



Extracción de interacciones regulatorias transcripcionales de bacterias a partir de literatura biomédica utilizando inteligencia artificial.

Ali-Berenice Posada-Reyes

Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Cuautitlán,
Laboratorio de Microbiología, Inmunología y Salud Pública, Estado de México, México.
ORCID: 0000-0003-3165-6099

Alfredo Varela-Vega

Universidad Nacional Autónoma de México, Centro de Ciencias Genómicas,
Programa de Genómica Computacional, Morelos, México.
ORCID: 0000-0003-4857-4997

Sara Berenice Martínez-Luna

Universidad Nacional Autónoma de México, Centro de Ciencias Genómicas,
Programa de Genómica Computacional, Morelos, México.
ORCID: 0009-0001-1329-6344

Carlos-Francisco Méndez-Cruz

Universidad Nacional Autónoma de México, Centro de Ciencias Genómicas,
Programa de Genómica Computacional, Morelos, México.
ORCID: 0000-0002-2549-1614

Recepción: 29 de octubre de 2024.

Aceptación: 21 de marzo de 2025.

Septiembre 2025 • número de revista 13 • DOI: 10.22201/dgtic.26832968e.2025.13.49

Extracción de interacciones regulatorias transcripcionales de bacterias a partir de literatura biomédica utilizando inteligencia artificial

Resumen

Las redes de regulación transcripcional (TRNs) de bacterias brindan una visión global de los mecanismos de respuesta de éstas a los cambios en su ambiente. El estudio de estas redes permite ampliar el conocimiento biológico y derivar en investigaciones con implicaciones clínicas o farmacéuticas. Sin embargo, la reconstrucción de TRNs se hace tradicionalmente de forma manual mediante un proceso demandante y costoso de curación de artículos científicos. En este trabajo, describimos la aplicación de aproximaciones de inteligencia artificial (IA), específicamente la afinación de *transformers* pre-entrenados BERT, para la extracción automática de TRNs de la literatura. Con 1562 oraciones de entrenamiento de la bacteria *Escherichia coli*, comparamos seis arquitecturas tipo BERT. El mejor modelo afinado obtuvo resultados significativos (F1-score: 0.8685, coeficiente de correlación de Matthews: 0.8163). Con este modelo, extrajimos correctamente el 82% de una TRN de la bacteria *Salmonella* utilizando 264 artículos completos. El factor de transcripción PhoP fue relevante en la red por obtener el mayor valor de conexiones (*degree*=180) por lo que analizamos biológicamente su comunidad de genes. Este trabajo muestra el uso de la IA para facilitar la extracción de conocimiento biológico que podría ser utilizado por estudios futuros en el área biomédica.

Palabras Clave: Inteligencia artificial, procesamiento de lenguaje natural, extracción de información, extracción de interacciones regulatorias, regulación transcripcional.

Extraction of transcriptional regulatory interactions of bacteria from biomedical literature using artificial intelligence

Abstract

Transcriptional regulatory networks (TRNs) from bacteria provide a global view of their mechanisms of response to changes in the environment. The study of these networks allows to expand biological knowledge and leads to research with clinical or pharmaceutical implications. However, the reconstruction of TRNs is traditionally done manually through a demanding and expensive process of curation of scientific articles. In this work, we describe the application of Artificial Intelligence (AI) approaches, specifically the fine-tuning of BERT pre-trained transformers, for the automatic extraction of TRNs from biomedical literature. With 1562 training sentences from the Escherichia coli bacterium, we compared six BERT architectures. The best fine-tuned model obtained significant results (F1-score: 0.8685, Matthews correlation coefficient: 0.8163). With this model, we correctly extracted 82% of a TRN from the Salmonella bacterium using 264 full articles. The PhoP transcription factor was relevant in the network because it obtained the highest value of connections (degree=180) so we biologically analyzed its gene community. This work shows the use of AI to facilitate the extraction of biological knowledge that could be used for future studies in the biomedical field.

Keywords: Artificial intelligence, Natural language processing, Information extraction, Regulatory interaction extraction, Transcriptional regulation.

1. Importancia de las redes de regulación transcripcional de bacterias

El genoma de las bacterias es controlado por diferentes mecanismos de regulación que permiten o inhiben la expresión de sus genes en un proceso denominado regulación de la transcripción. Así, con la expresión de sus genes, las bacterias pueden responder a señales internas o ambientales [1]. Un ejemplo de lo anterior es lo que ocurre en la bacteria *Escherichia coli* (*E. coli*), en la que altas concentraciones de oxígeno funcionan como una señal para activar la respiración aeróbica mediante la activación o represión de ciertos genes [2]. Estos mecanismos reguladores pueden ser representados por las denominadas redes de

regulación transcripcional (*transcriptional regulatory networks*, TRNs), que describen la organización global de la regulación transcripcional de un organismo y muestran características jerárquicas y modulares gracias a la representación en red [3].

Las bacterias se adaptan a los cambios de su entorno principalmente a través de la acción de proteínas denominadas factores de transcripción (*transcription factors*, TFs), que activan o reprimen la expresión de genes blanco para inducir mecanismos dentro de la bacteria y así adaptarse a su entorno [3]. Precisamente, el conjunto de interacciones regulatorias (TF-gen) de la transcripción en una bacteria conforman una TRN bacteriana [4]. El estudio de TRNs de bacterias es relevante para el entendimiento de diversos mecanismos, por ejemplo, se ha descrito que la TRN de *E. coli* desempeña un papel importante en la supervivencia de esta bacteria para colonizar el intestino grueso dentro de sus hospederos [5].

Reportar el conocimiento derivado del estudio de las TRNs es crucial, ya que podemos descubrir las adaptaciones que tienen las bacterias a los desafíos ambientales. Por ejemplo, desde la mirada de la evolución, Dalldorf y colaboradores [6] descubrieron en experimentos de laboratorio que *E. coli* puede afrontar desafíos ambientales a corto plazo a través de la modulación genética de las TRNs a pesar de mutaciones aplicadas a los TFs [6]. Una utilidad relevante del estudio de TRNs de bacterias en el área de la biomedicina es la identificación de mutaciones reguladoras en cepas de bacterias con asociación a resistencia a fármacos. Tal es el caso del estudio de Praski y colaboradores del 2017 [7], que se encargaron de identificar mutaciones en el gen *marR* en aislados clínicos de *E. coli* resistentes a ciprofloxacina. El gen *marR* codifica la proteína MarR, que es un TF represor que regula la transcripción de diversos genes encargados del mecanismo de resistencia a fármacos. En su estudio, reportan las correlaciones de las mutaciones en *marR* y la susceptibilidad al fármaco ciprofloxacina [7].

Otros estudios han descrito las modificaciones en la regulación de genes de bacterias de importancia clínica, responsables de infecciones en humanos o animales, y su asociación con implicaciones importantes en la gravedad de una enfermedad. Un ejemplo de lo anterior es el reportado por Shah en el 2013 [8], quien describió el transcriptoma global de la bacteria *Salmonella enterica* serovar Enteritidis en cepas de aves de corral. La *Salmonella* es una bacteria causante de gastroenteritis bacteriana en todo el mundo, transmitida por alimentos. En su trabajo, Shah estudió 252 genes expresados de manera diferencial en cepas con baja

patogenicidad en comparación con su expresión en cepas con alta patogenicidad y comprobó que los genes con baja patogenicidad mostraron un perfil transcripcional único que se caracterizó por la expresión reducida de genes involucrados en la virulencia (genes fimbriales, de motilidad y asociados al estrés). Caracterizar este tipo de patrones de regulación ayudaría a diversos investigadores a identificar cepas de virulencia relevantes, que posteriormente podrían ser atenuadas y ser candidatas de vacunas contra la bacteria [8].

Todo lo anterior demuestra que reportar redes de regulación transcripcional de bacterias tiene implicaciones importantes para el conocimiento biológico, biomédico y clínico. Además, pueden tener aplicación en el área farmacéutica con el posible desarrollo de vacunas o la identificación de genes con funciones en virulencia que posteriormente sean blanco de fármacos.

2. Bases de datos y curación de redes de regulación transcripcional

Hoy en día, para tener acceso a TRNs, existen bases de datos de conocimiento biológico en plataformas digitales. Algunos ejemplos son: RegulonDB [9] y EcoCyc [10], que reportan la TRN de *E. coli* K-12, organismo modelo para estudios en microbiología; RegulomePA, que reporta la TRN de *Pseudomonas aeruginosa* PAO1 [11]; y SalmoNet2 [12], que reporta TRNs de varias cepas de *Salmonella*; entre otras bases de datos. Su objetivo es integrar el conocimiento de los diferentes mecanismos reguladores de las bacterias en redes cada día más completas. Sin embargo, a pesar de estos y otros esfuerzos, las TRNs publicadas hasta hoy aún están incompletas o son inexistentes para muchas bacterias [13]. Lo anterior puede deberse a que aún no se han identificado experimentalmente las interacciones de regulación para una gran cantidad de bacterias o, por otro lado, a que los esfuerzos sostenidos de curación de la literatura que extraiga sus TRNs, las organice y brinde acceso en base de datos se ha hecho en pocas bacterias [14].

Es así que uno de los retos más importantes para la publicación de TRN es su creación manual. A partir de enormes bases de datos de literatura biomédica, como PubMed [15], tradicionalmente se realiza una extracción manual de información de esa literatura, tarea que se define como *curación de literatura biomédica*. Ésta consiste en la identificación de las publicaciones pertinentes y en la extracción de la información, de manera selectiva, para su

despliegue en forma organizada en bases de datos [16]. Dicha curación manual permite obtener la información de forma muy precisa, pero presenta la desventaja de ser demandante y requerir mucho tiempo [17]-[18]. Además, enfrenta el reto de una acelerada producción científica [19]. Dadas estas limitaciones y retos, desde hace tiempo, ha sido indispensable la adopción de aproximaciones automáticas para facilitar e impulsar el trabajo de curación de literatura biomédica [20]-[21].

El presente trabajo va en esa dirección y busca facilitar la creación de TRNs de bacterias usando aproximaciones novedosas de IA, como son los *transformers* pre-entrenados bidireccionales, llamados BERT (*Bidirectional Encoder Representations from Transformers*). Por ello, hemos participado por dos años en los *Proyectos de investigación en Inteligencia Artificial en el Espacio de Innovación UNAM-HUAWEI*, desarrollando modelos para aplicarlos en la extracción de conocimiento de literatura biomédica. Gracias al apoyo del *Espacio de Innovación UNAM-HUAWEI* de la *Alianza para Promover el Desarrollo de Capacidades Digitales en México*, en su convocatoria 2022, logramos obtener un modelo de IA para la extracción de TRNs usando artículos científicos. Esto se logró mediante la afinación (*fine-tuning*) y evaluación de modelos tipo BERT.

3. La IA en la extracción de conocimiento de literatura biomédica

Nuevas oportunidades se han creado para la curación de literatura biomédica gracias a las tecnologías de Inteligencia artificial (IA), en particular, aquellas basadas en métodos de aprendizaje profundo (*deep learning*). Un ejemplo de lo anterior es el chatbot de la empresa OpenAI, ChatGPT (*Chat Generative Pretrained Transformer*), que tiene la capacidad de escribir ensayos, responder preguntas e imitar una conversación humana, aunque presenta la desventaja de que, a menudo, la información que ofrece es incorrecta o irreproducible. Es posible que las actualizaciones continuas y el mejoramiento de las técnicas basadas en órdenes (*prompts*) ofrezcan el potencial para un análisis de conocimiento biomédico más preciso y confiable en el futuro [22].

Las herramientas de minería de texto basadas en IA aprovechan las metodologías del Procesamiento del lenguaje natural (*Natural Language Processing*, NLP) para facilitar el reconocimiento de entidades y la extracción de relaciones que se describen en la literatura biomédica. Por ejemplo, en la literatura publicada al respecto, el gen “BRCA1” es descrito

como una entidad relacionada con la enfermedad “cáncer” y es asociado con un mayor riesgo de la misma [23]. Mediante el NLP, se podrían extraer automáticamente de la literatura estas entidades (genes y enfermedades) y sus relaciones para aumentar el número de éstas relacionadas entre sí, reportadas en bases de datos de asociaciones gen-enfermedad.

Dos tareas típicas del NLP para la extracción de conocimiento de literatura biomédica son el reconocimiento de entidades nombradas (*Named Entity Recognition*, NER) y la extracción de relaciones (*Relation Extraction*, RE) [24]. La primera tarea consiste en identificar y etiquetar nombres de entidades biológicas o biomédicas, como proteínas, genes, fármacos o enfermedades, en un corpus de literatura biomédica. El NER tiene aplicación en la extracción de conceptos biológicos clave, lo que ayuda a construir ontologías y bases de conocimiento. La segunda tarea detecta si entidades presentes en la literatura tienen alguna relación, como las interacciones proteína-proteína, asociaciones gen-enfermedad, relaciones genotipo-fenotipo o interacciones farmacológicas (fármaco-enfermedad, fármaco-efecto secundario). Debe notarse que, además de texto, la literatura biomédica también contiene conocimiento valioso en forma de figuras, diagramas y rutas metabólicas sobre eventos moleculares que conducen a un proceso biológico o enfermedad [22]. En la figura 1, se pueden observar algunas aplicaciones de los métodos de IA basados en minería de texto.

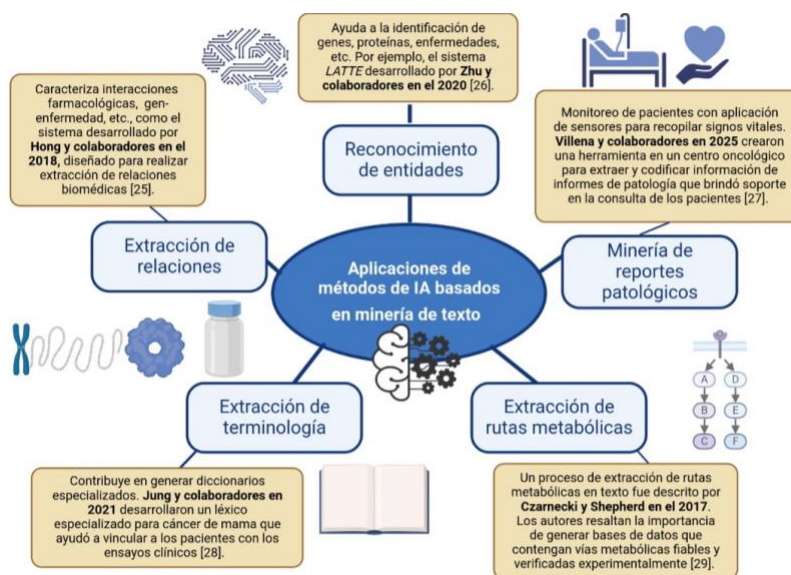


Fig. 1. Aplicaciones de los métodos de IA basados en minería de texto. En color azul, se muestran las áreas en las se puede aplicar la IA y, en color naranja, se describe un ejemplo desarrollado en la biología o medicina. Imagen creada con Biorender.com [30].

4. Desarrollo

4.1. Materiales y métodos

4.1.1. Transformers pre-entrenados tipo BERT

Las arquitecturas de redes de neuronas artificiales para NLP, entre ellas los *transformers* pre-entrenados bidireccionales BERT [31] y los generativos GPT (*Generative Pre-trained Transformer*) [32], están en pleno auge en la industria de la IA. Aunque los modelos GPT se han convertido en los más populares, los modelos BERT siguen siendo su principal competidor. La arquitectura BERT fue desarrollada en 2018 por Google [31]. En comparación con otras arquitecturas, presenta la ventaja de leer simultáneamente ambas direcciones del texto (bidireccionalidad). Leer de izquierda a derecha o de derecha a izquierda es importante porque puede ayudar al modelo a captar mejor el contexto de una palabra objetivo en una oración. Los modelos BERT son de acceso abierto, contrario a los GPT de la empresa OpenAI. El comprender las arquitecturas de los modelos y las características de cada uno puede ayudar a utilizarlos de mejor manera, por ejemplo, los modelos GPT son aprovechados para interactuar por medio de chatbots o bien utilizarse eficientemente para traducción. Por su parte, los modelos BERT se recomiendan para el entendimiento de lenguaje natural (*natural language understanding*), al igual que son recomendados para interpretar emociones u opiniones en texto (*sentiment analysis*) [33].

En su momento, un modelo BERT pre-entrenado con 800 millones de palabras del BookCorpus y 2500 millones de palabras de la Wikipedia en inglés fue lanzado a la comunidad [31]. Después, gracias a que fue una arquitectura abierta, diversos grupos de investigación entrenaron y publicaron variantes de este modelo, ya sea cambiando algún mecanismo de la arquitectura o aumentando los textos de entrenamiento. Sin embargo, la característica de estos *transformers* que más llamó nuestra atención fue su capacidad para ser afinado (*fine-tuning*) con pocos datos y para tareas específicas de NER o RE. El *fine-tuning* consiste entonces en aprovechar los pesos ya ajustados del modelo BERT durante la fase de pre-entrenamiento y sólo ajustar una parte de la arquitectura encargada de clasificar palabras (*tokens*) u oraciones de entrada, utilizando sólo pocos datos.

En este trabajo, comparamos la afinación de modelos de seis arquitecturas tipo BERT: BERT-base [31], BioBERT [34], BioLinkBERT [35], BioMegatron [36], BioRoBERTa [37] y LUKE [38]. Los modelos fueron afinados con búsqueda por *grid* de dos hiperparámetros:

batch size: 10, 16, 32, 64.

learning rate: 1e-5, 3e-5, 5e-5.

Para obtener los mejores hiperparámetros, utilizamos una estrategia de parada anticipada (*early stopping*) cuando la pérdida de entropía cruzada en el conjunto de datos de validación no mejorara en dos épocas (*epochs*). El modelo final fue creado uniendo los conjuntos de datos de entrenamiento y validación; luego fue evaluado con los datos de prueba y las métricas más usadas en problemas de clasificación: *F1-score*, *Precision*, *Recall* y el coeficiente de correlación de Matthews (*Matthews correlation coefficient*, MCC).

4.1.2. Datos de *E. coli* para entrenamiento, validación y prueba

Un conjunto de 1562 oraciones de literatura de *E. coli* K-12, en las que se describen interacciones de regulación transcripcional entre TFs y genes, se clasificó manualmente en 4 categorías de acuerdo al tipo de efecto regulatorio del TF sobre el gen: activación, represión, regulación y sin relación. Este conjunto de datos fue dividido en conjuntos para entrenamiento (999 oraciones), validación (250 oraciones) y prueba (313 oraciones). Estas oraciones fueron proporcionadas para nuestro estudio, sin clasificación en las cuatro categorías, por el equipo de curación de RegulonDB, la base de datos con la principal TRN disponible públicamente para *E. coli* [39]. RegulonDB comprende una gran colección de interacciones regulatorias seleccionadas manualmente durante décadas a partir de la literatura biomédica y, en su momento, su equipo de curación recopiló el conjunto de 1562 oraciones.

4.1.3. Datos de *Salmonella* para la extracción de una TRN

El equipo de curación de RegulonDB también nos proporcionó un conjunto de 3005 oraciones con interacciones de regulación de la bacteria *Salmonella enterica* serovar Typhimurium (*Salmonella*), que curó manualmente de 264 artículos de literatura biomédica. *Salmonella* es uno de los principales patógenos, a nivel mundial, que infectan humanos y animales [40] y su regulación transcripcional ha sido estudiada para enfrentar problemas de resistencia a antimicrobianos [41]. Este conjunto de oraciones fue utilizado para evaluar la

extracción automática de nuestro mejor modelo BERT, usando los mismos 264 artículos curados manualmente y comparando las interacciones curadas con las interacciones extraídas. Estos artículos fueron seleccionados por el equipo de RegulonDB, nosotros sólo los recopilamos y procesamos (ver sección *Evaluación de la extracción de una TRN de Salmonella*). Un análisis desde la perspectiva biológica de la TRN extraída también fue realizado.

4.2. Resultados y Discusión

4.2.1. Mejor modelo BERT afinado

Encontramos que el mejor modelo BERT afinado con las oraciones de *E. coli* fue de tipo LUKE con una entropía cruzada de 0.4024 y un *F1-Score Macro* de 0.9107, ambos medidos en el conjunto de validación. Es interesante que el modelo LUKE superara a los modelos BERT de dominio biomédico; LUKE es especializado en tarea (*task-specific*) y no en dominio (*domain-specific*). Consideramos que los mecanismos de auto-atención enfocados en entidades (*entity-aware self-attention*) del modelo LUKE fueron relevantes para lograr un mejor aprendizaje [38].

Este modelo logró un rendimiento significativo en el conjunto de datos de prueba con *Precision* de 0.8601 y *Recall* de 0.8788 (*F1-Score*: 0.8685), lo que indica que, cuando el modelo predice un tipo de regulación para una interacción, el 86% de las veces es correcto y que el 87% de las oraciones totales tendrá asignada correctamente su tipo de regulación. El modelo obtuvo un MCC de 0.8163. Este coeficiente indica la correlación entre los tipos de regulación de las oraciones del conjunto de prueba y los tipos predichos por el modelo; dado que este coeficiente va de -1 a +1, nuestro mejor modelo obtuvo una correlación sobresaliente.

Finalmente, resaltamos que el rendimiento de nuestro mejor modelo fue significativo a pesar del desbalance de datos en las categorías de regulación de las oraciones del conjunto de prueba (activación: 113, represión: 56, regulación: 44 y sin relación: 100). De hecho, la segunda categoría mejor predicha fue regulación (*F1-Score*: 0.8915), no obstante, era la categoría con menos oraciones. Todas las categorías obtuvieron un *F1-Score* arriba de 0.82.

4.2.2. Evaluación de la extracción de una TRN de *Salmonella*

Como mencionamos, el estudio nos permitió extraer automáticamente una TRN de la bacteria *Salmonella*, utilizando 264 artículos completos correspondientes a los datos curados que nos fueron proporcionados. Estos artículos fueron descargados en archivos PDF aprovechando el acceso abierto (*open-access*) o el acceso dado por la UNAM. Luego, los archivos PDF fueron convertidos en archivos de texto plano con una herramienta desarrollada por el Programa de Genómica Computacional del CCG-UNAM y separados en oraciones con la herramienta *Stanford CoreNLP* [42]. El reconocimiento de genes y TFs en estas oraciones fue realizado mediante una lista de ambas entidades, proporcionada por el equipo de RegulonDB. Después, el conjunto de oraciones fue ingresado al modelo LUKE para su clasificación en alguno de los tipos de interacción (activación, represión o regulación) o sin relación. La comparación de esta TRN con la curada manualmente mostró que nuestro modelo pudo extraer correctamente el 82% de la red curada (*Recall*: 0.8217). Además, en una revisión de 158 oraciones seleccionadas aleatoriamente de aquellas predichas erróneamente por el modelo, pudimos constatar que 40% de estas oraciones fueron en realidad bien predichas y pudimos extraer 25 interacciones ausentes en la curación manual. Nuestro proyecto es uno de los pioneros en obtener un enfoque exitoso basado en modelos BERT para extraer este tipo específico de interacciones [43].

4.2.3. Análisis biológico de la TRN de *Salmonella*

El estudio también nos permitió examinar, desde una mirada biológica, la TRN de *Salmonella* extraída por nuestro modelo. Utilizamos las 1826 interacciones regulatorias extraídas automáticamente para construir una red y visualizarla con Cytoscape [44]. Con este software, se hizo además un análisis que calculó la centralidad y el número de conexiones asociadas a los nodos de la red (*degree*). Observamos que el TF PhoP presentó el quinto valor mayor de centralidad (0.2642) y el valor más alto de *degree* (180), que indicó que PhoP presentó el mayor número de enlaces relevantes (*edges*) entre nodos en la red. Biológicamente, nos interesó seleccionar una comunidad relevante de la TRN para un análisis más profundo, por lo que la red del TF PhoP fue seleccionada. Esta comunidad (subred) contiene 90 nodos con 180 enlaces (Figura 2).

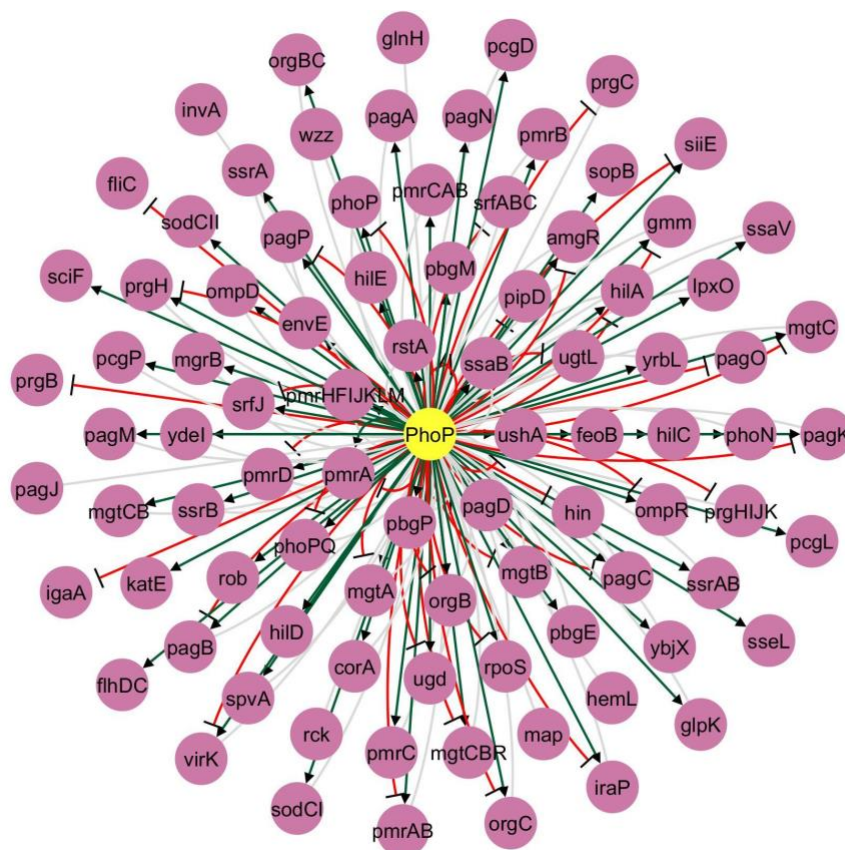


Fig. 2. Red de la comunidad del TF PhoP. Los nodos describen al factor de transcripción (amarillo) y genes regulados (rosa). Por otro lado, los enlaces (edges) se caracterizan por su efecto regulador: activación (verde), represión (rojo) y regulación (gris).

Ahora bien, la lista de genes de la comunidad de PhoP fueron los datos de entrada para el sistema PANTHER [45], con el fin de identificar sus procesos biológicos y características de función biológica mediante un análisis de sobrerrepresentación de elementos con la prueba exacta de Fisher. Al analizar los resultados arrojados por el sistema PANTHER, se encontraron dos procesos biológicos relevantes e interesantes: el primero asociado a la función de transportadores de membrana de iones de Magnesio, con un *P-value* asociado a la prueba de Fisher de $8.31e^{-05}$ y FDR de $3.88e^{-02}$; el segundo, a procesos de sistemas de transducción de señales de fósforo con un *P-value* de $1.20e^{-05}$ y FDR de $1.43e^{-02}$.

Lo siguiente que decidimos realizar fue describir la función de PhoP. Lo primero que identificamos es que esta proteína regula el 3% de los genes de *Salmonella* y pertenece al

sistema de dos componentes PhoP/PhoQ [46]. PhoP tiene influencia en la virulencia junto con la proteína sensora PhoQ y se activa por múltiples señales o variaciones ambientales, incluyendo niveles bajos de Mg^{2+} , ciertos péptidos antimicrobianos y ácidos grasos insaturados de cadena larga. PhoQ promueve el estado fosforilado de PhoP, en consecuencia, la fosforilación de PhoP puede tener efectos transcripcionales sobre los genes y puede modificar la transcripción en *Salmonella* [46]. En futuras aplicaciones, los bioinformáticos, biólogos y médicos podrían generar estudios directos en este TF en el área de farmacología, agronomía y clínica, entre otros.

4.2.4. Siguiendo pasos en el estudio

Como vimos, los avances recientes en las tecnologías de IA ofrecen el potencial de ayudar al proceso de curación e integración del conocimiento extraído de la literatura biomédica. Por ello, para la convocatoria 2023 de los *Proyectos de investigación en Inteligencia Artificial en el Espacio de Innovación UNAM-HUAWEI*, propusimos dar un avance más en el acceso y organización de conocimiento biomédico sobre la TRNs. El proyecto que estamos desarrollando ahora consiste en obtener un modelo BERT afinado para extraer automáticamente las condiciones en las que se dan las interacciones regulatorias bacterianas, usando miles de registros textuales tomados de la *Gene Expression Omnibus* (GEO) [47]. La GEO es un repositorio a gran escala de datos genómicos funcionales de tecnologías de alto rendimiento. Nuestro objetivo final es el reconocimiento y extracción de condiciones experimentales de crecimiento celular de miles de registros de la GEO usando el mejor modelo afinado. Recuperar las condiciones de crecimiento en las que se manejan los experimentos biológicos depositados en la GEO es crucial para contextualizar, reproducir y comparar la información de análisis posteriores.

Actualmente contamos con resultados preliminares en la extracción de las siguientes condiciones utilizadas en experimentos microbianos: Aireación, agitación, medio de cultivo, antecedentes genéticos, fases de crecimiento, suplementos del medio, densidad óptica, pH, temperatura y recipiente. Para llegar a estos resultados, hemos comparado tres arquitecturas tipo BERT: BERT-base, BioBERT y BioLinkBERT. Un modelo BERT-base afinado superó a los otros dos modelos de dominio biomédico (BioBERT y BioLinkBERT) logrando un *F1-score* relevante de 0.9424 (*Precision*: 0.9449, *Recall*: 0.9676) y un alto MCC de 0.9544 en un conjunto de datos de prueba. El valor obtenido del MCC mostró un rendimiento relevante a

pesar de la distribución desequilibrada de instancias en los diez diferentes tipos de condiciones. Nuestro enfoque se podrá eventualmente aplicar para identificar y anotar las condiciones de crecimiento de miles de experimentos de la GEO para respaldar los análisis posteriores de datos genómicos que realicen otros grupos de investigación [48].

5. Conclusión

Este trabajo presentó la extracción automática de interacciones regulatorias transcripcionales de dos bacterias de importancia biológica (*E. coli* y *Salmonella*) con el uso de aproximaciones de IA y mediante el procesamiento de literatura biomédica en el marco de los *Proyectos de investigación en Inteligencia Artificial en el Espacio de Innovación UNAM-HUAWEI*. Ya que nuestro trabajo resaltó la importancia del TF PhoP y sus interacciones regulatorias en *Salmonella*, nuestra aproximación puede ser de utilidad para el desarrollo futuro de nuevas aplicaciones en el área farmacéutica con estudios de resistencia farmacológica o en aplicaciones en el área industrial. Además, este trabajo deja claro el potencial de las aproximaciones actuales de IA, como los *transformers* pre-entrenados BERT, para la extracción de interacciones regulatorias transcripcionales de bacterias que aún carecen de una TRN curada y publicada. Nuestro estudio se suma a muchos otros que aprovechan las tecnologías asociadas a la IA para extraer, organizar y expandir el conocimiento biológico.

Financiamiento

Investigación realizada gracias al Programa UNAM-PAPIIT IN219523.

Agradecimientos

Nuestra gratitud por el soporte computacional, servicio y becas que otorgó el *Espacio de Innovación UNAM-HUAWEI* de la *Alianza para Promover el Desarrollo de Capacidades Digitales en México*. Nuestro agradecimiento al equipo de curación de RegulonDB y Sara Berenice Martínez Luna por los conjuntos de oraciones curadas de *E. coli* y *Salmonella*, respectivamente. Agradecemos enormemente a los revisores del manuscrito por sus comentarios, que ayudaron a mejorar la calidad y brindar robustez al artículo. También agradecemos a los Técnicos Académicos Víctor Del Moral y Alfredo Hernández por el soporte informático para realizar este estudio.

Referencias

- [1] X. Fang, A. Sastry, N. Mih, D. Kim, J. Tan, J.T. Yurkovich, C.J. Lloyd, Y. Gao, L. Yang, and B.O. Palsson, "Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities," in Proc. Natl. Acad. Sci., vol. 114, no. 38, pp. 10286-10291 Sep. 2017. Available: <https://doi.org/10.1073/pnas.1702581114>
- [2] J.A. Freyre-Gonzalez and L.G. Trevino-Quintanilla, "Analyzing regulatory networks in bacteria," Nat. Educ., vol. 3, no. 24, 2010. [Online]. Available: <https://www.nature.com/scitable/topicpage/analyzing-regulatory-networks-in-bacteria-14426192/> [Accessed Feb. 28, 2025].
- [3] E. Balleza, L.N. López-Bojorquez, A. Martínez-Antonio, O. Resendis-Antonio, I. Lozada-Chávez, Y.I. Balderas-Martínez, S. Encarnación, and J. Collado-Vides, "Regulation by transcription factors in bacteria: beyond description," FEMS Microbiol. Rev., vol. 33, no.1, pp. 133-151, Jan. 2009. Available: <https://doi.org/10.1111/j.1574-6976.2008.00145.x>
- [4] M.M. Babu and S. A. Teichmann, "Evolution of transcription factors and the gene regulatory network in Escherichia coli," Nucleic Acids. Res., vol. 31, no. 4, pp. 1234-1244, Feb. 2003. Available: <https://doi.org/10.1093/nar/gkg210>

- [5] K. Yamamoto, "The hierarchic network of metal-response transcription factors in *Escherichia coli*," *Biosci. Biotechnol. Biochem.*, vol. 78, no. 5, pp. 737-747, May 2014. Available: <https://doi.org/10.1080/09168451.2014.915731>
- [6] C. Dalldorf, Y. Hefner, R. Szubin, J. Johnsen, E. Mohamed, G. Li, J. Krishnan, A.M. Feist, B.O. Palsson, and D.C. Zielinski, "Diversity of Transcriptional Regulatory Adaptation in *E. coli*," *Mol. Biol. Evol.*, vol. 41, no. 11, p. msae240, Nov. 2024. Available: <https://doi.org/10.1093/molbev/msae240>
- [7] L. Praski Alzrigat, D.L. Huseby, G. Brandis, and D. Hughes, "Fitness cost constrains the spectrum of *marR* mutations in ciprofloxacin-resistant *Escherichia coli*," *J. Antimicrob. Chemother.*, vol. 72, no. 11, pp. 3016-3024, Aug. 2017. Available: <https://doi.org/10.1093/jac/dkx270>
- [8] D.H. Shah, "RNA sequencing reveals differences between the global transcriptomes of *Salmonella enterica* serovar enteritidis strains with high and low pathogenicities," *Appl. Environ. Microbiol.*, vol. 80, no. 3, Feb. 2014. Available: <https://doi.org/10.1128/AEM.02740-13>
- [9] H. Salgado, S. Gama-Castro, P. Lara, C. Mejia-Almonte, G. Alarcón-Carranza, A.G. López-Almazo, F. Betancourt-Figueroa, P. Peña-Loredo, S. Alquicira-Hernández, D. Ledezma-Tejeida, L. Arizmendi-Zagal, F. Mendez-Hernandez, A.K. Diaz-Gomez, E. Ochoa-Praxedis, L.J. Muñiz-Rascado, J.S. García-Sotelo, F.A. Flores-Gallegos, L. Gómez, C. Bonavides-Martínez, V.M. Del Moral-Chávez, A.J. Hernández-Alvarez, A. Santos-Zavaleta, S. Capella-Gutierrez, J. L. Gelpi, and J. Collado-Vides, "RegulonDB v12.0: a comprehensive resource of transcriptional regulation in *E. coli* K-12," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D255-D264, Jan. 2024. Available: <https://doi.org/10.1093/nar/gkad1072>
- [10] P. D. Karp, S. Paley, R. Caspi, A. Kothari, M. Krummenacker, P. Midford, L.R. Moore, P. Subhraveti, S. Gama-Castro, V. Tierrafria, P. Lara, L. Muñiz-Rascado, C. Bonavides-Martinez, A. Santos-Zavaleta, A. Mackie, G. Sun, T.A. Ahn-Horst, H. Choi, M.W. Covert, J. Collado-Vides, and I. Paulsen "The EcoCyc Database (2023)," *EcoSal Plus*, vol. 11, no. 1, pp. 1-22, Dec. 2023. Available: <https://doi.org/10.1128/ecosalplus.esp-0002-2023>

- [11] E. Galán-Vásquez, B.C. Luna-Olivera, M. Ramírez-Ibáñez, and A. Martínez-Antonio, "RegulomePA: a database of transcriptional regulatory interactions in *Pseudomonas aeruginosa* PAO1," *Database (Oxford)*, vol. 2020, p. baaa106, Dec. 2020. Available: <https://doi.org/10.1093/database/baaa106>
- [12] M. Olbei, B. Bohar, D. Fazekas, M. Madgwick, P. Sudhakar, I. Hautefort, A. Métris, J. Baranyi, R.A. Kingsley, and T. Korcsmaros, "Multilayered Networks of *Salmonella* Genus on a Molecular Level," *MSystems*, vol.7, no.4, p. e01493-21, Aug. 2022. Available: <https://doi.org/10.1128/msystems.01493-21>
- [13] J.M. Escorcía-Rodríguez, A. Tauch, and J.A. Freyre-González. "Abasy atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1228-1237, Jan. 2020. Available: <https://doi.org/10.1016/j.csbj.2020.05.015>
- [14] S.R. Davies, "Working in biocuration: contemporary experiences and perspectives," *Database (Oxford)*, vol. 2025, p. baaf003, Feb. 2025. Available: <https://doi.org/10.1093/database/baaf003>
- [15] National Center for Biotechnology Information (NCBI), "Pubmed Central," National Library of Medicine (US). [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/>. [Accessed Feb. 28, 2025].
- [16] S. Burge, T.K. Attwood, A. Bateman, T.Z. Berardini, M. Cherry, C. O'Donovan, L. Xenarios, and P. Gaudet. , "Biocurators and biocuration: surveying the 21st century challenges," *Database (Oxford)*, vol. 2012, p. bar059, Mar. 2012. Available: <https://doi.org/10.1093/database/baaf003>
- [17] J. Guallar, L. Codina and E. Abadal, "La investigación sobre curación de contenidos: análisis de la producción académica," *Ibersid.*, vol. 14, no. 1., pp. 13-22, Jun. 2020. Available: <https://repositori.upf.edu/handle/10230/52168>
- [18] I.S. for Biocuration: "Distilling data into knowledge," *PLoS Biol.*, vol. 16, no. 4, p. 2002846, Apr. 2018. Available: <https://doi.org/10.1371/journal.pbio.2002846>

- [19] L. Bornmann, R. Haunschild, and R. Mutz, "Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases," *Humanit. and Soc. Sci. Commun.*, vol. 8, no. 1, pp. 1-15, Oct. 2021. Available: <https://doi.org/10.1057/s41599-021-00903-w>
- [20] S. Ananiadou, P. Thompson, R. Nawaz, J. McNaught, and D. B. Kell, "Event-based text mining for biology and functional genomics," *Briefings Funct. Genomics*, vol. 14, no. 3, pp. 213-230, May. 2014. Available: <https://doi.org/10.1093/bfpg/elu015>
- [21] C.C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: success, failure and the future," *Briefings Bioinf.*, vol. 17, no. 1, pp. 132-144, Jan. 2016. Available: <https://doi.org/10.1093/bib/bbv024>
- [22] F. He, K. Liu, Z. Yang, M. Hannink, R.D. Hammer, M. Popescu, and D. Xu, "Applications of cutting-edge artificial intelligence technologies in biomedical literature and document mining," *Med. Rev.*, vol. 3, no. 3, pp. 200-204, Jun. 2023. Available: <https://doi.org/10.1515/mr-2023-0011>
- [23] M.A. Thorat and R. Balasubramanian, "Breast cancer prevention in high-risk women," *Best Pract. Res. Clin. Obstet. Gynaecol.*, vol. 65, pp. 18-31, May 2020. Available: <https://doi.org/10.1016/j.bpobgyn.2019.11.006>
- [24] S. Zhao, C. Su, Z. Lu, and F. Wang, "Recent advances in biomedical literature mining," *Briefings Bioinf.*, vol. 22, no. 3, May 2020. Available: <https://doi.org/10.1093/bib/bbaa057>
- [25] L. Hong, J. Lin, S. Li, F. Wan, H. Yang, T. Jiang, D. Zhao, and J. Zeng. "A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories". *Nat. Mach. Intell.*, vol. 2, no. 6, pp. 347-355, Jun. 2020. Available: <https://doi.org/10.1038/s42256-020-0189-y>
- [26] M. Zhun, B. Celikkaya, P. Bhatia, and C.K. Reddy, "Latte: Latent type modeling for biomedical entity linking," In *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 34, no. 05, pp. 9757-9764, Apr. 2020. Available: <https://doi.org/10.1609/aaai.v34i05.6526>
- [27] F. Villena, P. Báez, S. Peñafiel, M. Rojas, I. Paredes, and J. Dunstan, "Developing and Validating an Automatic Support System for Tumor Coding in Pathology Reports in

- Spanish," *JCO Clin. Cancer Inform.*, vol. 9, p. e2400124, 2025. Available: <https://doi.org/10.1200/CCI.24.00124>
- [28] E. Jung, H. Jain, A.P. Sinha, and C. Gaudioso, "Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis," *Health Informatics J.*, vol. 27, no. 1, Jan. 2021. Available: <https://doi.org/10.1177/1460458221989392>
- [29] J.M. Czarnecki and A.J. Shepherd, "Metabolic Pathway Mining," *Methods Mol. Biol.*, vol. 2, pp. 139-158, Nov. 2016. Available: https://doi.org/10.1007/978-1-4939-6613-4_8
- [30] Created in BioRender. Available: <https://www.BioRender.com/>
- [31] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, Association for Computational Linguistics, vol. 1, pp. 4171-4186, May. 2019. Available: <https://doi.org/10.18653/v1/N19-1423>
- [32] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," [Online]. 2018. Available: https://cdn.openai.com/researchcovers/languageunsupervised/language_understanding_paper.pdf [Accessed Feb. 28, 2025].
- [33] C. Mottes, "BERT vs GPT: Comparing the two most popular language models", 29 August, 2024. [Online]. Available: <https://blog.invgate.com/gpt-3-vs-bert> [Accessed Feb. 28, 2025].
- [34] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, Feb. 2020. Available: <https://doi.org/10.1093/bioinformatics/btz682>
- [35] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining Language Models with Document Links," In S. Muresan, P. Nakov, A. Villavicencio, Eds. *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland:

- Association for Computational Linguistics, vol. 1, pp. 8003–8016, Mar. 2022. Available: <https://aclanthology.org/2022.acl-long.551>
- [36] H.C. Shin, Y. Zhang, E. Bakhturina, R. Puri, M. Patwary, M. Shoeybi, and R. Mani, "BioMega-tron: Larger Biomedical Domain Language Model," In B. Webber, T. Cohn, Y. He, Y. Liu, Eds. *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, pp. 4700–4706, Oct. 2020. Available: <https://aclanthology.org/2020.emnlp-main.379>
- [37] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N.A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," In D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, pp. 8342–8360, Jul. 2020. Available: <https://aclanthology.org/2020.acl-main.740>
- [38] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention," In B. Webber, T. Cohn, Y. He, Y. Liu, Eds. *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, pp. 6442–6454, Nov. 2020. Available: <https://aclanthology.org/2020.emnlp-main.523>
- [39] V.H. Tierrafría, C. Rioualen, H. Salgado, P. Lara, S. Gama-Castro, P. Lally, L. Gómez-Romero, P. Peña-Loredo, A.G. López-Almazo, G. Alarcón-Carranza, F. Betancourt-Figueroa, S. Alquicira-Hernández, J.E. Polanco-Morelos, J. García-Sotelo, E. Gaytan-Núñez, C.F. Méndez-Cruz, L.J. Muñiz, C. Bonavides-Martínez, G. Moreno-Hagelsieb, J.E. Galagan, J.T. Wade, and J. Collado-Vides, "RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12," *Microb. Genom.*, vol. 8, no. 5, p. 000833, May 2022. Available: <https://doi.org/10.1099/mgen.0.000833>
- [40] A. Fàbrega and J. Vila, "*Salmonella enterica* serovar Typhimurium skills to succeed in the host: virulence and regulation," *Clin. Microbiol. Rev.*, vol. 26, no. 2, pp. 308–341, Apr. 2013. Available: <https://doi.org/10.1128/cmr.00066-12>

- [41] J.S., Gunn and S.I. Miller, "PhoP-PhoQ activates transcription of pmrAB, encoding a two-component regulatory system involved in *Salmonella typhimurium* antimicrobial peptide resistance," *J. Bacteriol.* vol. 178, no 23, pp. 6857–6864, Dec. 1996. Available: <https://doi.org/10.1128/jb.178.23.6857-6864.1996>
- [42] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in K. Bontcheva and J. Zhu, Eds. *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, Maryland: Association for Computational Linguistics, pp. 55-60, Jun. 2014. Available: <https://aclanthology.org/P14-5010/>
- [43] A. Varela-Vega, A.B. Posada-Reyes, and C.F. Méndez-Cruz, "Automatic extraction of transcriptional regulatory interactions of bacteria from biomedical literature using a BERT-based approach," *Database (Oxford)*, vol. 2024, p. baae094, Aug. 2024. Available: <https://doi.org/10.1093/database/baae094>
- [44] D. Otasek, J.H. Morris, J. Boucas, A.R. Pico, and B. Demchank, "Cytoscape automation: empowering workflow-based network analysis," *Genome Biol.*, vol. 20, pp. 1-15, Sep. 2019. Available: <https://cytoscape.org/>. [Accessed Feb. 28, 2025].
- [45] H. Mi, A. Muruganujan, J.X. Huang, D. Ebert, C. Mills, X. Gou, and P.D. Thomas, "Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)," *Nat. Protoc.*, vol. 14, no. 3, pp. 703-721, Feb. 2019. Available: <https://doi.org/10.1038/s41596-019-0128-8>
- [46] E.A. Groisman, A. Duprey, and J. Choi, "How the PhoP/PhoQ system controls virulence and Mg²⁺ homeostasis: lessons in signal transduction, pathogenesis, physiology, and evolution," *Microbiol. Mol. Biol. Rev.*, vol. 85, no. 3, p. e0017620, Aug. 2021. Available: <https://doi.org/https://doi.org/10.1128/membr.00176-20>
- [47] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets-update," *Nucleic. Acids. Res.*, vol. 41, pp. D991-D995, Nov. 2012. Available: <https://doi.org/10.1093/nar/gks1193>

- [48] A. Varela-Vega, M. Ávila-García, A. León Mendoza, A.B. Posada-Reyes, Víctor H. Tierrafría, C.F. Méndez-Cruz, and J. Collado-Vides, "Automatic annotation of growth conditions from *Escherichia coli* experiments deposited in the Gene Expression Omnibus," Presented at the 19th *Conf. on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2024)*, Benevento, Italy, 2024.