciones adicionales para evitar la ambigüedad.

Un enfoque habitual para hacerlo es especificar las Φ utilizando aproximaciones de máxima entropía (**MaxEnt**) coherentes con la información disponible sobre el sistema en forma de marginales. La teoría de la información proporciona un conjunto de criterios útiles para configurar funciones de distribución de probabilidad (**PDF**) sobre la base de un conocimiento parcial.

La estimación MaxEnt de un PDF es la estimación menos sesgada posible, dada la información. No es posible restringir el sistema a través de la especificación de todas las posibles $N-$ vías de potenciales cuando N es grande, por lo tanto, uno tiene que aproximarse a la estructura de interacción. De acuerdo con la literatura genómica actual, los tamaños de muestra de orden 10^2 (el tamaño máximo habitual disponible en la mayoría de los estudios actuales) son generalmente suficientes para estimar los marginales de 2 vías, mientras que los marginales de 3 vías (por ejemplo, las interacciones de los tripletes $i, j, k)$ g_i, g_j, g_k)) requieren muestras aproximadamente de un orden de magnitud más , un tamaño de muestra inalcanzable en las circunstancias actuales. Siendo este el caso, uno normalmente se enfrenta a un hamiltoniano de dos vías de la forma:

$$H^{aprox} = -\sum_{i}^{N} \Phi_i(g_i) - \sum_{i,j}^{N} \Phi_{i,j}(g_i, g_j). \tag{1.22}$$

Si consideramos un hamiltoniano de interacción bidireccional, se dice que todos los pares de genes i, j para los cuales $\Phi_{i,j} = 0$ no interactúan. Esto es cierto para

los genes que son estadísticamente independientes, $P(g_i, g_j) \approx P(g_i)P(g_j)$, pero también es válido para los genes que no tienen una interacción directa pero están conectados a través de otros genes, es decir, $\Phi_{i,j} = 0$ pero $P(g_i, g_j) \neq P(g_i)P(g_j)$.

1.3.2. Métodos teóricos de la información

Información mutua

Una medida teórica de la información que se ha utilizado con éxito para inferir interacciones de 2 vías en los GRN es la información mutua (MI). MI para un par de variables aleatorias α, y β se define como $I(\alpha, \beta) = H(\alpha) + H(\beta) - H(\alpha, \beta)$. Aquí H es la entropía teórica de la información (entropía de Shannon),$H(x) = -\langle \log p(x_i) \rangle = -\sum_i p(x_i) \log p(x_i)$. MI mide el grado de dependencia estadística entre dos variables aleatorias. De la definición se puede ver que $I(\alpha, \beta) = 0$ si y solo si α y β son estadísticamente independientes. La estimación del IM entre los perfiles de expresión génica en configuraciones experimentales de alto rendimiento típicas de la investigación actual en este campo es un desafío teórico y computacional de magnitud considerable. Una posible aproximación es el uso de estimadores. Bajo una aproximación del kernel gaussiano [15], [16], la JPD de una medición de 2 vías $\overrightarrow{X}_i = (x_i, y_i), \;\; i = 1, 2, ..., M$ está dada por:

$$f(\overrightarrow{X}) = \frac{1}{M} \sum_i \frac{G\left[\left(h^{-1} \mid \overrightarrow{X} - \overrightarrow{X}_i \mid\right)\right]}{h^2} \tag{1.23}$$

G es la densidad normal estándar bivariable y h es el ancho del kernel asociado. La información mutua podría ser evaluada como sigue:

$$I(\{x_i\}, \{y_i\}) = \frac{1}{M} \sum_i \log \frac{f(x_i, y_i)}{f(x_i) f(y_i)} \tag{1.24}$$

por lo tanto, se dice que dos genes con perfiles de expresión g_i y g_j para los cuales $I(g_i, g_j) \neq 0$ interactúan entre sí con una fuerza $I(g_i, g_j) \approx \Phi(g_i, g_j)$, mientras que dos genes para los cuales $I(g_i, g_j)$ es cero y se declaran no interactuando directamente dentro de las aproximaciones dadas. Dado que la MI es una reparametrización invariante , normalmente se calcula la información mutua normalizada. En este caso $I(g_i, g_j) \in [0, 1], \forall i, j$.

Campos aleatorios de Markov

Un campo aleatorio de Markov es un proceso aleatorio n-dimensional definido en una red discreta. Por lo general, la red es una cuadrícula bidimensional regular

en el plano, ya sea finita o infinita. Suponiendo que X_n es una cadena de Markov que toma valores en un conjunto finito,

$$P(X_n = x_n \mid X_k = x_k; k \neq n) =$$

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}; X_{n+1} = x_{n+1}) \tag{1.25}$$

Por lo tanto, la distribución condicional completa de X_n depende solo de los vecinos X_{n-1} y X_{n+1}: en el ajuste 2-D, si $(S = 1; 2; ...; N) \times (S = 1; 2; ...; N)$ este será el conjunto de puntos N^2 llamados sitios o estados.

Los modelos de campo aleatorio de Markov (**MRF**) se han aplicado en varios escenarios dentro de la configuración de biología molecular computacional, en general, análisis basados en red para datos genómicos. En el caso de los métodos de ingeniería inversa para la inferencia de red, un modelo de MRF podría establecerse de la siguiente manera:
Una asignación de estado arbitrario para un conjunto de genes se indicará por $x = (x_1, x_2, ..., x_p)$, aquí x_i es el estado de expresión (expresado de manera igual o diferencial, 0 o 1 respectivamente) del gen i, sea x^* el verdadero pero desconocido estado de expresión génica. Podemos interpretar esto como una realización particular de un vector aleatorio $X = (X_1, X_2, ..., X_p)$ donde X_i asigna un estado de expresión al gen i. Dejemos que y_i represente el nivel de expresión de mRNA[9] observado experimentalmente de los genes i y y el vector correspondiente, que aquí se interpreta como una realización particular de un vector aleatorio $Y = (Y_1, Y_2, ..., Y_n)$. Y_i en sí es un vector

$$y_i = (y_{i,1}, y_{i,2}, ..., y_{i,m}, y_{i,m+1}, y_{i,m+1}, ..., y_{i,m+n})$$

Este vector contiene m réplicas en una condición y n réplicas en la otra condición. La distribución conjunta de Y podría darse en términos de un MRF, para anotar esta probabilidad conjunta necesitamos conocer la dependencia/independencia condicional. La teoría de la información podría ser útil para determinar a partir de las distribuciones tales dependencias condicionales.
Para estudiar la robustez funcional en los GRN, Emmert-Streib y Dehmer [17] modelaron el procesamiento de la información dentro de la red como una cadena de Markov de primer orden y estudiaron la influencia de las perturbaciones de un solo gen en la comunicación global y asintótica entre los genes. Las diferencias se contabilizaron mediante una medida teórica de la información que permitió predecir los genes que son frágiles respecto a la eliminación de un solo gen. La medida

[9] Es el ácido ribonucleico que transfiere el código genético procedente del ADN del núcleo celular a un ribosoma en el citoplasma, es decir, el que determina el orden en que se unirán los aminoácidos de una proteína y actúa como plantilla o patrón para la síntesis de dicha proteína

teórica de la información utilizada para capturar el comportamiento asintótico del procesamiento de la información evalúa la desviación del estado no perturbado (o normal (n)) del estado perturbado (p) causado por la perturbación del gen k. Se utilizó la divergencia de entropía relativa o Kullback-Leibler (KL) para cuantificar esta desviación:

$$\mathcal{KL}_{i,k} = \mathcal{KL}\left[p_{i,k}^{p,\infty}; p_i^{n,\infty}\right] = \sum_m p_{i,k}^{p,\infty}(m) \log \frac{p_{i,k}^{p,\infty}(m)}{p_i^{n,\infty}(m)} \tag{1.26}$$

En la ecuación 22 las distribuciones estacionarias $p_{i,k}^{p,\infty}$ y $p_i^{n,\infty}$ están dadas por :

$$p_{i,k}^{p,\infty} = \lim_{t\to\infty} T^t p_i^0 \tag{1.27}$$

$$p_i^{n,\infty} = \lim_{t\to\infty} T_k^t p_i^0 \tag{1.28}$$

La cadena de Markov dada por T_k corresponde al proceso obtenido al perturbar el gen k en la red. Por medio de este modelo de cadena de Markov complementado con una medida de KL teórica de la información, Emmert-Streib y Dehmer pudieron estudiar el comportamiento asintótico de la red reguladora transcripcional de la levadura con respecto a la propagación de la información bajo la influencia de perturbaciones de un solo gen. Por lo tanto, no solo las propiedades de red estáticas (como la estructura) de las redes de regulación transcripcional, sino también las características dinámicas (como la solidez) se podrían analizar desde el punto de vista de TI. El estudio concluye que los genes eliminados destruyen algunas vías de comunicación y, por lo tanto, aún pueden tener un fuerte impacto en el procesamiento de la información dentro de la célula. Parece razonable suponer que cuanto más lejos esté el gen eliminado del gen de inicio cuanto menor será el impacto. Esta es una fuerte evidencia de que el procesamiento de la información a nivel de sistemas depende fundamentalmente del procesamiento de la información en un entorno local del gen que envía la información.

Desde la perspectiva del procesamiento de información, la conexión entre el cambio de información asintótica y la estructura de la red local representada por sus grados es interesante porque indica que un subgrafo local puede ser suficiente para estudiar el procesamiento de la información en la red general. Este hallazgo parece verdaderamente interesante porque permitiría reducir la complejidad computacional que surge al estudiar genomas grandes en la escala de un sistema. Desde el punto de vista del procesamiento de la información, se demostró que la conexión entre los cambios asintóticos de la información y la estructura de la red local en un subgrafo local puede ser suficiente para estudiar el procesamiento de la información en la red general.

Desigualdad en el procesamiento de datos

En ingeniería y teoría de la información, la desigualdad en el procesamiento de datos (**DPI**) es un teorema simple pero útil que establece que no importa qué proceso realice con algunos datos, no puede obtener más información (en el sentido de Shannon) del conjunto de datos que estaba allí al principio . En cierto sentido, proporciona un límite sobre cuánto se puede lograr con el procesamiento de la señal.

Más cuantitativamente, considere dos variables aleatorias, X e Y, cuya información mutua es $I(X, Y)$. Ahora considere una tercera variable aleatoria, Z, que es una función (probabilística) de Y solamente. El DPI establece que Z no puede tener más información sobre X que Y tiene sobre X; es decir $I(X; Z) \leq I(X; Y)$. Esta inecuación que nuevamente es una propiedad que la información de Shannon debería tener, se puede probar, por lo tanto,

$$I(X; Z) = H(X) - H(X \mid Z) \leq H(X) - H(X \mid Y, Z)$$

$$= H(X) - H(X \mid Y) = I(X; Z) \tag{1.29}$$

Demostración:

De lo anterior vemos que bastará demostrar que

$$-H(X \mid Z) \leq -H(X \mid Y, Z)$$

o de forma equivalente

$$H(X \mid Y, Z) \leq H(X \mid Z) \tag{1.30}$$

pero del resultado de la ecuación (11) en [4] se sabe que

$$H(Y \mid X) \leq H(Y) \tag{1.31}$$

de aquí se sigue que:

$$H(X \mid Z) \leq H(X) \tag{1.32}$$

y por transitividad de (1.29) con (1.31) se obtiene

$$H(X \mid Y, Z) \leq H(X) \tag{1.33}$$

y (1.32) tiene la misma forma de (1.30) por lo que si X es una variable objetivo y Y, Z un predictor, la adición de variables solo puede disminuir la incertidumbre

sobre el objetivo X. Con esto se explica la primera desigualdad en la expresión (1.28), y en cuanto a la última igualdad, se sigue porque

$$P_{X|YZ}(x \mid y, z) = P_{X|Y}(x \mid y)$$

como se puede verificar en [4].

■

Este mismo principio es aplicable al procesamiento de señales biológicas como el presente en las redes de regulación genética. El DPI es útil para cuantificar eficientemente las dependencias entre un gran número de genes. Un algoritmo desarrollado llamado **ARACNe** [1] desarrollado en el lenguaje C++ , elimina aquellas dependencias estadísticas que pueden ser de naturaleza indirecta. En las siguientes secciones se explicará con más detalle sus características, pues fue una herramienta crucial para obtener resultados importantes en este trabajo.

Capítulo 2

ARACNe

2.1. Antecedentes

Los fenotipos celulares están determinados por la actividad dinámica de grandes redes de genes co-regulados. Así, el análisis de los mecanismos de selección fenotípica requiere una explicación de las funciones de los genes individuales dentro del contexto de las redes en las que operan. Debido a que la expresión génica está regulada por proteínas, que son en sí mismas productos génicos, las relaciones estadísticas entre los niveles de abundancia de ARNm de genes, aunque no son directamente proporcionales a las concentraciones de proteínas activadas, deberían proporcionar una vía para descubrir mecanismos reguladores de genes. En consecuencia, el advenimiento de las tecnologías de microarrays de alto rendimiento (y la secuenciación de genoma más recientemente) para medir simultáneamente los niveles de abundancia de ARNm en todo un genoma ha generado mucha investigación destinada a utilizar estos datos para construir modelos conceptuales de red de genes para describir de manera concisa las influencias reguladoras que los genes ejercen entre sí [1].

La agrupación de perfiles de expresión génica[1] en todo el genoma proporciona un primer paso importante hacia este objetivo al agrupar genes que exhiben respuestas transcripcionales similares a diversas afecciones celulares y, por lo tanto, es probable que estén involucrados en procesos celulares similares. Sin embargo, la organización de genes en grupos corregulados proporciona una representación muy aproximada de la red celular. En particular, no puede separar las interacciones estadísticas que son **irreducibles** (es decir, directas) de las que surgen de las cascadas de interacciones transcripcionales que correlacionan la expresión de

[1] Consultar definición en el apéndice A

muchos genes que no interactúan. En términos más generales, como se aprecia en la física estadística, el **orden de largo alcance** (es decir, una alta correlación entre las variables que interactúan de forma no directa) puede resultar fácilmente de interacciones de corto alcance . Por lo tanto, las correlaciones, o cualquier otra medida de dependencia local, no pueden utilizarse como la única herramienta para la reconstrucción de redes de interacción sin supuestos adicionales [2].

En los últimos años, han surgido varios enfoques sofisticados para la ingeniería inversa de las redes celulares (también llamada deconvolución) a partir de los datos de expresión génica . Su objetivo es producir una representación de alta fidelidad de la topología de la red celular como un grafo, donde los genes se representan como vértices y están conectados por bordes que representan interacciones reguladoras directas. Los criterios para definir un borde, así como su interpretación biológica, siguen siendo imprecisos y varían entre las aplicaciones [1].

ARACNe [1](Algoritmo para la reconstrucción de redes celulares precisas), es un algoritmo teórico de información novedoso para la ingeniería inversa de redes transcripcionales a partir de datos de RNA-seq. ARACNe primero identifica la corregulación gen-gen estadísticamente significativa mediante información mutua, una medida de parentesco basada en la información teórica. Luego, elimina las relaciones indirectas, en las que dos genes se regulan conjuntamente a través de uno o más intermediarios, aplicando un elemento básico bien conocido de la teoría de la transmisión de datos, la 'desigualdad en el procesamiento de datos' (DPI). Por lo tanto, las relaciones incluidas en la red reconstruida final tienen una alta probabilidad de representar interacciones reguladoras directas o interacciones mediadas por modificadores postranscripcionales que son indetectables en los perfiles de expresión génica.

El objetivo de ARACNE no es recuperar todas las interacciones transcripcionales en una red genética, sino más bien recuperar algunas interacciones transcripcionales con alta confianza. El algoritmo puede aplicarse a redes arbitrariamente complejas de interacciones transcripcionales [40].

Las tecnologías de alto rendimiento han permitido la medición simultánea de las concentraciones de miles de especies moleculares en un sistema biológico, como mRNA, microRNA, proteínas y metabolitos. Como la dinámica de cada especie molecular está influenciada por la concentración de varias otras especies, se han desarrollado varios enfoques estadísticos para inferir relaciones funcionales dentro de grandes conjuntos de variables bioquímicas basadas en las correlaciones entre las modalidades de datos disponibles. En particular, los **perfiles de expresión génica**[2], que representan las concentraciones promedio de ARNm en una población celular, han surgido entre las mediciones de genoma más fácilmente disponibles para una variedad de organismos [42].

[2]Consultar definición en el apéndice A

Los métodos computacionales utilizados para inferir las interacciones bioquímicas a partir de los datos del perfil de expresión génica prometen dilucidar los mecanismos funcionales subyacentes a los procesos celulares, así como la identificación de objetivos moleculares de compuestos farmacológicos [42].

2.1.1. Aspectos matemáticos relevantes del algoritmo

Información mutua

La información mutua **MI** para un par de variables aleatorias x y y se define como:

$$I(x, y) = H(x) + H(y) - H(x, y), \tag{2.1}$$

donde $H(t)$ es la ya antes llamada entropía de Shannon dada por la ecuación (1.1). La informaciôm mutua mide el grado de dependencia estadística entre dos variables aleatorias. Sin embargo, si bien los coeficientes de correlación no son invariantes en reparametrizaciones y pueden ser cero incluso para variables manifiestamente dependientes, la MI es invariante en reparametrizaciones y es distinta de cero si y sólo si existe algún tipo de dependencia estadística [12].

Desigualdad en el procesamiento de datos

La desigualdad en el procesamiento de datos (**DPI**)[1] establece que si los genes g_1 y g_3 interactúan solo a través de un tercer gen, g_2 (i.e, si la red de interacción es $g_1 \longleftrightarrow ... \longleftrightarrow g_2 \longleftrightarrow ... \longleftrightarrow ... \longleftrightarrow g_3$) y no existe un camino alternativo entre g_1 y g_3, entonces:

$$I(g_1, g_3) \leq \text{mín}\,[I(g_1, g_2); I(g_2, g_3)] \tag{2.2}$$

Ejemplo a)

En la figura 2.1 podemos ver 4 genes conectados por medio de una cadena lineal.

Aunque los seis pares de genes (ya que hay 6 posibles combinaciones de interacción de los genes g_1, g_2, g_3, g_4) probablemente habrán determinado el valor información mutua, el DPI inferirá la ruta más probable de flujo de información. Por ejemplo, $\mathbf{g_1} \longleftrightarrow \mathbf{g_3}$ será eliminado porque $I(g_1, g_2) > I(g_1, g_3)$ y $I(g_2, g_3) > I(g_1, g_3)$. $\mathbf{g_2} \longleftrightarrow \mathbf{g_4}$ será eliminado porque $I(g_2, g_3) > I(g_2, g_4)$ y $I(g_3, g_4) > I(g_2, g_4)$. $\mathbf{g_1} \longleftrightarrow \mathbf{g_4}$ será eliminado de dos formas: primero, porque $I(g_1, g_2) > I(g_1, g_4)$ y $I(g_2, g_4) > I(g_1, g_4)$ y porque $I(g_1, g_3) > I(g_1, g_4)$ y $I(g_3, g_4) > I(g_1, g_4)$.

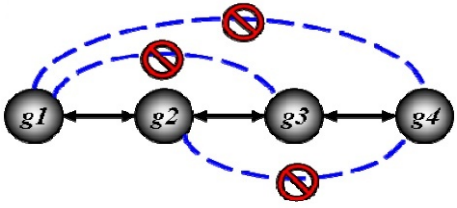

Figura 2.1: Ejemplo de cálculo DPI en una cadena lineal de 4 genes. Imagen tomada de [1].

Ejemplo b)

Ahora consideremos la imagen de la figura 2.2. Si las interacciones subyacentes forman un **árbol**[3] (y el MI puede medirse sin errores), ARACNE reconstruirá la red exactamente eliminando todas las interacciones candidatas falsas (líneas azules discontinuas) y conservando todas las interacciones verdaderas (líneas negras continuas).

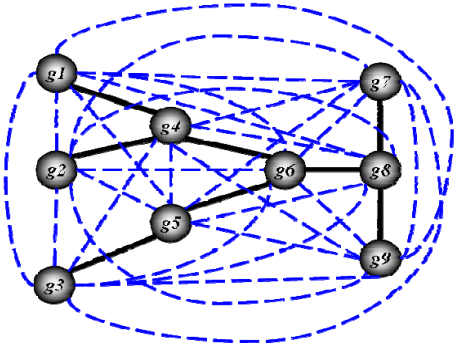

Figura 2.2: Ejemplo de cálculo DPI en una cadena lineal de 4 genes. Imagen tomada de [1].

Por lo tanto el menor de las tres MI's puede provenir solo de interacciones indirectas, y la comparación con el DPI puede identificar aquellos pares de genes para los cuales $\phi_{ij} = 0$ aunque $P(g_i, g_j) \neq P(g_i)P(g_j)$. En consecuencia, ARACNe comienza con un gráfico de red donde cada $I_{ij} > I_0$ [4] está representado por un

[3] Consultar definición en el apéndice A

[4] Valor umbral de información mutua para un valor p específico p_0 en la hipótesis nula de dos genes independientes[1]

gráfico (ij). El algoritmo luego examina cada triplete del gen para el cual las tres IM's son mayores que I_0 y elimina el borde con el valor más pequeño. Cada triplete se analiza independientemente de si sus bordes se han marcado para su eliminación mediante aplicaciones DPI anteriores a diferentes tripletes. Por tanto, la red reconstruida por el algoritmo es independiente del orden en que se examinan los tripletes.

Dado que este enfoque se centra solo en la reconstrucción de redes de interacción por pares, un par de genes mutuamente independientes, $I_{ij} < I_0$, nunca estarán conectados por un borde. Por lo tanto, las interacciones representadas por potenciales de orden superior para los cuales los potenciales pares correspondientes son cero no se recuperarán.

En el apéndice A de [1] se enuncian y demuestran 3 teoremas que especifican las condiciones bajo las cuales ARACNe reconstruirá la red exactamente.

2.1.2. Algoritmo

ARACNE genera una red transcripcional en dos pasos computacionales. Primero, los pares de genes que muestran respuestas transcripcionales correlacionadas se identifican midiendo el MI entre sus perfiles de expresión de ARNm. Podría decirse que MI es la mejor medida de la correlación estadística en un entorno no lineal [41]. Los elementos clave en este paso son la determinación de los parámetros para el cálculo del MI (es decir, el ancho del núcleo del estimador) y del umbral de MI para la independencia estadística. En el segundo paso, ARACNE elimina aquellas dependencias estadísticas que podrían ser de naturaleza indirecta, como entre dos genes que están separados por pasos intermedios en una cascada transcripcional. Dichos genes probablemente tendrán perfiles de expresión correlacionados, lo que resultará en un alta MI, y de lo contrario podrían seleccionarse como genes candidatos que interactúan. Las interacciones indirectas se eliminan aplicando una propiedad bien conocida de MI llamada la desigualdad de procesamiento de datos (**DPI**). Dado un **TF**[5], la aplicación del DPI, bajo supuestos apropiados, generará así predicciones sobre qué otros genes podrían ser sus objetivos transcripcionales directos o sus reguladores transcripcionales ascendentes. Después de este paso, se pueden aplicar algunos procedimientos adicionales de filtrado y posprocesamiento. El resultado final es una matriz de interacciones candidatas, también llamada matriz de adyacencia, que se puede utilizar para una mayor visualización y análisis de la red, tal como se analizará en la sección de resultados en el presente trabajo.

[5] Consultar definición en el apéndice A

2.1.3. Parallel-ARACNe

Dada la gran cantidad de datos que contiene la matriz de expresion obtenida con los scripts que se encuentran en el repositorio https://github.com/CSB-IG/ARACNE-multicore, y usando los los archivos *manifest* descargados del GDC Data Portal del NIH especificados en el apéndice B, fue necesario usar una nueva versión de ARACNe optimizada para hacer cálculos mediante un *cluster* . Esta versión se encuentra en https://github.com/CSB-IG/parallel-aracne. En este repositorio se especifican los pasos que deben seguirse para poder hacer el cálculo. Es de especial importancia resaltar que esta versión solo calcula la diagonal superior de la matriz de adyacencia por lo cual se tuvo que transformar al formato requerido para los análisis que se mostrarán más adelante. El formato que tiene el archivo de salida de ARACNe es una tabla como la que se muestra en la figura 2.3, en esta podemos observar además de lo antes dicho, solo un fragmento del data set original, pues este último esta compuesto por 16748 filas y 16748 columnas. Además de que la diagonal está compuesta por unos, y el resto de entradas con ceros y números, por lo que fue necesario hacer limpieza de datos para obtener una matriz de adyacencia(de ceros y unos) con ceros en la diagonal. Estos aspectos serán tratados a detalle al principio del capítulo 4.

```
data.head()
```

	ENSG00000000003	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000938	ENSG00000001036	ENSG00000001167	ENSG0000000
0	1.0	0.103197	0.152306	0.107439	0.151863	0.050207	0.047916	0.152306
1	NaN	1.000000	0.120830	0.128810	0.121863	0.076382	0.064962	0.120830
2	NaN	NaN	1.000000	0.051569	0.107554	0.057486	0.038834	0.334130
3	NaN	NaN	NaN	1.000000	0.032182	0.069682	0.042741	0.051569
4	NaN	NaN	NaN	NaN	1.000000	0.214658	0.029944	0.107554

5 rows × 16748 columns

Figura 2.3: Header del Data Set de la matriz de adyacencia de información mutua (archivo de salida de ARACNe)

Capítulo 3

Teoría de grafos

En las últimas décadas se han desarrollado una gran cantidad de técnicas algebraicas para poder describir la topología de diferentes tipos de grafos. Es por ello que en este capítulo se revisarán los conceptos que nos serán de mayor utilidad para poder realizar con éxito los análisis de datos que se han propuesto en el capítulo 4, los cuales tiene como base principal el *data set* de la figura 2.3 del capítulo anterior.

3.1. La matriz de adyacencia

Polinomio característico y espectro de un grafo

Definición 3.1.1: Sea $G = G(V, E)$ un grafo no dirigido es un par (V, E), donde V es un conjunto cuyos elementos son llamados vértices $V(G) = \{1, ..., n\}$ y E es un conjunto de vértices emparejados, cuyos elementos se denominan aristas $E(G) = \{e_1, ..., e_n\}$. Por otro lado, la matriz de adyacencia $A(G)$ de G es la matriz cuadrada de orden n cuyas entradas son

$$A_{ij} = \left\{ \begin{array}{c} 1, \text{ si } \{(v_i, v_j) \in E\} \text{ para } v_i, v_j \in V; \\ 0, \text{ en otros casos} \end{array} \right\}$$

$A(G)$ es una matriz real formada por unos y ceros. Una matriz de adyacencia es simétrica , esto es, para toda entrada i, j, $A_{ij} = A_{ji}$. Esta propiedad refleja el hecho de que una arista es representada como un par no ordenado de vértices $e = (v_i, v_j) = (v_j, v_i)$ [43]. A partir de ahora trataremos con **grafos no dirigidos**[1],

[1] Un grafo no dirigido es aquel en el que todas sus aristas son bidireccionales.

por lo que será bueno tomar en cuenta esta particularidad para todos los análisis subsecuentes.

Definición 3.1.2: El **polinomio característico** de la matriz de adyacencia $A(G)$ de un gráfico G, es decir, $\det(\lambda I - A(G))$, se denota por $p_G(\lambda)$, se dice que λ es un **valor propio del grafo** G cuando λ es una raíz de $p_G(\lambda)$. Si $A(G)$ tiene valores propios diferentes $\lambda_1 > \cdots > \lambda_s$ con multiplicidades[2] iguales, respectivamente, a $m(\lambda_1), \cdots, m(\lambda_s)$ el **espectro del gráfico** G, denotado por el $\text{spect}(G)$, se define como la matriz $2 \times s$, donde la primera línea consiste en los valores propios distintos de $A(G)$ dispuestos en orden decreciente y el segundo, por sus respectivas multiplicidades algebraicas. Es decir, escribimos

$$\text{spect}(G) = \begin{bmatrix} \lambda_1 & \cdots & \lambda_s \\ m(\lambda_1) & \cdots & m(\lambda_s) \end{bmatrix}.$$

El valor propio más grande de G se llama **índice de G** y se denota por $\text{ind}(G)$.
Ejemplo: Consideremos el siguiente grafo

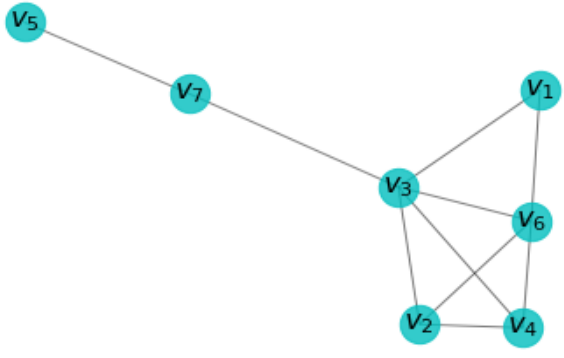

Figura 3.1: Grafo asociado a la matriz 7×7 de adyacencia dada por la ecuación (3.1). Lo denotaremos por $G_{7\times 7}$.

cuya matriz de adyacencia es:

[2]La multiplicidad algebraica de un valor propio λ de A es el orden de λ como cero del polinomio característico de A.

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}. \tag{3.1}$$

Su polinomio característico es $\lambda^7 - 10\lambda^5 - 10\lambda^4 + 11\lambda^3 + 14\lambda^2 - 2\lambda$ y su espectro es:

$$\text{spect}(G_{7\times 7}) = \begin{bmatrix} 3.41 & -1.94 & -1.37 & -0.55 & 0.35 & 1.11 & -1.00 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

por lo tanto $\text{ind}(G) = 3.41$. Este índice cobrará mucha relavancia en el análisis topológico de los grafos al usar el análisis por centralidad de vector propio y con la conectividad algebraica como se verá en el capítulo 4.

Debemos tener en cuenta que la suma de las entradas de cada fila de la matriz de adyacencia de un gráfico es igual al grado del vértice correspondiente.La siguiente proposición es un primer ejemplo de cómo las propiedades algebraicas de las matrices asociadas con ellas describen algunas propiedades estructurales de los gráficos.

Proposición 3.1.1: Sea G un grafo con n vértices y m aristas y sea

$$p_G(\lambda) = \lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + ... + a_{n-1}\lambda + a_n$$

el polinomio característico de G. Entonces los coeficientes de $p_G(\lambda)$ satisfacen:

(i) a_1 =0;
(ii) a_2 =-m;
(iii) $a_3 = -2t$, donde t es el número de triángulos en la gráfica.

Dada la relevancia de estos resultados, se dará una demostración a cada uno de estos puntos:

Demostración:

Para cada $i \in \{1, 2, ..., n\}$, se sabe del álgebra lineal que el número $(-1)^i a_i$ es la suma de los menores principales de A los cuales tienen i columnas y filas(se puede consultar una demostración de este hecho en [7]). Así:

(i) debido a que los elementos de la diagonal son todos cero en A, entonces todos sus menores de una fila y una columna son iguales a cero, de lo cual se deduce que $a_1 = 0$.
(ii) Un menor principal con dos filas y columnas, y que tiene una entrada distinta de cero, debe tener la forma:

$$\begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}$$

Hay uno de esos menores para cada par de vértices adyacentes de G, y cada uno tiene un valor -1. Por lo tanto $(-1)^2 a_2 = -1.|E| = (-1)m$,donde m es el número de aristas de G y $|E|$ es la cardinalidad del conjunto de aristas del grafo, lo que implica que $(-1)^2 a_2 = (-1)m$ y por lo tanto

$$a_2 = (-1)m$$

, dando el resultado.

(iii) Solo hay tres posibilidades para submatrices principales distintas de cero de $A(G)$ con 3 filas y 3 columnas; sus determinantes son:

$$\begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{vmatrix}, \quad \begin{vmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{vmatrix}, \quad \begin{vmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{vmatrix} \tag{3.2}$$

y, de estos, el único que no es cero es el último (cuyo valor es 2). Este menor principal corresponde a tres vértices mutuamente adyacentes, es decir, un triángulo. Entonces $(-1)^3 a_3 = 2t$ donde t es el número de triángulos de G. Por lo tanto, $a_3 = -2t$, como se quería ■

Proposición 3.1.2: El número de cadenas de longitud l que conectan el vértice v_i al vértice v_j en un gráfico G viene dado por el orden de entrada (i, j) de la matriz A^l, donde $A = A(G)$ es la matriz de adyacencia de G .

Ejemplo: Al tomar la matriz dada por la expresión (3.1), que representa el grafo de la figura (3.1), y elevarla al cuadrado obtenemos la matriz:

$$A^2 = \begin{bmatrix} 2 & 2 & 1 & 2 & 0 & 1 & 1 \\ 2 & 3 & 2 & 2 & 0 & 2 & 1 \\ 1 & 2 & 5 & 2 & 1 & 3 & 0 \\ 2 & 2 & 2 & 3 & 0 & 2 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 2 & 3 & 2 & 0 & 4 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 2 \end{bmatrix} \tag{3.3}$$

en la cual se puede obervar que la entrada $A^2_{3,6} = 3$; según lo anteriormente dicho, representa el número de cadenas de longitud 2 que conectan a los vértices v_3 y v_6 como puede observarse en la siguiente figura, en donde se ha hecho uso del script **PlotGraphWithLabeledVertices** que se puede encontrar en el repositorio [47] observamos que como se había predicho antes hay tres cadenas de longitud 2

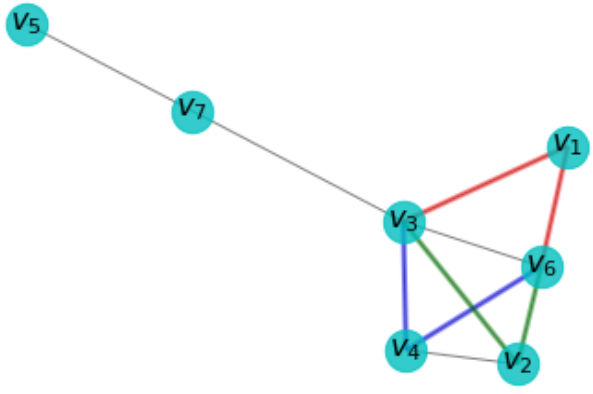

Figura 3.2: Grafo asociado a la matriz de adyacencia (3.1) etiquetada con los vertices $v_1, ..., v_7$

posibles que conectan los vértices v_3 y v_6; la primera cadena (en color azul), va de v_3 a v_4 y luego esta a v_6; la segunda (en color rojo), va de v_3 a v_1 y luego esta a v_6; finalmente la tercera (en color verde) posibilidad es v_3 a v_2 y luego a v_6.

Además es importante resaltar que la diagonal de la segunda potencia de la matriz de adyacencia, nos brinda otra útil información sobre la topología del grafo: cada entrada de la diagonal, en orden ascendente, muestra la cantidad de bordes que inciden sobre cada vértice, es decir, la primera entrada de la diagonal de (3.3), indica que en el vértice v_1 inciden dos bordes; la segunda entrada de la diagonal, indica que en el vértice v_2 inciden 3 bordes; la tercera entrada de la diagonal, indica que en el vértice v_3 inciden 5 bordes, y así sucesivamente hasta el vértice 7 en ese orden. Esto se puede corroborar fácilmente viendo el grafo de la figura 3.2. Otros ejemplos útiles pueden encontrarse en [8].

Corolario 3.1.1: Sea G un gráfo con n vértices y m aristas y sean $\lambda_1, ...\lambda_n$ sus valores propios. Entonces:

(i) Si T_l es el número de cadenas cerradas de longitud l en G entonces $T_l = \text{tr}(A^l) = \sum_{i=1}^{s} \lambda_i^l$. En particular:

(ii) La suma de los cuadrados de valores propios es el doble del número de aristas, es decir, $T_2 = \text{tr}(A^2) = 2m$

(iii) La suma de los cubos de valores propios es seis veces el número t de triángulos, es decir, $T_3 = \text{tr}(A^3) = 6t$

Cuando se usan grafos para modelar redes, un aspecto relevante es resaltar los vértices más importantes, es decir, los vértices más influyentes en la red. Esta influencia o importancia del vértice depende del tipo de relación modelada, representada por los bordes del gráfico, y se evalúa mediante medidas de centralidad.

Centralidad de vector propio

Definición 3.1.3: La **centralidad del vector propio** [44] del vértice v_i del grafo G es la coordenada i-ésima x_i del vector propio no nulo $x = [x_1...x_n]^T$ asociado con el **índice de G**, λ_1(valor propio más grande) del grafo asociado, es decir, es el número

$$x_i = \frac{1}{\lambda_1} \sum_{j=1}^{n} a_{ij} x_j,$$

donde a_{ij} son las entradas de su matriz de adyacencia.

La centralidad de vector propio mide la influencia de un nodo en una red. Fue propuesta por Phillip Bonacich [44] en 1972, y corresponde al vector propio principal[3] de la matriz de adyacencia del grafo analizado. Intuitivamente, los nodos que poseen un valor alto de esta medida están conectados a muchos nodos que a su vez están bien conectados; por lo tanto, son buenos candidatos para distribuir información. Los nodos más centrales desde esta perspectiva corresponden a centros de grandes grupos donde se aglutina la mayor cantidad de información.

En general habrá varios valores propios para los cuales existe una solución al problema de vector propio. Sin embargo hay un requerimiento adiconal que debe cumplirse; las entradas de los vectores propios deben ser positivos y por el teorema de Perron-Frobenius implica que solo los mayores valores propios conducen a la medida de centralidad deseada como lo demuestra Newman [46]. El vector propio principal puede ser calculado con el *power method* [45].

[3] Aquél vector propio asociado al valor propio más grande en valor absoluto

3.2. La matriz laplaciana

La matriz laplaciana de un grafo y, en particular, sus valores propios están fuertemente relacionados con el número de vértices y al grado máximo de vértices [48]. Se presentarán las propiedades básicas de esta matriz a continuación.

3.2.1. Conceptos preliminares

Definición 3.3.1: Sea D la matriz diagonal de los grados de los vértices de un gráfico G (es decir, la matriz D tal que $D_{ii} = d(vi))$[4] y sea A la matriz de adyacencia de G. Al restarle a la matriz de grado, la matriz de adyacencia, obtenemos la siguiente matriz :

$$L = D - A, \tag{3.4}$$

la cual se llama matriz laplaciana o laplaciana del gráfico G. Cuando sea necesario, usaremos $L(G)$ en lugar de L.

Por ejemplo, al tomar una submatriz de la matriz de adyacencia obtenida con ARACNe al usar los datos obtenidos en el GDC Data Portal del National Cancer Institute [49], correspondientes a datos de *cervix-uteri* de tejido tumoral, donde cada entrada de esta matriz representa la información mutua entre cada par de genes presentes en la muestra, al tomar una submatriz de 5×5 se ve que tiene la forma :

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Que se asocia al grafo de la figura 3.3. Además la matriz de grado está dada por:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

y al efectuar la operación $D - A$ obtendremos la matriz laplaciana dada por:

[4] Grado del vértice v_i, esto es, el número de aristas (bordes) incidentes sobre él.

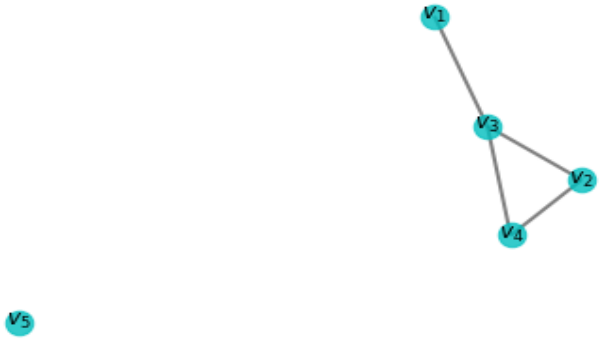

Figura 3.3: Grafo asociado a la matriz de adyacencia.

$$L = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 2 & -1 & -1 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & -1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Definición 3.3.1: El espectro laplaciano de un grafo G, denotado por $\xi(G)$, es el vector cuyos elementos son todos los valores propios de L ordenados de manera no creciente. Por lo tanto, si $\mu_1 \geq \cdots \geq \mu_n$ son los valores propios de L entonces

$$\xi(G) = (\mu_1, \cdots, \mu_n) \tag{3.5}$$

Ejemplo: Para ejemplificar esto tomemos además del grafo de la figura 3.4, el grafo de la figura 3.5 mostrado a continuación:

cuya matriz laplaciana asociada está dada por:

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \tag{3.6}$$

Al usar la herramienta de Python linalg de la extensión numpy, podemos calcular rápidamente los valores propios de una matriz vista como un "array"; en este caso, obtenemos para la anterior matriz laplaciana la siguiente matriz de línea:

Figura 3.4: Grafo asociado a la matriz de adyacencia.

$$\xi(G2) = (4\ ,\ 1\ ,\ -1.51\ ,\ 3\ ,\ 2\ ,\ 0\ ,\ 0) \tag{3.7}$$

y para la matriz laplaciana asociada al grafo dado por la figura 3.4 se tendrá el siguiente espectro laplaciano:

$$\xi(G1) = (4\ ,\ 1\ ,\ -1.51\ ,\ 3\ ,\ 0\ ). \tag{3.8}$$

Resulta que el número de componentes conectados[5] de cada gráfico coincide exactamente con la multiplicidad de 0, que es un valor propio que aparece en ambos casos . Estos hechos se aplican a todos los grafos(no dirigidos), como se muestra a continuación.

Definición 3.3.2: Sea G un grafo. La matriz de incidencia β respecto a una orientación dada es aquella cuyas entradas son:

$$\beta_{ij} = \left\{ \begin{array}{c} +1,\ \text{si } v_i \text{ es el vértice dónde llega } e_j; \\ -1,\ \text{si } v_i \text{ es el vértice desde el cual comienza } e_j; \\ 0,\ \text{en otros casos} \end{array} \right\}.$$

Se puede demostrar que $L = \beta\beta^T$ [20]. De ello se deduce que la matriz laplaciana L es una matriz positiva semi-definida que tiene, por lo tanto, todos sus valores propios mayores o iguales a cero.

Proposición 3.3.2: Sean $\mu_1 \geq \mu_2 \geq ... \geq \mu_n$ los valores propios de la matriz Laplaciana L de un grafo G [6]. Entonces:

[5]En la teoría de grafos , un componente conectado de un grafo no dirigido es un subgrafo en el que dos vértices cualesquiera están conectados entre sí por caminos , y que no está conectado a ningún vértice adicional en el supergráfico.

(i) $\mu_n = 0$ con vector propio asociado $\mathbf{1} = [1, 1, ..., 1]^T$;
(ii) G es conexo si y sólo si $\mu_{n-1} > 0$;

3.2.2. Conectividad algebraica

Definición 3.3.3: El segundo valor propio más pequeño del Laplaciano de G, μ_{n-1}, se llama **conectividad algebraica** del gráfico G y ahora se denotará con $a(G)$. El valor propio más grande del laplaciano de G, μ_1, se llama **índice laplaciano** de G.

La conectividad algebraica juega un papel fundamental en el estudio de un grafo. Recientemente se ha demostrado que los grafos con alta o baja conectividad algebraica (en comparación con el grado máximo) tienen propiedades importantes en varias aplicaciones [50].

Definición 3.3.4: La **conectividad de vértice** de un grafo, denotado por $k(G)$, es el número más pequeño de vértices que, cuando se eliminan, hacen que el grafo se desconecte.

Definición 3.3.5: La **conectividad de bordes**, denotada por $k'(G)$, es el número más perqueño de bordes que, cuando se eliminan, hacen que el grafo se desconecte. La conectividad algebraica y la conectividad de vértices y bordes se relacionan de acuerdo al siguiente resultado, probado por Fiedler [28].

Proposición 3.3.3: Si G no es el grafo completo[6] entonces $a(G) \leq k(G) \leq k'(G)$

Dado que la conectividad de borde es menor o igual que el grado mínimo de un grafo, podemos reescribir la proposición anterior como $a(G) \leq k(G) \leq k'(G) \leq \delta(G)$

3.3. La utilidad de la "matriz laplaciana sin signo"

La matriz laplaciana sin signo, a través del espectro e invariantes derivados de ella, aunque no es suficiente para permitir la caracterización de grafos, parece garantizar la existencia de un número mucho mayor de grafos que se pueden caracterizar por su espectro que el de matriz Laplaciana convencional, que a su vez parece ser más eficiente que la matriz de adyacencia para ayudar a realizar esta tarea [6].

Definición 3.4.1 : La matriz laplaciana sin signo de un gráfico G viene dada por $Q = D + A$ donde D es la matriz diagonal cuyas entradas son los grados

[6] Un grafo completo es un grafo simple donde cada par de vértices está conectado por una arista.

de sus vértices y A es su matriz de adyacencia. El polinomio característico de la matriz laplaciana sin signo se denota por $p_Q(\lambda)$ y sus valores propios se denotan por $q_1 \geq q_2 \geq ... \geq$, siendo q_1 el índice de Q.

La matriz laplaciana sin signo $Q(G)$ es simétrica y tiene entradas no negativas, como la matriz de adyacencia. Además, si G está conectado, esta matriz también es irreducible. Entonces podemos usar el Teorema de Perron-Frobenius (consúltese el apéndice D para más detalles), en relación con la matriz $A(G)$, obteniendo que, para grafos conexos, q_1 es un valor propio simple. También notamos que $q_1 = 0$ si, y solo si, G es un gráfico sin aristas.

Es posible expresar el **número de aristas de un grafo** en función de uno de los **coeficientes** de $p_Q(\lambda)$, como nos dice la siguiente proposición [6].

Proposición 3.4.1: El número de aristas de un grafo G con n vértices es igual a $\frac{-p_1}{2}$, donde p_1 es el coeficiente de λ^{n-1} en el polinomio característico de Q.

Ejemplo: Con la función que se ha creado en *Python*: **pol_char_from_signless_lap(mat,n)** donde el parámetro **mat** es un *data frame* que representa una matriz de adyacencia, en nuestro caso de información mutua y el parámetro **n** definirá las dimensiones de la submatriz que tomaremos, ya que el tamaño de la matriz original es demasiado grande. Así, al hacer el cálculo para la sub-matriz $A_{8\times 8}$ que es la matriz de adyacencia de información mutua [47] se obtuvieron sus coeficientes del polinomio característico como se ve en la figuran 3.5.

```
In [16]: inicio = time.perf_counter()

         polinomio = pol_char_from_signless_lap(A,8)

         final = time.perf_counter()

         print("Los coeficientes del polinomio característico son", polinomio)

         print("El tiempo de ejecución fue de ", final-inicio,"segundos")

         Los coeficientes del polinomio característico son [ 1.0000e+00 -2.8000e+01  3.2700e+02 -2.07
         80e+03  7.8430e+03 -1.7954e+04
           2.4266e+04 -1.7648e+04  5.2760e+03]
         El tiempo de ejecución fue de  0.049802171000010276 segundos
```

Figura 3.5: Polinomio característico de la matriz laplaciana sin signo asociada a la matriz de adyacencia A 8×8 para datos UNT

vemos inmediatamente que el coeficiente $p_1 = -28$ que al usar el resultado de la proposición 3.4.2 vemos que el número de aristas N_e es igual a $N_e = -(-28)/2 = 14$. Usando la herramienta de *networkx* podemos dibujar el grafo asociado a la

matriz de adyacencia A usada en la función de la figura 3.5, con lo cual se obtiene el grafo de la figura 3.6.

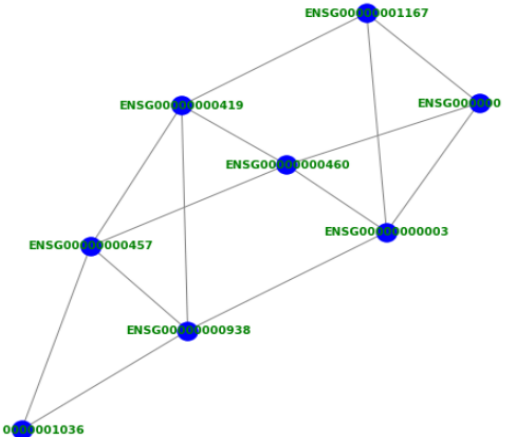

Figura 3.6: Sub-grafo de la matriz 8×8 para datos UNT. Los vértices están etiquetados por sus *Gene IDS* con los que podemos identificar a cada gen.

Visualmente podemos corroborar que hay 14 aristas tal como lo predice la proposición. Este resultado es muy poderoso sobre todo si queremos darnos una idea de cuántas interacciones estadísticas tendrán un grupo mucho más grande de genes que en el caso anterior: Recuérdese que en este grafo, cada vértice representa un gen y cada arista representa la información mutua entre cada par de genes. Hay genes que sólo interactúan con uno más y hay otros que interactúan con dos o más genes, habiéndolos aquellos que interactúan con muchos más que el resto, lo cuál puede darnos informadción relevante sobre qué genes son los que tienen una mayor influencia sobre los demás, aquellos que regulan de manera más significativa que el resto.

En el anterior ejemplo, se pueden contar incluso "a ojo", la cantidad de aristas que tiene el grafo, pero ¿qué pasa si se nos presenta el problema de analizar un grafo como el de la figura 3.8? Este corresponde a tomar la interacción solo de los primeros 80 genes del Data Frame original (el cual contiene en total la interacción de 16748 genes).

Si quisiéramos ver como en el caso anterior "a ojo" cuántas interacciones hay en total en este grafo sería imposible, y por eso el resultado de la proposición 3.4.2 es tan útil. Usando nuevamente la función **pol_char_from_signless_lap(mat,n)** para este caso obtenemos que el coeficiente $p_1 = -2692$ que al usar el resultado de la proposición 3.4.2 vemos que el número de aristas N_e es igual a $N_e =$

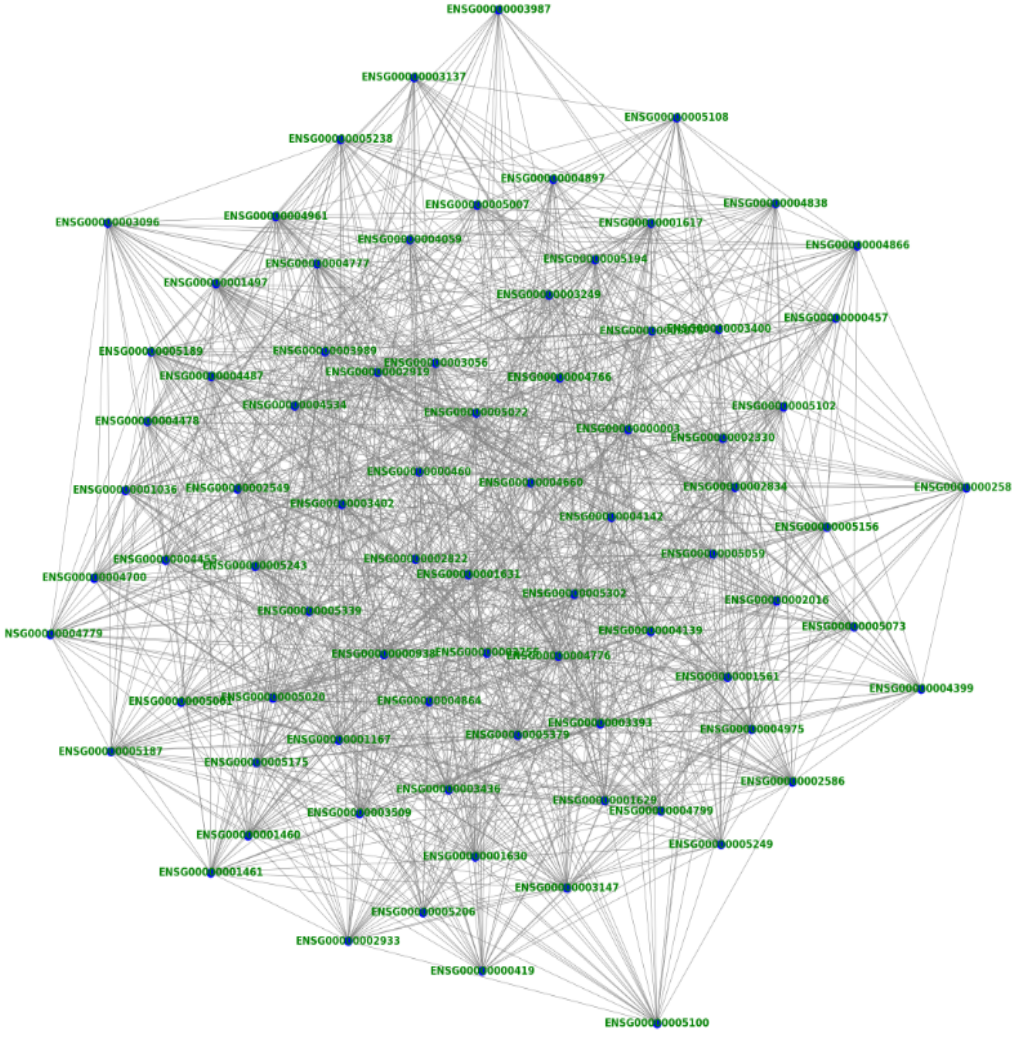

Figura 3.7: Sub-grafo de la matriz de adyacencia 80×80 para datos UNT

$-(-2692)/2 = 1346$.

Una vez establecidos los principios teóricos sobre los cuales se lleva a cabo el análisis espectral de grafos, podemos describir todo el proceso algorítmico[7] que

[7] Véase el apéndice F

se llevó a cabo para hacer el análisis de datos de se describirá en el capítulo 4. El siguiente capítulo es la parte fundamental del presente trabajo, pues recoge la contribución original respecto a los temas tratados asta aquí.

Capítulo 4

Resultados y discusión

En el presente capítulo se analizará el tejido de los casos de cáncer cuyo sitio primario[1] está ubicado en *Uterus nos*, por sus siglas en inglés significa *not otherwise specified* [19], es decir, se refiere a los datos de cáncer en el útero en circunstancias distintas a las demás consideradas: *cervix-uteri* y *corpus-uteri*. Para extraer estos datos, primero fue necesario hacer la segunda consulta mostrada en el apéndice F en el *GDC Data Portal*. De esta consulta se descarga un archivo de tipo *manifest* el cual debe someterse a todo un proceso de control de calidad que puede verse en [51]. Este código tuvo que modificarse ligeramente para poder ser funcional para nuestros datos, pero en esencia puede funcionar para cualquier tipo de archivos *manifest* descargados de [49]. Al final de este proceso de control de calidad se obtienen matrices de expresión génica, que servirán de *input* para la versión de ARACNe usada en este trabajo [52].

Además en nuestro caso, los datos **UNS** corresponderán con los de **tejido sano** y los datos **UNT** a los de **tejido tumoral**. Primero será de gran utilidad identificar aquellos genes que tienen mayor relevancia estadística al ser reguladores de una mayor cantidad de genes que los demás, proporcionándonos información sobre qué genes son los que están activos y cuáles inactivos via los factores de transcripción. La primera y más sencilla medida de **centralidad** [2] que será usada será la de centralidad de grado, pues nos proporciona información sobre la cantidad de aristas que inciden sobre un véretice en particular[3]. Luego, se usará una medida de centralidad más sofisticada: la centralidad de vector propio, que

[1] El tejido de origen de un tumor metastásico.

[2] La centralidad en un grafo puede ser entendida como una medida o un valor que posee un vértice dentro de un grafo. Este al ser evaluado en una escala, determina su relevancia dentro del grafo y permite comparar o contrastar dicho vértice con otros.

[3] Véase el capítulo 3.

hace suposiciones adicionales sobre las interacciones entre vértices. Con los resultados anteriores se hará un contraste con la evidencia de la literatura oncológica[4] comparando nuestros resultados con los de esta.

Prácticamente todo el análisis de datos, así como el análisis espectral tienen como punto de partida a la matriz de adyacencia. En el repositorio [47] podrá encontrarse el código con los cálculos que se mostrarán a continuación.

4.1. Análisis de la matriz de grado

Lo primero que debe hacerse es leer el archivo que contiene los datos de la matriz de adyacencia de información mutua, a partir del cual se hace limpieza de datos para tener la matriz de adyacencia en una forma conveniente con la cual se puedan hacer cálculos.

La matriz de adyacencia, por definición, debe contener solo ceros y unos, sin embargo, el archivo de salida de ARACNe nos proporciona una matriz de ceros y números, por lo cual es necesario fijar un criterio a partir del cual por medio de un valor umbral convirtamos esos ceros y números en ceros y unos. El criterio utilizado consiste en calcular el valor promedio por cada columna usando el valor de cada una de las entradas y guardar todos esos promedios en un array que es nuestro caso contendrá 16748 elementos sobre los cuales se vuelve a promediar para obtener una umbral a partir de cual obtener ceros y unos. Este umbral tiene un valor de 0.033064 para los datos de UNT y de 0.085677 para los datos de UNS[5].

A partir de esto es posible construir la matriz de grado, que como se ha visto, es una matriz diagonal que contiene información sobre el grado de cada vértice , es decir, el número de bordes unido a cada vértice. La construcción de esta matriz de grado tiene un doble objetivo, primero; dilucidar qué genes tienen mayor relevancia estadística dentro de la red , y segundo; para construir la matriz laplaciana a partir de la cual se hará el análisis espectral de grafos.

Esta matriz se obtiene con la función `degreeMat_fromAdj(mat,n)`. Adicionalmente se usan otras tres funciones: `get_genes(data,n)`, `get_elem_in_diag(Mat)` y `func(a,N)` [47]. Con la primera, se obtienen los nombres de los genes compuestos por caracteres; con la segunda, se obtienen todos los elementos de la diagonal de la matriz de grado y se guardan en un *array* y finalmente; con la tercera, se ordenan de mayor a menor los elementos de ese *array* de elementos de la diagonal.

A continuación se muestra dos tablas de 20 diferentes corridas, 10 para los datos UNS y 10 para los datos de UNT. Para hacer los cálculos se hizo uso de

[4] Rama de la medicina especializada en el diagnóstico y tratamiento del cáncer. Incluye la oncología médica (uso de quimioterapia, terapia con hormonas y otros medicamentos para tratar el cáncer), la radioncología (uso de radioterapia para tratar el cáncer) y la oncología quirúrgica (uso de cirugía y otros procedimientos para tratar el cáncer).

[5] Véase el apéndice F para más detalles.

funciones del tipo `func(get_elem_in_ diag(degreeMat_from_Adj(A,n)),n)` [47] con valores de $n = 1000, 2000, \cdots, 10000$. Esto es una primera aproximación, pues si se quisiera hacer una análisis completo tendría que hacerse el cálculo con $\mathbf{n = 16748}$ pues es la **cantidad de filas y columnas** de la que está compuesta la matriz de adyacencia original.

Valor de n	Genes
1000	237, 658, 296
2000	1222, 658, 237
3000	1222, 658, 237
4000	1222, 3696, 3103
5000	1222, 3696, 3103
6000	5341, 1222, 3696
7000	6244, 5341, 1222
8000	6244, 5341, 1222
9000	6244, 5341,1222
10000	6244, 9242, 5341

Cuadro 4.1: Genes con mayor grado para datos de UNS. Siendo n el número de filas y columnas de submatrices tomadas de la matriz de adyacencia completa. Las etiquetas de los genes con mismo color se tratan de un gen repetido, esto se hizo con el fin de tener una mejor visualización de como van cambiando la relevancia del gen en términos de centralidad para valores de n cada vez mayores.

Obsérvese que en al cuadro 4.1 los genes con más relevancia estadística en la red (esto es, vía la información mutua, en el supuesto de que el valor de la información mutua fuera diferente de 0 entre un gen G_1 y un gen G_2 significaría que G_1 comparte información con G_2 y G_1 determinaría el valor de G_2 y viceversa) son los genes 6344, 9242, 5341, 1222, 3696, 3103, 658, 296 y el 237. Sin embargo, es relevante notar que mientras más alto sea el valor de n, algunos genes que antes eran los más relevantes dejan de serlo en corridas con alto valor de n. Así al tomar más datos y haciendo que los cálculos pasen de ser aproximados a tener el valor más preciso, vemos que hay genes que aparecen como los más relevantes y que antes ni si quiera figuraban en los primeros puestos. Esto se debe a que localmente hay genes que tienen un alto grado, pero que al aumentar el valor de n la cantidad máxima de interacciones que pudieran tener aumenta considerablemente y esto hace que relevancia local pase a segundo plano al crecer el grafo en número de vértices; aunado al hecho de que al tomarse solo subgrafos se están omitiendo una gran cantidad de interacciones posibles para genes en particular. Esto se hace claro al recordar que el número máximo de aristas que puede tener un grafo es:

Valor de n	Genes
1000	465, 221, 555
2000	465, 1716, 365
3000	465, 2457, 2005
4000	465, 3742, 365
5000	465, 3742, 365
6000	465, 3742, 365
7000	465, 3742, 365
8000	465, 7948, 3742
9000	465, 7948, 3742
10000	465, 7948, 7261

Cuadro 4.2: Genes con mayor grado para datos de UNT. Siendo n el número de filas y columnas de submatrices tomadas de la matriz de adyacencia completa. Las etiquetas de los genes con mismo color se tratan de un gen repetido, esto se hizo con el fin de tener una mejor visualización de como va cambiando la relevancia del gen en términos de centralidad para valores de n cada vez mayores.

$$e_{max} = \frac{n(n-1)}{2}$$

por lo tanto si n=1000, que corresponde a la primera línea del cuadro 4.1, el numero máximo de aristas sería de $e_{max} = 499\ 000$, y por otro lado, con tan solo aumentar el número de vértices en 1000 se obtiene que para $n = 2000$ el número de aristas máximo será de $e_{max} = 1\ 999\ 000$.

Otro aspecto interesante sobre estos datos es que en los datos de cáncer el gen con posición 465 en todas las corridas figura como el más relevante al tener el mayor grado en todas ellas, algo que claramente no ocurre en los datos del tejido sano, que más bien siguen una tendencia un poco más irregular conforme más preciso es el cálculo.

Este hecho tan simple refleja que los genes con mayor relevancia estadística en tejido tumoral tienen una influencia más significativa sobre el resto de su red respectiva, que los genes con mayor relavancia estadística en tejido sano en su propia red. Es crucial resaltar que el solo poner los primeros tres genes con mayor grado en los cuadros 4.1 y 4.2 es completamente arbitrario, pues pudimos haber tomado los primeros 10 o primeros 100, etc. Se decidió hacerlo así con el fin de hacer más sencillo este primer análisis y porque hasta ahora no tenemos un criterio adicional que nos permita asegurar con fundamentos sólidos el por qué los primeros 3, 10, 100, etc., serán los más importantes a tomar en cuenta.

Adicionalmente podemos identificar los Ensembl IDs de estos genes, pues tene-

mos su ubicación dentro del *array* que se obtiene con la función `get_genes(data,n)`. Este resultado se resume en el cuadro 4.3.

Datos	Ensembl IDs (prefijo ENSG)	Gene Symbols
UNS	00000137807, 00000163808, 00000130635	KIF23, KIF15, COL5A1
	00000077721, 00000114529, 00000108821	UBE2A, C3orf52, COL1A1
	00000055917, 00000011021, 00000013810	PUM2, CLCN6, TACC3
UNT	00000037042, 00000010318, 00000048162	TUBG2, PHF7, NOP16
	00000092148, 00000023330, 00000103512	HECTD1, ALAS1, NOMO1
	00000100412, 00000114956, 00000152284	ACO2, DGUOK, TCF7L1
	00000144857	BOC

Cuadro 4.3: Genes con mayor relevancia estadística para datos UNT, y UNS, según la matriz de grado para valores de n=10 000 de los cuadros 4.1 y 4.2. Se muestran los Ensembl IDs y su conversión a notación de *gene symbols* aprobada por el *HUGO Gene Nomenclature Committee*. Para hacer la conversión se usó la herramienta [53]. Para hacer una consulta en [53] se toma por ejemplo el primer gen de UNS: 00000137807 y se le agrega el prefijo ENSG, quedando como ENSG00000137807, que es lo que debe ponerse en el buscador de [53]; este arrojara su gene symbol asociado: KIF23.

Para obtener los valores de la segunda columna anteriores se usó `get_genes(data,n)[l]` donde l son las diferentes posiciones de los genes dadas por los cuadros 4.1 y 4.2, dentro de la lista de genes .

En cuanto a los valores de la tercera columna del cuadro 4.3, corresponden a los *Gene Symbols*, son nombres que identifican a moléculas específicas que tienen características especiales en la red, además de que son útiles pues pueden remitirnos a bases de datos que contienen información que son de relevancia para los biólogos moleculares del cáncer y oncólogos, como en la base de datos de [53] en donde incluso se pueden encontrar referencias a artículos clave que describen el gen y/o sus productos, o son particularmente relevantes para su nomenclatura y/o función.

Este método nos permite tener una visión confiable, rápida de hacer y siendo bastante intuitiva de entender, sin embargo, surge una pregunta natural al respecto: ¿qué sucede si hay vértices de la red que a pesar de tener un valor pequeño de bordes conectados a él, tiene una mayor cantidad de bordes asociados a vértices con un alto valor en su grado? Existe la posibilidad entonces de que el último algoritmo[6] que involucra solo a la matriz de grado esté omitiendo muchos genes que

[6] Véase el apéndice F

pueden ser igualmente influyentes o aún más que los presentados en los cuadros 4.3. Afortunadamente existe una medida que arregla esta limitación de la matriz de grado: la centralidad de vector propio, que será analizada a continuación.

4.2. Análisis con la centralidad de vector propio

La centralidad de vector propio es una visión más sofisticada de la centralidad: un gen con pocas conexiones (bordes en la red) podría tener una centralidad de vector propio muy alta si esas pocas conexiones estuvieran con otras muy bien conectadas. La centralidad del vector propio permite que las conexiones tengan un valor diferente al encontrado con la matriz de grado, de modo que conectarse a algunos vértices tiene un mayor impacto sobre la transferencia de información dentro de la red, que conectarse a otros.

Primero tomemos una sub-grafo de los datos UNT correspondientes a la matriz de adyacencia, tomando una sub-matriz de 45 por 45.

Ahora, consideramos un vector con 45×1 valores, uno para cada vértice o nodo en el grafo. En este caso, hemos utilizado la centralidad grado de cada vértice, extrayendo los valores de la diagonal de la matriz de grado, para ello usamos el archivo EigVecCentAnalysisUNT del repositorio [47], usando:

```
M45=mat_shape(A,45).dot(get_elem_in_diag(degree_mat_from_adj(A,45)))
M45_2=np.squeeze(np.asarray(M45))
```

con lo cual obtenemos el producto[7] entre la matriz de adyacencia asociada al grafo y la lista con los valores de la diagonal de la matriz de grado, obteniéndose otra lista con 45 entradas. En el primer elemento del vector resultante se recogen los valores de cada vértice al que está conectado el primer vértice (en este caso, el segundo, tercero, cuarto, etc. hasta llegar al cuadragésimo quinto, esto sólo si existe un borde entre el primer vértice y aquellos) y el valor resultante es la suma de los valores de cada uno de estos vértices adyacentes.

En otras palabras, lo que se logra con el producto entre la matriz de adyacencia y el vector de grados, es reasignar a cada vértice la suma de los **valores de sus vértices vecinos**.

Al comparar ambos resultados vemos que el gen con posición 22 es el décimo gen más relevante en la red según el criterio de la matriz de grado(al observar el primer output de la figura 4.2), sin embargo al compararlo con el segundo criterio del producto de la matriz de adyacencia por el vector de grado, al ver cómo se "difunde" la centralidad de grado, vemos que este mismo gen 22 es el sexto más relevante dentro de la red(el segundo ouput en la figura 4.2).

[7] Véase el apéndice F

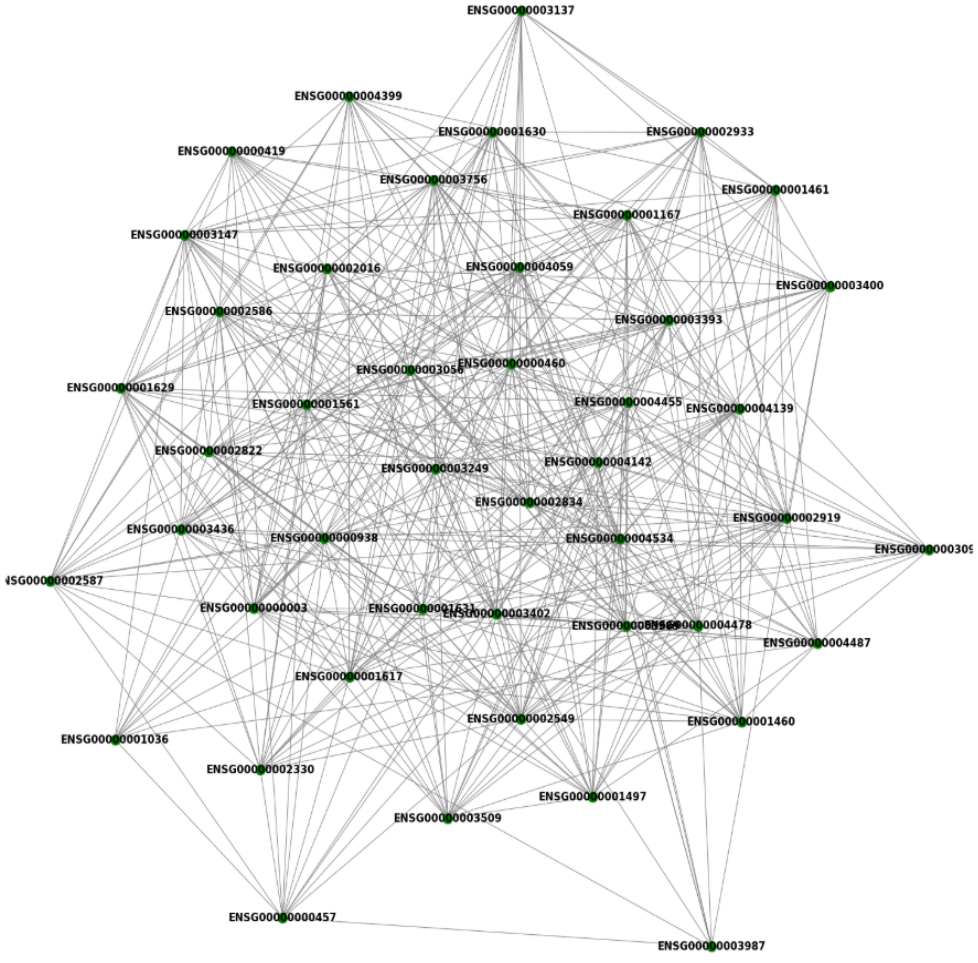

Figura 4.1: Sub-grafo de datos UNT correspondiente a una matriz de adyacencia de 45×45. Imagen tomada de [47]

En este caso conviene hacer una comparación entre el primer y segundo método usando la cantidad suficiente de genes como para no tener errores de memoria[8] en la ejecución. Así, se ejecutan los scripts de tal manera que se usen la mayor cantidad de genes en el proceso. Recuérdese que la cantidad total de genes involu-

[8] Consúltese el apéndice C

```
func(get_elem_in_diag(degree_mat_from_adj(A,45)),45)
```

```
array([24, 44,  3, 10, 31, 28, 14, 17, 27, 22, 36, 42, 39,  4, 37,  0, 32,
       34,  6, 41, 20, 11, 12, 18, 29,  9, 23, 33,  7, 15, 21, 38,  8, 19,
       30, 16, 43, 13,  1, 40, 25, 26,  2,  5, 35], dtype=int64)
```

```
func(M45_2,45)
```

```
array([24, 44, 17, 10,  3, 22, 28, 36, 14, 37, 31, 42, 27, 32,  0, 34,  6,
       41, 39, 20,  4, 29, 12,  9, 11, 18, 15, 33,  7, 23, 21, 16, 30,  8,
        1, 38, 19, 40, 13, 43, 25, 26,  2,  5, 35], dtype=int64)
```

Figura 4.2: Los out put muestran la comparación de centralidad de genes entre el método de la matriz de grado y el producto entre la matriz de adyacencia por el vector de grado respectivamente.

crados es de 16748, mientras más nos aproximemos a usar ese total, más precisas serán nuestras predicciones.

Análisis de la centralidad de vector propio para UNS y UNT

Una vez entendido el ejemplo visto anteriormente podemos hacer análisis submatrices mucho más grandes. En los siguientes análisis, nuevamente n se refiere a la cantidad de filas y columnas tomadas de la matriz de adyacencia original.

Comenzaremos analizando los resultados mostrados en el cuadro 4.4. En él podemos obervar que para n=1000, los primeros 10 genes más relevantes según ambos criterios, son exactamente los mismos en el mismo orden, pero a partir de n=2000 la diferencia entre ambos criterios es notable. Hay casos en los que, por ejemplo, para $n = 3000$ el **gen 1481** (recuérdese que este número representa la posición que tiene el gen en el *header* del *data set* original con el que se empiezan a hacer todos los cálculos, y con el cual podemos identificar después de qué gen se trata en específico) pasa de ser el sexto gen más relvante según el criterio de la matriz de grado a ser el cuarto más relevante, según el criterio de la centralidad de vector propio. De manera opuesta, ahora para n=2000 se puede osbervar que el **gen 1366** pasa de ser el cuarto gen más relevante, según el criterio de la matriz de grado, a ser el octavo gen más relevante según el criterio de la centralidad de vector propio. Con este pequeño ejemplo ya se puede atisbar la utilidad de la centralidad de vector propio; puede mostrarnos por un lado, genes que antes parecían no jugar un papel muy importante dentro de la red como otros; y por el otro, genes que antes parecían los más relevantes, en realidad pueden no serlo.

Dada la gran cantidad de datos que se analizan, siempre es arriesgado sacar conclusiones de una muestra tan pequeña de los datos; por ello sería pertinente

Valor de n	Análisis con la matriz de grado	Análisis de centralidad de vector propio
1000	237, 658, 296, 213, 436, 244, 861, 77, 592, 486	237, 658, 296, 213, 436 244, 861, 77, 592, 486
2000	1222, 658, 237, 1366, 436, 1481, 296 , 861 , 32, 213	1222, 237, 658, 1481, 436, 296, 861, 1366, 32, 213
3000	1222, 658, 237, 2370, 2124 1481, 861, 2149, 32, 811	1222, 237, 2124, 2370, 1481 658, 861, 32, 2149, 2369
4000	1222, 3696, 3103,2370 , 1481 658, 2124, 237, 1366,3529	1222, 3696, 3103, 2124, 2370 3269, 1481, 237, 3529, 861
5000	1222, 3696, 3103,2124 , 237, 658, 1481, 4103, 3529, 2370	1222, 3696, 2124, 3103, 4192 237, 3269, 1481, 3529, 4103
6000	5341, 1222, 3696, 2124, 5762 3103, 658, 237, 4427, 1481	5341, 1222, 3696, 2124, 3103 5762, 4192, 237, 1481, 3269

Cuadro 4.4: Comparación entre el método de la matriz de grado y la centralidad de vector propio para diferentes subgrafos de la matriz de adyacencia para los datos de UNS. Los valores están ordenados de mayor a menor. Las etiquetas coloreadas corresponden a genes que han cambiado de posición de relevancia al usar otro método de medida de centralidad.

después crear un algoritmo que cuantifique la diferencia que hay entre los datos proporcionados por ambos métodos. Sin embargo, dado que estamos haciendo un análisis un poco más cualitativo y muy específico, en este caso, los 10 primeros genes más relavantes dentro de la red según ambos criterios, podemos distinguir una diferencia sutil de los datos del cuadro 4.5 (UNT) respecto a los del cuadro 4.4 (UNS) y es que para datos UNT y n= 1000, 4000, ambos métodos muestran el mismo orden en los genes, y si vemos los otros casos de n (n=2000,3000,5000,6000) podemos ver que los genes que aparecen y el orden en el que se encuentran sí varían en cuanto a sus posiciones. Este último hecho se observa también para los datos de UNS.

De forma más específica podemos obtener nuevamente los Ensembl IDs para n=6000, tanto para los datos de UNS como para los de UNT, además de sus Gene symbols, tal como se hizo en el análisis de la matriz de grado. Estos resultados pueden observarse en el cuadro 4.6.

Dada la clara relevancia del gen TUBG2 tanto en el análisis de la matriz de grado como en la centralidad de vector propio, y por el comportamiento que tiene en particular este gen, el cual puede ser apreciado en el cuadro 4.2 (recuérdese que TUBG2 tiene asociado el número de gen 465) ,resulta pertinente hacer un análisis más exhaustivo de este y los otros genes con un comportamiento similar.

Valor de n	Análisis con la matriz de grado	Análisis de centralidad de vector propio
1000	465, 221, 555, 365, 732, 373, 346, 514, 66, 166	465, 221, 555, 365, 732 373, 346, 514, 66, 166
2000	465, 1716, 365, 1540, 166, 1467 , 358, 1430, 1299, 1631	465, 1716, 365, 1540, 358 , 1430, 1467, 1299, 166, 1631
3000	465, 2457, 2005, 1716, 2096, 365 , 2285, 166, 1631, 1467	465, 2457, 2005, 1716, 365, 2096, 2285, 1631, 166 , 1299
4000	465, 3742, 365, 3766, 2457, 3866, 1540, 894, 3067, 3383	465, 3742, 365, 3766, 2457, 3866, 1540, 894, 3067, 3383
5000	465, 3742, 365, 3766, 3866, 2457, 3241, 1540, 894, 3383	465, 3742, 365, 3766, 3866, 3241, 1540, 2457, 1631, 3383
6000	465, 3742, 365, 3766, 5436, 3866, 2457, 1631, 3241, 3067	465, 3742, 365, 5436, 3766, 3866, 1631, 3241 ,2457, 3067

Cuadro 4.5: Comparación entre el método de la matriz de grado y la centralidad de vector propio para diferentes subgrafos de la matriz de adyacencia para los datos de UNT. Los valores están ordenados de mayor a menor. Las etiquetas coloreadas corresponden a genes que han cambiado de posición de relevancia al usar otro método de medida de centralidad.

4.3. Comparación con la literatura oncológica

Los números por sí mismos no nos pueden brindar información útil si no contamos con un resultado experimental con el cual podamos comparar nuestras predicciones numéricas, por ello se describirá una comparativa con resultados hallados vía métodos experimentales que serán de especial utilidad para encontrarle sentido a estas predicciones. El estudio de cada gen individual en diferentes estudios podría ser una ardua tarea, sin embargo presentamos los casos más relevantes, pero cada análisis particular puede extenderse al estudio de una mayor cantidad de genes.

4.3.1. La importancia del gen TUBG2

Los microtúbulos están involucrados en varios procesos celulares, incluido el mantenimiento de la forma celular, la motilidad[9] celular y la formación del huso

[9]Facultad de moverse que tiene la materia viva como respuesta a ciertos estímulos.

Datos	Ensembl IDs (prefijo ENSG)	Gene Symbols
UNS	00000077721, 00000011021, 00000055917	UBE2A, CLCN6, PUM2
	00000086589, 00000033178, 00000013810	RBM22, UBA6, TACC3
	00000066468, 00000082512, 00000003436	FGFR2, TRAF5, TFPI
	00000010244	ZNF207
UNT	00000037042, 00000114956, 00000023330	TUBG2, DGUOK , ALAS1
	00000131171, 00000115170, 00000115839	SH3BGRL, ACVR1, RAB3GAP1
	00000090061, 00000110172, 00000103512	CCNK, CHORDC1, NOMO1
	00000108561	C1QBP

Cuadro 4.6: Genes con mayor relevancia estadística para datos UNT, y UNS según la centralidad de vector propio para n=6000.

mitótico [10]. En las células, la nucleación[11] de los microtúbulos está estrictamente regulada y los sitios de nucleación se denominan centros organizadores de microtúbulos (MTOC). Como un importante MTOC en células de mamíferos, el centrosoma regula la formación de la red de microtúbulos citoplásmicos durante la interfase y define el número y la posición de los polos del huso durante la mitosis. A nivel molecular, una proteína altamente conservada, la γ-tubulina, sirve como sitio de unión para el dímero α/β -tubulina. En los mamíferos, la γ-tubulina, una proteína clave en la nucleación de microtúbulos, está codificada por dos genes, **TUBG1** y **TUBG2**. La γ-tubulina se localiza en los centrosomas y es capaz de nuclear microtúbulos. Se ha demostrado además que las anomalías del centrosoma son una de las principales características del cáncer y conducen a la inestabilidad cromosómica y la progresión maligna. La sobreexpresión de γ-tubulina también se ha informado en algunos cánceres y, por lo tanto, se cree que su expresión está estrechamente relacionada con la oncogénesis [23]. En el experimento de Tsubasa Ohashi [23] para examinar si alguna línea celular de cáncer humano[12] expresa γ-tubulin2 (TUBG2) además de γ-tubulin1, buscaron detectar la expresión proteica de cada γ-tubulina mediante la técnica SDS-PAGE[13] y usando transferencia Western encontraron que γ-tubulin2 se expresa ectópicamente en algunas líneas de celulas cancerosas.

[10] Es el conjunto de microtúbulos que brotan de los centriolos durante los procesos de división celular, sea mitosis (huso mitótico) o meiosis (huso acromático o meiótico).

[11] La nucleación se utiliza para finalizar la fase crítica en el montaje de una estructura polimérica, como un microtúbulo.

[12] Ver apéndice A.

[13] Es una técnica ampliamente utilizada en bioquímica, genética, biología molecular y ciencia forense para separar las proteínas de acuerdo a su movilidad electroforética.

Adicionalmente, en el experimento de Yun Niu[25], debido a la evidencia que existía previa a este artículo sobre anomalías centrosómicas en varios tipos de cáncer buscaron determinar si las disfunciones del centrosoma ocurren en la secuencia de hiperplasia ductal atípica (**ADH**) -carcinoma del cáncer de mama-. Como las α y γ-tubulinas son los componentes estructurales de los centrosomas, mediante métodos experimentales determinaron la expresión de proteínas de α y γ-tubulinas. Encontraron que el ARNm de γ-tubulina se expresaba cada vez más a partir de tejido mamario normal (NBT) a ADH, carcinoma ductal in situ (DCIS) y carcinoma ductal infiltrativo (IDC), respectivamente, con las mayores expresiones en DCIS. Al final sus resultados demostraron que las aberraciones del centrosoma pueden desempeñar un papel clave en la etapa inicial de la tumorogénesis mamaria.

En contraste con el actual trabajo, es indudable la utilidad de los métodos matemáticos, y en específico del análisis espectral, pues dados los resultados a los que se llegaron, se determinó, para el caso específico de datos de tejido tumoral RNA-seq, cuyo sitio primario es "uterus-nos", que dada la matriz de adyacencia de información mutua, mediante el análisis tanto de la **matriz de grado** como de la **centralidad de vector propio**, el gen TUBG2 siempre mostraba una **alta conectividad** en las sub-gráficas de la gráfica completa asociada a la matriz de adyacencia original y un mayor **flujo de información** respectivamente, sin importar qué tan grandes o pequeñas fueran estas, en términos del valor de n(véase el primer párrafo de la sección 4.1). El método del análisis espectral puede hacer que grandes cantidades de datos puedan ser analizadas con relativa rapidez y obtener información que puede serle útil a los expertos de la oncología y biólogos moleculares del cáncer, y llegar a resultados análogos a los que se podrían encontrar con los métodos convencionales experimentales. Así, podemos corroborar que los métodos computacionales proporcionan resultados que son congruentes con trabajos publicados anteriormente.

4.3.2. La importancia del gen ALAS1

Según los cuadros 4.3 y 4.6 ALAS1 es el quinto gen con mayor conectividad y el tercer gen con mayor flujo de información respectivamente. ALAS1(delta-aminolevulinato sintasa) codificada por el gen ALAS1 es la enzima que limita la velocidad de la biosíntesis de hemo, que participa en numerosas funciones celulares y que según Zhao Y. [26] tiene un efecto significativo sobre el cáncer de pulmón de células no pequeñas (**CPCNP**). Los productos de degradación del hemo [14] se expresan en gran medida en los tejidos tumorales y desempeñan un papel importante en el desarrollo del tumor. En el experimento de Yalei Zhao [26] tras la

[14] El grupo hemo es un grupo prostético que forma parte de diversas proteínas, entre las que destaca la hemoglobina.

inhibición de la actividad de ALAS1, la capacidad de proliferación, formación de colonias y migración de células de CPCNP se redujo significativamente.

Adicionalmente en dicho trabajo encontraron que la expresión de ALAS1 se incrementó en los tejidos del **CRC** (cáncer colorrectal). Por lo tanto, plantearon la hipótesis de que ALAS1 juega un papel fundamental en el desarrollo y la metástasis del CRC. Este último hecho se puede resumir en la gráfica mostrada en la figura 4.3.

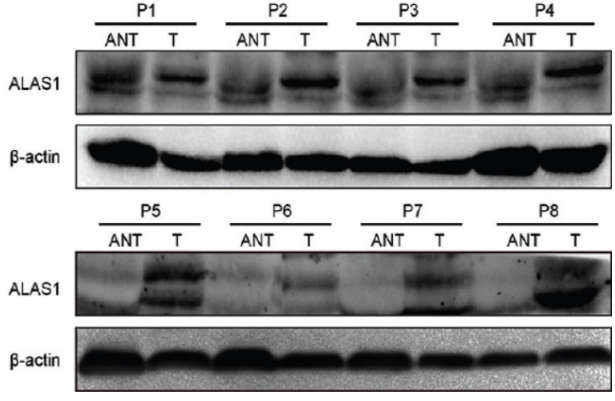

Figura 4.3: Western blot(véase apéndice el A) de ocho pacientes con CCR(cáncer colorectal) emparejados. Imagen tomada de [26].

En ella puede observarse que la presencia o ausencia, así como la densidad de banda está correlacionada con la presencia o ausencia de la proteina en cuestión. Una banda más gruesa está relacionada con una mayor señal lo cual hace referencia a una mayor concentración. Además, para hacer western blot se requiere de un control de expresión o proteína de referencia, es decir, es necesario comparar la expresión de nuestra proteína con alguna otra que se exprese de manera continua o constitutiva. En este caso podemos obervar la expresión de ALAS1 en 8 diferentes pacientes con CRC. En cada paciente aparecen un par de de gráficas en donde se muestran dos diferentes tipos de tejido. **ANT** significa tejido colónico normal adyacente, mientras que la letra T hace referencia al tejido tumoral. Así, comparando ambas bandas, la de ANT y T para cada uno de los pacientes, es fácil observar que en todos los casos la proteína ALAS1 se expresa en mayor cantidad en la banda correspondiente al tejido tumoral, pues en todos los casos las bandas en T son más gruesas que en ANT mostrando por lo tanto una mayor expresión de ALAS1.

Otra forma de representación gráfica usada en el ámbito de los estudios del cáncer es la famosa representación gráfica de Kaplan- Meier mostrada en la figura

4.4. En ella se puede observar que la probabilidad de supervivencia promedio fue

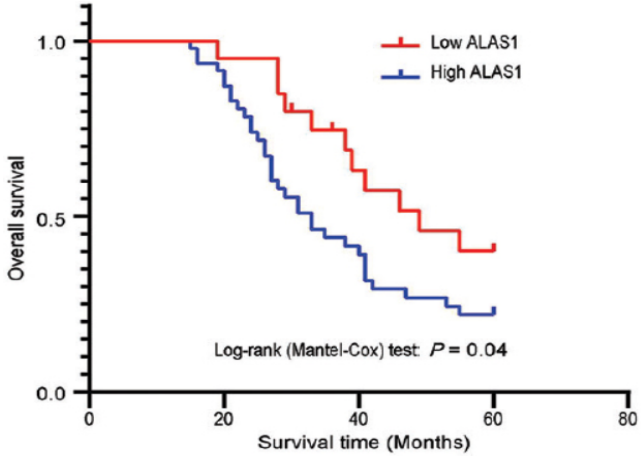

Figura 4.4: Los pacientes con alta expresión de ALAS1 mostraron tiempos de supervivencia reducidos en comparación con los pacientes con baja expresión de ALAS1. Imagen tomada de [26].

mayor para los pacientes que mostraron una **menor expresión** de la proteína ALAS1, y caso contrario, en los pacientes en los que se mostraba una menor probabilidad de supervivencia se mostraba una **mayor expresión** de la proteína ALAS1. Por ejemplo, según la gráfica de la figura 4.4, para pacientes con una mayor expresión de ALAS1 (línea azul) aproximadamente el 30 por ciento de ellos, logró un tiempo de supervivencia de 40 meses, sin embargo para pacientes con menor expresión de la proteína ALAS1 (línea roja) aproximadamente el 60 por ciento de ellos logró el mismo tiempo se supervivencia de 40 meses.

Dentro de estos experimentos también se transfectó si-RNA (ARN pequeño de interferencia) en células HCT116 (que representa células de CRC)con el objetivo de anular ALAS1. Después de la transfección durante 48 horas, el análisis vía western blot mostró que la expresión de ALAS1 en las células HCT116 se redujo significativamente, esto lo podemos observar en los gráficos western blot de la figura 4.5. En ellos se puede observar que el ARN siRNA-ALAS1#2 fue mas efectivo que el siRNA-ALAS1#1 para reducir la expresión de ALAS1 en células HCT116, pues el grosor de las bandas se vio reducido al usar siRNA-ALAS1#2 en comparación con siRNA-ALAS1#1.

Con lo anterior podemos ver cómo existen en la actualidad métodos experimentales con los cuales inhibir los efectos malignos que tienen ciertos tipos de genes en las células sanas y que provocan la aparición o proliferación de ciertos

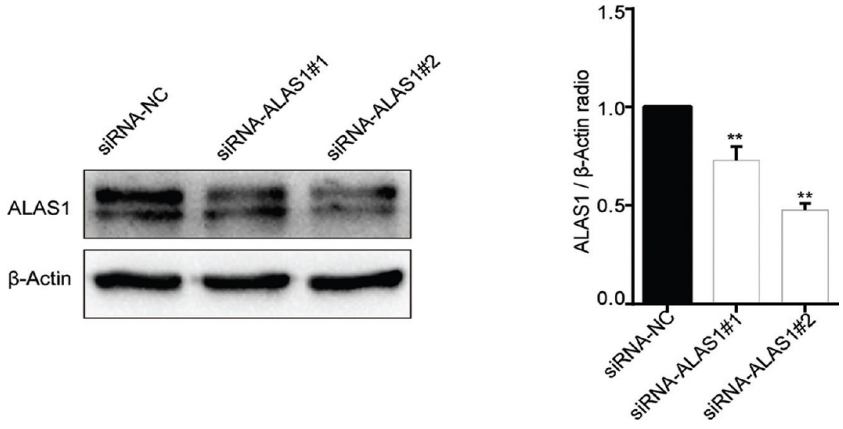

Figura 4.5: Gráficos western blot de la eficacia de si-ARN en la eliminación de ALAS1 en células HCT116. Imagen tomada de [26].

tipos de cáncer. Es por eso que es tan útil saber cuáles son los genes subyacentes en el desarrollo de un cáncer específico, y ahí es donde entra la relevancia de los métodos espectrales aplicados en las redes de regulación genética en cáncer, el objetivo presente de esta tesis.

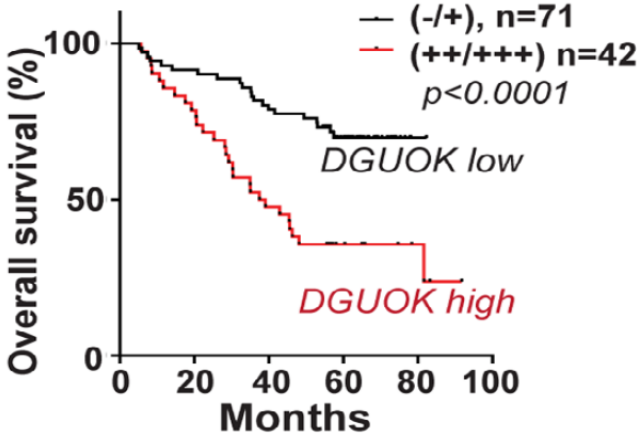

Figura 4.6: La correlación entre los niveles de expresión de DGUOK y la tasa de supervivencia global en pacientes con adenocarcinoma de pulmón. Imagen tomada de [27].

4.3.3. La importancia del gen DGUOK

Según los cuadros 4.3 y 4.6, DGUOK es el octavo gen con mayor conectividad y el segundo gen con mayor flujo de información respectivamente. Es pertinente entonces, como en los casos anteriores, consultar artículos en donde se ponga de manifiesto la influencia de este gen en particular para la proliferación de cierto tipo de cáncer, pues no es coincidencia que según los genes relevantes que favorecen al cáncer cervico-uterino vía el análisis espectral de grafos descrito hasta ahora en el presente capítulo, también aparezcan en otros estudios importantes en dónde son genes desicivos en diferentes tipos de cáncer como en el de mama en el caso de TUBG2 o el cáncer colorectal en el caso del gen ALAS1.

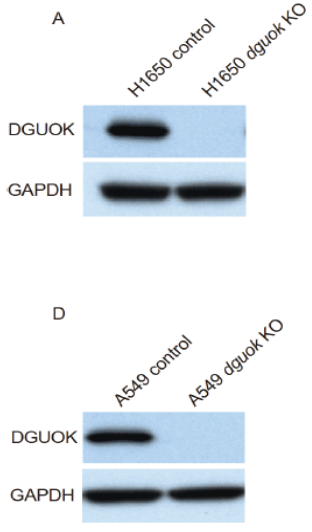

Figura 4.7: A: Western blot mostró que dguok estaba completamente eliminado en H1650;D: Western blot mostró que dguok fue eliminado por completo en A549. Imagen tomada de [29].

Se cree que un subconjunto de células de ciclo lento con propiedades similares a las de las células madre denominadas células madre cancerosas (CSC) son responsables del inicio del tumor y la recurrencia local o metastásica en cáncer de pulmón. La fosforilación oxidativa [15] mitocondrial (OXPHOS) es crucial para la autorrenovación de CSC en cáncer de pulmón, glioblastoma y leucemia. El ADN

[15] Proceso metabólico que utiliza energía liberada por la oxidación de nutrientes para producir adenosina trifosfato (ATP).

mitocondrial humano (mtADN) codifica 13 proteínas de la cadena respiratoria mitocondrial esenciales para OXPHOS, además la replicación del mtADN es esencial para que la célula reponga las mitocondrias dañadas y mantenga la funcionalidad mitocondrial. DGUOK es crucial para la fosforilación[16] de profármacos análogos de nucleósidos antivirales y antileucémicos como la forodesina y la nelarabina.

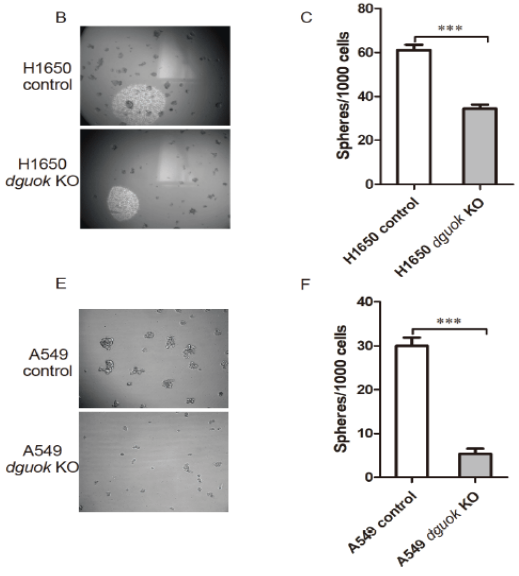

Figura 4.8: B: La eliminación de DGUOK inhibió la formación de esferas de células H1650; C: Los datos cuantifican que el knockout de DGUOK inhibe la formación de esferas celulares H1650 ; E: La desactivación de DGUOK inhibió la formación de esferas de células A549 ; F: Los datos cuantificaron que la desactivación de DGUOK inhibía la formación de esferas de células A549. Imagen tomada de [29].

Shengchen Lin y su equipo [27] demostraron que sus datos indican que DGUOK se sobreexpresa en pacientes con adenocarcinoma de pulmón y los niveles de expresión de DGUOK, se correlacionan fuertemente con la supervivencia de pacientes con adenocarcinoma de pulmón.

Los pacientes con alta expresión de DGUOK tienen una supervivencia global más corta en comparación con aquellos con baja o ninguna expresión de DGUOK, esto puede verse de forma más clara en la gráfica de la figura 4.6.

En otro estudio llevado a cabo por Rui Sun [29] se seleccionaron las células pulmonares H1650 y A549 (células de cáncer de pulmón de células no pequeñas

[16] Es la adición de un grupo fosfato a cualquier otra molécula

) como sujetos experimentales, y se eliminó el gen DGUOK vía la tecnología CRISPR/Cas9 . A través de la verificación de transferencia Western (Figura 4.7 A y D), se encontró que DGUOK no pudo detectar la señal en las células knockout[17], lo que indica que el experimento construyó con éxito el knockout DGUOK. Además de esto, Rui Sun y su equipo a través del experimento de formación de esferas, encontraron que el knockout de DGUOK redujo significativamente la formación de esferas de células (Figura 4.8. B, C, E y F). Esto indica que la falta de DGUOK afecta la autorrenovación de las células madre cancerosas. Habilidad, que indica que DGUOK tiene una relación importante con la aparición y desarrollo de tumores.

En este mismo estudio, con un microscopio confocal láser de inmunotinción se tomaron fotografías y los resultados mostraron que la pérdida de DGUOK provocó que las mitocondrias se acortaran y fragmentaran, y que las mitocondrias se concentraran alrededor del núcleo (Figura 4.9). Adicionalmente, con el fin de verificar el efecto de la pérdida de DGUOK sobre la energía mitocondrial, se realizó el Seahorse Assay [18] y se encontró que en las células H1650, la pérdida de DGUOK inhibía significativamente la tasa de consumo de oxígeno (OCR) de las mitocondrias. Esto indica que DGUOK desempeña un papel clave en la regulación del metabolismo respiratorio mitocondrial de las células del cáncer de pulmón. También muestra que DGUOK está involucrado en la regulación de la función mitocondrial, y que la desactivación genética DGUOK puede influir en la aparición de tumores al inhibir el metabolismo respiratorio de las mitocondrias.

En el repositorio mencionado al inicio de este capítulo, en los archivos AnalysisOf_Degree Matrix_UNT_2.0, eig_vec_cent_ analysis_UNT_2.0, AnalysisOf_Degree Matrix_UNS_2.0 y eig_vec_cent_analysis_UNS_2.0 se encuentran todos los cálculos hechos para esta sección con todo detalle.

4.4. Análisis de la conectividad algebraica

Como se discutió en el capítulo 3, la conectividad algebraica juega un papel altamente relevante en el estudio de un grafo. A continuación se mostrarán diferentes valores que se obtuvieron para $a(G)$ y $\delta(G)$[19], tanto para tejido sano como para tejido tumoral , recuerdese que estos son: la conectividad algebraica y el menor grado respectivamente. Recuérdese además que el valor de n no es más que el número de filas y columnas que se toman de la matriz original, es decir, si n =1000 estaremos analizando una submatriz de la matriz de adyacencia de 1000 filas por 1000 columnas.

[17] Células a las que se les ha suprimido la expresión de un gen específico.
[18] Los analizadores Seahorse miden la tasa de consumo de oxígeno (OCR)
[19] Véase la sección 3.3.3

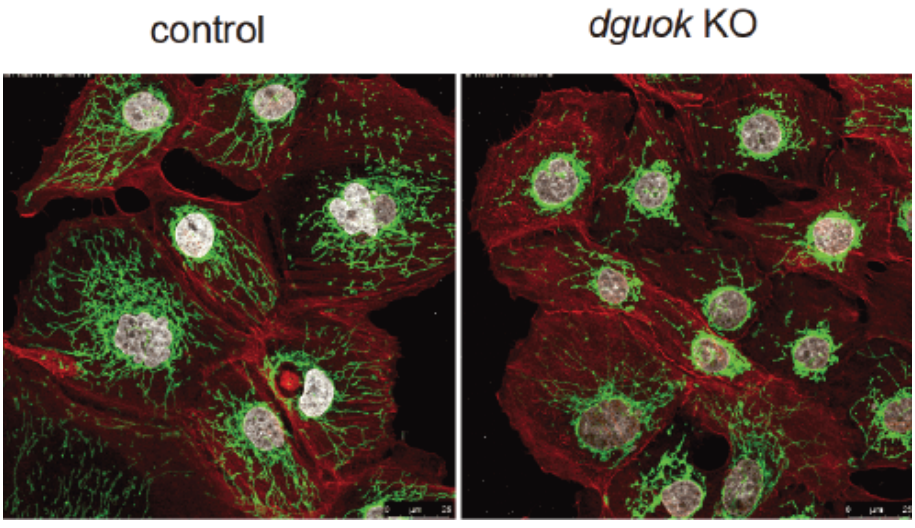

Figura 4.9: La deleción de DGUOK afecta la morfología mitocondrial. Imagen tomada de [29].

Ahora debemos recurrir a la proposición 3.3.4 de la sección 3.3.2, recordando el resultado $a(G) \leq k(G) \leq k'(G) \leq \delta(G)$ y con ellos podremos analizar los cuadros 4.4 a 4.7. En ellos notamos, inmediatamente, otra diferencia en la topología de la red entre aquella que corresponde al tejido sano y aquella que corresponde al tejido tumoral.

Conectividad algebrica en sub-grafos de datos UNT

Tomando números redondos en los cuadros 4.7 y 4.8 observamos que la diferencia entre los valores de $a(G)$ y $\delta(G)$ para cada valor de n siempre es de 2. Por ejemplo, para n=1000 observamos que se cumple la desigualdad:

$$284 \leq k(G) \leq k'(G) \leq 286$$

para n=2000 se cumplirá la desigualdad

$$588 \leq k(G) \leq k'(G) \leq 590$$

esta tendencia continua para los demás valores de n.

Valor de n	Valor de $a(G)$
1000	284.49
2000	588.31
3000	895.21
4000	1212.13
5000	1516.10
6000	1772.31

Cuadro 4.7: Conectividad algebraica para diferentes subgrafos de la matriz de adyacencia para los datos de UNT

Valor de n	Valor de $\delta(G)$
1000	286
2000	590
3000	897
4000	1214
5000	1518
6000	1774

Cuadro 4.8: Menor grado para diferentes subgrafos de la matriz de adyacencia para los datos de UNT

Usando las propiedades de orden y el hecho de que $k(G), k'(G) \in \mathbb{N}$, podemos determinar los valores tanto de $k(G)$ como de $k'(G)$. Por ejemplo, tomemos para n= 1000 $k(G) = 284$ se deberá cumplir la desigualdad

$$284 \leq k'(G) \leq 286$$

pero $k'(G) \in \mathbb{N}$, esto implica que necesariamente $k'(G) = 285$. Usando este mismo procedimiento con los demás valores de n, se obtuvieron los datos del cuadro 4.11.

Usando la misma lógica, supongamos ahora el otro caso posible en el que $k(G) \neq 284$, esto implicaria que necesariamente $k(G) = 285$ ya que $k(G) \in \mathbb{N}$ y por lo tanto debe cumplirse la desigualdad

$$285 \leq k'(G) \leq 286$$

lo cual es imposible, pues no existe un número natural tal que cumpla esta última condición de desigualdad.

Valor de n	Valor de $a(G)$
1000	168.52
2000	291.13
3000	437.18
4000	589.15
5000	738.17
6000	866.11
7000	1027.07

Cuadro 4.9: Conectividad algebraica para diferentes subgrafos de la matriz de adyacencia para los datos de UNS

Valor de n	Valor de $\delta(G)$
1000	170
2000	292
3000	438
4000	590
5000	739
6000	867
7000	1028

Cuadro 4.10: Menor grado para diferentes subgrafos de la matriz de adyacencia para los datos de UNS

Conectividad algebrica en sub-grafos de datos UNS

Nuevamente, tomando números redondos, observamos que, en este caso, la diferencia entre $a(G)$ y $\delta(G)$ para cada valor de n siempre es de 1. Por ejemplo, comencemos tomando n=1000 y veamos que se cumple la desigualdad:

$$169 \leq k(G) \leq k'(G) \leq 170$$

para n=2000 se cumplirá la desigualdad:

$$291 \leq k(G) \leq k'(G) \leq 292$$

Aquí podemos hacer un análisis de orden sencillo que arrojará un dato interesante. Nuevamente dado que $k(G), k'(G) \in \mathbb{N}$, tomando la primera de las dos desigualdades anteriores, la única forma en que se puede cumplir la desigualdad $k(G) \leq k'(G)$, es que $k(G) = 169$ y $k'(G) = 170$. De manera que para los datos de **tejido sano** vemos que se cumple el hecho particular de que

Valor de n	k(G)	k'(G)
1000	284	285
2000	588	589
3000	895	896
4000	1212	1213
5000	1516	1517
6000	1772	1773

Cuadro 4.11: Valores de conectividad de vértice ($k(G)$) y de borde ($k'(G)$) para diferentes valores de n para los datos UNT

$$a(G) = k(G) \tag{4.1}$$

y

$$\delta(G) = k'(G) \tag{4.2}$$

esto es, la conectividad algebraica es igual a la conectividad de vértice, y el menor de grado es igual a la conectividad de borde. Este hecho se puede resumir en la siguiente tabla:

Valor de n	k(G)	k'(G)
1000	169	170
2000	291	292
3000	437	438
4000	589	590
5000	738	739
6000	866	867
7000	1027	1028

Cuadro 4.12: Valores de conectividad de vértice ($k(G)$) y de borde ($k'(G)$) para diferentes valores de n para los datos UNS

que al comprarar con los cuadros 4.9 y 4.10 vemos que realmente se cumplen las expresiones en (4.1) y (4.2).

Intuitivamente, la conectividad de vértices nos proporciona una idea de la cantidad total de genes que son realmente importantes dentro de la red, sin los cuales, la fenomenología de la regulación génica no sería apreciada. Esta información es útil para saber por ejemplo, en los cálculos hechos en la sección 4.1, qué cantidad de genes tomar como los más relevantes pues, en un inicio solo se tomaron los

tres genes más relevantes(aquellos que tenían un mayor grado), esto se hizo por dos razones; la primera, que aún no se conocía un criterio para seleccionar aquellos genes que serían dominantes; y segundo, que al hacer este tipo de cálculos en Jupyter Notebook, la salida en realidad contiene miles de resultados, pero en este caso se muestran *outputs* como las de la figura 4.10.

```
In [24]: func(get_elem_in_diag(degreeMat_from_Adj(A,10000)),10000)
Out[24]: array([ 465, 7948, 7261, ..., 5143, 9116, 8300])
```

Figura 4.10: Array con los valores de la diagonal de la matriz de grado, ordenados de mayor a menor, según su posición.

Con la función `get_genes(data,n)`[ ][47] que hemos contruido podemos obtener las etiquetas del gen con los números de la lista de la figura 4.10.

Así, con el concepto de la conectividad de vértices , podemos mejorar la confiabilidad de los cálculos hechos con la matriz de grado.

4.5. Análisis de la matriz laplaciana sin signo

En el repositorio ya antes citado se encuentran otros dos archivos más: SignlessLapMatAnalysys_UNS.ipynb y SignlessLapMatAnalysys_UNT.ipynb en donde se hace el análisis del polinomio característico de la matriz laplaciana sin signo usando la proposición 3.4.2 , la cual nos proporciona información sobre el número de aristas que contiene el grafo, así podemos saber que porcentaje de interacción diferencia a los datos de UNS y UNT y cuantificar la participación global de los genes en el proceso de regulación genética.

En las siguientes tablas se resumen los resultados obtenidos de estos análisis, donde n será el número de filas y columnas tomadas de la matriz de adyacencia original de 16748 $\times$ 16748 y $N_{a,s}$ será el número de aristas calculado para los datos de tejido sano UNS, y $N_{a,t}$ para los datos de tejido tumoral UNT.

Al ver los resultados de las anteriores tablas podemos vislumbrar nuevamente un contraste entre ambos casos; se pone de manifiesto la antes subyacente inestabilidad genética, característica del cáncer, pues ahora poseemos varios resultados que nos hacen reconocer la evidente diferencia entre ambos conjuntos de datos.

Los oncólogos clínicos y moleculares afirman que, el crecimiento y la supervivencia de las células tumorales, que constituyen las principales características del cáncer, pueden verse fuertemente restringidas por la inactivación de un solo oncogén [21]. De ahí la importancia del análisis de la conectividad algebraica y en este caso, del número de aristas total del grafo, que podría proporcionarnos

Valor de n	$N_{a,s}$
1000	205,660
2000	822,904
3000	1,847,626
4000	3,284,893
5000	5,105,175
6000	7,323,408
7000	9,968,445
8000	13,017,554

Cuadro 4.13: Número de aristas para datos de UNS para diferentes valores de n

Valor de n	$N_{a,t}$
1000	210,363
2000	842,622
3000	1,892,339
4000	3,367,148
5000	5,258,967
6000	7,557,680
7000	10,284,562
8000	13,413,487

Cuadro 4.14: Número de aristas para datos de UNT para diferentes valores de n

un criterio más, para depurar información arrojada por el análisis de la matriz de grado y de la centralidad de vector propio, y elegir aquel conjunto de genes más destacados dentro de la red.

Ajuste no lineal de datos de número de aristas para datos UNS y UNT

Dado el alto costo computacional que se requiere para hacer los cálculos mostrados en los cuadros 4.13 y 4.14, debe pensarse en alternativas para reducirlo . Uno de ellos es el uso de métodos estadísticos para predecir resultados futuros de la muestra ; esto mediante el uso del ajuste no lineal de curva, que puede hacerse fácilmente con ayuda del método **curve_fit** de la librería **SciPy** de Python. Por ejemplo, para los datos de UNS , graficando además de los datos del cuadro 4.13, otros 42 puntos que fueron calculados necesariamente para tener un conjunto de datos con los cuales poder hacer estadística, se obtuvo la gráfica de la figura 4.11.

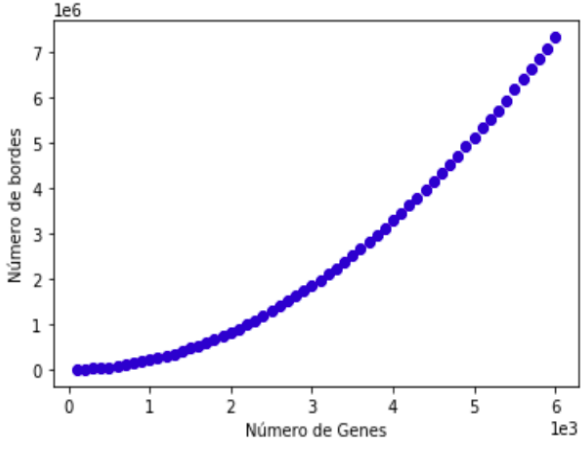

Figura 4.11: Número de bordes calculados, según el número de genes tomados para datos UNS

Dada la forma de la gráfica se propuso ajustar la curva con una función del tipo

$$f(x) = ax^2 - bx^3 + bx^4; \;\; a, b \in \mathbb{R} \tag{4.3}$$

con curve_fit podemos determinar el valor de a y b de tal forma que se ajusten a

los datos de la figura 4.11. Se encontró que:

$$a = 2.064390 \times 10^{-1} \tag{4.4}$$

$$b = -8.608490 \times 10^{-11} \tag{4.5}$$

con estos valores ya es posible dibujar un ajuste no lineal a la curva, este puede apreciarse en la figura 4.12.

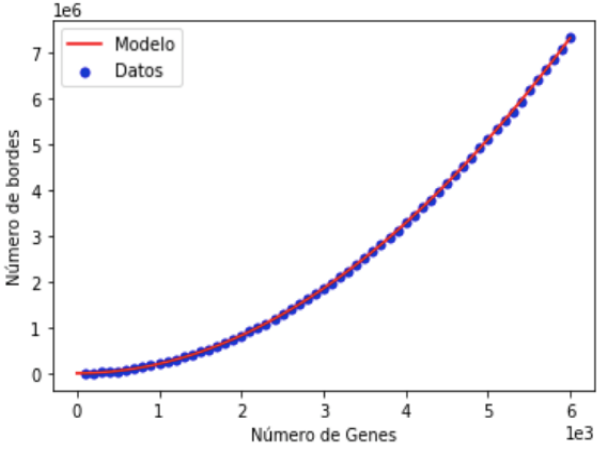

Figura 4.12: Número de bordes calculados, según el número de genes tomados para datos UNS

por si no bastara la evidencia gráfica de que la función dada por la expresión (4.3) junto con los parámetros de (4.4) y (4.5), son un buen ajuste no lineal al *data set*, siempre está el uso del estadístico $\mathbf{R^2}$. Este se puede calcular con **r2_score** de la librería **sklearn**. Se obtiene para este caso que:

$$R^2 = 0.999587 \tag{4.6}$$

Con lo anterior, se observa que el ajuste del modelo a la variable que estamos intentando explicar es bastante bueno, pues explica el 99.9587 % a la variable real.

Para los datos de UNT se obtienen resultados análogos, estos se presentan a continuación.

Ahora, además de los datos del cuadro 4.14, se grafican otros 42 puntos para poder hacer estadística, y con los cuales se pudo hacer la gráfica de la figura 4.13.

Usando nuevamente la función (4.3) como modelo y usando curve_fit se determinan los valores de a y b para los datos de UNT los cuales fueron:

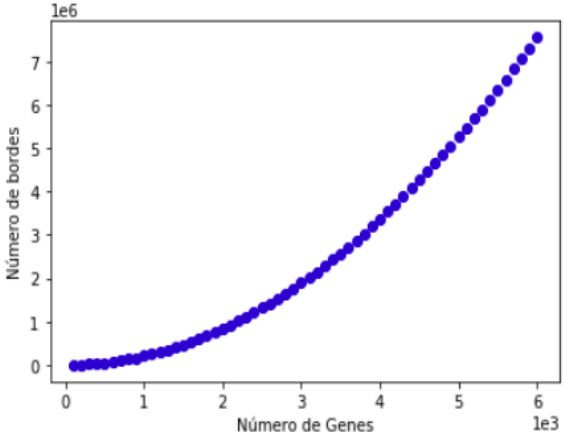

Figura 4.13: Número de bordes calculados, según el número de genes tomados para datos UNT

$$a = 2.109089 \times 10^{-1} \tag{4.7}$$

$$b = -2.909001 \times 10^{-11} \tag{4.8}$$

con estos valores ya es posible dibujar un ajuste no lineal a la curva. Este puede apreciarse en la figura 4.14.

Para este caso, el vcalor de R^2 fue:

$$R^2 = 0.999591 \tag{4.9}$$

Con ambos modelos, ahora es posible dibujar dos gráficas comparativas tomando los 16748 genes, esto es, el total de genes con los que se obtuvo la matriz de adyacencia.

Específicamente, para el caso de UNT se predice que usando el total de datos de 16748 genes, el grafo obtenido asociado a la matriz de adyacencia tendrá 56 870 398.97 bordes. Por otro lado, para los datos de UNS se predicen 51 132 651.97 bordes. Entre ambos resultados existe una diferencia colosal de 5 737 747.54 bordes. Esta diferencia se hace notar en la divergencia de la gráfica roja respecto a la azul en la figura 4.15 a medida que se toman cada vez más genes del total hasta acercarnos al cien por ciento de ellos. Con esto es evidente notar que la red global para los datos UNT es mucho más compleja que la de UNS.

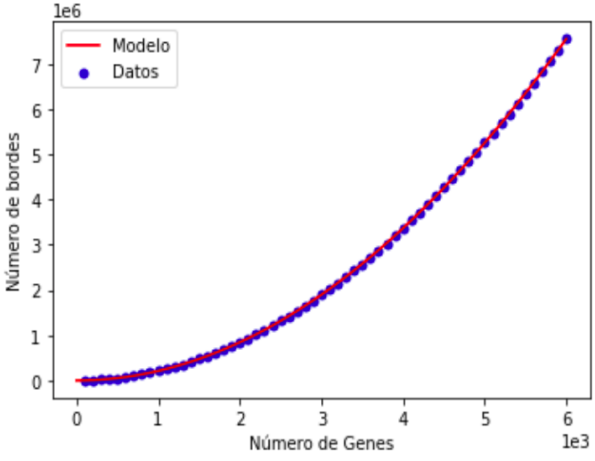

Figura 4.14: Número de bordes calculados, según el número de genes tomados para datos UNT

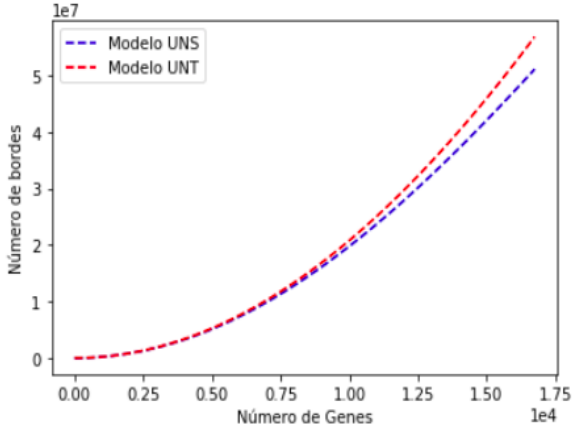

Figura 4.15: Comparación entre los ajustes no lineales obtenidos par los datos de UNS y UNT para el número de bordes

Capítulo 5

Conclusiones y perspectivas

En este trabajo se pudo verificar que la matriz de grado, a pesar de tener una base teórica muy sencilla e intuitiva resulta muy útil como un primer método para comprender de manera general la topología de la red de la matriz de adyacencia. Se pudo corroborar un notable contraste en los genes con mayor grado entre ambos tipos de datos (UNT y UNS), siendo el aspecto más relevante el que en los datos UNT se observó una alta relevancia del gen en la posición 465, el gen ENSG00000037042 (TUBG2), pues para distintos valores de n, se observó un claro dominio de este gen sobre los demás para cada ejecución realizada , algo que claramente no sucede en los datos de tejido sano UNS, pues cada cierto valor de n, el gen con mayor grado cambiaba.

Respecto al análisis con la centralidad de vector propio, es muy enriquecedora pues nos permitie observar que la matriz de grado solo sirve como una primera aproximación al estudio de la cantidad de conexiones que tiene un gen con los demás. La centralidad de vector propio hizo posible observar que puede haber genes que aparentemente, con el método de la matriz de grado, tenían mayor centralidad y resultaron después pasar a segundo plano, y de manera dual, estaban otros que aparentemente tenían baja centralidad y resultaron ser más relevantes en términos de centralidad.

Tanto el análisis de la matriz de grado como el de la centralidad de vector propio nos permitieron identificar genes que en otros estudios han mostrado ser cruciales en la génesis y proliferación de diferentes tipos de cáncer, como se muestra en las secciones 4.3.1, 4.3.2 y 4.3.3. En estos estudios previos no solamente se demostró la influencia que tienen esos genes, si no que además, mediante tecnologías de "edición" de genoma (CRISPR/Cas9) o la reducción de expresión génomica mediante ARN de interferencia, lograron inhibir la expresión de estos genes, observando una clara reducción en el crecimiento de tumores. Con esta evidencia,

podría un experto en estas técnicas experimentales usar nuestros resultados para utilizar estos métodos en muestras de células de cáncer cervico-uterino cuyo sitio primario es *Uterus NOS*, y poder inhibir el efecto de los genes antes estudiados.

En cuanto al análisis con la conectividad algebraica, al determinar los valores de esta y los grados menores para cada subgrafo se pudieron determinar, usando además axiomas de orden, los valores de $k(G)$ y $k'(G)$, estos son la conectividad de vértice y de borde respectivamente. Esto se hizo para ambos conjuntos de datos, UNT y UNS. En particular se encontró un resultado interesante respecto a los datos de tejido sano UNS, para este caso se encontró que la conectividad algebraica es igual a la conectividad de vértice y el menor grado es igual a la conectividad de borde, esto se aprecia en las igualdades (4.1) y (4.2).

Finalmente, el análisis de la matriz laplaciana sin signo proporciona una idea muy clara de la cantidad de bordes que involucra el analizar un sub-grafo determinado. Nuevamente este análisis a pesar de ser muy general, nos provee información sobre la topología de la red. En este punto ya es innegable la oculta inestabilidad genética en la red de UNT, pues la diferencia que existe en la cantidad total de aristas de la red completa, entre los datos de tejido sano y tumoral, según la predicción arrojada por el ajuste no lineal de datos de números de aristas, es de casi 6 000 000 de aristas, poniendo de manifiesto la mayor complejidad de la red de UNT respecto a la de UNS.

La teoría de grafos y en específico la teoría espectral de grafos, proporciona un marco computacional para modelar una variedad de conjuntos de datos, incluidos los que surgen de la genómica. Las redes de genes se pueden representar como gráficos de nodos (vértices) e interacciones (bordes) que pueden tener diferentes pesos. Este trabajo puede servir como pauta para crear un nuevo software que sea capaz de analizar grandes bases de datos y en el caso más idóneo debería ser lo suficientemente flexible como para analizar y visualizar redes con distintas interpretaciones en áreas como la biología, computación, economía y la sociología. En este trabajo, todos los cálculos numéricos fueron realizados en *Python*, siendo de especial relevancia el hecho de que todo el código detrás de los cálculos y visualizaciones son minimalistas, una de las grandes bondades de *Python*, pero al mismo tiempo eficientes ya que aun los más complicados cálculos se llevaron a cabo en un tiempo aún razonable.

Aún hay grandes interrogantes en lo que a la teoría de grafos respecta, y en cuánto a la matematización de ciencias como la biología, aún falta mucho por hacer, pero dado el avance vertiginoso que se ha dado en áreas como la genómica y los resultados que de ella emanan, la inversión de capital humano con científicos de diferente formación en esta área es de vital importancia para mejorar la salud a nivel mundial.

En la época de oro de la física, por las condiciones históricas que regían el mundo, a muchos científicos de la talla de A. Einstein y W.Heisenberg se les delegó

la tarea de encontrar nuevos y mejores métodos para destruir la naturaleza, en lugar de aprovecharla para mejorar el equilibrio entre el hombre y esta. Hoy y bajo el contexto de una pandemia que está arrasando con cientos de miles de vidas, es necesario que los esfuerzos de matemáticos, ingenieros, biólogos, físicos, informáticos y un largo etcétera, se unan para resolver los problemas más urgentes, que aquejan a tantas millones de personas alrededor del mundo.

Apéndice

Apéndice A

Glosario

Affected-sib-pair (ASP)

Definición: Los modelos affected-sib-pair son un enfoque popular para la detección de loci genéticos vinculados a un gen de una enfermedad cuando se desconoce el modo de herencia. Los métodos para el análisis de los datos affected-sib-pair generalmente estiman una función del alelo o haplotipo esperado que comparte idéntico por descendencia (IBD) en un locus marcador en los pares afectados

Alelos

Definición: Un alelo es cada una de las dos o más versiones de un gen. Un individuo hereda dos alelos para cada gen, uno del padre y el otro de la madre. Los alelos se encuentran en la misma posición dentro de los cromosomas homólogos. Si los dos alelos son idénticos, el individuo es homocigoto para este gen. En cambio, si los alelos son diferentes, el individuo es heterocigoto para este gen. Aunque el término alelo fue usado originariamente para describir variaciones entre los genes, ahora también se refiere a las variaciones en secuencias de ADN no codificante (es decir, que no se expresan).

Alfabeto finito

Definición: Un alfabeto es un conjunto finito, no vacío de símbolos. Convencionalmente, usamos el símbolo $\sum$ para un alfabeto. Alfabetos comunes incluyen:

1.- $\sum = \{0, 1\}$, el alfabeto binario.

2.- $\sum = \{a, b, ..., z\}$, el conjunto de todas las letras minúsculas [11].

Análisis univariado

Definición: Se describen las características de una variable por vez.También se le conoce como Estadística Descriptiva.

Árbol

Definición: Un árbol es un grafo simple unidireccional G que satisface la condición de que dos vértices cualesquiera de G están conectados por un único camino simple.

Factor de transcripción TF

Definición: En biología molecular, un factor de transcripción (TF) (o factor de unión a ADN específico de secuencia) es una proteína que controla la tasa de transcripción de información genética de ADN a ARN mensajero, al unirse a una secuencia de ADN específica. La función de los TF es regular (activar y desactivar) los genes para garantizar que se expresen en la célula correcta en el momento adecuado y en la cantidad correcta durante toda la vida de la célula y el organismo.

Falsos positivos

Definición: El error de tipo I o falso positivo, es el error que se comete cuando el investigador rechaza la hipótesis nula siento esta verdadera en la población.

Gráfico acíclico dirigido

Definición: Es un grafo dirigido que no tiene ciclos; esto significa que para cada vértice v, no hay un camino directo que empiece y termine en v.

Hibridación

Definición: La hibridación de ácidos nucleicos (ADN o ARN) es un proceso por el cual se combinan dos cadenas de ácidos nucleicos antiparalelas y con secuencias de bases complementarias en una única molécula de doble cadena, que toma la estructura de doble hélice, donde las bases nitrogenadas quedan ocultas en el interior. Esto hace que si irradiamos la muestra con la longitud de onda a la que absorben estas bases (260 nm), la absorción de energía será mucho menor si la cadena es doble que si se trata de la cadena sencilla, ya que en esta última los dobles enlaces de las bases nitrogenadas, que son las que captan la energía, están totalmente expuestos a la fuente emisora de energía.

Identical by descent (IBD)

Definición: Un segmento de ADN es idéntico por estado (IBS) en dos o más individuos si tienen secuencias de nucleótidos idénticas en este segmento. Un segmento de IBS es idéntico por descendencia (IBD) en dos o más individuos si lo han heredado de un ancestro común sin recombinación, es decir, el segmento tiene el mismo origen ancestral en estos individuos. Los segmentos de ADN que son IBD son IBS por definición, pero los segmentos que no son IBD aún pueden ser IBS debido a las mismas mutaciones en diferentes individuos o recombinaciones que no alteran el segmento.

Información

Definición: Sea E un suceso que puede presentarse con probabilidad $P(E)$. Cuando E tiene lugar, decimos que hemos recibido:

$$I(E) = \log \frac{1}{P(E)} \quad \text{unidades de información}$$

Si introducimos el logaritmo de base 2 , la unidad correspondiente se denomina bit (unidad binaria)

$$I(E) = \log_2 \frac{1}{P(E)} \quad \text{bits}$$

Líneas de células cancerosas

Definición: Células cancerosas que continúan dividiéndose y creciendo con el tiempo, bajo ciertas condiciones en un laboratorio. Las líneas de células cancerosas se utilizan en la investigación para estudiar la biología del cáncer y para probar tratamientos contra el cáncer.

Locus/Loci

Definición: 'Locus' es el término que usamos para decir dónde está localizado en un cromosoma un gen específico. Así que realmente es la ubicación física de un gen o de un polimorfismo del ADN en un cromosoma. Y es algo así como la dirección de una calle para la gente. Una de las formas en que podemos pensar cuando estamos hablando de genes y cromosomas, es que podemos comparar un cromosoma con un país, una región de un cromosoma tal vez sería una ciudad y, a continuación, si nos situamos en un área muy específica, que es el locus, sería equivalente a, por ejemplo, la dirección de una persona, su calle. Y esa es la dirección de ese gen. Y una cosa importante a recordar, el plural de 'locus' es 'loci', no 'locuses'.

Microarreglo o Micromatriz

Definición: Un chip de ADN (del inglés DNA microarray) es una superficie sólida a la cual se une una colección de fragmentos de ADN. Las superficies empleadas para fijar el ADN son muy variables y pueden ser de vidrio, plástico e incluso de silicona. Los chips de ADN se usan para analizar la expresión diferencial de genes, y se monitorizan de manera simultánea los niveles de miles de ellos. Su funcionamiento consiste, básicamente, en medir el nivel de hibridación entre la sonda específica (probe, en inglés), y la molécula diana (target), y se indican generalmente mediante fluorescencia y a través de un análisis de imagen, lo cual indica el nivel de expresión del gen.

Perfiles de expresión génica

Definición: El perfil de expresión génica es la medida de la actividad de miles de genes simultáneamente, para crear una imagen global de la función celular. Estos perfiles pueden, por ejemplo, distinguir entre las células que se están dividiendo activamente, o mostrar cómo las células reaccionan a un tratamiento en particular. Muchos experimentos de este tipo analizan un genoma completo simultáneamente, es decir, cada gen presente en una célula en particular.

Probabilidad clásica

Definición: La probabilidad p de que suceda un evento S de un total de n casos posibles igualmente probables es igual a la razón entre el número de ocurrencias h de dicho evento (casos favorables) y el número total de casos posibles n .

$$p = \text{Prob}(S) = \frac{h}{n}$$

RNA-Seq

Definición: ('secuenciación de ARN'), también llamado Secuenciación del Transcriptoma Entero para Clonación al Azar, utiliza la secuenciación masiva (NGS) para revelar la presencia y cantidad de ARN, en una muestra biológica en un momento dado. Así, la RNA-Seq se usa para analizar cambios en el transcriptoma.

Western blot

Definición: Un Western Blot se utiliza a veces para diagnosticar enfermedades. En el laboratorio a menudo queremos medir si una proteína específica se expresa en una muestra. Podemos hacer esto tomando el material de la muestra y correrlo en un gel, y luego transferir las proteínas resueltas en una pieza especial de una

membrana - de papel, si se quiere - y luego se expone el papel a una sonda que contine un anticuerpo contra la proteína específica de interés. Debido a que el anticuerpo está marcado con una molécula que podremos visualizar, podemos preguntarnos si la proteína de interés se expresa en esta muestra y tener una idea de la concentracion, así como la composicion y el tamaño es la proteína.

Apéndice B

Consultas hechas en el Data Portal - National Cancer Institute

Fecha de consulta: 19/11/2019
Controles :

Para *cervix uteri* la consulta fue: cases.primary_site in[“cervix uteri”] and cases.samples.sample_type in [“solid tissue normal”] and files.analysis.workflow_type= “HTSeq - Counts”

Para *uterus nos* la consulta fue: cases.primary_site in[“uterus,nos”] and cases.samples.sample_type in [”Solid Tissue Normal”] and files.analysis.workflow_type= “HTSeq - Counts”

Para *corpus uteri* la consulta fue: cases.primary_site in[“corpus uteri”] and cases.samples.sample_type in [“Solid Tissue Normal”] and files.analysis.workflow_type= “HTSeq - Counts”

Tumores:

Para *cervix uteri* la consulta fue: cases.primary_site in[”cervix uteri”] and files.analysis.workflow_type= “HTSeq - Counts”

Para *uterus nos* la consulta fue: cases.primary_site in[“uterus, nos”] and files.analysis.workflow_type= ”HTSeq - Counts”

Para *corpus uteri* la consulta fue: cases.primary_site in[”corpus uteri”] and files.analysis.workflow_type= “HTSeq - Counts”

Apéndice C

Un pequeño detalle técnico

Se usaron dos equipos diferentes para realizar la ejecución de los scripts. Las especificaciones técnicas se enuncian a continuación:

MacBook Air (principios de 2014)

Sistema operativo macOS Mojave
Intel Core i5 dual core de 1.4 GHz (Turbo Boost de hasta 2.7 GHz) con 3 MB de caché L3 compartido
4 GB de memoria integrada LPDDR3 de 1600 MHz
128 GB almacenamiento en flash basado en PCIe de 128 GB

PC armada.

Sistema operativo Windows 10 / Linux(Ubuntu)
Intel Core i5-8400 CPU 2.80GHz (6 CPUs)
16 GB RAM
SSD de 256 GB y HDD de 1 TB

En el caso de las ejecuciones que se hicieron en Mac, se pudo hacer el cálculo enunciado en la sección 4.1, con un valor de hasta n=10000, sin embargo en el segundo equipo no fue posible hacerlo si no hasta n = 8000, a pesar de que esta última tiene mejores especificaciones técnicas. Por esta razón se monitoreo el consumo de memoria RAM y memoria de almacenamiento. Se observó que al parecer Jupyter Notebook está mejor optimizado para sistemas operativos Mac Os, pues al hacer las ejecuciones de los scripts toma memoria de almacenamiento para no saturar la memoria RAM y evitar así errores de memoria mientras se hacen los cálculos, por otro lado, tanto en Ubuntu como en Windows, Jupyter Notebook

consume grandes cantidades de memoria RAM causando incluso el congelamiento de la pantalla y la imposibilidad de revertir el proceso, dejando como única opción reiniciar el equipo.

Así, Jupyter Notebook, en Mac OS, toma memoria de almacenamiento cuando la memoria RAM se ve revasada y tanto en Windows como en Ubuntu solo consume RAM y cuando esta se usa al máximo, simplemente detiene la ejecución y arroja un error de memoria.

Apéndice D

Teoremas y demostraciones relevantes

Teorema de Perron-Frobenius

Definición: Decimos que una matriz simétrica A es irreducible si no hay permutación ρ tal que:

$$A_\rho = \begin{bmatrix} X & Z \\ O & Y \end{bmatrix}$$

donde X e Y son matrices cuadradas. De lo contrario, se dice que A es una matriz reducible.

Lema: Si la matriz A no es negativa y A es irreducible, entonces $(I+A)^{n-1} > 0$, o equivalentemente, $I + A + A^2 + ... + A^{n-1} > 0$

Proposición: Si G es un gráfico conectado a n vértices y A es su matriz de adyacencia, entonces A es irreducible.

Teorema de Perron-Frobenius. Suponga que A es una matriz no negativa, irreducible y con valores propios $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. Entonces:

(i)$\lambda_1 > 0$ y hay un vector propio asociado positivo;

(ii) $\lambda_1 > \lambda_2$;

(iii) $|\lambda_i| \leq \lambda_1$, para todo $i \in \{1, 2, ..., n.\}$

$$H\left[u(Y)\right] = \log|\mathcal{Y}|$$

Demostración:

Dada la desigualdad de Shannon-Gibbs: para toda $P = (p_1, p_2, ..., p_n) \in \Delta_n$ y $Q = (q_1, q_2, ..., q_n) \in \Delta_n$ se tiene que:

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i, \tag{D.1}$$

la igualdad se cumple si y sólo si $p_i = q_i$, $\forall i$ o si y sólo si $P = Q$ (siempre que $q_i = 0$ para algún i, el correspondiente p_i es también cero).

Como p es cualquier función de densidad de probabilidad sobre $\{x_1, ..., x_n\}$, con $p_i = p(x_i)$. Dejando $q_i = 1/n$ (que es la densidad de probabilidad unforme para el caso discreto) para todo i,

$$-\sum_{i=1}^{n} p_i \log q_i = \sum_{i=1}^{n} p_i \log n = \log n, \tag{D.2}$$

lo anterior es válido pues por los axiomas de Kolmogorov se sabe que $\sum_{i=1}^{n} p_i = 1$. Por transitividad de (1) con (2) tendremos que:

$$H\left[u(Y)\right] = H(p_1, p_2, ..., p_n) \leq H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}) = \log n$$

la anterior desigualdad será igualdad cuando $p_i = \frac{1}{n}$ en el lado izquierdo de la desigualdad, por lo tanto:

$$H\left[u(Y)\right] = \log n \tag{D.3}$$

pero n se puede escribir en términos de la cardinalidad de $\mathcal{Y}$ como $|\mathcal{Y}| = n$ por lo que (D.3) se puede reescribir como

$$H\left[u(Y)\right] = \log |\mathcal{Y}|$$

■

Apéndice E

Cómo usar ARACNE

Preparando el archivo de entrada

El primer paso para usar ARACNE es importar datos [39]. Actualmente, ARACNE solo lee archivos de texto delimitados por TAB en un formato particular, que se describe a continuación. Dichos archivos se pueden crear y exportar en cualquier programa de hoja de cálculo estándar, como Microsoft Excel.

Por convención, las entradas de ARACNE se pueden representar como tablas donde las filas representan variables (por ejemplo, ProbeSets [1] en el conjunto de datos de Affymetrix **GEP** (Gene Expression)) y las columnas representan muestras u observaciones (por ejemplo, un solo experimento de microarrays). Hay una regla general que se aplica a todas las entradas de la tabla: ningún carácter TAB debe estar contenido en ninguna entrada, ya que causará problemas de análisis al programa. Usando un conjunto de datos de Affymetrix GEP como ejemplo, un archivo de entrada de ARACNE de muestra se parecería a lo siguiente (Figura 2.1)

Cada fila de variable tiene un identificador único (en verde) y una anotación (en naranja) que siempre van en la primera y la segunda columna respectivamente. Aquí estamos usando el ID de ProbeSet de Affymetrix como el identificador. El campo de anotación para cada variable tiene un uso no trivial en el programa ARACNE: si varias variables tienen el mismo campo de anotación (coincidencia por cadena, distingue entre mayúsculas y minúsculas), se tratarán como duplicados entre sí, por lo que no se computará una IM entre ellos. Si una anotación no está disponible para una variable, use la cadena '- - -' en el campo correspondiente. Para un conjunto de datos GEP de Affymetrix, se pueden usar los símbolos de los genes HUGO o los identificadores de Entrez Gene para los campos de anotación.

[1] En matrices de expresión, un probeset es una secuencia corta de ADN que se dirige a una región corta de una transcripción

ColHeader1	ColHeader2	SampleName1	SampleName1	…
Description				
…				
Description				
AffyProbeId1	ProbeAnnot1	3.6	0.5	2.8
AffyProbeId2	ProbeAnnot2	4.5	9.8	5.6
…	…	…	…	…

Figura E.1: Formato de archivo de entrada de muestra para ARACNE. Imagen tomada de [54].

Debido a que varios ProbeSets de Affymetrix a veces se pueden asignar al mismo símbolo genético, el IM no se computará entre dichos Sondeos, a menos que ambos símbolos genéticos no estén disponibles.

Cada columna de muestra tiene una etiqueta (en azul) que siempre está en la primera fila. Estas etiquetas pueden ser cualquier cadena que describa una muestra, condiciones experimentales, tipos de células, etc. La primera y segunda columnas de la primera fila (en rojo) pueden contener texto arbitrario, por ejemplo "AffyID" y "Annotation".

Puede haber un número arbitrario de filas insertadas después de la primera fila (sombreadas en amarillo). Deben tener el mismo número de columnas que el resto de la tabla, y la primera columna debe usar la etiqueta "Descripción"(distingue entre mayúsculas y minúsculas). El programa ignorará esas líneas, pero se pueden usar para almacenar información adicional sobre cada muestra, como las variables clínicas.

Las celdas restantes en la tabla contienen datos para la variable y muestra apropiadas. Por ejemplo, el '3.6' en la fila correspondiente a ' AffyProbeId1 ' y la columna 3 significa que el valor de expresión observado para ' AffyProbeId1 ' en 'SampleName1' fue 3.6.

El archivo de salida

Antes de pasar al uso del programa ARACNE, primero introduzcamos el formato de su salida, que se mencionará con frecuencia en las siguientes secciones. De manera predeterminada, el programa emitirá los resultados en un archivo con la extensión .adj, que representa un archivo de matriz de adyacencia (o archivo ADJ). El archivo ADJ contiene una representación de lista de adyacencia de la matriz completa, en la que solo se representan las interacciones inferidas. Para continuar con el ejemplo que usamos en la Figura 1, se muestra un archivo ADJ de muestra en la Figura (2.4).

>	Input file	input_file.exp				
>	ADJ file	adjacency_matrix.adj				
>	Output file	Output_file.adj				
>	Algorithm	Accurate				
>	Kernel width	0.15				
>	No. bins	6				
>	MI threshold	0.065				
>	MI P-value	1e-7				
>	DPI tolerance	0.15				
>	Correction	0				
>	Subnetwork file					
>	Hub probe					
>	Control probe					
>	Condition					
>	Percentage	0.35				
>	TF annotation	tf_list.dat				
>	Filter mean	50				
>	Filter CV	0.3				
AffyProbeId1	**AffyProbeId2**	**0.08**	**AffyProbeId5**	**0.15**	**…**	
AffyProbeId2	**AffyProbeId1**	**0.08**	**AffyProbeId3**	**0.22**	**…**	
…	**…**	**…**	**…**	**…**	**…**	

Figura E.2: Salida de muestra ARACNE. Imagen tomada de [54].

Las primeras 18 líneas (fijas) del archivo ADJ registran todos los parámetros utilizados por el programa para generar el archivo. Todos comienzan con un carácter “>”, para que puedan ser analizados fácilmente por cualquier lenguaje de scripting.

El resto de las filas están delimitadas por TAB y contienen todas las interacciones inferidas por ARACNE. La primera columna (en verde) es siempre el identificador de la variable cuyas interacciones se informan en la fila. El resto de las entradas en cada fila consisten en el identificador (en naranja) - pares de valores de MI. Por ejemplo, la fila correspondiente a “AffyProbeId1” puede leerse como sigue: la MI(Información Mutua) entre “AffyProbeId1” y “AffyProbeId2” es 0.08, y la MI entre “AffyProbeId1” y “AffyProbeId5” es 0.15, etc. Las interacciones se almacenan simétricamente Por lo tanto, la interacción entre “AffyProbeId1” y “AffyProbeId2” también se informa en la fila correspondiente a “AffyProbeId2”. Cada fila puede tener un número diferente de entradas, dependiendo del número de interacciones que tenga una variable. Las variables que no tienen interacciones inferidas por ARACNE estarán ausentes del archivo de salida.

Ejecutando ARACNE en la línea de comando

La sintaxis de la línea de comandos para el programa ARACNE es la siguiente:

En MacOs Mojave se uso la siguiente línea de comandos:

```
cd Desktop/ARACNE

./aracne2.macosx
```

Los primeros solo sirven para acceder a los archivos que se encuentran localizados en la carpeta ARACNE que a su vez se encuentran en el escritorio del ordenador. La última línea de código sirve para correr el programa ARACNE usando el archivo de MACOSX llamado "aracne2.macosx". Al no poner ningún archivo de entrada, el compilador mostrará un texto de ayuda. Por otro lado si incluimos un archivo de entrada con extensión .exp con el formato descrito al principio de esta sección, por ejemplo:

```
/aracne2.macosx -i /Users/alex/Desktop /ARACNE.src/ARACNE/test
/arraydata10x336.exp -k 0.15 -t 0.04 -e 0.1
```

donde los valores de "-k", "-t" y "-e" indican el ancho del kernel, el umbral de la IM y la tolerancia del DPI respectivamente.

Si no ponemos una ruta y nombre del archivo de salida, el programa lo generará de manera automática. Así se obtendrá el archivo de salida

```
/Users/alex/Desktop/ARACNE.src/ARACNE/test/arraydata10x336_k0.15_t0

.04_e0.1.adj
```

Si el usuario desea enfocarse solo en una variable particular o en un subconjunto de variables en el conjunto de datos, las opciones "–h" y "–s" se pueden usar para reconstruir las interacciones de la red solo alrededor de la (s) variable (s) de interés. En este caso, el archivo ADJ de salida contendrá solo las filas correspondientes a estas variables. La opción "–s" debe ir seguida de un archivo que enumera las variables en consideración. Utilizando nuevamente el conjunto de datos GEP de Affymetrix como ejemplo, el formato del archivo se muestra en la Figura (2.5).

Para la ingeniería inversa de la red de interacción transcripcional utilizando datos GEP, el conocimiento de todos los TF en el conjunto de datos puede guiar el programa para aplicar el DPI de una manera más precisa, como se ilustra en la Figura 4.

El nodo en azul representa el TF(factor de transcripción) de interés; 'nTF'

```
AffyProbeId_1<\n>
AffyProbeId_2<\n>
AffyProbeId_3<\n>
...
```

Figura E.3: Formato del archivo especificado por la opción "–s" o "–l"

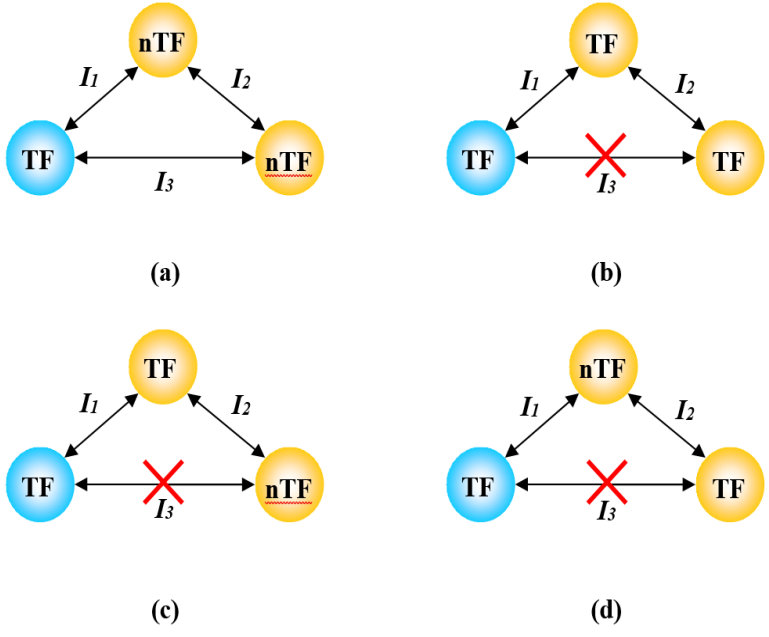

Figura E.4: DPI integrado con la información de anotación TF. Imagen tomada de [54].

significa un gen distinto de los TF. En todos los paneles, suponga $I_1 > I_2 > I_3$. Sin información de anotación de TF, DPI siempre eliminará el borde con I_3. Sin embargo, si sabemos qué genes codifican los TF, los paneles (a) - (d) muestran todas las combinaciones posibles de anotación de nodo. En el esquema (b) - (d) la implementación de DPI no se ve afectada; sin embargo, en el esquema (a) el borde con I_3 estará protegido de la eliminación, ya que el DPI está diseñado para eliminar las interacciones indirectas mediadas por dos interacciones transcripcionales, y la interacción entre dos "nTF"no puede ser transcripcional.

La lista de todos los genes anotados como TF en el conjunto de datos se puede almacenar en un archivo en el mismo formato que en la Figura (2.5) y se puede proporcionar al programa ARACNE mediante la opción “–l”.

Apéndice F

Algoritmos correspondientes a los resultados del capítulo 4

Cálculo del umbral

La matriz de adyacencia, por definición, debe contener solo ceros y unos, sin embargo, el archivo de salida ARACNe nos proporciona una matriz de ceros y números (y unos en la diagonal que deben ser reemplazados por ceros), por lo que es necesario determinar un criterio a partir del cual mediante un valor umbral convertir esos ceros y números en ceros y unos. El criterio que se utilizó consiste en calcular el valor promedio de cada columna utilizando el valor de cada entrada y guardar todos esos promedios en un arreglo que en este caso contendrá 16748 elementos sobre los cuales se promedia nuevamente para obtener un valor umbral a partir del cual obtener ceros y unos.

Sea A_{ij} la matriz de adyacencia. El primer promedio se realiza sobre las columnas y se calcula como:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} A_{ij} \;\; ; i = 1, ..., N$$

Una vez que se ha determinado este valor, podemos calcular el valor umbral T_v como:

$$T_v = \frac{1}{N}\sum_{i=1}^{N} \mu_i$$

Así, obtenemos 2 valores umbral; uno para el tejido sano ($T_{v,h}$) y otro para el tejido canceroso ($T_{v,c}$), estos son:

$$T_{v,h} = 0{,}085677 \quad T_{v,c} = 0{,}033064$$

Con estos valores se reescribió la matriz de adyacencia con el siguiente criterio:

$$\mathcal{A}_{ij} = \left\{ \begin{matrix} 1, \text{ si } A_{ij} > T_v \\ 0, \text{ cualquier otro caso} \end{matrix} \right\}$$

Matrtiz de grado

Con la matriz de adyacencia y la matriz laplaciana asociada podemos obtener una matriz de grados, que se define como:

$$\mathcal{D}_{ij} = \left\{ \begin{matrix} \deg(v_i), \text{ si } i = j \\ 0, \text{ cualquier otro caso} \end{matrix} \right\}$$

además,

$$\mathcal{D}_{ij} = \mathcal{L}_{ij} + \mathcal{A}_{ij}$$

Extrajimos elementos diagonales de $\mathcal{D}_{ij}$ y los asignamos dentro de un vector. Estos elementos se ordenaron de mayor a menor grado. La salida no es el valor del grado del vértice, sino la etiqueta numérica asociada con el gen dentro del conjunto de datos original. Dentro del código [47] podemos encontrar una función que permite obtener *Ensembl gene IDs* cuyo parámetro es la etiqueta numérica del gen y el número de *Ensembl gene IDs* a obtener.

Centralidad de vector propio

Sea a_{ij} una submatriz de $\mathcal{A}_{ij}$ con $i, j = 1, \cdots, n$. Ahora consideremos un vector v_j (que le llamaremos **vector de grados**) con $j = 1, \cdots, n$ cuyos elementos son los grados de los vértices de la submatriz a_{ij}. Entonces, se calculó:

$$v_i = a_{ij} v_j, \;\; i = 1, \cdots, n$$

Lo que se logra obtener con el producto entre la matriz de adyacencia y el vector de grados es reasignar a cada vértice la suma de los valores de sus vértices vecinos.

Los análisis hechos en [47] mostraron que los resultados obtenidos usando la noción anterior y usando la definición de centralidad de vector propio:

$$x_i = \frac{1}{\lambda} \sum_k a_{ki}\, x_k,$$

(donde $\lambda \neq 0$ es una constante) cuya forma matricial es:

$$\lambda x = xA$$

y calculando el valor propio más grande y su vector propio asociado, usando el *power method*, muestran resultados equivalentes respecto a los valores de centralidad de vector propio tanto para datos de tejido sano como tumoral.

Análisis de la conectividad algebraica

Para este análisis es necesario tener la matriz laplaciana $\mathcal{L}_{ij}$ asociada con los datos. Un algoritmo para calcular valores propios está disponible en la biblioteca *SciPy* de *Python*, este es el método `.eig()` de la clase `linalg`. Esto nos da los valores propios de la matriz dentro de una *array*. Sea λ este *array*, se calculó un índice de intercambio s como

$$s = i + \text{argmin}(\lambda_i, \lambda_i + 1, ..., \lambda_i + n)$$

para $i = 0, 1, ..., |\lambda|$. Además argmin(.) operando sobre un $\mathbf{v}$ *array* significa que obtenemos la entrada del valor mínimo de todas las entradas de $\mathbf{v}$. Y finalmente se calculó:

$$(\lambda_i, \lambda_s) = (\lambda_s, \lambda_i)$$

En resumen, los valores se intercambian para que el valor más pequeño de la matriz λ se coloque en el primer lugar de la matriz y los demás se coloquen en orden ascendente. Y finalmente, se devuelve el valor de λ. Con esta forma es fácil obtener el segundo valor propio más pequeño.

Análisis de la matriz laplaciana sin signo

Realizamos el análisis del polinomio característico de la matriz laplaciana sin signo mediante la proposición 3.4.1, que nos da información sobre el número de aristas contenidas en el grafo, para saber qué porcentaje de interacción difiere entre datos UNS y UNT y cuantificar la participación general de los genes en el proceso de regulación de genes. Este cálculo se realizó para 60 submatrices que representan subconjuntos de genes de la matriz completa que contiene (16748,16748) filas y columnas. Primero se creó una función que permite obtener un polinomio característico a partir de una matriz laplaciana sin signo. Con esta función a través de un bucle, se creó una lista que contiene diferentes valores de coeficiente correspondientes al término λ^{n-1} del polinomio característico para diferentes subconjuntos de genes, luego otra función se encarga de multiplicar cada elemento del coeficiente lista por $-\frac{1}{2}$ para finalmente obtener el número de aristas de los grafos asociados con las submatrices en cuestión [47].

Bibliografía

[1] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006 Mar 20;7 Suppl 1(Suppl 1):S7. doi: 10.1186/1471-2105-7-S1-S7. PMID: 16723010; PMCID: PMC1810318.

[2] Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005 Apr;37(4):382-90. doi: 10.1038/ng1532. Epub 2005 Mar 20. PMID: 15778709.

[3] Oscar Rojo , Ricardo Soto. The spectra of the adjacency matrix and Laplacian matrix for some balanced trees. ELSEVIER Linear Algebra and its Applications 403 (2005) 97–117. https://doi.org/10.1016/j.laa.2005.01.011.

[4] Enrique Hernández Lemus and Claudia Rangel-Escareño. The Role of information theory in gene regulatory network inference. ISBN: 978-1-62100-325-0. 2011 Nova Science Publishers, Inc.Editors: P. Deloumeaux et al, pp. 137-184.

[5] Daniel A. Spielman. Spectral Graph Theory Lecture 1.Introduction. 2015. http://www.cs.yale.edu/homes/spielman/462/462schedule.html

[6] N. Abreu, R. Del-Vecchio, V. Trevisan, C. Vinagre.Teoria Espectral de Grafos - Uma Introdução IIIo Colóquio de Matemática da Região Sul. http://mtm.ufsc.br/coloquiosul/notas_minicurso_6.pdf.

[7] Peter Lancaster, Department of Mathematics and Statistics Peter Lancaster, Miron Tismenetsky. The Theory of Matrices: With Applications.Academic Press, 1985. ISBN 0124355609, 9780124355606.

[8] Olsen, Mika. Notas de la UEA Matemáticas discretas II. 2017. Universidad Autónoma Metropolitana, Unidad Cuajimalpa. ISBN: 978-607-28-1097-6.

[9] Archana S. Iyer, Hatice U. Osmanbeyoglu and Christina S. Leslie. Computational methods to dissect gene regulatory networks in cancer. ELSEVIER. Current Opinion in Systems Biology 2017, pp 115-122. https://doi.org/10.1016/j.coisb.2017.04.004.

[10] https://www.gob.mx/salud/acciones-y-programas/informacion-estadistica

[11] John E. Hopcroft. Pearson. Automata Theory,Languages, and Computation 3rd Edition. 2007 Pearson Education, Inc. ISBN 0-321-45536-3.

[12] Thomas M. Cover.Joy A. Thomas A. John Wiley & Sons. Elements of information theory. 2006 John Wiley and Sons, Inc.

[13] Ilya Nemenman. Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106.Information theory, multivariate dependence, and genetic network inference. arXiv:q-bio/0406015.

[14] Jacek Majewski, Hao Li, Jurg Ott Am J Hum Genet. The Ising Model in Physics and Statistical Genetics. 2001 Oct; 69(4): 853–862. Published online 2001 Aug 20. doi: 10.1086/323419 PMCID: PMC1226070.

[15] Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: detecting and evaluating dependencies between variables. Bioinformatics. 2002;18 Suppl 2:S231-40. doi: 10.1093/bioinformatics/18.suppl_2.s231. PMID: 12386007.

[16] Luis Rodriguez Ojeda. Construcción de Kernels y funciones de densidad de probabilidad. Departamento de Matemáticas, ESPOL. https://www.dspace.espol.edu.ec/bitstream/123456789/25019/1/CONSTRUCCION_DE_KERNELS_Y_FUNCIONES_DE_DENSIDAD_DE_PROBABILIDAD.pdf

[17] Emmert-Streib F, Dehmer M. Information processing in the transcriptional regulatory network of yeast: functional robustness. BMC Syst Biol. 2009 Mar 19;3:35. doi: 10.1186/1752-0509-3-35. PMID: 19298671; PMCID: PMC2679710.

[18] Chaitankar V, Ghosh P, Perkins EJ, Gong P, Deng Y, Zhang C. A novel gene network inference algorithm using predictive minimum description length approach. BMC Syst Biol. 2010 May 28;4 Suppl 1(Suppl 1):S7. doi: 10.1186/1752-0509-4-S1-S7. PMID: 20522257; PMCID: PMC2880413.

[19] C L Percy, J W Horm, J L Young, Jr, and A J Asire. Uterine cancers of unspecified origin–a reassessment. Public Health Rep. 1983. Public Health Rep. 1983 Mar-Apr; 98(2): 176–180. PMCID: PMC1424417. PMID: 6856742.

[20] Rachel Quinlan.Niall Madden. Notas del curso: Advanced Linear Algebra. NUI Galway. http://www.maths.nuigalway.ie/~rquinlan/linearalgebra/

[21] Enrique Hernández-Lemus.Cancer a complex disease. A complex path(way) to cancer phenomenology. CopIt-arXives Publishing Open Access with an Open Mind 2018, pp.3-26.

[22] Chen DB, Yang HJ. Comparison of gene regulatory networks of benign and malignant breast cancer samples with normal samples. Genet Mol Res. 2014 Nov 11;13(4):9453-62. doi: 10.4238/2014.November.11.10. PMID: 25501155.

[23] Ohashi T, Yamamoto T, Yamanashi Y, Ohsugi M. Human TUBG2 gene is expressed as two splice variant mRNA and involved in cell growth. FEBS Lett. 2016 Apr;590(8):1053-63. doi: 10.1002/1873-3468.12163. Epub 2016 Apr 6. PMID: 27015882.

[24] http://www.ensembl.org/biomart/martview/

[25] Yun Niu, Tieju Liu, Gary M. K. Tse, Baocun Sun, Ruifang Niu, Hiu-ming Li, Hui Wang, Yi Yang, Xue Ye, Ying Wang, Qi Yu1 and Fei Zhang1. Increased expression of centrosomal α, γ-tubulin in atypical ductal hyperplasia and carcinoma of the breast. Cancer Sci. 2009 Apr;100(4):580-7. doi: 10.1111/j.1349-7006.2008.01075.x. Epub 2009 Feb 2. PMID: 19215229.

[26] Zhao Y, Zhang X, Liu Y, Ma Y, Kong P, Bai T, Han M, Li B. Inhibition of ALAS1 activity exerts anti-tumour effects on colorectal cancer in vitro. Saudi J Gastroenterol. 2020 May-Jun; 26(3): 144–152. Published online 2020 Apr 6. doi: 10.4103/sjg.SJG_477_19. PMCID: PMC7392291. PMID: 32270771.

[27] Lin S, Huang C, Sun J, Bollt O, Wang X, Martine E, Kang J, Taylor MD, Fang B, Singh PK, Koomen J, Hao J, Yang S. The mitochondrial deoxyguanosine kinase is required for cancer cell stemness in lung adenocarcinoma. EMBO Mol Med. 2019 Dec;11(12):e10849. doi: 10.15252/emmm.201910849. Epub 2019 Oct 21. PMID: 31633874; PMCID: PMC6895611.

[28] Miroslav Fiedler.Algebraic connectivity of graphs.Czechoslovak Mathematical Journal ,1973. https://dml.cz/bitstream/handle/10338.dmlcz/101168/CzechMathJ_23-1973-2_11.pdf

[29] Rui Sun, YuanZhao Hu, RongLei Shang, WeiYu Bai, JianWei Sun. Role and mechanism of mitochondrial deoxyguanosine kinase in lung cancer tumorigenesis. SCIENTIA SINICA Vitae 2019.http://lib.cqvip.com/Qikan/Article/Detail?id=7002613682

[30] De Martino A, De Martino D. An introduction to the maximum entropy approach and its application to inference problems in biology. Heliyon. 2018 Apr 13;4(4):e00596. doi: 10.1016/j.heliyon.2018.e00596. PMID: 29862358; PMCID: PMC5968179.

[31] Wei Z, Li H. A Markov random field model for network-based analysis of genomic data. Bioinformatics. 2007 Jun 15;23(12):1537-44. doi: 10.1093/bioinformatics/btm129. Epub 2007 May 5. PMID: 17483504.

[32] Liang KC, Wang X. Gene regulatory network reconstruction using conditional mutual information. EURASIP J Bioinform Syst Biol. 2008;2008(1):253894. doi: 10.1155/2008/253894. PMID: 18584050; PMCID: PMC3171392.

[33] Jang IS, Margolin A, Califano A. hARACNe: improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. Interface Focus. 2013 Aug 6;3(4):20130011. doi: 10.1098/rsfs.2013.0011. PMID: 24511376; PMCID: PMC3915831.

[34] Dougherty J, Tabus I, Astola J. Inference of gene regulatory networks based on a universal minimum description length. EURASIP J Bioinform Syst Biol. 2008;2008(1):482090. doi: 10.1155/2008/482090. PMID: 18437238; PMCID: PMC3171396.

[35] Zhang X, Zhao J, Hao JK, Zhao XM, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic Acids Res. 2015 Mar 11;43(5):e31. doi: 10.1093/nar/gku1315. Epub 2014 Dec 24. PMID: 25539927; PMCID: PMC4357691.

[36] Yu, D., Lim, J., Wang, X. et al. Enhanced construction of gene regulatory networks using hub gene information. BMC Bioinformatics 18, 186 (2017). https://doi.org/10.1186/s12859-017-1576-1

[37] Mousavian Z, Kavousi K, Masoudi-Nejad A. Information theory in systems biology. Part I: Gene regulatory and metabolic networks. Semin Cell Dev Biol. 2016 Mar;51:3-13. doi: 10.1016/j.semcdb.2015.12.007. Epub 2015 Dec 14. PMID: 26701126.

[38] Leonardo E R. Medidas de distinguibilidad entre distribuciones de probabilidad. Aspectos teóricos y aplicaciones al estudio de las series temporales. Tesis Doctoral (**p.24**). Facultad de Matemática Astronomía, Física y Computación. Universidad Nacional de Córdoba. Marzo 2020. https://rdu.unc.edu.ar/handle/11086/16528

[39] ARACNE Manual. An Algorithm for the Reconstruction of Accurate Cellular Networks. https://manual-guide.com/manu/2786/index.html

[40] Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A. Reverse engineering cellular networks. Nat Protoc. 2006;1(2):662-71. doi: 10.1038/nprot.2006.106. PMID: 17406294.

[41] Mutual information: a measure of dependency for nonlinear time series, Physica A: Statistical Mechanics and its Applications, Volume 344, Issues 1–2, 2004, Pages 326-329, ISSN 0378-4371, https://doi.org/10.1016/j.physa.2004.06.144.

[42] Lee WP, Tzou WS. Computational methods for discovering gene networks from expression data. Brief Bioinform. 2009 Jul;10(4):408-23. doi: 10.1093/bib/bbp028. PMID: 19505889.

[43] Maarten van Steen. Graph Theory and Complex Networks An Introduction. 2010 Maarten van Steen. ISBN-10 : 9081540610. https://www.researchgate.net/publication/267804733_Graph_Theory_and_Complex_Networks_An_Introduction

[44] Bonacich, P. Factoring and weighting approaches to clique identification. Journal of Mathematical Sociology 2 (1): 113-120.(1972).

[45] https://math.unice.fr/~frapetti/CorsoF/cours4part1.pdf

[46] M. E. J. Newman. The mathematics of networks. M. E. J. Newman. http://www-personal.umich.edu/~mejn/papers/palgrave.pdf

[47] https://github.com/Alejandro1848/SGT-in-cancer-GRN/

[48] William N. Anderson, Thomas D. Morley Eigenvalues of the Laplacian of a graph Linear Multilinear Algebra, 18 (2) (1985), pp. 141-145. https://doi.org/10.1080/03081088508817681.

[49] https://portal.gdc.cancer.gov/

[50] de Haan W, van der Flier WM, Wang H, Van Mieghem PF, Scheltens P, Stam CJ. Disruption of functional brain networks in Alzheimer's disease: what can we learn from graph spectral analysis of resting-state magnetoencephalography? Brain Connect. 2012;2(2):45-55. doi: 10.1089/brain.2011.0043. Epub 2012 Jun 11. PMID: 22480296.

[51] https://github.com/josemaz/lung-mirnas

[52] https://github.com/CSB-IG/parallel-aracne

[53] https://www.genenames.org/

[54] Adam A Margolin, Kai Wang1, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman, Andrea Califano. Reverse engineering cellular networks. 2006 Nature Publishing Group. Published online 27 June 2006; doi:10.1038/nprot.2006.106.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00098
Matrícula: 2182800793

Propiedades espectrales de redes de regulación genética en cáncer.

Con base en la Legislación de la Universidad Autónoma Metropolitana, en la Ciudad de México se presentaron a las 16:00 horas del día 28 del mes de mayo del año 2021 POR VÍA REMOTA ELECTRÓNICA, los suscritos miembros del jurado designado por la Comisión del Posgrado:

DR. DAVID PHILIP SANDERS
DR. RICARDO MARCELIN JIMENEZ
DR. LEONARDO DAGDUG LIMA

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (FISICA)

DE: ALEJANDRO JUAREZ TORIBIO

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

APROBAR

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

ALEJANDRO JUAREZ TORIBIO
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTE

DR. DAVID PHILIP SANDERS

VOCAL

DR. RICARDO MARCELIN JIMENEZ

SECRETARIO

DR. LEONARDO DAGDUG LIMA

El presente documento cuenta con la firma –autógrafa, escaneada o digital, según corresponda- del funcionario universitario competente, que certifica que las firmas que aparecen en esta acta – Temporal, digital o dictamen- son auténticas y las mismas que usan los c.c. profesores mencionados en ella