Diseño e implementación de una estrategia integrativa para la detección de nuevos módulos genéticos y nuevos genes asociados al inicio y desarrollo del cáncer colorrectal

J.D. Arce Rentería*
N. Ibagón Rivera*

*Pontificia Universidad Javeriana Cali, Colombia**
juanarce01@javerianacali.edu.co
nicolasibagon99@javerianacali.edu.co

Resumen

El advenimiento de las tecnologías ómicas, el desarrollo de técnicas computacionales basadas en el aprendizaje de máquina aplicado a sistemas biológicos y la integración de ambos paradigmas en modelos matemáticos, ha permitido avanzar en el entendimiento causal de enfermedades complejas como el cáncer. En este sentido, desde de una perspectiva sistémica, el uso de redes biológicas y la representación de sistemas moleculares como genes, proteínas y sus dinámicas de interacción, ha permitido realizar una aproximación a los sistemas biológicos desde la teoría de grafos. Desde esta perspectiva, en los últimos años se han desarrollado una gran variedad de estrategias, las cuales, desde la teoría de grafos, han contribuido al entendimiento del proceso deletéreo que conduce a la enfermedad y, equitativamente, a identificar nodos clave de la red los cuales podrían estar relacionados con diferentes tipos de enfermedades, como lo sería el c \'ancer. En el presente trabajo, integramos distintos tipos de información biológica asociada a la comprensión genética del origen y desarrollo de la enfermedad, acoplándola al mapa más detallado de interacción proteína-proteína que existe. Posteriormente, realizamos análisis fundamentales sobre medidas clásicas de la topología de la red construida, que fueron útiles para identificar elementos claves de la red. Asignamos pesos a los nodos y a las aristas de la red según la información biológica, lo cual fue un procedimiento fundamental para priorizar elementos de la red (proteínas) asociadas al cáncer y específicamente al cáncer colorrectal. Con base en dicha información y con la red construida, implementamos algoritmos de modularidad para identificar comunidades específicas que pudieran estar específicamente asociadas al desarrollo de cáncer colorrectal, y finalmente implementamos algoritmos de caracterización de comunidades no sobrelapantes y estrategias específicas de aprendizaje de máquina para encontrar potenciales proteínas asociadas al cáncer colorrectal.

1. Introducción

Las redes de interacción proteína-proteína están definidas por el contacto físico existente entre un par o un grupo de proteínas específicas. En la representación gráfica de la red, los nodos representan las proteínas y los vínculos o aristas entre ellas, la interacción física entre las mismas. En este sentido el producto directo de las tecnologías ómicas puede representarse en redes de interacción gen-gen, gen-proteína o proteína-proteína, cuya caracterización topológica y posterior análisis biológico permiten el entendimiento de un fenómeno particular desde una perspectiva holística y sistémica por medio de la caracterización de patrones funcionales derivados de la organización de la red. Teniendo estas consideraciones en mente, el objetivo de esta investigación es diseñar una estrategia integrativa que utilice distintos tipos de información y que acople la teoría de grafos con métodos de aprendizaje de máquina, para identificar elementos y módulos genéticos asociados al origen o progresión del cáncer de colon.

2. Fundamentación Teórica

- **Biología de sistemas** La aplicación directa de la perspectiva sistémica a la biología se conoce como biología de sistemas y define que la funcionalidad celular surge como consecuencia de la interacción precisa y coordinada de los distintos componentes celulares. Las funciones específicas que desempeña una célula particular son propiedades emergentes, derivadas de la dinámica de interacción de los componentes celulares. Por lo tanto el entendimiento de las dinámicas que determinan la funcionalidad celular sólo cobra sentido en un contexto holístico y no pueden ser accedidas mediante un entendimiento aislado de los distintos componentes del sistema [1].

- **El cáncer como red de interacción** En el campo de la biología, y más específicamente en la biología molecular, se han llevado diversos estudios los cuales proponen distintos puntos de vista para poder estudiar esta enfermedad, entre ellos se encuentra la célula vista como un conjunto de interacciones moleculares, entre las cuales se encuentran las interacciones proteína-proteína, las cuales determinan el interactoma funcional de la célula en la que se determinan dichas interacciones [2].

- **Algoritmos de detección de comunidades en cáncer** En los últimos años el problema computacional en detección de comunidades ha recibido una atención considerable. Detectar comunidades nos brinda la oportunidad de analizar el comportamiento que poseen los nodos por medio de sus interacciones, con esto se espera poder encontrar comunidades las cuales compartan características comunes, atributos o inclusive relaciones funcionales que puedan brindar un mejor entendimiento de su funcionamiento. En el contexto de las redes biológicas se espera poder encontrar comunidades las cuales puedan brindar información asociadas a las interacciones que poseen las proteínas en un interactoma humano, o al mismo tiempo en el estudio de enfermedades como lo sería el cáncer se pueden detectar genes los cuales puedan estar relacionados a la carcinogénesis. [3]

- **Aprendizaje de máquina y métodos de aprendizaje de máquina aplicados al entendimiento del cáncer** El aprendizaje de máquina es una rama de la inteligencia artificial la cual se centra en el uso de datos y algoritmos con el fin de aprender comportamientos y replicarlos. Gracias al aprendizaje de máquina se han podido llevar a cabo varios estudios y aplicar una gran variedad de técnicas con los cuales poder diagnosticar el cáncer en las personas. Al mismo tiempo también se puede llegar a predecir factores como variables clínicas, así como parámetros histológicos que pueden hacer parte de los conjuntos de datos de entrada, utilizados para implementar las estrategias de aprendizaje de máquina [4].

3. Resultados

Construcción del mapa de interacción proteína-proteína

El interactoma humano utilizado se configur· a partir de 21 bases de datos p¼blicas que compilan datos de interacciones prote²na-prote²na (PPI) obtenidos experimentalmente. Para establecer los pesos de las conexiones en la red de interacci· n prote²na-prote²na, utilizamos informaci· n preexistente sobre las prote²nas incluidas en el grafo.

Descripición topológica de la red construída

Se realizó una clasificación de los nodos de la red por su peso funcional y por su número de conexiones totales. Se calculó el peso total de los nodos, teniendo en cuenta la sumatoria total del peso de las aristas de conexión. Se estableció una correlación entre el grado de los nodos del interactoma y su peso en todo el interactoma, y luego solo con las proteínas asociadas al proceso de inestabilidad genómica y las proteínas implicadas en la aparición del cáncer de colon.

Identificación de comunidades sobre la red construída

Con el objetivo de determinar los patrones de distribución de las distintas proteínas asociadas a las categorías funcionales consideradas, y obtener comunidades de proteínas modulares, es decir, se usaron tres algoritmos: (i) Algoritmo de propagación de etiquetas. (ii) Algoritmo de comunidades modulares codiciosas. (iii) Algoritmo de comunidades de Louvain. Posteriormente, para obtener comunidades que se sobrelapen unas a otras se usaron dos algoritmos: (i) IPCA y (ii) El algoritmo de Angel.

Predicción de proteínas asociadas a cáncer colorrectal

Para la implementación de los modelos de aprendizaje de máquina en la red, se realizó primero un rebalanceo de las clases con el método SMOTE y en segundo lugar con el método G-SMOTE. Finalmente, se implementaron los mismos modelos a través de la estrategia PU Learning con el objetivo de mejorar los resultados del aprendizaje. Para cada estrategia se aplicaron los mismos cuatro modelos: (i) Regresión logística, (ii) Bosques aleatorios (random forest), (iii) K Vecinos más cercanos y (iv) Análisis discriminante lineal.

A continuación se muestra el módulo propuesto resultante de los algoritmos de identificación de comunidades, y una lista de 14 genes propuestos a partir de los resultados de los modelos de aprendizaje de máquina.

Id NCBI	Nombre Oficial	Número de conexiones
472	ATM	249
641	BLM	72
672	BRCA1	705
675	BRCA2	85
701	BUB1B	95
999	CDH1	155
1029	CDKN2A	176
9401	RECQL4	42
83990	BRIP1	56
2072	ERCC4	34
2132	EXT2	22
2175	FANCA	86
2177	FANCD2	75
84464	SLX4	31
2956	MSH6	59
11200	CHEK2	135
4221	MEN1	71
4683	NBN	79
5395	PMS2	39
5889	RAD51C	22

Id NCBI	Nombre Oficial	Número de conexiones
79728	PALB2	43
55215	FANCI	24
129563	DIS3L2	6
6794	STK11	196
64324	NSD1	67
7157	TP53	762
7248	TSC1	158
7428	VHL	195
7486	WRN	42
7508	XPC	87

Figura 1: Módulo que podría asociarse causalmente al origen o desarrollo de cáncer de colon.

Id NCBI	Nombre oficial	Número de conexiones	Asociación
220082	CBY2	131	Supresión de proliferación celular
100499483	CCDC180	22	No se encontró artículo relacionado
378708	CENPS	30	No se encontró artículo relacionado
2355	FOSL2	160	Metastasis
51659	GINS2	13	Proliferación celular y apoptosis
2961	GTF2E2	109	Proliferación celular, prognosis de cáncer
8968	H3C7	327	Firma genética de CRC
3727	JUND	534	Proliferación celular
79786	KLHL36	19	No se encontró artículo relacionado
4656	MYOG	66	factor de transcripción en rabdomiosarcoma
388677	NOTCH2NLA	368	Cáncer de ovario
267004	PGBD3	21	No se encontró artículo relacionado
4250	SCGB2A2	77	Gen común de inestabilidad genómica
4904	YBX1	830	Inhibición de apoptosis en cáncer colorrectal

Figura 2: Genes propuestos sin categoría funconal asociada.

4. Discusión y conclusiones

A partir de los resultados del análisis topológico de la red, podemos notar que existen proteínas que no necesariamente poseen un alto número de conexiones pero que participan en diversos tipos de cáncer. Por otro lado, notamos que el uso de algoritmos de comunidades no demostró ser del todo adecuada para la predicción de proteínas, a excepción del algoritmo GMC, del cual resultó el módulo de proteínas propuesto en la figura 1.

En el caso de los resultados mostrados por el aprendizaje de máquina, PU Learning demostró éxito para la predicción de proteínas asociadas a cáncer colorrectal. Esto se evidencia en valores altos de exhaustividad (recall). A partir de esta estrategia se propuso una lista de 79 proteínas candidatas nuevas que podrían participar en el desarrollo de cáncer colorrectal. Experimentación en el futuro. Las demás estrategias no tuvieron resultados confiables en entrenamiento para considerarlos en el resultado final.

Se sugiere el uso de un nuevo set de datos con un mayor número de proteínas asociadas a cáncer colorrectal o la inclusión de datos de entrenamiento nuevos con el fin de mejorar la predicción de nuevas proteínas candidatas mediante el uso de PU Learning.

Referencias

[1] F. Bruggeman and H. Westerhoff, "The nature of systems biology," *Trends in microbiology,* vol. 15, pp. 45–50, 02 2007.

[2] A. L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," jan 2011.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communi- ties in large networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, p. P10008, oct 2008.

[4] IBM, "What is machine learning?." https://www.ibm.com/cloud/learn/machine-learning. Accessed: 2021-05-22.