PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Exploración de métodos de clasificación de proteínas repetidas basado en su información estructural utilizando aprendizaje de máquina

TRABAJO DE INVESTIGACIÓN PARA OBTENER EL GRADO ACADÉMICO DE BACHILLER EN CIENCIAS CON MENCIÓN EN INGENIERÍA INFORMÁTICA

Luiggi Gianpiere Tenorio Ku

20151204

Asesor: Dra. Layla Hirsh Martinez

Lima, julio de 2020

Resumen

En la actualidad, existen métodos complejos para la clasificación e identificación de proteínas repetidas a partir de su estructura, los cuales implican un uso intenso y costoso de recursos computacionales. Debido a ello, en el presente trabajo de investigación se busca explorar soluciones alternativas y complementarias a otros sistemas en la etapa de clasificación de proteínas repetidas con técnicas del área de estudio de aprendizaje de máquina. Estas técnicas son conocidas por ser efectivas y rápidas para la sistematización de varios procedimientos de clasificación, segmentación y transformación de datos con la condición de que se disponga de una cantidad considerable de datos. De esa forma, en consecuencia de la cantidad de datos estructurales que se han generado en los últimos años en el ambito de las proteínas y las proteínas repetidas, es posible utilizar técnicas de aprendizaje de máquina para la clasificación de las mismas. Por ello, en este trabajo, a partir de un análisis a los datos que se poseen en la actualidad y una revisión sistemática de la literatura, se proponen posibles soluciones que utilizan aprendizaje de máquina para la clasificación automatizada y rápida de proteínas repetidas a partir de su estructura. De estas posibles soluciones, se concluye que es posible la implementación de un clasificador con múltiples entradas utilizando información de los ángulos de torsión y distancia entre aminoácidos de una proteína, la cual va a ser implementada y evaluada en un trabajo futuro.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Capítulo 1. Generalidades

1.1 Problemática

En los últimos años, diversos estudios han demostrado el rol fundamental de las proteínas repetidas dentro de la naturaleza. Estas proteínas se caracterizan por tener segmentos duplicados de aminoácidos en su secuencia (Kajava, 2012). Además, estos segmentos de secuencias y estructuras duplicadas ocurren en el 14% de todas las proteínas y actúan como mecanismos de construcción modular de nuevas proteínas, los cuales involucran una rápida evolución y, por lo tanto, una mejor adaptación a nuevos entornos (Marcotte, Pellegrini, Yeates, & Eisenberg, 1999).

Muchos de los estudios realizados en este tipo de proteínas se han debido a su relevancia dentro de varios procesos biológicos como la salud, el desarrollo neuronal y la ingeniería de proteínas (Di Domenico et al., 2014). Esta relevancia se debe a las propiedades funcionales de las proteínas repetidas que no solo involucran su alta frecuencia dentro del universo de todas la proteínas conocidas, sino que también su habilidad de enlazar distintas proteínas y los roles estructurales que puede tomar (Andrade, Perez-Iratxeta, & Ponting, 2001).

Además, se sabe que la estructura y las propiedades funcionales de las proteínas repetidas se preservan incluso bajo la presencia de altas divergencias en sus subsecuencias correspondientes a las unidades de repetición (Pellegrini, 2015) e incluso se presenta una gran cantidad de mutaciones (Andrade et al., 2001). Entonces, para darle un sentido funcional a miles de secuencias de proteínas se requiere un enfoque sistemático e información de la estructura de la misma (Kajava, 2012). Un buen factor para predecir la función de una proteína es la configuración tridimensional de la cadena de aminoácidos que, de hecho, es más confiable que su secuencia porque está de tres a diez veces mejor conservada en la naturaleza (Illergård, Ardell, & Elofsson, 2009).

Por otro lado, por la aparición de nuevos procesos experimentales, la cantidad de proteínas recientemente descubierta, pero posiblemente redundantes, se incrementa rápidamente (Zacharaki, 2017), como se puede observar en la Ilustración 1 con el crecimiento exponencial del Protein Data Bank. Sin embargo, la cantidad de anotaciones de propiedades funcionales

que las describen continúa siendo limitada (Zacharaki, 2017). Dentro del universo de las proteínas, las proteínas repetidas aún son “materia oscura”. Como muestra de ello, sus dominios se encuentran dentro de los grupos menos caracterizados de secuencias de proteínas en el proteoma humano (Mistry et al., 2013). Por lo tanto, el entendimiento sobre proteínas repetidas con respecto a su estructura, función y evolución representa un reto considerable (Andrade et al., 2001) en especial en la identificación de regiones de repetición dentro de una proteína, la cual puede estar en un estado altamente degenerado (Di Domenico et al., 2014).

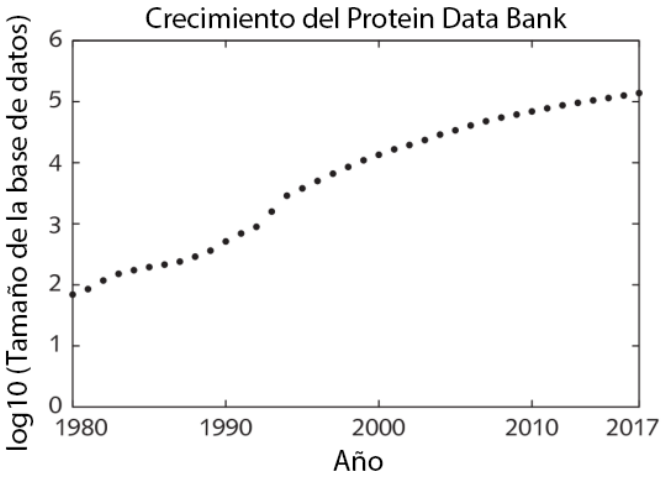

Ilustración 1: Crecimiento de "the world-wide Protein Data Bank" de los últimos 37 años en base logarítmica. Adaptado de (Lesk, 2019).

Dentro del universo de las proteínas repetidas, se ha desarrollado una base de datos de anotaciones basada en la estructura de la proteína, llamada RepeatsDB (Di Domenico et al., 2014). Esta base de datos busca expandir el conocimiento de las proteínas repetidas utilizando métodos del estado del arte para generar anotaciones de alta calidad de forma sistemática (Di Domenico et al., 2014). Actualmente, RepeatsDB 2.0, la base de datos actualizada, provee de más de 5400 proteínas repetidas con anotaciones, el cual supera en gran escala a su antecesor que solo poseía alrededor de 300 manualmente curadas (Paladin et al., 2017). Para realizar esta actualización se extrajeron proteínas candidatas a tener regiones de repetición con el método conocido como RAPHAEL (Walsh et al., 2012) para luego confirmar la presencia de unidades de repetición con ReUPred (Hirsh, Piovesan, Paladin, & Tosatto, 2016).

La nueva versión de RepeatsDB incluye una clasificación por subclases y definición de unidades de repetición de los cuales más del 60% de las anotaciones han sido validadas por expertos en el área (Paladin et al., 2017). Estas subclases pertenecen a una clasificación basada en su estructura tridimensional que facilita el entendimiento de la relación entre secuencia, estructura, función y mecanismos de evolución de este tipo de proteínas (Kajava, 2012).

Con la presencia de RepeatsDB, se evidencia que el número de proteínas con anotaciones funcionales es adecuadamente grande para permitir el entrenamiento de modelos de aprendizaje de máquina para generalizar estas anotaciones funcionales a proteínas nuevas, automáticamente (S. Amidi, Amidi, Vlachakis, Paragios, & Zacharaki, 2017). Además, son muy pocos los métodos existentes para la identificación de proteínas repetidas que lidien con el problema de clasificación de su estructura (Hirsh et al., 2016) y los métodos existentes poseen un tiempo de ejecución promedio de varios minutos para una sola cadena de proteína (Hirsh, Paladin, Piovesan, & Tosatto, 2018). Asimismo, la ejecución de estos métodos sobre PDB Data Bank, el banco de datos que posee información estructural de más de 110 mil proteínas individuales en la actualidad (Burley et al., 2019) que corresponde a analizar más de 400 mil cadenas de proteínas, es una tarea que se realiza de forma periódica en consecuencia de los nuevos descubrimientos que se realizan en esta área. Esta tarea implicara que estos procesos realicen una gran cantidad ejecuciones y generen una gran cantidad de archivos que, en consecuencia, se traducen en grandes costos de procesamiento y almacenamiento.

Por ello, en este proyecto se propone una solución basada en métodos de aprendizaje de máquina para la preclasificación de proteínas repetidas en base a su estructura con el objetivo de reducir el número de posibilidades a evaluar para que otros métodos utilicen esta información de tal modo que se incremente el número de anotaciones disponibles de proteínas repetidas y esto se realice en un menor tiempo.

1.2 Objetivos

1.2.1 Objetivo general

El objetivo general del proyecto es identificar las proteínas repetidas almacenadas en el PDB Data Bank y clasificar las proteínas repetidas según su estructura. Todo esto por medio de la generación de un modelo de aprendizaje de máquina con la capacidad de detectar la presencia de regiones repetidas dentro de una cadena proteica.

1.2.2 Objetivos específicos

O 1. Definir representaciones de datos válidas para la extracción de características estructurales de una cadena de proteína

O 2. Desarrollar y validar un método de aprendizaje de máquina que identifique proteínas repetidas y las clasifique en su respectiva clase y subclase

O 3. Implementar un servicio web que permita caracterizar una cadena de proteína

1.2.3 Resultados esperados

R 1. Procedimiento para la obtención de una representación de datos a partir de una cadena de proteínas, utilizando información estructural (O1)

R 2. Conjunto de representaciones de datos de cadenas de proteínas (O1)

R 3. Conjunto de clasificadores por subclase de proteína (O2)

R 4. Reporte de predicciones hechas a todo el PDB Data Bank (O2)

R 5. Servicio web que utiliza los clasificadores (O3)

1.2.4 Mapeo de objetivos, resultados y verificación

Tabla 1 Resultados esperados, metas físicas y medios de verificación del objetivo específico 1.

Objetivo 1: Definir representaciones de datos válidas para la extracción de características estructurales de una cadena de proteína			
Resultado	**Meta física**	**Medio de**	**Indicador**

		verificación	objetivamente verificable
R1: Procedimiento para la obtención de una representación de datos a partir de una cadena de proteínas, utilizando información estructural	Software	Código fuente Reporte de resultados de pruebas de transformación de datos	Pruebas aprobadas al 100%
R2: Conjunto de representaciones de datos de cadenas de proteínas	Archivos conteniendo la representación de datos de cada cadena proteica de PDB Data Bank	Reporte de resultados de pruebas de integridad de datos	Pruebas aprobadas al 100%

Tabla 2 Resultados esperados, metas físicas y medios de verificación del objetivo específico 2. Elaboración propia

Objetivo 2: Desarrollar y validar un método de aprendizaje de máquina que identifique proteínas repetidas y las clasifique en su respectiva clase y subclase			
Resultado	**Meta física**	**Medio de verificación**	**Indicador objetivamente verificable**
R3: Conjunto de clasificadores por subclase de proteína	Componentes de software	Reporte de precisión y exhaustividad del conjunto de datos de prueba de cada	Pruebas de rendimiento aprobadas y verificadas al 100%

		clasificador.	
R4: Reporte de predicciones hechas a todo el PDB Data Bank	Lista de cadenas de proteínas repetidas con su subclase respectiva	Reporte de las pruebas aleatorias sobre las subclases	Aprobación por parte de un especialista en el dominio

Tabla 3 Resultados esperados, metas físicas y medios de verificación del objetivo específico 3. Elaboración propia

Objetivo 3: Implementar un servicio web que permita caracterizar una cadena de proteína			
Resultado	**Meta física**	**Medio de verificación**	**Indicador objetivamente verificable**
R5: Servicio web que utiliza los clasificadores	Software	Reporte de pruebas unitarias del funcionamiento del servicio URL	Pruebas unitarias aprobadas al 100%

1.3 Herramientas y métodos

En esta sección se van a indicar las herramientas y métodos por cada objetivo dentro del trabajo a realizar. Además, se describirá cada una de ellas con el aporte que brinda en el desarrollo del mismo.

1.3.1 Herramientas y métodos por utilizar

En la Tabla 4, se relaciona los resultados esperados con sus respectivas herramientas y métodos a utilizar dentro del proyecto.

Tabla 4 Resultados esperados con sus respectivas heramientas y métodos a utilizar. Elaboración propia.

Resultado esperado	Herramientas por utilizar	Métodos por utilizar
R1: Procedimiento para la obtención de una representación de datos a partir de una cadena de proteínas, utilizando información estructural	• Python • Jupyter Notebook • Conda • Git y Gitlab • RCSB PDB Data Bank Download Tool	• Revisión sistemática de literatura
R2: Conjunto de representaciones de datos de cadenas de proteínas	• Python • Conda • RCSB PDB Data Bank Download Tool • RepeatsDB 2.0	(Se obtiene a partir de la ejecución del Resultado 1)
R3: Conjunto de clasificadores por subclase de proteína	• Python • Jupyter Notebook • Conda • Google Colab • Git y Gitlab	• Adam • Validación cruzada k-fold • Métricas de evaluación

	• Tensorflow • Keras	
R4: Reporte de predicciones hechas a todo el PDB Data Bank	• Python • Conda • Tensorflow • Keras	(Se obtiene a partir de la aplicación del Resultado 3)
R5: Servicio web que utiliza los clasificadores	• Python • Git y Gitlab • Amazon Web Services	• Transferencia de Estado Representacional (REST)

1.3.2 Descripción de herramientas y métodos

A continuación, se describirán las herramientas y métodos mencionados en la sección anterior.

Herramientas

- **Python** (Python Software Foundation, 2017)
  Python es un lenguaje de programación de alto nivel de múltiples paradigmas. Se caracteriza por ser simple y fácil de aprender, así como trabajar con módulos y paquetes, los cuales fomentan modularidad, reutilización y reduce los costos de mantenimiento de código. Adicionalmente, su interprete y extensa librería está libre para su distribución como código abierto o de forma binaria sin ningún tipo de costo para plataformas mayores. Además, posee una larga colección de herramientas, de las cuales muchas son necesarias para el desarrollo del trabajo.

- **Conda** (Continuum Analytics, 2017)
  Es un gestor de paquetes y entornos de programación de código abierto. Por ello, en él se puede instalar, ejecutar y actualizar paquetes con sus dependencias rápidamente. Además, crea, guarda, carga y cambia entre entornos de programación con facilidad lo que permite mantener el orden de diferentes grupos de paquetes que se necesiten a lo largo del proyecto.

- **Jupyter Notebook** (Pérez & Granger, 2007)

  Es una aplicación web de código abierto que permite crear documentos que pueden contener código, texto, ecuaciones y visualizaciones en tiempo real. Dentro del proyecto se va a utilizar para escribir y ejecutar código en el lenguaje de programación Python lo que permitirá visualizar resultados parciales fácilmente.

- **Git y Gitlab** (Torvalds & Hamano, 2010)

  Git es un sistema de control de versiones de archivos gratuito y de código abierto. Este es fácil de utilizar y permite un seguro y rápido mantenimiento de archivos. Adicionalmente, Gitlab es una plataforma de desarrollo colaborativo que utiliza el sistema de versiones de Git que, además, permite la integración continua del proyecto. Esta última plataforma va a ser utilizada para controlar el progreso del mismo.

- **RCSB PDB Data Bank Download Tool** (Burley et al., 2019)

  La base de datos de dominio público que cuenta con información de estructuras tridimensionales de proteínas, posee una herramienta adicional para realizar descargas masivas de la información mencionada. Esta será utilizada para obtener información estructural de todas las proteínas disponibles para su análisis posterior.

- **RepeatsDB 2.0** (Paladin et al., 2017)

  Es una base de datos que cuenta con información estructural de las proteínas repetidas. Además, provee la posición de inicio y fin de las unidades de repetición, su clasificación y referencias a otras bases de datos. Será utilizado en el proyecto para obtener la información estructural de las proteínas a analizar.

- **Google Colab** (Carneiro et al., 2018)

  Es un servicio en la nube basado en Jupyter Notebook para difundir la educación e investigación del aprendizaje de máquina. Además, este provee acceso unidades de GPU de gran potencia configuradas para el desarrollo de aprendizaje profundo sin ningún costo. Estas últimas van a ser utilizadas para el entrenamiento de los clasificadores de aprendizaje de máquina del proyecto.

- **Tensorflow** (Abadi et al., 2016)

  Es un sistema de aprendizaje de máquina que opera a gran escala. Este utiliza grafos de flujo de datos para representar computación, estado compartido y operaciones que mutan el estado. Además, permite utilizar múltiples dispositivos computacionales para acelerar el procesamiento computacional con facilidad. Este sistema es utilizado mayormente como marco de trabajo para aprendizaje de máquina que es lo que se va a realizar dentro del proyecto.

- **Keras** (Ketkar & Ketkar, 2017)

  Es una librería que provee bloques de construcción abstractos para construir redes de aprendizaje profundo, los cuales son construidos, principalmente, utilizando Tensorflow. De esa forma, permite la construcción rápida y fácil de entender de prototipos de redes de aprendizaje profundo. Además, facilita el entrenamiento, uso y visualización de los mismos. Dentro del proyecto, se va a utilizar como herramienta de construcción y prueba de los clasificadores.

Métodos

- **Revisión sistemática de literatura**

  Con el objetivo de encontrar representaciones de datos utilizadas para los procedimientos de clasificación de proteínas, se va a realizar una revisión sistemática de literatura basada en las pautas establecidas por en la metodología de B. Kitchenham (Kitchenham & Charters, 2007) que, posteriormente, se va a evaluar dentro de este documento.

- **Adam**

  Con el fin de entrenar a la red convolucional profunda construida a partir de las herramientas ya mencionadas, se va a utilizar el método Adam de optimización estocástica. Este método es computacionalmente eficiente, necesita de pocos requisitos de memoria, es fácil de ajustar los híper-parámetros del algoritmo y es adecuado para problemas con gran cantidad de datos o parámetros (Kingma & Ba, 2014) por lo que apropiado para el entrenamiento de redes convolucionales profundas.

- **Métricas de evaluación**

  Con el fin de verificar la eficacia en los resultados de la clasificación se utilizarán las siguientes métricas en el contexto de la clasificación de proteínas:

  Precisión:

  Esta medida se define como el número de proteínas correctamente clasificadas de cierta subclase sobre el número total de muestras de proteínas clasificadas en dicha subclase.

$$Precision = \frac{\#\ de\ proteinas\ clasificadas\ correctamente\ de\ subclase\ i}{\#\ de\ proteinas\ clasificadas\ como\ subclase\ i}$$

  Exhaustividad:

  Esta medida se define como el número de proteínas correctamente clasificadas de cierta subclase sobre el número real de muestras de proteínas que pertenecen a dicha subclase.

$$Exhaustividad = \frac{\#\ de\ proteinas\ clasificadas\ correctamente\ de\ subclase\ i}{\#\ de\ proteinas\ de\ subclase\ i}$$

  F-score:

  El F-score es una métrica que utiliza las métricas de precisión y exactitud para obtener una ponderación harmónica de ambas métricas. Además, utiliza un parámetro β para definir la importancia de una métrica sobre la otra.

$$Fscore = \frac{(\beta^2 + 1)\ \times\ Precision\ \times\ Exhaustividad}{(\beta^2\ \times\ Precision) + Exhaustividad}$$

- **Validación cruzada k-fold**

  Es un método utilizado dentro de la evaluación y selección de un modelo en la cual se itera cruzando a lo largo de fases de entrenamiento y validación en rondas sucesivas

como se muestra en la para obtener el rendimiento promedio de las mismas (Raschka, 2018).

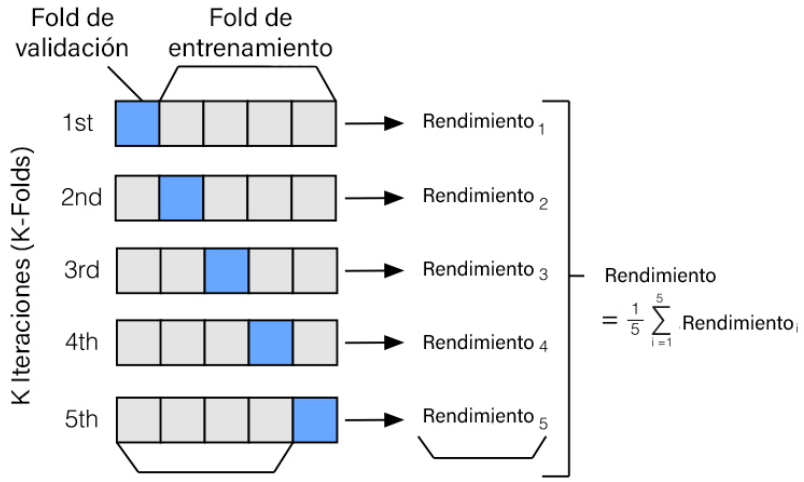

Ilustración 2 Validación cruzada k-fold con 5 folds. Se puede observar que se evalúa el rendimiento con 5 diferentes conjuntos de datos (Fold de validación y entrenamiento). Adaptado de (Raschka, 2018).

Este método se va a utilizar para evaluar el rendimiento de los clasificadores dentro del proyecto.

- **Transferencia de Estado Representacional (REST)**

  Es un estilo de arquitectura de software para crear servicios web llamados RESTful. Este estilo provee restricciones que, al ser aplicadas, enfatizan la escalabilidad de las interacciones entre componentes, generalidad de las interfaces, independencia al desplegar componentes y la creación de componentes intermedios para reducir latencia en la interacción, reforzar la seguridad y encapsular sistemas heredados (Fielding, 2000). Dentro del proyecto, se va a considerar este estilo para el desarrollo del servicio web.

1.4 Alcances y limitaciones del proyecto

1.4.1 Alcance del proyecto

El presente proyecto enfocado al área de bioinformática tiene como finalidad pre clasificar las proteínas repetidas considerando la sub clasificación con el fin reducir el número de posibilidades a analizar por otros métodos para la identificación de unidades de repetición.

Asimismo, no se va a tomar en cuenta algunas clases de proteínas repetidas debido a la falta de información acerca de sus estructuras tridimensionales la cual es fundamental para identificar sus unidades de repetición. En específico, no se van a considerar las clases homo repetidas I y II de las cinco clases expuestas en el capítulo 2.1.2.

Adicionalmente, dentro de las clases a considerar, la evaluación de los resultados del proyecto se va a medir a partir de la información obtenida de las clases de las que se posee más información de su estructura.

Además, el alcance se ve definido por los datos disponibles en el PDB Data Bank a la fecha de agosto de 2019.

1.4.2 Limitaciones del proyecto

Debido a que el proyecto va a utilizar los datos del RCSB PDB Data Bank, este va a estar limitado a la disponibilidad de la información estructural de las proteínas que contenga el mismo. En otras palabras, a pesar de que se tenga información secuencial de millones de proteínas, solo se va a limitar a analizar las proteínas que posean información estructural dentro de este banco de datos.

1.4.3 Riesgos del proyecto

Riesgo identificado	**Probabilidad**	**Impacto**	**Mitigación**	**Contingencia**

Ausencia del asesor	0.3	0.6	Apoyo de investigadores que trabajan en temas similares	Apoyo de expertos del área
Dificultades con la curva de aprendizaje de temas biológicos	0.3	0.5	Llevar cursos virtuales de biología y bioquímica y bioinformática	Consultar con expertos del tema. Se cuenta con el contacto de toda la comunidad de expertos del área.
Falta de acceso al conjunto de datos de proteínas	0.1	0.7	Comunicación con las organizaciones a cargo de las bases de datos de proteínas para informarles sobre el proyecto.	La asesora de tesis es autora de estas bases de datos por lo que se podría acceder a estas por medio de ella.

Capítulo 2. Marco Conceptual

2.1 Conceptos de biología molecular

2.1.1 Proteínas

Las proteínas son largos polímeros de aminoácidos que constituyen una larga fracción de las células en la naturaleza. Algunas de estas funcionan como enzimas que catalizan reacciones químicas, como elementos estructurales que brindan rigidez a la estructura molecular, como receptores de señales enviadas al interior de la célula o como transportadores de sustancias específicas dentro y fuera ellas. De esta forma, las proteínas son las más versátiles de todas las biomoléculas (Nelson & Cox, 2008).

Para entender la estructura de las proteínas, estas son ordenadas de forma jerárquica en cuatro niveles de organización (Ver Ilustración 3):

- La **estructura primaria** indica los enlaces covalentes de los amino caídos que conforman cada una de las cadenas polipeptídicas de la proteína. De esta forma, el elemento más importante dentro de la estructura primaria es conocido como la secuencia de aminoácidos (Nelson & Cox, 2008).
- La **estructura secundaria** se refiere a los particularmente estables arreglos de aminoácidos que generan patrones estructurales recurrentes. Algunas estructuras secundarias conocidas son hélice α y lámina β (Nelson & Cox, 2008).
- La **estructura terciaria** describe todos los aspectos de los pliegues tridimensionales del polipéptido (Nelson & Cox, 2008).
- La **estructura cuaternaria** ocurre cuando la proteína posee dos o más subunidades de polipéptidos. Por lo tanto, este nivel hace referencia al arreglo espacial de estas subunidades (Nelson & Cox, 2008).

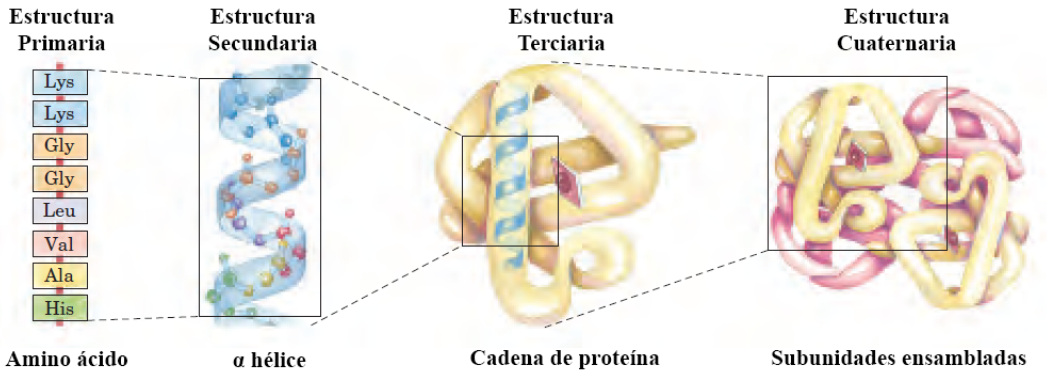

Ilustración 3: Estructura jerárquica de las proteínas. La estructura primaria representa una secuencia de aminoácidos La estructura secundaria corresponde a una α hélice. La estructura terciaria muestra una sola cadena de proteínas y la estructura cuaternaria a un conjunto de cadenas. Adaptado de (Nelson & Cox, 2008).

2.1.2 Proteínas repetidas

Las proteínas repetidas se caracterizan por poseer una estructura tridimensional que contiene regiones repetitivas, también conocidas como unidades de repetición, y se encuentran presentes en gran cantidad dentro de la naturaleza (Kajava, 2012). Además, estas están comprendidas en cinco clases (Ver Ilustración 4):

- La **clase I** son proteínas y péptidos cuyas unidades de repetición se componen de uno o dos aminoácidos y forman diferentes tipos de cristales de tamaño ilimitado, los cuales son dañinos para los organismos vivientes (Kajava, 2012).
- La **clase II** incluye dos de las mayores estructuras fibrosas dentro del universo de las proteínas, que son el colágeno y las alfa-hélices. Estas estructuras se caracterizan por poseer unidades de repetición de tres o cuatro aminoácidos y por poseer estructuras estabilizadas por la interacción entre cadenas (Kajava, 2012).
- La **clase III** son estructuras alargadas en donde cada una de las unidades de repetición dependen de otra para mantener su estructura. Estas se pueden distinguir en dos grupos: las estructuras solenoides y no solenoides. Las **estructuras solenoides** se componen de unidades de repetición de cinco a aproximadamente cuarenta aminoácidos de largo cuyos pliegos están basados en espirales solenoidales dentro de la cadena del polipéptido mientras que las **estructuras no solenoides** poseen cualquier otra estructura (Kajava, 2012).

- La **clase IV** son estructuras cerradas donde cada una de las unidades de repetición dependen de otra para mantener su estructura y pueden tener un número ilimitado de unidades de repetición sin ningún límite en cuanto a su crecimiento (Kajava, 2012).
- La **clase V** está compuesta por estructuras suficientemente largas como para generar pliegos de forma independiente y están compuestas, usualmente, por cincuenta a aproximadamente sesenta aminoácidos (Kajava, 2012).

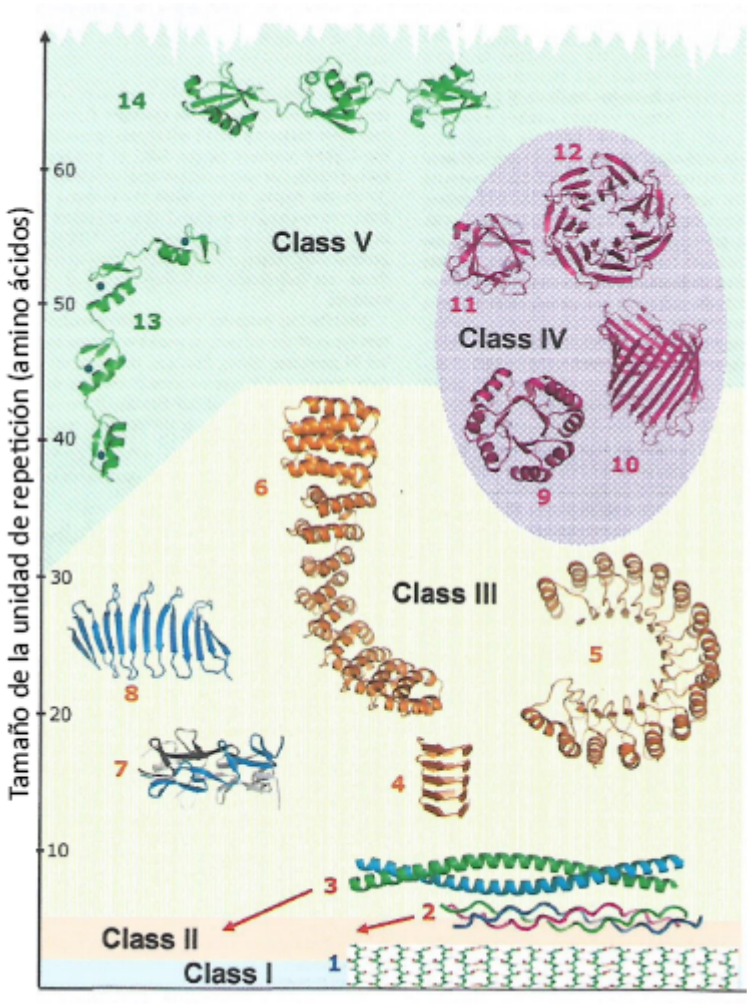

Ilustración 4: Clasificación estructural de proteínas repetidas. En este se muestran todas las clases (5 tipos) y subclases (14 tipos) diferentes divididas por tamaño de unidad de repetición. Adaptado de (Kajava, 2012).

2.1.3 Gráfico de Ramachandran

Una forma de visualizar los ángulos de torsión es a partir del gráfico de Ramachandran introducido por Gopalasamudram Ramachandran en 1963. Por convención, los ángulos de

torsión se describen como los ángulos resultantes de las rotaciones del Alpha carbono (Cα) de un aminoácido con respecto al enlace N-Cα como ángulo Φ (phi) y ψ (psi) cuando es con respecto al enlace Cα-C. Además, estos se representan con valores entre -180° a 180° grados sexagesimales. De esta forma, cada aminoácido de la cadena es ubicado dentro de un plano bidimensional dando como resultado el gráfico de Ramachandran (Nelson & Cox, 2008).

Como ya se mencionó, la estructura secundaria se refiere a una subsección de la cadena del polipéptido que forma patrones de pliegos conocidos como la hélice α y la lámina β. Según observaciones, los aminoácidos correspondientes a la hélice α poseen ángulos de torsión ψ de -45° a 50° y ángulo Φ de -60° y cada hélice contiene aproximadamente 3.6 aminoácidos. Por otro lado, las láminas β poseen ángulos ψ cercanos a 140° y ángulos Φ cercanos a -130° (Nelson & Cox, 2008). Estas regiones se pueden observar en la Ilustración 5.

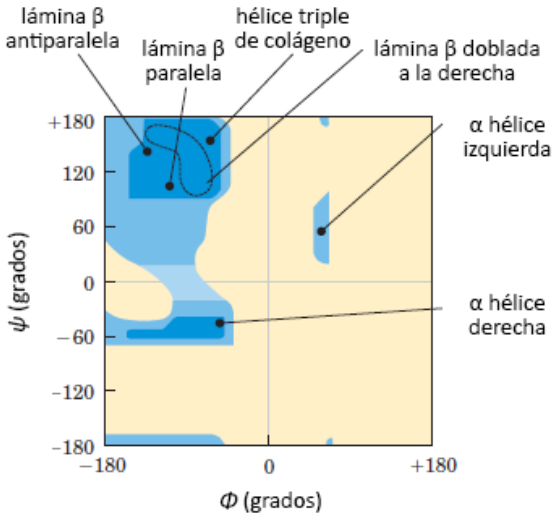

Ilustración 5 Gráfico de Ramachandran con regiones correspondientes a patrones de estructura secundaria. En este se muestran las diferentes estructuras secundarias conocidas con su región coloreada de azul con intensidad relacionada a su densidad. Adaptado de (Nelson & Cox, 2008).

2.2 Conceptos de ciencias de la computación

2.2.1 Aprendizaje de máquina

El aprendizaje de máquina es un área de estudio enfocada en cómo detectar patrones automáticamente a partir de un conjunto de datos y luego utilizar dichos patrones para

predecir datos futuros o realizar otras tareas de decisión bajo contextos inciertos (Murphy, 2012).

Los métodos de aprendizaje de máquina se dividen en varios tipos, como el aprendizaje supervisado y no supervisado (Murphy, 2012).

En el **aprendizaje supervisado**, el objetivo es aprender patrones bajo un conjunto de pares de datos de entrada y salida llamado conjunto de entrenamiento. Usualmente, cada entrada del conjunto de entrenamiento es descrito como un vector de n dimensiones que describen las características de un objeto (Ej. peso, altura de una persona). De forma similar, cada salida se puede expresar como una variable nominal o categórica perteneciente a un conjunto finito de posibilidades (Ej. Sexo masculino o femenino), o como una variable escalar con valor real (Ej. El salario de una persona). Si esta salida pertenece a una variable nominal o categórica, el problema que se está resolviendo es uno de clasificación; en cambio, si la variable es escalar con valor real, se está resolviendo un problema de regresión (Murphy, 2012).

En la clasificación, si el número de clases mutuamente exclusivas es igual a 2, el problema corresponde a uno de clasificación binaria. Asimismo, si el número de clases mutuamente exclusivas es mayor a 2, trata de una clasificación múltiples clases. Por otro lado, si las clases no son mutuamente exclusivas, el problema de clasificación es de múltiples etiquetas (Murphy, 2012).

Por otro lado, el **aprendizaje no supervisado** solo recibe datos de entrada y el objetivo es encontrar patrones interesantes en los datos. Este no corresponde a un problema bien definido ya que no expresa qué tipos de patrones se buscan desde un principio y no se utiliza una métrica para saber la eficiencia del método (Murphy, 2012).

2.2.2 Redes convolucionales profundas

Las redes convolucionales profundas están diseñadas para procesar datos que se encuentran representados en forma de múltiples arreglos, por ejemplo, imágenes a color compuestas por

tres arreglos bidimensionales que contienen la intensidad de color en su respectivo canal (Lecun, Bengio, & Hinton, 2015).

Existen cuatro ideas de las redes convolucionales que toman ventaja de las señales naturales: Conexiones locales, pesos compartidos, reducción de muestreo y el uso de varias capas. Las primeras secciones de la red están compuestas de dos tipos de capas: capas convolucionales y de reducción de muestreo (Lecun et al., 2015), como se muestra en la Ilustración 6.

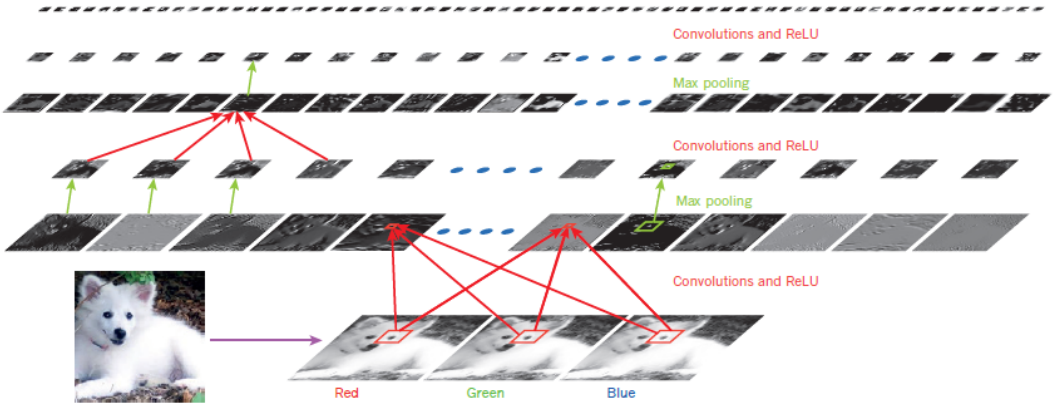

Ilustración 6 Red convolucional típica aplicada a una imagen de tres canales (RGB). Cada una de las imágenes rectangulares corresponde a un mapa de características. La primera capa (abajo para arriba) corresponde a los canales RGB de la imagen. Las capas posteriores representan mapas de características que son obtenidas por operaciones de convolución y reducción de muestreo. Adaptado de Lecun (2015)

Las unidades dentro de la capa convolucional están organizadas en mapas de características que son arreglos multidimensionales calculados a través de un conjunto de pesos llamado banco de filtros. De esa forma, el rol de este tipo de capas es detectar características locales de la capa anterior, mientras que el rol de las capas de reducción de muestreo es combinar semánticamente características similares en una. Una capa típica de reducción de muestreo calcula el máximo de una sección local en un mapa de características. Así, estas capas son apiladas en varias fases de las cuales, posteriormente, entran dentro de capas totalmente conectadas (Lecun et al., 2015).

Capítulo 3. Estado del Arte

3.1 Revisión y discusión

3.1.1 Método de revisión de la literatura

En la presente investigación, se realizó una revisión sistemática de literatura basada en la metodología de B. Kitchenham (Kitchenham & Charters, 2007). De esta manera, se identificaron las publicaciones realizadas con respecto a los estudios y proyectos realizados para la clasificación de proteínas en general.

Como parte de la planificación de la revisión sistemática, se busca formular preguntas a resolver en base a los criterios de población, intervención, comparación, salida y contexto del modelo PICOC (Petticrew & Roberts, 2006). Estos criterios se muestran en la siguiente te tabla:

Tabla 5 Tabla PICOC

Criterio	**Explicación**	**Descripción**
Población	¿Quién?	Proteínas
Intervención	¿Qué o cómo?	Aplicación de métodos de aprendizaje de máquina a partir de la estructura de la proteína
Comparación	¿Comparado con qué?	-
Salida	¿Qué queremos lograr?	Clasificación de proteínas por clase y subclase
Contexto	¿En qué circunstancias?	Académico

Lo que se buscó dentro de esta revisión sistemática, fueron métodos de aprendizaje de máquina aplicados a la estructura de las proteínas en general para implementarlos dentro del grupo de las proteínas repetidas. Esto se debe a que, en una primera revisión, no se encontró este tipo de aplicaciones dentro de esta clase de proteínas. Además, se va a restringir la

búsqueda con el fin de encontrar métodos que puedan ser adaptados para detectar la presencia de unidades repetidas dentro una cadena de proteína, siendo esta la unidad mínima analizar.

A partir de ello, se plantearon las siguientes preguntas que se buscan responder:

- ¿Qué métodos se han utilizado previamente para clasificar diferentes tipos de proteínas con información del Protein Data Bank?
- ¿Cómo los datos han sido utilizados para los modelos de clasificación de aprendizaje de máquina?
- ¿Qué métricas se utilizaron para evaluar el rendimiento del clasificador?
- ¿Qué resultados se obtuvieron al evaluar el método?

3.1.2 Palabras clave y cadena de búsqueda

Basándonos en la tabla PICOC, para la búsqueda de investigaciones previas se utilizaron las siguientes palabras claves:

- Protein structure
- Classification
- Neural network
- PDB Data File

En base a ello, se obtuvo la siguiente cadena de búsqueda:

> (**protein** AND **structure** AND **classification** AND **neural** AND **network** AND **PDB**)

3.1.3 Fuentes de información

Debido a que las preguntas planteadas dentro de la revisión sistemática implican métodos y conceptos pertenecientes al área de ciencias de la computación que son aplicados en biología se realizó la búsqueda de la cadena mencionada dentro de una base de datos científica enfocada en este tipo de investigaciones, llamada PeerJ. Además, se apoyará la búsqueda con Web of Science que es otra gran base de datos que posee información de publicaciones de otros repositorios.

3.1.4 Estrategia de extracción

Para la selección de artículos en las bases de datos se identificaron criterios de selección y criterios de exclusión.

Criterios de Selección

- Métodos para clasificar diferentes tipos de proteínas basadas en la estructura.
- Métodos aplicados con anotaciones a nivel de cadena o a la proteína en su totalidad.
- Artículo perteneciente al campo de ciencia de la computación, bioinformática o biología computacional.

Criterios de Exclusión

- Artículos de más de 5 años de antigüedad
- Métodos para clasificar diferentes tipos de proteínas que utilicen su información secuencial.
- Capítulos o artículos de libros.

3.1.5 Selección de estudios

Una vez definidos los criterios de extracción de información se procedió a la búsqueda dentro de las fuentes de información y se seleccionaron los siguientes estudios para el estado del arte:

Tabla 6 Artículos encontrados en la revisión sistemática

Fuente	Cantidad de artículos de la búsqueda	Cantidad de artículos seleccionados	Cantidad de artículos repetidos
PeerJ	17	2	0
Web of Science	9	1	4

Como se mostró en la tabla anterior, se han seleccionado solo los artículos que cumplen con los criterios establecidos de selección e inclusión.

Tabla 7 Artículos seleccionados para ser incluidos en el estado del arte.

N°	Título	Referencia
1	Prediction of protein function using a deep convolutional neural network ensemble	Zacharaki, 2017
2	EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation	A. Amidi et al., 2018
3	Prediction of Enzyme Function Based on Three Parallel Deep CNN and Amino Acid Mutation	Gao et al., 2019

3.1.6 Revisión de estudios

En esta sección se presenta la revisión de investigaciones recientes con diferentes métodos para la clasificación de diferentes tipos de proteínas, así como la representación de los datos de entrada y el proceso de evaluación del método mencionado.

Prediction of protein function using a deep convolutional neural network ensemble (Zacharaki, 2017)

En este trabajo, se extrajeron mapas de características de la información estructural de la proteína en forma de distribución local (por aminoácido) de ángulos de torsión y de distancia entre tipos de aminoácidos. Luego, cada mapa de características es introducido dentro de una red convolucional profunda para la predicción de su función y las salidas son ensambladas por medio de dos métodos de ensamblado: máquina de vectores de soporte o un clasificador k-NN basado en la correlación de las características. Además, dos arquitecturas diferentes son exploradas, una utilizando una red convolucional para todos los canales, como se puede observar en la Ilustración 7, y otra para cada canal.

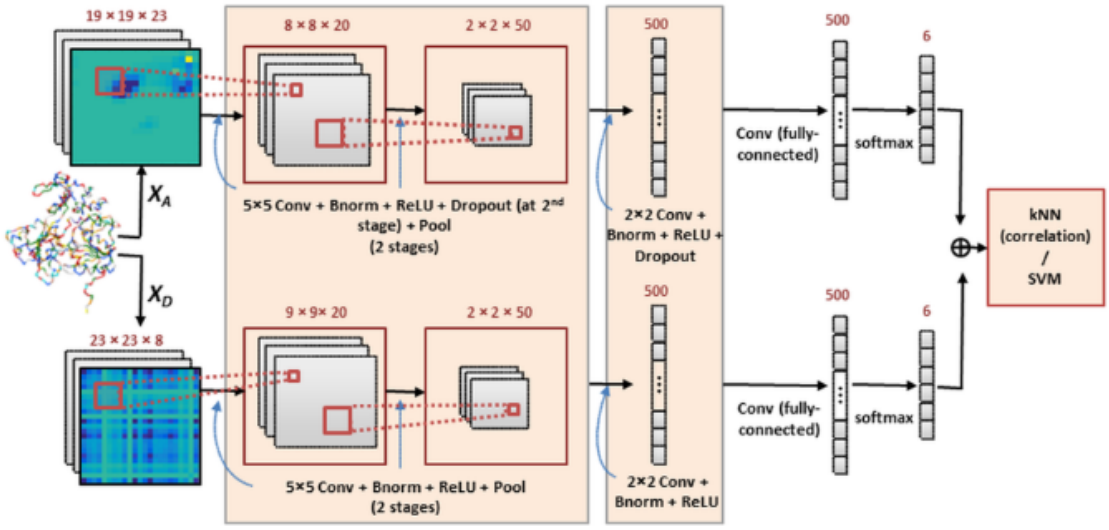

Ilustración 7 Arquitectura de redes convolucionales profundas con ensamblado para clasificación de proteínas. Cada conjunto de datos es introducido a una red convolucional (área en rojo) compuesta por capas de convolución bidimensional, normalización batch, activación ReLU y dropout. Finalmente, los resultados son combinados por un clasificador kNN y SVM, Adaptado de (Zacharaki, 2017).

Adicionalmente, como métrica de evaluación se utilizó el área bajo la curva (AUC) de ROC. Dentro de este estudio, por medio de validación cruzada en enzimas de fusión individual extraídas del PDB Data Bank se logró clasificar correctamente el 90.1% de todas, demostrando una mejora sustancial con respecto a métodos anteriores que solo lograban un 70% de clasificaciones correctas.

Esta investigación propone que, actualmente, la predicción automática de funciones de proteínas basada en su estructura ya se puede realizar de forma efectiva y rápida con la presencia de grandes conjuntos de datos de proteínas.

EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation (A. Amidi et al., 2018)

Este estudio, presenta una red convolucional tridimensional profunda para la clasificación de enzimas basado en una estructura espacial de voxels como se observa en la Ilustración 8. Además, se realizó una inspección complementaria de las propiedades espaciales de la estructura de datos planteada.

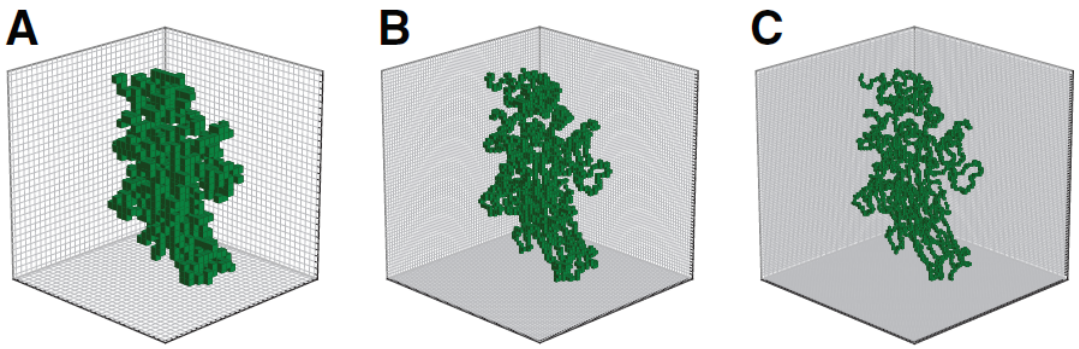

Ilustración 8 Estructura espacial de voxels de una proteína con diferentes tamaños de grilla. (A) grilla de 32x32x32. (B) grilla de 64x64x64. (C) grilla de 96x96x96. Adaptado de (A. Amidi et al., 2018).

Antes de la generación del espacio de voxels, se implementaron varios métodos para la curación de la estructura de la proteína. Estos incluyen métodos para completar la estructura espacial, escalar el tamaño, ajustar la orientación y una aumentación aleatoria. En cuanto al modelo, la arquitectura está compuesta, principalmente, de capas convolucionales tridimensionales y capas completamente conectadas como se muestra en la Ilustración 9.

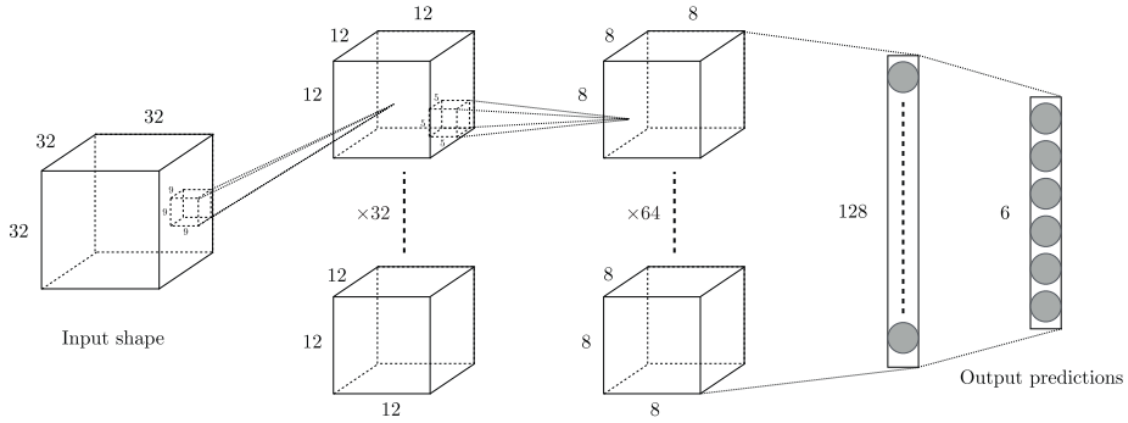

Ilustración 9 Arquitectura EnzyNet compuesta de capas de convolución tridimensionales de 32 y 64 canales. Posteriormente, la salida es introducida en una red completamente conectada. Adaptado de (A. Amidi et al., 2018).

Por otro lado, los resultados de la investigación fueron medidos con varias métricas como precisión, exhaustividad, exactitud y f1-score. Además, se indica que se logró clasificar correctamente el 78.4% de las proteínas.

Prediction of Enzyme Function Based on Three Parallel Deep CNN and Amino Acid Mutation (Gao et al., 2019)

Este trabajo reutiliza la arquitectura planteada por Zacharaki, 2017 pero, adicionalmente, incluye una tercera estructura de datos basada en la matriz de puntajes de posición específica (PSSM) que es generada mediante el método PSI-BLAST (Altschul & Koonin, 1998). Esta arquitectura paralela se puede apreciar en la Ilustración 10.

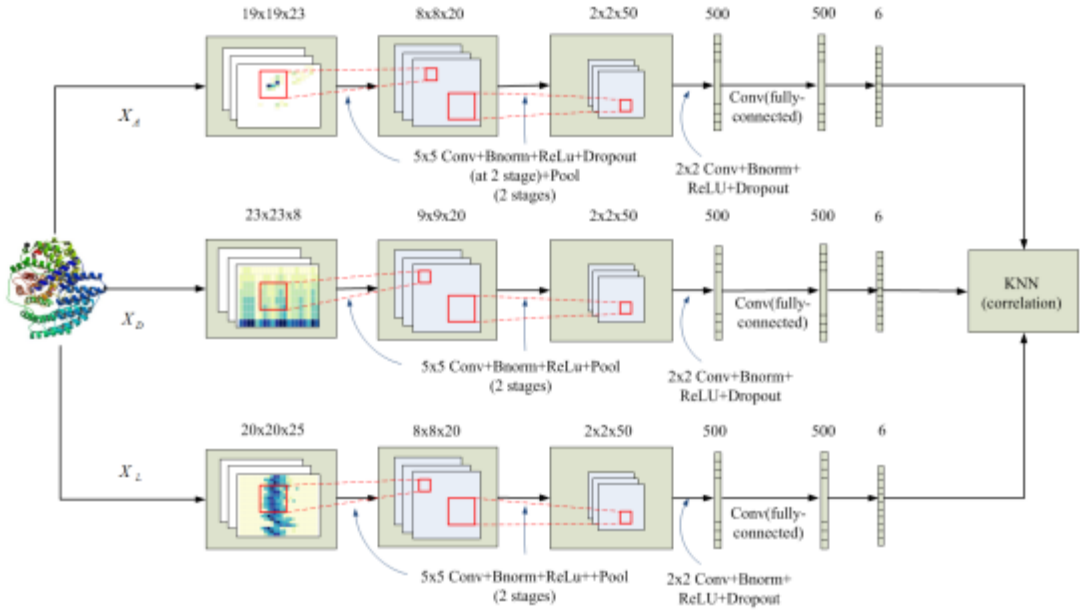

Ilustración 10 Arquitectura de tres redes convolucionales paralelas. Arquitectura similar a la planteada en la Ilustración 7. Adicionalmente, incluye un conjunto de datos generado por el PSI-BLAST (Altschul & Koonin, 1998). *Adaptado de* (Gao et al., 2019).

Los resultados del método fueron medidos con la métrica de ROC AUC. Además, se logró clasificar correctamente el 92.34% de las enzimas, obteniendo una pequeña mejora con respecto al método en el que está basado esta estructura.

3.2 Conclusiones

Dentro de los métodos mencionados para la clasificación de proteínas, se pueden sugerir varias representaciones de datos obtenidas a partir de la información estructural de la proteína. Dentro de estas representaciones podemos distinguir la importancia de la

identificación de la estructura secundaria de la proteína por el gráfico de Ramachandran (Ramachandran et al., 1963) o por el método PSI-BLAST (Altschul & Koonin, 1998).

Así mismo, métodos como EnzyNet proponen observar la estructura tridimensional de la proteína en su totalidad, pero los resultados de dicho método se ven dependientes por los métodos de rotación y escala dentro del preprocesamiento de la información estructural de la proteína. Por lo tanto, representaciones independientes a dichos factores como la arquitectura paralela propuesta por Zacharaki (2017) no se ven afectadas y tienden a obtener resultados más significativos. Además, es posible seguir agregando otras representaciones de la proteína como lo sugirió Gao (2019) al agregar el PSSM obtenido por PSI-BLAST (Altschul & Koonin, 1998).

En conclusión, debido al análisis realizado en esta revisión sistemática se va a adaptar la arquitectura paralela por Zacharaki (2017) dentro del proyecto para identificar la presencia de regiones de repetición y construir un clasificador con la capacidad de distinguir entre clases y subclases de una proteína repetida con el fin de reducir el número de posibilidades a analizar por ReUPred (Hirsh et al., 2016). Adicionalmente, se utilizarán los métodos de evaluación mencionados dentro de la revisión de estudios junto con el método de evaluación cruzada para medir el rendimiento del clasificador.

Referencias

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Kudlur, M. (2016). TensorFlow: A System for Large-Scale Machine Learning. *In 12th USENIX Symposium on Operating Systems Design and Implementation*.

Altschul, S. F., & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST - A tool for discovery in protein databases. *Trends in Biochemical Sciences*, *23*(11), 444–447. https://doi.org/10.1016/S0968-0004(98)01298-5

Amidi, A., Amidi, S., Vlachakis, D., Megalooikonomou, V., Paragios, N., & Zacharaki, E. I. (2018). EnzyNet: Enzyme classification using 3D convolutional neural networks on spatial representation. *PeerJ*, *2018*(5), 1–18. https://doi.org/10.7717/peerj.4750

Amidi, S., Amidi, A., Vlachakis, D., Paragios, N., & Zacharaki, E. I. (2017). Automatic single- and multi-label enzymatic function prediction by machine learning. *PeerJ*, *2017*(3), 1–16. https://doi.org/10.7717/peerj.3095

Andrade, M. A., Perez-Iratxeta, C., & Ponting, C. P. (2001). Protein repeats: Structures, functions, and evolution. *Journal of Structural Biology*, *134*(2–3), 117–131. https://doi.org/10.1006/jsbi.2001.4392

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., … Zardecki, C. (2019). RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, *47*(D1), D464–D474. https://doi.org/10.1093/nar/gky1004

Carneiro, T., Da Nobrega, R. V. M., Nepomuceno, T., Bian, G. Bin, De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, *6*, 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767

Continuum Analytics. (2017). Conda — conda 4.7.12.post130+ac9f5ee7 documentation. Retrieved October 30, 2019, from https://docs.conda.io/projects/conda/en/latest/index.html

Di Domenico, T., Potenza, E., Walsh, I., Gonzalo Parra, R., Giollo, M., Minervini, G., … Tosatto, S. C. E. (2014). RepeatsDB: A database of tandem repeat protein structures. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1175

Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures* (University of California, Irvine). Retrieved from https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Gao, R., Wang, M., Zhou, J., Fu, Y., Liang, M., Guo, D., & Nie, J. (2019). Prediction of Enzyme Function Based on Three Parallel Deep CNN and Amino Acid Mutation. *International Journal of Molecular Sciences*, *20*(11). https://doi.org/10.3390/ijms20112845

Hirsh, L., Paladin, L., Piovesan, D., & Tosatto, S. C. E. (2018). RepeatsDB-lite: A web server for unit annotation of tandem repeat proteins. *Nucleic Acids Research*, *46*(W1), W402–W407. https://doi.org/10.1093/nar/gky360

Hirsh, L., Piovesan, D., Paladin, L., & Tosatto, S. C. E. (2016). Identification of repetitive units in protein structures with ReUPred. *Amino Acids*, *48*(6), 1391–1400. https://doi.org/10.1007/s00726-016-2187-2

Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. *Proteins: Structure, Function and Bioinformatics*, *77*(3), 499–508. https://doi.org/10.1002/prot.22458

Kajava, A. V. (2012). Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*, *179*(3), 279–288. https://doi.org/10.1016/j.jsb.2011.08.009

Ketkar, N., & Ketkar, N. (2017). Introduction to Keras. *Deep Learning with Python*, 97–111. https://doi.org/10.1007/978-1-4842-2766-4_7

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. 1–15. Retrieved from http://arxiv.org/abs/1412.6980

Kitchenham, B., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. In *Technical report* (Vol. 2).

Laurens van der Maaten, G. H. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research 9*, 2579–2605. https://doi.org/10.1007/s10479-011-0841-3

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lesk, A. M. (2019). *Introduction to bioinformatics* (5th, illustr ed.). New York: Oxford

University Press.

Marcotte, E. M., Pellegrini, M., Yeates, T. O., & Eisenberg, D. (1999). A census of protein repeats. *Journal of Molecular Biology*, *293*(1), 151–160. https://doi.org/10.1006/jmbi.1999.3136

Mistry, J., Coggill, P., Eberhardt, R. Y., Deiana, A., Giansanti, A., Finn, R. D., … Punta, M. (2013). The challenge of increasing Pfam coverage of the human proteome. *Database*, *2013*, 1–11. https://doi.org/10.1093/database/bat023

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Retrieved from https://books.google.com.pe/books?id=NZP6AQAAQBAJ

Nelson, D. L., & Cox, M. M. (2008). *Lehninger Principles of Biochemistry* (7th ed.; W. H. Freeman, Ed.). New York: W. H. Freeman.

Paladin, L., Hirsh, L., Piovesan, D., Andrade-Navarro, M. A., Kajava, A. V., & Tosatto, S. C. E. (2017). RepeatsDB 2.0: Improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Research*, *45*(D1), D308–D312. https://doi.org/10.1093/nar/gkw1136

Pellegrini, M. (2015). Tandem repeats in proteins: Prediction algorithms and biological role. *Frontiers in Bioengineering and Biotechnology*, *3*(SEP), 1–8. https://doi.org/10.3389/fbioe.2015.00143

Pérez, F., & Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing. *Computing in Science and Engineering*, *9*(3), 21–29. Retrieved from http://ipython.org

Petticrew, M., & Roberts, H. (Eds.). (2006). *Systematic Reviews in the Social Sciences*. https://doi.org/10.1002/9780470754887

Python Software Foundation. (2017). What is Python? Executive Summary | Python.org. Retrieved October 29, 2019, from Python Software Foundation website: https://www.python.org/doc/essays/blurb/

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

Raschka, S. (2018). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. Retrieved from http://arxiv.org/abs/1811.12808

Torvalds, L., & Hamano, J. (2010). Git: Fast version control system. Retrieved November 3, 2019, from https://git-scm.com/

Walsh, I., Sirocco, F. G., Minervini, G., Di Domenico, T., Ferrari, C., & Tosatto, S. C. E. (2012). RAPHAEL: Recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*, *28*(24), 3257–3264. https://doi.org/10.1093/bioinformatics/bts550

Zacharaki, E. I. (2017). Prediction of protein function using a deep convolutional neural network ensemble. *PeerJ Computer Science*, *3*, e124. https://doi.org/10.7717/peerj-cs.124

Anexos

Anexo A: Plan de Proyecto

Resultado	**Fecha de inicio**	**Fecha de fin**
Objetivo 1: Definir una representación de datos válida para la extracción de características estructurales de una proteína		
R1: Procedimiento para la obtención de una representación de datos a partir de una cadena de proteínas, utilizando información estructural	03/02/2020	23/02/2020
R2: Conjunto de datos de cadenas de proteínas repetidas	24/02/2020	17/05/2020
Objetivo 2: Desarrollar y validar un método de aprendizaje de máquina que identifique proteínas repetidas y las clasifique en su respectiva clase y subclase		
R3: Conjunto de clasificadores por clase y subclase de proteína	16/03/2020	05/04/2020
R4: Reporte de predicciones hechas a todo el PDB Data Bank	06/04/2020	30/05/2020
Objetivo 3: Implementar un servicio web que permita caracterizar una cadena de proteína		
R5: Servicio web que utiliza los clasificadores	06/04/2020	03/05/2020