CÓMO LOS GRAFOS Y LA INTELIGENCIA ARTIFICIAL PUEDE TRANSFORMAR LA BUSQUEDA DE METABOLICOS

HOW GRAPHS AND ARTIFICIAL INTELLIGENCE CAN TRANSFORM THE HUNT FOR METABOLITES

Reynold Osuna González*
Guillermo De Ita Luna

ISSN 2448-5829

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68

https://orcid.org/0009-0002-5228-7317
https://orcid.org/0000-0001-7948-8253

Fecha de recepción 5/02/2024
fecha de revisión 2/12/2024
fecha de publicación 20/01/2025

Facultad de Ciencias de la Computación, Edif. CCO1 – 14 Sur y Av. Sn. Claudio, C.U.
Doctorado en Ingeniería del lenguaje y del conocimiento
Benemérita Universidad Autónoma de Puebla
reynold.osuna@alumno.buap.mx *
guillermo.deita@correo.buap.mx

Resumen

Ante la posibilidad de que la vasta cantidad de información genética de un gran número de microorganismos se encuentre subutilizada, el presente trabajo explora el potencial de combinar la bioinformática con técnicas de inteligencia artificial para descubrir microorganismos capaces de producir metabolitos de interés. Se discute el uso de grafos de conocimiento para representar relaciones biológicas, el graph embedding para transformar su información en representaciones espaciales, y el clustering para identificar patrones en secuencias genéticas obtenidas mediante BLAST. Se destaca la importancia de estas herramientas en la búsqueda de soluciones innovadoras para desafíos sociales, como la degradación de contaminantes o el control de plagas. Además, se resalta el papel crucial de la inteligencia artificial en acelerar la comprensión y el aprovechamiento del vasto conocimiento biológico disponible en la actualidad. Este enfoque integrado ofrece nuevas oportunidades para explorar y comprender el universo biológico, así como para desarrollar aplicaciones prácticas en campos como la medicina y la biotecnología.

Palabras clave: Grafos de conocimiento, Bioinformática, Inteligencia Artificial, BLAST

Abstract

Given the possibility that the vast amount of genetic information from numerous microorganisms is underutilized, this paper explores the potential of combining bioinformatics with artificial intelligence techniques to discover microorganisms capable of producing metabolites of interest. The use of knowledge graphs to represent biological relationships, graph embedding to transform their information into spatial representations, and clustering to identify patterns in genetic sequences obtained through BLAST are discussed. The importance of these tools in the search for innovative solutions to social challenges, such as the degradation of pollutants or pest control, is emphasized. Additionally, the crucial role of artificial intelligence in accelerating the understanding and utilization of the vast biological knowledge available today is highlighted. This integrated approach offers new opportunities to explore and understand the biological universe, as well as to develop practical applications in fields such as medicine and biotechnology.

Keywords: Knowledge graphs, Bioinformatics, Artificial Intelligence, BLAST

ISSN 2448-5829

Año 11, No. 31, 2025, pp. 56 - 68

RD-ICUAP

Introducción

La Inteligencia Artificial (IA), se enfoca en el estudio del comportamiento inteligente que es logrado por medios computacionales, y donde la representación del conocimiento y el razonamiento desempeñan un papel sumamente importante(Brachman & Levesque, 2004).

Con la gran cantidad de información que la humanidad genera constantemente, las ciencias de la computación se han convertido en un faro que ilumina el camino hacia la comprensión y el aprovechamiento de estos vastos datos. Para ser procesada por una computadora, la información debe contar con una representación adecuada, lo mismo que el conocimiento, siendo ambos temas fundamentales para las ciencias de la computación, existiendo ya diversas propuestas al respecto.

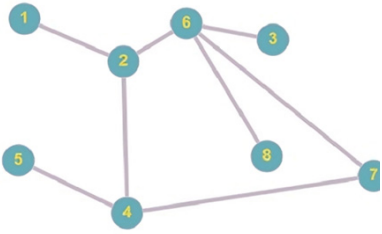


Figura 1. G: V(G) = {1,2,3,4,5,6,7,8}, E(G)={<1,2>,<2,4>, <2,6>,<3,6>,<4,5>,<4,7>,<6,7>,<6,8>}

Una forma de representación de la información ampliamente utilizada en diversas ciencias es la de grafo (Figura 1). Particularmente, los grafos aparecen de forma natural en el estudio de la biología, puesto que tradicionalmente, sistemas biológicos complejos son modelados mediante entidades biológicas interconectadas formando grafos de conocimiento biológico (Figura 2) sobre los que pueden aplicarse técnicas de exploración de grafos para su análisis y tareas predictivas (Mohamed et al., 2021).

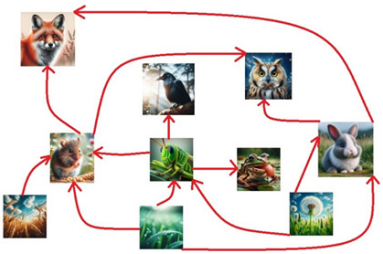


Figura 2. Ejemplo de red trófica (Fuente: Elaboración propia)

Entre las formas de representar conocimiento, una de gran uso y relevancia actual es a través de grafos de conocimiento. En 2012, con el anuncio de que utilizaría grafos de conocimiento, Google atrajo la atención de otras grandes empresas tecnológicas (Hogan et al., 2021) que utilizan esta representación para no sólo almacenar información, sino para realizar consultas y procesar la información que se encuentra en forma implícita, así como para descubrir relaciones entre entidades que no han sido aún declaradas o descubiertas.

En la era del big data, el clustering y los grafos de conocimiento se han convertido en herramientas indispensables para aprovechar y estructurar las grandes cantidades de información generada. Por un lado, las técnicas de clustering permiten agrupar datos similares y descubrir patrones en conjuntos de datos, mientras que, por el otro lado, los grafos de conocimiento permiten conectar y representar conocimientos de manera computacionalmente viable. La combinación de ambas técnicas potencia el aprovechamiento de grandes cantidades de información y ha encontrado aplicaciones en diversos campos como la bioinformática, donde se analizan datos biológicos para comprender relaciones complejas.

Grafos de Conocimiento: La Revolución de la Representación

La representación adecuada de la información es un pilar fundamental en las ciencias de la computación. En este

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

contexto, los grafos de conocimiento (Knowledge Graphs o KG, figura 3) se vislumbran como una herramienta poderosa. Si bien es posible encontrar referencias hacia ellos desde 1972, fue Google quienes, en 2012, al anunciar el uso de grafos de conocimiento, atrajo la atención de otros gigantes tecnológicos, marcando el comienzo de una era donde estos grafos no solo almacenan información, sino que también facilitan consultas y procesan datos de manera implícita, descubriendo relaciones entre entidades (Hogan et al., 2021).

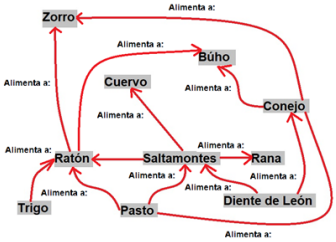


Figura 3. Grafo de conocimiento basado en la red trófica de la figura 2

Un grafo de conocimiento está dado por KG = < E, R, T >, donde E (entity) y T (tail) representan al conjunto de entidades y R (relation) representa al conjunto de relaciones y las aristas en R conectan dos nodos para formar una tripleta (h, r, t).

En la tripleta (h, r, t) se encuentra implícita una direccionalidad de la relación, que parte de la entidad "cabeza" (head) hacia la entidad "cola" (tail), así que puede ser deducido que una característica de un grafo de conocimiento es que se trata de un grafo dirigido.

Siguiendo la definición del grafo de conocimiento basada en tripletas, puede definirse también al razonamiento sobre grafos de conocimiento como el proceso por el cuál, siguiendo un camino relacional P, se genera una tripleta (h, r, t) tal que h E, r R, t T, (h, r, t) KG.

Las aplicaciones de los KG son vastas, desde generar nuevo conocimiento hasta respaldar decisiones. La detección de comunidades dentro de un grafo es una aplicación intrigante, desafiando algoritmos de agrupamiento no supervisado. Por ejemplo, la detección de comunidades dentro de un grafo puede ser vista como un proceso de agrupamiento no supervisado: el reto de esta tarea es inferir estructuras latentes de comunidad a partir de únicamente un grafo G = (V,E) como entrada, y puede tener múltiples aplicaciones en la vida real (Hamilton, 2020) como el descubrimiento de módulos funcionales en redes o grafos de interacción genética (Agrawal et al., 2018) o el descubrimiento de grupos de usuarios fraudulentos en redes de transacciones financieras (S. Pandit et al., 2007).

Agrupamientos: Explorando las Similitudes y Descubriendo Patrones

El agrupamiento o clustering es una herramienta esencial de aprendizaje automático para explorar grandes cantidades de datos, permitiendo descubrir patrones y relaciones. La tarea básica de clustering se enuncia como la partición de un conjunto de data points en grupos donde los miembros sean tan similares como sea posible (Reddy, 2014). Cuando se trata de grafos de conocimiento, se presenta un desafío único: la representación de un grafo no se adapta directamente a puntos de datos, llevando a dos enfoques. El primero consiste en diseñar algoritmos de agrupamiento para trabajar directamente sobre grafos, mientras que el segundo se enfoca en transformar la información del grafo en una representación espacial (embedding) antes de aplicar algoritmos de clustering. El embedding del grafo de conocimiento se vuelve esencial para esta transformación, ya que, aunque existen algoritmos de clustering para grafos, pueden resultar computacionalmente costosos.

El tomar un conjunto de objetos y dividirlo en subconjuntos o grupos más pequeños es una tarea cotidiana que se ha llevado al campo de la computación. El análisis por clúster divide los datos en grupos que son significativos, útiles o ambos, siendo una técnica de aprendizaje no supervisado, mediante la cual se agrupan objetos o datos en conjuntos denominados clústeres, basándose en su similitud y cuyas

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

aplicaciones se extienden a ramas como la recuperación de información, estudio del clima, medicina, negocios y la biología (Tan et al., 2006).

En la literatura es posible encontrar múltiples técnicas de agrupamiento, dado que no existe una técnica universal, en su lugar, hay diferentes propuestas con diferentes desempeños y especializadas para una gran variedad de escenarios (Tan et al., 2019), sin embargo, es posible definir ciertas técnicas básicas.

Por ejemplo, el algoritmo básico llamado k-means consiste en elegir un número inicial de k centroides, siendo este un parámetro especificado por el usuario y que corresponde al número de clústeres deseados. El algoritmo entonces asigna a cada punto el centroide más cercano. Posteriormente, el algoritmo actualiza el centroide de cada grupo formado, con base en los puntos pertenecientes al grupo, para volver a asignar a cada punto el centroide más cercano. Este proceso se repite hasta que ya no sea posible reasignar los centroides asociados a cada punto, lo que por necesidad implica, que los centroides ya no pueden ser modificados.

Para calcular la cercanía de un punto a un centroide es necesario utilizar una función de proximidad o distancia, siendo determinada su naturaleza por el tipo de espacio utilizado, existiendo múltiples posibilidades. Cuando la representación de los datos está dada en un espacio euclidiano, es común el uso de la distancia euclidiana o la distancia de Manhattan como funciones de proximidad.

También es utilizada una función objetivo que permita medir la calidad de los clústeres obtenidos. Suponiendo un espacio de representación euclidiano, utilizando una función de proximidad para calcular la cercanía a los centroides, es posible utilizar la suma del error cuadrático (SSE por sus siglas en inglés, definida en la ecuación 1 tal que, dados los clústeres obtenidos en cada iteración de k-means, puede compararse la calidad de cada aproximación comparando sus respectivos SSE's: entre menos sea su valor, mayor es la calidad de los clústeres, dado que cada punto se encuentra más cercano a su respectivo centroide (Tan et al., 2006).

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} distancia(c_i, x)^2$$

Ecuación 1. Suma del error cuadrático

Siendo x un objeto, Ci el i-ésimo clúster, ci el centroide de Ci, y k el número de clústeres (Tan et al., 2006).

Por último, para elegir los centroides de cada clúster se requiere también una forma matemática de hacerlo, que, en el mismo caso de espacios euclidianos planteado previamente, la media es la medida que minimiza el SSE y está dada por la ecuación 2

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

Ecuación 2. Cálculo de centroidos

Siendo x un objeto, Ci el i-ésimo clúster, ci el centroide de Ci, y mi el número de objetos en el i-ésimo clúster (Tan et al., 2006). En la figura 3 podemos apreciar datos agrupados por medio del algoritmo k-means teniendo como entrada k = 3.

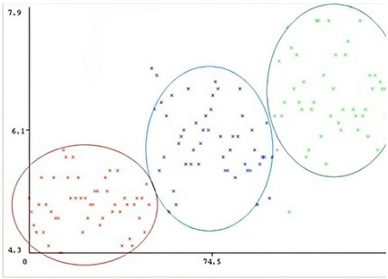


Figura 4. Ejemplo de datos agrupados por medio de k-means (k = 3)

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

Es importante destacar que otras medidas de distancia además de la euclidiana pueden ser utilizadas, como la similitud del coseno o distancia de Manhattan.

k-means tiene como ventaja ser un algoritmo que no demanda grandes recursos de almacenamiento y cuya complejidad en el tiempo es del orden lineal con respecto al número total de datos a agrupar.

DBSCAN en el Mundo del Agrupamiento de Datos

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) es un algoritmo de agrupamiento que busca zonas densas. DBSCAN adopta un enfoque único al realizar el agrupamiento basado en la densidad de los datos. Visualicemos nuestro conjunto de datos como un terreno donde algunas regiones son más densas que otras. Este algoritmo identifica áreas de alta densidad, separándolas de las áreas con baja densidad de datos. La magia comienza al determinar el vecindario de cada punto en el conjunto de datos, formado por todos los puntos dentro de un radio específico (llamado "epsilon"). La densidad local alrededor de cada punto es calculada, y aquellos puntos cuya densidad supera un valor crítico (denominado "minPts") se convierten en centros de clústeres, mientras que su vecindario conforma dicho clúster. Los puntos que quedan fuera de estos clústeres se consideran ruido en los datos.

Una característica sobresaliente de DBSCAN es que los parámetros épsilon y minPts son especificados por el usuario, y a diferencia de algoritmos como k-means, DBSCAN no requiere como entrada el número de clústeres esperados. Esto otorga mayor flexibilidad al proceso, aunque esta libertad implica que el determinar los valores de épsilon y minPts cobra gran importancia, siendo parámetros que determinan si los resultados obtenidos son o no significativos.

La técnica básica para definir épsilon y minPts consiste en realizar un análisis del comportamiento de la distancia de un punto hasta k número de vecinos cercanos, valor denominado k-dist. Lo esperable para puntos que pertenecen a un clúster es que su valor k-dist sea pequeño siempre que este no sea mayor al del tamaño del clúster, con pequeñas variaciones dada la naturaleza aleatoria de la distribución de los puntos. Sin embargo, para puntos que no están en un clúster, como aquellos puntos considerados ruido, el valor de k-dist sea relativamente grande. Por lo tanto, si se calcula k-dist para todos los puntos del conjunto de datos, se ordenan de menor a mayor y se grafiquen, debería observarse un cambio abrupto en el valor k-dist que corresponda a un valor de épsilon viable. Si se toma entonces este punto de cambio abrupto como valor para épsilon y el valor de k como el valor de minPts, los puntos cuya k-dist sea menor a épsilon serán marcados como parte del clúster, mientras que otros puntos serán marcados como ruido (Tan et al., 2006).

DBSCAN tiene una complejidad temporal del orden O(m2) en el peor de los casos, donde m es número de puntos que conforman el conjunto de datos, pero si se limita a espacios de pocas dimensiones, es posible mejorar su complejidad al orden de O(m · log m), mientras que su complejidad espacial, incluso para espacios de muchas dimensiones corresponde al orden de O(m). Si bien DBSCAN es capaz de encontrar clústeres que a k-means le resulta imposible identificar, no es un algoritmo tan eficaz cuando los clústeres poseen densidades con grandes variaciones y cuando los datos pertenecen a espacios con muchas dimensiones(Tan et al., 2006).

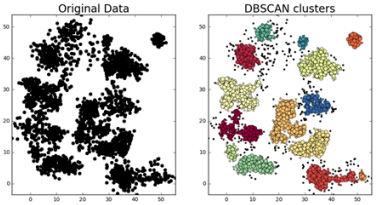


Figura 5 Ejemplo de datos agrupados por medio de DBSCAN (referencia aquí: Ernst, 2017)

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

Es posible utilizar aprendizaje automático para obtener representaciones del grafo de conocimiento, es decir, llevar a cabo un embedding al grafo para aplicar diversos algoritmos de agrupamiento sobre la nueva representación (Hamilton, 2020).

Embeddings de Grafos: Transformando Dimensiones y Preservando Conocimiento

El embedding de grafos emerge como una herramienta eficaz para abordar los desafíos del análisis de grafos. Esta técnica transforma un grafo en un espacio de baja dimensionalidad, preservando sus estructuras (Cai et al., 2017). El 'Graph embedding' apunta a representar el grafo mediante vectores de baja dimensionalidad, facilitando tanto el análisis de grafos como el aprendizaje de representaciones. Formalmente, para un grafo G = (V, E), aprender el embedding de sus nodos implica codificar todos los nodos del grafo en dos formas: vectores de puntos determinísticos o distribuciones probabilísticas estocásticas. Estas representaciones preservan las propiedades de la estructura del grafo de manera óptima. El pairwise similarity en el espacio embebido latente facilita la aproximación de la similitud de nodos en el espacio original (Xu, 2020).

Dado un grafo G = (V,E), la tarea de aprender el embedding de sus nodos puede ser formulada matemáticamente como aprender una proyección φ tal que todos los nodos del grafo (V = {v1, v2, v3, ..., vn}, donde n = |V|) pueden ser codificados en dos formas diferentes de embedding (desde un espacio de alta dimensionalidad hacia uno de baja dimensionalidad) (M. Xu, 2020); una forma de embedding de los nodos son vectores de puntos determinísticos (φ= {zi L | i = 1,2,...,n}), mientras que la otra forma son mediante distribuciones probabilísticas estocásticas (φ= {ℙi φ(ϕ, ϕ) | i = 1,2,...,n}) siendo el vector medio ϕ L/2, y la matriz de covarianza φ L/2xL/2.

Además, las propiedades de la estructura del grafo se preservan de manera óptima en el espacio embebido. Con este fin, pairwise similarity (por ejemplo, producto punto ϕTϕ) en el espacio embebido latente facilita una aproximación de la similitud de nodos (Sim(vi,vj)) correspondiente en el espacio original, es decir, (Sim(vi,vj)) ~ ϕTϕ para embedding de grafos basado en vectores; específicamente, Sim(vi,vj) ~ ϕTϕ para embedding de grafo basado en distribuciones gaussianas, y donde Sim es una función de similitud predefinida (Xu, 2020).

Al aplicar embedding a grafos de conocimiento y conjuntarlo con clustering, es posible potenciar el aprovechamiento de bastas cantidades de información, misma que en esta era del Big Data, es generada a un ritmo superior del que puede ser analizada. Tener éxito en la aplicación conjunta de ambas técnicas podría acelerar la comprensión, mejorar la capacidad de análisis y aumentar el aprovechamiento del vasto conocimiento con el que se cuenta en la actualidad.

Entre las ciencias que pueden llegar a manejar grandes cantidades de información, al punto que resulta impráctico el análisis de ella sin apoyo computacional, y que gracias a ese apoyo, se ha logrado grandes avances; por ejemplo, la biología molecular, que recurriendo al uso de técnicas computacionales ha sido capaz de extraer información valiosa para el entendimiento de los procesos bioquímicos que se llevan a cabo a nivel celular, dando paso así, al surgimiento del campo de la bioinformática.

Bioinformática

La bioinformática es un campo interdisciplinario en el que se conjuntan la biología, la informática y la estadística con el objetivo de analizar datos biológicos y comprender complejas relaciones entre ellos. Esta disciplina se vale de técnicas computacionales y de Inteligencia Artificial (IA) para procesar, analizar e interpretar grandes cantidades de datos biológicos como son las secuencias de ADN y proteínicas presentes en los organismos (Notredame & Claverie, 2007).

Una de las líneas de investigación en la microbiología y la biología molecular consiste en el estudio de microorganismos

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

que puedan ser utilizados para el control de plagas, descontaminación de suelos o de los que se puedan obtener nuevos medicamentos. Por medio de pruebas experimentales en laboratorio, es posible encontrar organismos capaces de producir metabolitos (nombre dado a los productos del metabolismo), que puedan ser utilizados. Como por ejemplo, para degradar materiales peligrosos, o que sean capaces de producir antibióticos, tal y como sucede en el ejemplo clásico de la penicilina, metabolito compuesto por el hongo penicillium y que le sirve de defensa contra microorganismos invasores, por medio de un efecto llamado antagonismo.

Explorando el Universo Biológico a Través de la Bioinformática

De la intersección entre la biología, la informática y la estadística, emerge un fascinante campo de estudio: la bioinformática. Este emocionante dominio científico se dedica a desentrañar los misterios biológicos mediante el análisis de datos y la comprensión de complejas relaciones entre ellos. Empleando herramientas computacionales y técnicas de Inteligencia Artificial (IA), la bioinformática nos sumerge en un vasto océano de información biológica, como las secuencias de ADN y proteínas que se encuentran en los organismos.

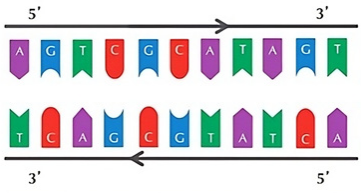
Una de las áreas de investigación más apasionantes en microbiología y biología molecular implica el estudio de microorganismos con potencial para el control de plagas, la descontaminación ambiental o el desarrollo de nuevos medicamentos. A través de experimentos de laboratorio, se exploran organismos capaces de producir metabolitos, sustancias vitales para diversas aplicaciones, desde la degradación de materiales peligrosos hasta la producción de antibióticos, tal como lo demostró el clásico ejemplo de la penicilina, un metabolito elaborado por el hongo Penicillium para defenderse contra invasores microbianos mediante un fenómeno conocido como antagonismo.

En el ámbito de la bioinformática, los grafos de conocimiento pueden ser herramientas esenciales para representar entidades y relaciones biológicas, como genes, proteínas e interacciones moleculares. El agrupamiento o clustering, por su parte, permite identificar patrones y relaciones entre secuencias genéticas y proteicas, ayudando a desentrañar relaciones evolutivas y estructuras de sistemas biológicos complejos. En este vasto campo, se busca entender las intrincadas relaciones entre genes, proteínas y el entorno de los organismos.

La secuenciación del ADN de diversos organismos es llevada a cabo día a día por gran cantidad de biólogos y biotecnólogos en todo el mundo, enriqueciendo rápidamente nuestro conocimiento genético. El resultado de este proceso es un genoma, un mapa detallado de los genes presentes en un organismo, que busca identificar la función específica de cada gen en la producción de proteínas.

El ADN, esa asombrosa molécula de la vida, está compuesta tan solo cuatro tipos de nucleótidos -adenina (A), citosina (C), guanina (G) y timina (T)- que forman dos cadenas complementarias, cada una siendo de miles de nucleótidos, formando la base de la información genética. Gracias a la bioinformática, hemos logrado secuenciar genomas completos, descifrando la secuencia de nucleótidos y determinando la composición de cada gen.

Es necesario recalcar el hecho de que dichas cadenas son complementarias: la adenina se une exclusivamente a la timina, mientras que la citosina se empareja únicamente con la guanina. En la figura 6, podemos apreciar un diagrama de bloques que representa un fragmento de ADN, una ventana a la maravillosa complejidad de la vida que la bioinformática nos ayuda a descifrar.



RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

Figura 6. Cadenas complementarias de ADN. Las cadenas que forman son idénticas, pero con sentido inverso (Compeau & Pevzner, 2015).

Una aplicación práctica de la biología molecular es la búsqueda de compuestos o metabolitos que nos proporcionen alguna utilidad, siendo un metabolito un compuesto químico que se produce a través del proceso metabólico de un organismo. Así, un metabolito de interés puede consistir en una sustancia química que ataque a otros organismos, lo que derive en identificación de nuevos fármacos o que funcionen como control de plagas.

A partir de la información genómica disponible, es deseable identificar organismos con características que el investigador considere responsables de la producción del metabolito de interés. Sin embargo, realizar comparaciones directas y de forma exhaustiva entre la información genómica de cada organismo resulta extenuante, y muchas veces impráctico. Esto es debido a que el número de microorganismos que pueden poseer genes de interés es extenso, y a pesar de contar con los mismos grupos de genes, habrá microorganismos que expresen el metabolito deseado y otros que no lo hagan, puesto que los genes no son el único factor que influye en la producción de este. También se debe considerar que los genes pueden tener pequeñas variaciones en su composición, debido a pequeñas mutaciones, sin que el cambio llegue a afectar la función del gen. Es decir, que un mismo gen permite cierta variación antes de que se convierta en un gen diferente. Por si fuera poco, podrían existir microorganismos que, contando con solo una parte de los genes identificados y relacionados con el metabolito, sean capaces de producirlo.

Una vez que se identifiquen organismos de interés basándose únicamente en la presencia de genes, será necesario realizar experimentos en laboratorio para corroborar que efectivamente se produzca el metabolito deseado. Siendo primordial limitar la selección de microorganismos a aquellos que presenten un alto grado de similitud, misma que va más allá de una simple elección de ellos en base de la presencia completa de los genes.

Comparar las secuencias de ADN y de proteínas entre individuos de la misma especie o entre especies, ayuda a identificar genes responsables de funciones similares entre ellos. Así como a detectar mutaciones, sin embargo, comparar cadenas de miles de caracteres en búsqueda de coincidencias parciales o exactas puede resultar en una tarea extenuante y temporalmente costosa. Siendo un problema previamente estudiado en las ciencias de la computación, llamado en ocasiones coincidencia de patrones o string matching en inglés, en el que se busca encontrar una determinada cadena de caracteres, ya sea de forma aproximada o exacta, dentro de una cadena de caracteres de mayor tamaño, existiendo diversas propuestas algorítmicas para su solución, pero en el ámbito de la biotecnología, la propuesta denominada Basic Local Alignment Search Tool (BLAST) es la que ha sido, desde su publicación en 1990, un algoritmo estándar para la comparación de cadenas.

Basic Local Alignment Search Tool (Altschup et al., 1990)

Basic Local Alignment Search Tool (BLAST) es una herramienta bioinformática ampliamente utilizada para comparar secuencias de ADN, ARN o proteínas con bases de datos biológicas para encontrar regiones similares. BLAST se utiliza para identificar similitudes funcionales entre secuencias y para inferir la función de secuencias desconocidas, publicada originalmente en 1990 por Stephen F. Altschul, Warren Gish, Webb Miller, Egene W. Myers y David J. Lipman.

Para hacer uso de ella, se comienza proporcionando una secuencia (secuencia de consulta o query), ya sea de ADN o de una proteína, que se desee comparar con secuencias conocidas ya almacenadas en una base de datos. y BLAST encuentra regiones locales de alta similitud (subcadenas alineadas) entre la query y las cadenas contenidas en la base de datos, permitiendo una respuesta rápida y eficiente, comparada contra el enfoque

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

de fuerza bruta de buscar coincidencias exactas entre cadenas.

BLAST ocupa principalmente un algoritmo de alineación de pares que se encarga de hacer esta comparación de la query con todas las secuencias contenidas en la base de datos, dividiendo la secuencia de consulta en palabras de longitud w, para buscar coincidencias exactas de estas palabras a lo largo de las secuencias de la base de datos. Una vez llevada a cabo esta tarea, BLAST extiende las coincidencias encontradas en ambas direcciones para identificar regiones más largas de similitud, aun cuando ya no sean coincidencias exactas, hasta que se alcanza una región donde la similitud cae por debajo de un umbral T.

Posteriormente, se lleva a cabo una fase de evaluación, en la que se calculan puntuaciones a estas regiones de similitud, tomando en cuenta las coincidencias exactas, las similitudes y las diferencias entre las bases o aminoácidos alineados y mediante estadísticas se determina si las similitudes encontradas son significativas o son dadas por mero azar, con lo que se filtran los resultados mediante umbrales de significado para reducir los falsos positivos.

Entre las aplicaciones existentes en el campo de la biotecnología para KGE se encuentran la predicción de interacciones entre fármacos y su objetivo (drug-target interactions o DTI) y la predicción de efectos secundarios debidos al uso de múltiples fármacos al mismo tiempo por parte de un paciente. Identificar las interacciones entre fármacos y su objetivo (DTI) tiene el potencial de reducir enormemente el campo de búsqueda por medicamentos candidatos en el tratamiento de una enfermedad, razón por la cual es importante contar con métodos computacionales de predicción eficientes (Chen et al., 2018), siendo un área de investigación activa por al menos los últimos 10 años. El flujo de procesamiento para aplicaciones cuyo objetivo sea DTI puede ser dividido en tres etapas: preprocesamiento de los datos de entrada, entrenamiento de un modelo subyacente en base a un conjunto de reglas de aprendizaje y por último un modelo predictivo. Así mismo clasifica los métodos de aprendizaje máquina para predicción DTI en métodos supervisados y semisupervisados. Otra aplicación biotecnológica consiste en el estudio de la similitud de medicamentos, ya que propiedades estructurales, moleculares y biológicas similares a menudo se relacionan con indicaciones o efectos secundarios similares (Ma et al., 2018).

Integración de Grafos de Conocimiento, Embedding y Clustering para el Descubrimiento de Metabolitos

Una gran cantidad de trabajos bioinformáticos están enfocados hacia aplicaciones médicas y a la secuenciación de cadenas de ADN o RNA, habiendo un rango amplio de aplicaciones sobre las cuales aplicar métodos de clustering en búsqueda de nuevos descubrimientos, como es, el de identificación de cepas bacterianas capaces de producir metabolitos de interés.Por una parte, los grafos de conocimiento permiten representar relaciones entre genes, proteínas y metabolitos de una manera completa y estructurada, facilitando la comprensión de las interacciones y vías metabólicas involucradas en la producción de metabolitos específicos.Por otra parte, las técnicas de graph embedding permiten transformar las secuencias genéticas en espacios de baja dimensionalidad, preservando similitudes y relaciones entre secuencias, lo que puede ayudar a la identificación de patrones característicos que sean relevantes para la producción de metabolitos.Utilizando técnicas de clustering al resultado del graph embedding de un grafo de conocimiento, es posible agrupar secuencias genéticas similares, permitiendo identificar cepas bacterianas o incluso otros organismos que compartan características genéticas que se relacionen con la producción del metabolito que se busca obtener.Aplicando estas técnicas, se pueden priorizar los organismos con mayor potencial para llevar a cabo experimentos prácticos en laboratorio, seleccionando aquellos organismos con mayores probabilidades de producir sustancias de uso médico o industrial.

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

Conclusiones

En conclusión, la convergencia de la bioinformática, la inteligencia artificial aplicada a grafos de conocimiento, ofrece un gran potencial para la exploración y comprensión del inmenso universo microbiológico. La representación por medio de grafos de conocimiento de secuencias genéticas obtenidas mediante BLAST, junto con la aplicación de técnicas de embedding y clustering, permiten desentrañar complejas relaciones entre genes, proteínas y metabolitos, pudiendo acelerar el descubrimiento de nuevos compuestos biológicamente activos e impulsando avances en diversos campos, como la medicina, biotecnología e incluso la conservación del medio ambiente.

Utilizando la inteligencia artificial en nuestra búsqueda de tesoros microbianos, podemos acercarnos cada vez más al descubrimiento de soluciones innovadoras para desafíos a los que nos enfrentamos como sociedad, como pueden ser la degradación de contaminantes o el control de plagas, aprovechando los metabolitos obtenidos de

forma sostenible a partir de microorganismos.

Declaración de privacidad

Los datos personales facilitados por los autores a RD-ICUAP se usarán exclusivamente para los fines declarados por la misma, no estando disponibles para ningún otro propósito ni proporcionados a terceros.

Declaración de no conflicto de intereses

Los autores declaran que no existe conflicto de interés alguno

Agradecimientos

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico otorgado para la presente realización de mis estudios de doctorado (Número de Beca 833591).

Referencias

Altschup, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. In J. Mol. Biol (Vol. 215).

Brachman, R. J., & Levesque, H. J. (2004). Knowledge Representation and Reasoning (1st Edition). Morgan Kaufmann.

Cai, H., Zheng, V. W., & Chang, K. C.-C. (2017). A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. http://arxiv.org/abs/1709.07604

Chen, R., Liu, X., Jin, S., Lin, J., & Liu, J. (2018). Machine learning for drug-target interaction prediction. In Molecules (Vol. 23, Issue 9). MDPI AG. https://doi.org/10.3390/molecules23092208

Hamilton, W. L. (2020). Graph Representation Learning (Vol. 14, Issue 3).

Hogan, A., Blomqvist, E., Cochez, M., D'Amato, C., Melo, G. De, Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A. C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys, 54(4). https://doi.org/10.1145/3447772

Ma, T., Xiao, C., Zhou, J., & Wang, F. (2018). Drug Similarity Integration Through Attentive Multi-view Graph Auto-Encoders. http://arxiv.org/abs/1804.10850

Mohamed, S. K., Nounu, A., & Nováček, V. (2021). Biological applications of knowledge graph embedding models. Briefings in Bioinformatics, 22(2), 1679–1693. https://doi.org/10.1093/bib/bbaa012

Notredame, C., & Claverie, J.-M. (2007). Bioinformatics for dummies (Second Edition). Wiley Publishing, Inc.

Reddy, A. •. (2014). DATA CLUSTERING DATA CLUSTERING Algorithms and Applications Chapman & Hall/CRC Data Mining and Knowledge Discovery Series Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

S. Pandit, D. Chau, S. Wang, & C. Faloutsos. (2007). NetProbe: A fast and scalable system for fraud detection in online acution networks. 16th International Conference on World Wide Web (WWW '07), 201–210. https://doi.org/https://doi.org/10.1145/1242572.1242600

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to Data Mining, 2/e (2nd ed.).

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining.

Xu, M. (2020). Understanding graph embedding methods and their applications. http://arxiv.org/abs/2012.08019

RD-ICUAP Año 11, No. 31, 2025, pp. 56 - 68 ISSN 2448-5829