//UCIENCIA/23
CONVENCIÓN CIENTÍFICA INTERNACIONAL

Temática: Bases de datos en Bioinformática

BASE DE DATOS DE SITIOS DE FOSFORILACIÓN

DATABASE OF PHOSPHORYLATION SITES

Keren Sánchez Padrón [1*], Jamilet Miranda Navarro[2], Ricardo Bringas Pérez[3]

[1] Universidad de las Ciencias Informáticas, Carretera de San A. de los Baños kilómetro 2½, Torrens, Boyeros, La Habana, Cuba. kerensanchezpadron@gmail.com
[2] Centro de Ingeniería Genética y Biotecnología, PO BOX 6162, La Habana, Cuba. jamilet.miranda@cigb.edu.cu
[3] Centro de Ingeniería Genética y Biotecnología, PO BOX 6162, La Habana, Cuba. bringas@cigb.edu.cu

[*] Autor para correspondencia: kerensanchezpadron@gmail.com

Resumen

Actualmente, el grupo de Bioinformática del Centro de Ingeniería Genética y Biotecnología (CIGB) cuenta con una herramienta para la construcción, análisis y visualización de redes biológicas. Esta aplicación se basa en la información biológica proporcionada por una base de datos integradora, llamada SysBiomics, que contiene información sobre genes, proteínas, interacciones proteína-proteína y proteína-ADN, ontologías de genes y vías metabólicas de múltiples fuentes de datos públicas. Sin embargo, carece de datos sobre modificaciones postraduccionales. Debido a la importancia del almacenamiento de información de fosforilación para ayudar a los investigadores a encontrar datos biológicos relevantes, la presente investigación tiene como propósito desarrollar una base de datos que contenga información de fosforilación, que sirva como nuevo módulo de SysBiomics. La metodología seleccionada para llevar a cabo este proceso fue la XP (del inglés, *Extreme Programming*), se seleccionaron dos fuentes para la integración de datos, se aplicaron técnicas de modelado de datos para obtener el diseño de la base de datos y se realizaron diferentes casos de pruebas a nivel de unidad para comprobar su correcto funcionamiento. Como resultado se obtuvo un repositorio de datos que contiene información actualizada de sitios de fosforilación que puede incorporarse a la base de datos SysBiomics.

Palabras clave: base de datos, bioinformática, fosforilación, integración

Abstract

Currently, the Bioinformatics group of the Center for Genetic Engineering and Biotechnology (CIGB) has a tool for the construction, analysis and visualization of biological networks. This application is based on the biological information provided by an integrative database, called SysBiomics, which contains information on genes, proteins, protein-protein and protein-DNA interactions, gene ontologies, and metabolic pathways from multiple public data

sources. However, it lacks data on post-translational modifications. Due to the importance of storing phosphorylation information to help researchers to find relevant biological data, the purpose of this investigation is to develop a database containing phosphorylation information, which serves as new module of SysBiomics. The methodology selected to carry out this process was XP (Extreme Programming), two sources were selected for data integration, data modeling techniques were applied to obtain the design of the database and different test cases at the unit level to verify its correct operation. As a result, a data repository containing updated information on phosphorylation sites was obtained that can be incorporated into SysBiomics database.

Keywords: *bioinformatics, database, integration, phosphorylation*

Introducción

Teniendo en cuenta la importancia de la fosforilación en la comprensión de los sistemas biológicos de las proteínas y su relevancia para el diseño de fármacos, el desarrollo de los métodos experimentales para determinar estos sitios se ha intensificado en los últimos años. Estos han generado gran cantidad de información que, a su vez, ha propiciado la aparición de bases de datos para su almacenamiento, clasificación y distribución (Yang et al., 2021).

Las bases de datos dedicadas al almacenamiento de información de fosforilación ayudan a los investigadores a encontrar datos biológicos relevantes, al integrar el contenido de múltiples estudios y ponerlos a disposición en un formato legible e interactivo. Además, se han vuelto trascendentales para proporcionar la infraestructura necesaria para la investigación sobre esta modificación, desde la preparación de datos hasta su extracción (Luo et al., 2019).

En los últimos tiempos el desarrollo de la biología ha alcanzado un nivel elevado, en lo que han jugado un papel fundamental las ciencias de la computación y la información, las matemáticas y la estadística, las que se han fusionado en la disciplina conocida hoy como Bioinformática. Esta ciencia proporciona los aspectos fundamentales a tener en cuenta a la hora de realizar análisis a los sistemas biológicos y a la vez está relacionada con los procesos informáticos que garantizan el almacenamiento de los datos generados en dichos sistemas.

Los nuevos modelos para la integración de datos, las especificaciones estándar para el intercambio de datos y el desarrollo de nuevas herramientas para la visualización y el análisis de datos son cruciales y representan una de las

tareas más desafiantes para los bioinformáticos (Shen et al., 2022). En este contexto, la industria biotecnológica cubana ha mostrado numerosos avances en la producción de este tipo de *software*, en instituciones como el Centro de Ingeniería Genética y Biotecnología (CIGB), el cual cuenta con una Bases de Datos Integradora o Data Warehouse (DW), llamada SysBiomics, que fue desarrollada por el grupo de Bioinformática del centro. Esta base de datos, administrada por PostgreSQL, combina de forma, no redundante, datos biológicos provenientes de múltiples fuentes de datos públicas, y tiene como principales dominios de información las interacciones moleculares y los procesos biológicos moleculares, entre otros (Martin et al., 2010). Sin embargo, el CIGB no cuenta con un repositorio de datos propio que contenga información sobre sitios de fosforilación, para la extensión de nuevos dominios de información biológica. A fin de solucionar este problema se define como objetivo general de la investigación: Desarrollar una base de datos que contenga información de fosforilación, a partir de la integración las bases de datos públicas y estudios que contienen esta información.

Materiales y métodos

Flujo de trabajo general

En el flujo de trabajo general que se llevó a cabo en esta investigación, la selección de diferentes fuentes de datos que contienen información de sitios de fosforilación, a partir del análisis de varios repositorios, fue el primer paso que se realizó, el cual permitió aumentar el dominio de conocimiento e identificar las posibles entidades, atributos, tipos de datos y relaciones que contribuirán a un mejor diseño de la base de datos (Konjevoda and Štambuk, 2021). El segundo paso fue la realización del diseño, para este se empleó el modelo relacional con las normas postuladas en 1970 por Edgar Frank Codd (Codd, 1970), el cual se dividió en tres etapas: diseño conceptual, diseño lógico y diseño físico de la base de datos. En la fase de implementación se realizan la serie de procedimientos almacenados o funciones capaces de realizar todas las tareas de la entrada de datos independiente de la fuente de datos. El sistema finaliza con la obtención de diferentes casos de pruebas que permiten su validación. Sin embargo, un aspecto a tener en cuenta es que la metodología XP es flexible, si la validación arroja resultados desfavorables permite volver al paso o fase de codificación y realizar los cambios pertinentes, hasta que los resultados sean favorables.

Análisis y selección de fuentes de datos

Un primer punto a abordar fue la revisión de diferentes fuentes de datos de eventos de fosforilación que podíamos disponer. Luego de revisar diversas fuentes, se utilizó PhosphositePlus (https://www.phosphosite.org/) como fuente de datos, para garantizar de esta manera la futura actualización de la base de datos. En la tabla 1 se muestran los nombres de los ficheros tabulados pertenecientes al conjunto de datos de sitios de modificación de PhosphositePlus y, de estos, los dos ficheros seleccionados para la base de datos.

Tabla 1. Ficheros presentes en PhosphoSitePlus

Ficheros PhosphoSitePlus	Ficheros a emplear
Acetylation_site_dataset	
Disease-associated_sites	
Kinase_Substrate_Dataset	x
Methylation_site_dataset	
O-GalNAc_site_dataset	
O-GlcNAc_site_dataset	
Phosphorylation_site_dataset	x
Phosphosite_PTM_seq	
Phosphosite_seq	
PTMVar	
Regulatory_sites	
Sumoylation_site_dataset	
Ubiquitination_site_dataset	

Del fichero Kinase_Substrate_Dataset se seleccionaron, a partir de las necesidades de la investigación, el siguiente conjunto de datos:

- GENE: contiene el símbolo oficial del gen.
- KINASE: contiene la proteína quinasa que fosforila al sitio.
- KIN_ACC_ID: a cada entrada de enzima kinasa se le asigna el número de acceso único, que se denomina ‘código de acceso primario’ (Primary *accession*) de UniProtKB.

- KIN_ORGANISMO: contiene el nombre del organismo asociado a esa quinasa.
- SUBSTRATO: nombre la proteína sustrato sobre el que actúa la enzima correspondiente a esa entrada.
- SUB_GENE_ID: identificador del gene que codifica ese sustrato.
- SUB_ACC_ID: código de acceso primario de UniProtKB del sustrato.
- SUB_GEN: nombre del gene que codifica ese sustrato.
- SUB_ORGANISMO: organism asociado a ese sustrato.
- SUB_MOD_RSD: sitio de fosforilación, contiene el residuo y la posición del sitio que está fosforilado.
- SITE_+/-7_AA: el fosfopéptido predicho con 7 aminoácidos aguas arriba y 7 aminoácidos aguas abajo alrededor del residuo modificado. contiene la secuencia que rodea el PTM (+/- 7 AA)

De Phosphorylation_site_dataset, se seleccionaron:

- GENE: símbolo oficial del gen que codifica a esa proteína.
- PROTEIN: nombre de la proteína fosforilada.
- ACC_ID: número de acceso de la proteína.
- MOD_RSD: sitio de fosforilación.
- ORGANISM: organismo.
- SITE_7_AA: secuencia alrededor del sitio.

PhosphoELM (http://phospho.elm.eu.org/) también se integró a la base de datos. En la tabla 2 se muestran los nombres de los ficheros pertenecientes a esta base de datos y, en esta ocasión, se seleccionó un solo fichero que contenía la información para todos los organismos, el resto se descartó ya que contenía la misma información, pero asociada a una especie específica, como se puede deducir de sus nombres.

Tabla 2. Ficheros pertenecientes a PhosphoELM

Ficheros PhosphoELM	Fichero a emplear
phosphoELM_Caenorhabditis_latest	
phosphoELM_Drosophila_latest	
phosphoELM_all_2015-04	x
phosphoELM_vertebrate_latest	

Del este fichero, se seleccionaron los siguientes datos de su conjunto de información:

- Acc: número de acceso único.
- Sequence: la secuencia de la proteína fosforilada.
- Position: posición del sitio de fosforilación.
- Code: residuo que está siendo fosforilado.
- PMids: número de referencia de PubMed.
- Species: organismos asociados a una entrada.

Diseño de la base de datos

El diseño de la base de datos como se muestra se dividió en tres etapas. Primeramente, se realizó diseño conceptual empleando un diagrama ER. Para transformar el modelo conceptual obtenido en la fase anterior en un modelo lógico que será el resultado de esta fase; en primer lugar, se revisó el modelo conceptual para asegurarnos de que está libre de algunos errores tipificados e identificables, como no se encontraron errores, seguido de esto, como el SGBD en el que se quiere implementar la base de datos es de tipo relacional se transformó el esquema conceptual resultante, en el modelo lógico (conjunto de relaciones con sus atributos, claves primarias y claves foráneas), y para finalizar esta etapa, se aplicó la teoría de la normalización de Codd al modelo lógico. Para modelar la base de datos es importante esta fase de normalización. Los atributos y valores luego de normalizar según la primera forma normal quedan atómicos, es decir, están en su forma mínima, primera forma normal (1FN). Además, las tablas contienen una clave primaria no nula, para garantizar la identificación y la relación entre los datos y tablas. Finalmente, a partir del esquema lógico se desarrolló el diseño físico de la base de datos con Visual Paradigm 8.0.

Flujo de trabajo de implementación

Una vez identificadas las fuentes a incorporar, y culminado el diseño de la base de datos, se procedió a diseñar un esquema para su incorporación al módulo de fosforilación que posteriormente va a formar parte de SysBiomics. El flujo de este esquema, incluye un primer paso, en el cual, para procesar los datos primarios de las fuentes, se desarrolló en python un programa analizador (parser), que se encarga de seleccionar la información de interés a incorporar a la base de datos y crea ficheros tabulados con estos datos procesados. Cada fuente de datos cuenta con un programa parser en particular que crea ficheros csv (valores separados por comas) con la información mínima definida para todas las fuentes. Estos ficheros procesados se encuentran en subdirectorios específicos de cada fuente.

En un segundo paso, luego de crear la base de datos en PostgreSQL, se crea una estructura temporal compuesta por dos tablas temporales, cuya nomenclatura se definió como tmp_nombre, que poseen una estructura similar a la de los ficheros tabulados creados. Hasta aquí, el proceso de integración de datos obtiene información de fuentes externas, pero en el proceso de identificación de entidades, comienzan a almacenarse en función de un identificador único de SysBiomics, lo cual facilita la posterior entrada en las tablas con estructura permanentes. En esta fase se implementan la serie de procedimientos almacenados o funciones, en su mayoría implementados en pl/pgSQL, capaces de realizar todas las tareas de la entrada de datos independiente de la fuente de datos, o sea, se llenan las tablas definitivas con sus respectivas relaciones que conformarán la BD.

Resultados y discusión

Diseño de la base de datos

Diseño conceptual:

En esta base de datos el modelo ER consta de cuatro entidades principales (Gene, Protein, Enzyme y PTM), una entidad débil (Transcript, depende de Gene) y otras entidades (XRef, PTM_Type, Taxonomy). En la tabla 3 se puede ver la descripción general de cada entidad presente en el modelo Entidad-Relación.

Tabla 3. Entidades presentes en la base de datos

Nombre	Descripción
Gene	Almacena los genes que codifican las diferentes proteínas fosforiladas.
Protein	Almacena las proteínas fosforiladas existentes en la base de datos.
Transcrip	Almacena los transcritos (diferentes variantes de ARNs) de un gen.
xRef	Contiene las referencias externas de entidades.
Taxonomy	Organismos contenidos en la BD.
PTM	Modificaciones postraduccionales.
PTM_Type	Tipos de modificaciones postraduccionales.
Enzyme	Contiene información de enzimas

Diseño lógico:

A continuación, se puede observar el esquema lógico de la base de datos, donde los atributos con subrayado simple representan las llaves primarias de la entidad correspondiente y los que poseen subrayado doble representan las llaves foráneas. Como se puede apreciar, aparece una nueva tabla en el modelo, la cual se creó debido a que la cardinalidad entre las entidades Enzyme y PTM es de muchos a muchos.

- Gene (Gene_id, Name, NCBI_GeneID, xRef_id, Taxonomy_id)
- Transcript (Transcript_id, Gene_id)
- Protein (Protein_id, Accession, Name, Status, xRef_id, Taxonomy_id, Transcript_id)
- Enzyme (Enzyme_id, Type, xName_EC, Protein_id)
- PTM (PTM_id, Site, Sequence, Protein_id, PTM_Type_id)
- PTM_Type (PTM_Type_id, Name)
- xRef (xRef_id, URL, Name, Description)
- Taxonomy (id, Name)
- Enzyme_PTM (Enzyme_id, PTM_id)

Diseño físico

En la figura 1 se muestra el diseño físico de la base de datos propuesta.

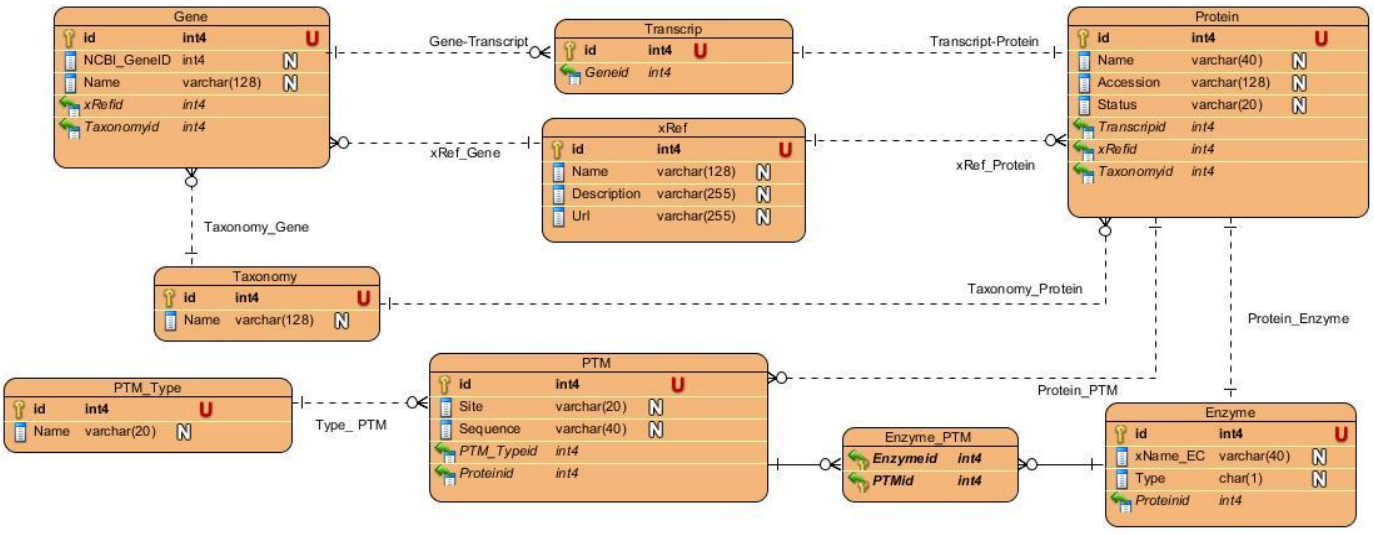

Figura 1. Diseño físico de la Base de Datos

Como se puede apreciar, el diseño contiene un total de 9 tablas, la tabla Enzyme_PTM se genera debido a la relación de muchos a muchos de las entidades que la forman. SysBiomics cuenta con las tablas Gene, Transcript, Protein, Taxonomy y xRef, sin embargo, Enzyme, PTM y PTM_type son las nuevas tablas que formarán parte del esquema phospho.

Implementación de la base de datos

A partir de las tablas temporales y el proceso de identificación de entidades se llenaron las tablas definitivas con sus respectivas relaciones. Para diseñar las tablas se aplicó una técnica de normalización de manera que se evitara la redundancia en los datos almacenados. Siguiendo el esquema de trabajo desarrollado en esta investigación, en la figura 2, se evidencia el proceso de implementación seguido.

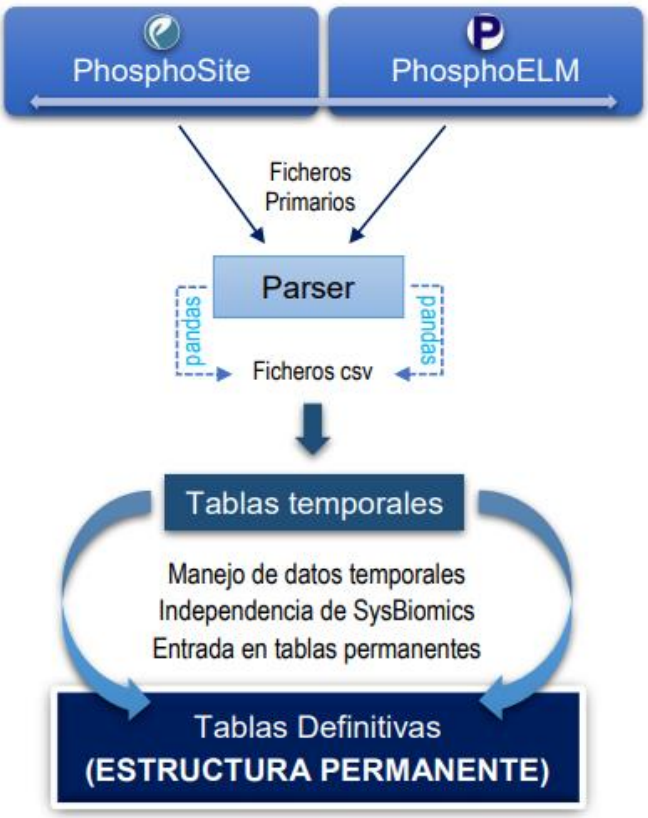

Figura 2. Modelo de implementación de la base de datos.

Principales funciones implementadas:

Manejo de datos temporales:

1. create_tmp_tables: Se crean las estructuras de las tablas temporales para el almacenamiento de los datos primarios de las fuentes de fosforilación.
2. fill_tmp_tables: Procedimiento que, dada la fuente de datos como parámetro, inserta los datos primarios de la fuente dada en las tablas temporales definidas anteriormente.
3. drop_tmp_tables: Procedimiento para borrar las tablas temporales.
4. del_tax_tmp: se mantienen solamente los datos de fosforilación en los datos primarios de organismos seleccionados.

Independencia de SysBiomics:

5. sysbiomic_*: Procedimientos para garantizar la independencia de la base de datos en caso de que no se reciban de SysBiomics, o sea, las tablas de las entidades Genes, Proteínas, xRef y Organismo (Taxonomy).
6. sysbiomic_create_tables: Crea las estructuras de las tablas mencionadas anteriormente.
7. sysbiomic_fill_xref: se cargan los datos fundamentales de las referencias externas declaradas en el fichero Xref.txt.
8. sysbiomic_fill_taxonomy: se almacenan los organismos que tienen datos biológicos en SysBiomics.
9. sysbiomic_fill_tables: Se llenan el resto de las tablas como Genes, Proteínas a partir de los ficheros temporales de los datos de fosforilación.

Entrada en tablas permanentes:

10. Identification_tmp_tables: Proceso de identificación de las entidades en la estructura temporal que sirve de preparación y facilita a posterior entrada de datos en las tablas permanentes.
11. create_phospho_tables: se crean las tablas del esquema phospho.
12. create_source_info_tables: se crean las tablas del esquema source_info.
13. create_relationships: se crean las relaciones entre tablas que garantizan las reglas de integridad referencial.
14. fill_perm_tables: Se llenan las tablas permanentes.

Para incluir estas nuevas funcionalidades a SysBiomics se cuenta con un fichero (.sql) que contiene los comandos con la definición de esquemas, tablas y funciones requeridas para la incorporación de los datos de fosforilación. Estos comandos se adicionarán a la creación actual de la BD SysBiomics, lo cual genera un esquema phospho con el contenido descrito en el trabajo. Luego, se realizan llamados a las funciones implementadas como parte del proceso

de actualización. Esto ocurre siguiendo las diferentes etapas descritas: carga de datos en tablas temporales e identificación de entidades y llenado de tablas permanentes. Los procedimientos para creación y llenado de tablas de SysBiomics se ejecutan dependiendo de su disponibilidad. Una vez terminado este proceso las tablas permanentes como ptm, ptm_type, enzyme, entre otras incluidas en el diseño, contendrán la información que integra múltiples fuentes de datos y que podrá ser consultada a través de un API o directamente por otras aplicaciones clientes como BisoGenet.

Estadísticas de la base de datos

La versión actual de esta base de datos (Versión 1.0, febrero de 2023) consta de un total de 390589 sitios de fosforilación de siete organismos que fueron incorporados. Y se reportan sitios en un total de 45138 proteínas, además, se identificaron 784 proteínas quinasas vinculadas a PTMs. Como era de esperar la mayoría de los sitios corresponden a proteínas humanas, por ser las más estudiadas en trabajos de fosfoproteómica (Paulo and Schweppe, 2021). Una vez completado el llenado de tablas y establecidas las relaciones entre estas, es posible realizar un gran número de consultas sobre el contenido de la base de datos como, por ejemplo: número de sitios de fosforilación que contiene una proteína, número de sitios fosforilados por una quinasa determinada, número de quinasas que fosforilan una proteína, etc. Cualquier interrogación que filtre por el organismo, identificador de una proteína o quinasa, sitio de fosforilación puede ser formulada. Hasta el momento se ha incorporado a la base de datos información de 784 enzimas, donde se tienen 427 de humanos, 248 de ratón y 109 de rata.

Discusión

El estudio de la fosforilación de proteínas ha ido en aumento en la medida en que, por un lado, se ha demostrado su rol en el desarrollo de enfermedades y, por otro lado, se han desarrollado las tecnologías proteómicas que permiten determinar las diferencias en el grado de fosforilación entre diferentes condiciones. Recientemente, en los estudios de infección del virus SARS-CoV-2 (causante de la Covid-19), se evidenció la importancia de la fosforilación de las proteínas para el ciclo de vida de este virus, una vez que infecta una célula. En estudios de fosfoproteómica de los efectos de la infección por SARS-CoV-2 de células humanas se demostró que ocurre una activación de la actividad proteína-quinasa de la enzima CK2 y el uso de inhibidores de esta quinasa resultó en un potente efecto antiviral (Bouhaddou et al., 2020).

Este papel de las quinasas ha sido igualmente reportado para otros virus. Estos mecanismos intervienen en el secuestro de proteínas implicadas en funciones celulares que son entonces utilizadas para aumentar la replicación y propagación viral (Laure et al., 2020). Fueron estos antecedentes, unido el desarrollo previo por el grupo de Bioinformática del CIGB del software BisoGenet para la construcción y análisis de redes biológicas, y en particular de una base de datos integradora de información de genes y proteínas, SysBiomics, lo que llevó a diseñar esta investigación con el objetivo de integrar información de fosforilación al desarrollo de software para la Biología de Redes. Esta base de datos es punto inicial para el estudio de las relaciones de genes, proteínas y enzimas vinculados a la fosforilación para lograr aprovechar los datos biológicos de SysBiomics.

El estudio de los antecedentes teóricos, la modelación y el análisis de las herramientas disponibles propiciaron lograr el objetivo propuesto en este trabajo. Fue necesario hacer un estudio de las fuentes de datos disponibles, de la base de datos SysBiomics y sus diferentes entidades y relaciones para el diseño e implementación de una base de datos que contiene modificaciones post-traducionales. El diseño tuvo en cuenta los requerimientos para el desarrollo de aplicaciones como BisoGenet, que en versiones futuras debe mostrar la información incorporada en este trabajo.

Se puede concluir que la presente investigación generó una base de datos de fosforilación de proteínas que contiene información actualizada. Una vez incorporada esta información a SysBiomics de conjunto con información de interacciones proteína-proteína, regulación de genes por factores de transcripción, eventos de silenciamiento por microRNAs y otras informaciones, enriquecerá la base de conocimientos para el desarrollo de aplicaciones para la Biología de redes.

Conclusiones

- A partir de la elaboración del marco teórico de la investigación, se trazó un flujo de trabajo para diseñar e implementar una base de datos que contenga información de sitios de fosforilación.
- Se dispone de una fuente de datos de fosforilación actualizada que puede incorporarse a la base de datos SysBiomics. El vínculo de estos datos con anotaciones funcionales de genes y proteínas, con múltiples fuentes de interacciones proteína - proteína y con rutas biológicas servirá como base de conocimiento para el desarrollo de aplicaciones de la Biología de Redes.

Para futuras investigaciones y como continuidad de la actual, se recomienda:

- Incorporar otras fuentes de datos como, por ejemplo, iPTMnet, mediante el acceso al API con que cuenta esta fuente de datos de modificaciones postraduccionales.
- Dar continuidad a esta investigación a partir de la integración de nuevos tipos de modificaciones postraduccionales.

Referencias

Bouhaddou, M., Memon, D., Meyer, B., White, K. M., Rezelj, V. V., Correa Marrero, M., Polacco, B. J., Melnyk, J. E., Ulferts, S., Kaake, R. M., Batra, J., Richards, A. L., Stevenson, E., Gordon, D. E., Rojc, A., Obernier, K., Fabius, J. M., Soucheray, M., Miorin, L., … Krogan, N. J. (2020). The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell*, *182*(3), 685-712.e19. https://doi.org/10.1016/j.cell.2020.06.034

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, *13*(6), 377-387. https://doi.org/10.1145/362384.362685

Konjevoda, P., & Štambuk, N. (2021). Relational model of the standard genetic code. *Biosystems*, *210*, 104529. https://doi.org/10.1016/j.biosystems.2021.104529

Laure, M., Hamza, H., Koch-Heier, J., Quernheim, M., Müller, C., Schreiber, A., Müller, G., Pleschka, S., Ludwig, S., & Planz, O. (2020). Antiviral efficacy against influenza virus and pharmacokinetic analysis of a novel MEK-inhibitor, ATR-002, in cell culture and in the mouse model. *Antiviral Research*, *178*, 104806. https://doi.org/10.1016/j.antiviral.2020.104806

Luo, F., Wang, M., Liu, Y., Zhao, X.-M., & Li, A. (2019). DeepPhos: Prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, *35*(16), 2766-2773. https://doi.org/10.1093/bioinformatics/bty1051

Martin, A., Ochagavia, M. E., Rabasa, L. C., Miranda, J., Fernandez-de-Cossio, J., & Bringas, R. (2010). BisoGenet: A new tool for gene network building, visualization and analysis. *BMC Bioinformatics*, *11*(1), 91. https://doi.org/10.1186/1471-2105-11-91

Paulo, J. A., & Schweppe, D. K. (2021). Advances in quantitative high-throughput phosphoproteomics with sample multiplexing. *PROTEOMICS*, *21*(9), 2000140. https://doi.org/10.1002/pmic.202000140

Shen, W., Song, Z., Zhong, X., Huang, M., Shen, D., Gao, P., Qian, X., Wang, M., He, X., Wang, T., Li, S., & Song, X. (2022). Sangerbox: A comprehensive, interaction-friendly clinical bioinformatics analysis platform. *IMeta*, *1*(3), e36. https://doi.org/10.1002/imt2.36

Yang, H., Wang, M., Liu, X., Zhao, X.-M., & Li, A. (2021). PhosIDN: An integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information. *Bioinformatics*, *37*(24), 4668-4676. https://doi.org/10.1093/bioinformatics/btab551