# PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## Escuela de Posgrado


Herramienta integrada para la curación de proteínas repetidas

Tesis para obtener el grado académico de Magíster en Informática que presenta:

***Manuel Alberto Bezerra Brandao Corrales***

Asesora:

***Dra. Layla Hirsh Martinez***

Lima, 2023

## Informe de Similitud

Yo, **Layla Hirsh Martinez**, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, **asesora de la tesis** titulada **Herramienta integrada para la curación de proteínas repetidas**, del autor **Manuel Alberto Bezerra Brandao Corrales**, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de **25%**. Así lo consigna el reporte de similitud emitido por el software Turnitin el **02/06/2023**.
- He revisado con detalle dicho reporte y la **Tesis**, y no se advierten indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.
- Cabe precisar que el mayor porcentaje de similitud que asciende a 13% corresponde a la tesis para la obtención del grado de Ingeniero Informático del autor en mención y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

**Lima, 3 de junio del 2023.**

| Apellidos y nombres de la asesora:<br>**Layla Hirsh Martinez** | |
|---|---|
| DNI:40329236 | Firma  |
| ORCID: 0000-0002-8215-6716 | |

## Dedicatoria

A mi abuelita, que desde el cielo me guía y cuida mi día a día.

A mi madre, por su amor infinito, cuyo corazón veo reflejado en cada logro de mi vida.

A mis tías, por llenar mi vida con el cariño y la alegría que las caracterizan.

A mi asesora la Dra. Layla Hirsh, por creer en mí y ayudar a potenciarme con cada reto profesional y académico que me propongo.

## Resumen

A finales de los años 1990, se identificó un conjunto de proteínas caracterizadas por tener patrones repetidos en su secuencia, lo que produce una estructura tridimensional repetitiva (Marcotte et al., 1999). Se han clasificado al menos 14% de proteínas encontradas en la naturaleza como repetidas, y presentan un rol crítico en procesos biológicos como la comunicación celular y el reconocimiento molecular (Brunette et al., 2015; Marcotte et al., 1999). Existe un creciente interés en el estudio de las proteínas repetidas debido a sus pliegues estructurales estables, una alta conversación evolutiva y un amplio repertorio de funciones biológicas (Chakrabarty & Parekh, 2022). Además, se estima que una de cada tres proteínas humanas son consideradas repetidas (Jorda & Kajava, 2010).

La identificación, clasificación y curación de regiones de repetición en proteínas es un proceso complejo que requiere del procesamiento manual de expertos, gran capacidad computacional y tiempo. Existen diversos avances recientes y relevantes que aplican modelos de aprendizaje automático para la predicción de estructura tridimensional de proteínas y la predicción de clasificación de proteínas repetidas. Este tipo de aplicaciones resultan útiles para este proceso de curación. No obstante, a pesar de que este tipo de software son de libre acceso y de código abierto, no se cuenta con un servicio integrado que contemple las herramientas y bases de datos que soporten la investigación en proteínas repetidas.

Por estos motivos, en este proyecto de investigación de plantea, diseña y desarrolla un servicio web integrado para la curación de proteínas repetidas. Con este objetivo, se ha considerado la integración con la base de datos de estructuras terciarias del Protein Data Bank (PDB) y la base de datos de predicciones de estructuras tridimensionales AlphaFold. Asimismo, se ha utilizado un modelo de redes neuronales que permite predecir la probabilidad de clasificación en cada clase de proteína repetida. Finalmente, con esta predicción, se implementó una mejora al algoritmo ReUPred para volver más eficiente el proceso de identificación de regiones y unidades de repetición.

Este servicio ha sido desplegado utilizando computación en la nube en la página bioinformática.org de la cual es parte el laboratorio de investigación en Bioinformática de la Pontificia Universidad Católica del Perú. Este servicio permite que los investigadores no requieran contar con alta capacidad de procesamiento computacional para el proceso de curación de proteínas repetidas e integra los resultados totales obtenidos.

# Índice general

## Índice de figuras

# Índice de tablas

# Capítulo 1.Generalidades

## 1.1 Problemática

En esta sección se procederá a presentar un árbol de problemas, con sus respectivos problemas causas y problemas efectos. Asimismo, se describe la problemática en la que se contextualiza el problema a abordar.

### 1.1.1 Árbol de Problemas

A continuación, en la Figura 1 se presenta el árbol de problemas, en el cual se muestran los problemas causas de la problemática seleccionada y los efectos que implica mediante

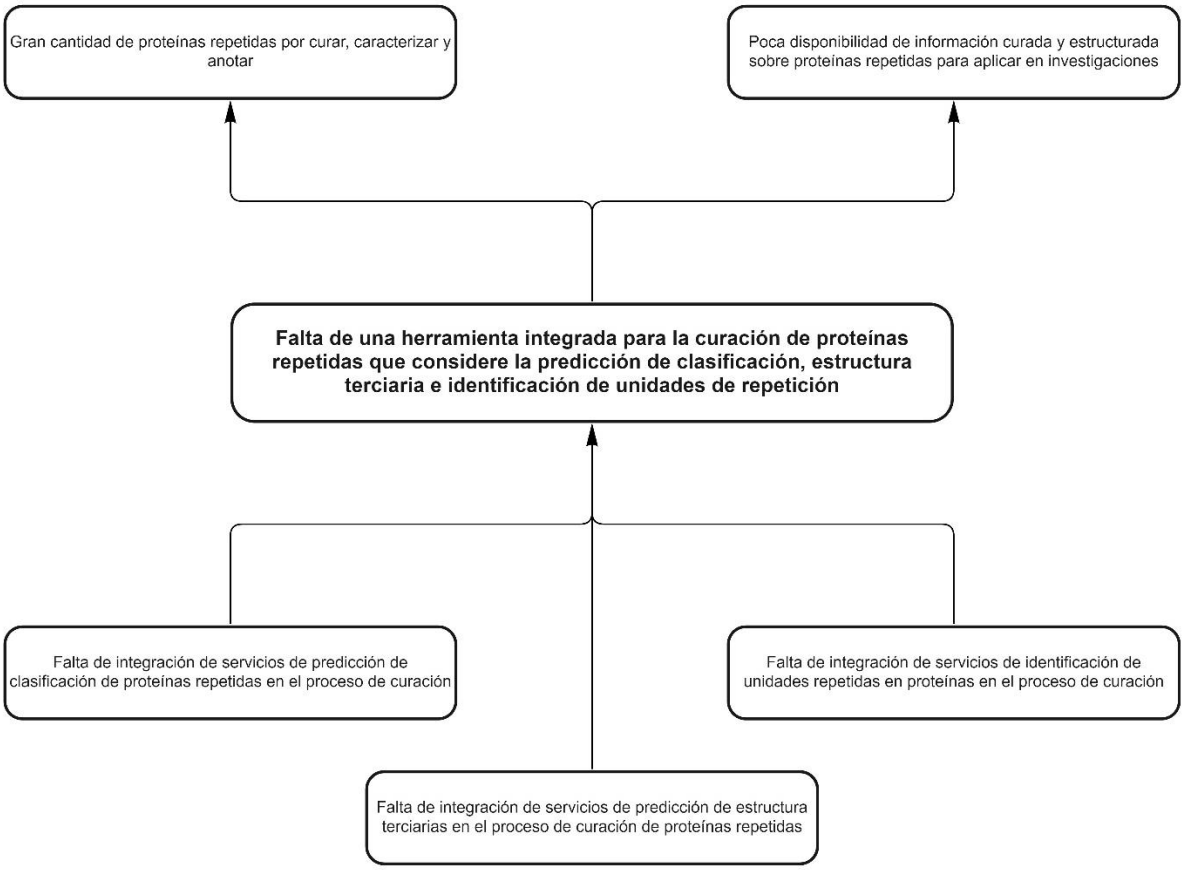



*Figura 1 Árbol de problemas: Parte central, problema principal- Zona superior, problemas efectos. Zona inferior, problemas causas.*

un diagrama.

### 1.1.2 Descripción

La bioinformática se origina hace más de 50 años, cuando las computadoras de oficina aún no existían y el ADN aún no podía ser secuenciado (Can, 2014). A finales de la década de los años 1950, se publicó la secuencia de aminoácidos de la insulina, convirtiéndose en la primera estructura primaria obtenida mediante los avances en la determinación estructural de proteínas mediante cristalografía (Gauthier et al., 2019). Este logro incentivó a desarrollar métodos más eficientes para el estudio de proteínas (Gauthier et al., 2019).

A finales de los años 1990, se identificó un conjunto de proteínas caracterizadas por tener patrones repetidos en su secuencia, lo que produce una estructura tridimensional repetitiva (Marcotte et al., 1999). Se han clasificado al menos 14% de proteínas encontradas en la naturaleza como repetidas, y presentan un rol crítico en procesos biológicos como la comunicación celular y el reconocimiento molecular (Brunette et al., 2015; Marcotte et al., 1999).

Existe un creciente interés en el estudio de las proteínas repetidas debido a sus pliegues estructurales estables, una alta conversación evolutiva y un amplio repertorio de funciones biológicas (Chakrabarty & Parekh, 2022). Además, se estima que una de cada tres proteínas humanas son consideradas repetidas (Jorda & Kajava, 2010). La clasificación y predicción de unidades de repetición es una tarea complicada que actualmente se realiza mediante la curación manual de expertos (Hirsh et al., 2016). Este proceso no tiene una definición formal, pero puede describirse como el esfuerzo realizado para la clasificación, identificación de regiones junto a unidades de repetición y la anotación de funciones, clases estructurales y familia de una proteína (Hirsh et al., 2016, 2017; Muroya Tokushima & Hirsh, 2022; Palomino & Hirsh, 2021; Pedraza & Hirsh, 2019; Tenorio Ku & Hirsh, 2021; Urbina & Hirsh, 2021).

Para la curación de proteínas repetidas, los investigadores acceden a servicios que permitan realizar estas tareas. Por ejemplo, se requiere consultar secuencias (The UniProt Consortium, 2019) y estructuras (Berman et al., 2000) de proteínas en bases de datos (Di Domenico et al., 2014). Asimismo, los investigadores podrían apoyarse en servicios que faciliten la curación, como predictores de clasificación de proteínas repetidas (Muroya Tokushima & Hirsh, 2022; Tenorio Ku & Hirsh, 2021), métodos de inferencia de estructuras terciarias (Palomino & Hirsh, 2021) o programas para la predicción de

unidades de repetición (Hirsh et al., 2016; Jumper et al., 2021; Pedraza & Hirsh, 2019; Tunyasuvunakool et al., 2021; Varadi et al., 2022).

A pesar de que este tipo de servicios existen y están disponibles para cualquier usuario, realizando una revisión del estado del arte[1] respecto a la curación de proteínas repetidas, no se cuenta con una herramienta integrada que permita realizar este proceso. Por todo esto, al existir una falta de integración de servicios para la predicción de clasificación de proteínas repetidas, la inferencia de estructuras tridimensionales y la detección de unidades de repetición, se afirma que existe la falta de una herramienta integrada para la curación de proteínas repetidas.

Esta falta ha causado que el proceso de curación de proteínas repetidas sea más engorroso y complicado, haciendo que los investigadores requieran perder tiempo y energía en sus labores. Como consecuencia, existe un número grande de proteínas repetidas por categorizar y anotar (Chakrabarty & Parekh, 2022; Di Domenico et al., 2014). Asimismo, no tener disponible la información curada y estructurada de una gran cantidad de proteínas repetidas dificulta su aplicación en proyectos de investigación relacionados a la biotecnología.

## 1.2 Objetivos

En esta sección del capítulo se presentan el objetivo general, los objetivos específicos y los resultados esperados en este proyecto de investigación.

### 1.2.1 Objetivo general

El objetivo general de este proyecto es desarrollar una herramienta que permita predecir la estructura de una proteína repetida e identificar sus unidades de repetición.

### 1.2.2 Objetivos específicos

- O1. Integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura terciaria.
- O2. Incorporar servicios de predicción de estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación
- O3. Implementar un servicio web para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición.

[1] Desarrollada en el Capítulo 3. Estado del arte de este documento.

### 1.2.3 Resultados esperados

O1. Integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura terciaria.
   R1. Descripción de un servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria.
   R2. Integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria.

O2. Incorporar servicios de predicción de la estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación.
   R3. Descripción de un servicio de predicción de estructura de una proteína basada en su secuencia.
   R4. Integración del servicio de predicción de estructura de una proteína basada en su secuencia.

O3. Implementar un servicio web integrado para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición.
   R5. Descripción de un servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición.
   R6. Integración del servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición.
   R7. Desarrollo del servicio web para predecir la estructura de una proteína repetida e identificar sus unidades de repetición siguiendo los lineamientos de usabilidad para servicios web bioinformáticos en el diseño de interfaz gráfica de usuario.

### 1.2.4 Mapeo de objetivos, resultados y verificación

A continuación, en las tablas 1, 2 y 3 se presentan el medio de verificación e indicador objetivamente verificable para cada resultado, organizado por cada objetivo.

*Tabla 1 Mapeo de resultados del objetivo específico 1*

| **O1.** Integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura terciaria. | | |
|---|---|---|
| **Resultado** | **Medio de verificación** | **Indicador objetivamente verificable** |
| R1. Descripción de un servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria. | • Informe de descripción de servicio<br>• Modelo de datos del servicio a utilizar | Aprobación general en una evaluación cualitativa de juicio experto en bioinformática. |
| R2. Integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria. | • Informe de despliegue e integración del servicio<br>• Propuesta de casos de prueba del servicio integrado<br>• Informe de ejecución de casos de prueba propuestos del servicio integrado | 100% casos de pruebas propuestos exitosos. |

*Tabla 2 Mapeo de resultados del objetivo específico 2*

| **O2.** Predecir la estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación. | | |
|---|---|---|
| **Resultado** | **Medio de verificación** | **Indicador objetivamente verificable** |
| R5. Descripción de un servicio de predicción de estructura de una proteína basada en su secuencia. | • Informe de descripción de servicio<br>• Modelo de datos del servicio a utilizar | Aprobación general en una evaluación cualitativa de juicio experto en bioinformática. |
| R6. Integración del servicio de predicción de estructura de una proteína basada en su secuencia. | • Informe de despliegue e integración del servicio<br>• Propuesta de casos de prueba del servicio integrado<br>• Informe de ejecución de casos de prueba propuestos del servicio integrado | 100% casos de pruebas propuestos exitosos. |

*Tabla 3 Mapeo de resultados del objetivo específico 3*

| **O3.** Implementar un servicio web integrado para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición. | | |
|---|---|---|
| **Resultado** | **Medio de verificación** | **Indicador objetivamente verificable** |
| R7. Descripción de un servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición. | • Informe de descripción de servicio<br>• Modelo de datos del servicio a utilizar | Aprobación general en una evaluación cualitativa de juicio experto en bioinformática. |
| R8. Integración del servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición. | • Informe de despliegue e integración del servicio<br>• Propuesta de casos de prueba del servicio integrado<br>• Informe de ejecución de casos de prueba propuestos del servicio integrado | 100% casos de pruebas propuestos exitosos. |
| R9. Desarrollo del servicio web para predecir la estructura de una proteína repetida e identificar sus unidades de repetición siguiendo los lineamientos de usabilidad para servicios web bioinformáticos en el diseño de interfaz gráfica de usuario. | • Diagrama de flujo del proceso del servicio integrado<br>• Requerimientos funcionales y no funcionales del servicio<br>• Mockups navegables del diseño de interfaces<br>• Código fuente del servicio desarrollado<br>• Propuesta de casos de prueba del servicio integrado desarrollado<br>• Informe de ejecución de casos de prueba propuestos del servicio integrado<br>• Informe de adaptación de lineamientos y evaluación de usabilidad | 100% de casos de pruebas propuestos exitosos.<br>Aprobación general en una evaluación cualitativa de juicio experto en bioinformática. |

## 1.3 Métodos, procedimientos y herramientas

Para desarrollar este proyecto de tesis, se considerarán diversos métodos, procedimientos y herramientas con los cuales se busca obtener los diversos resultados esperados.

Para el despliegue de los servicios para la predicción de clasificación, estructura terciaria y unidades de repetición se seguirán los procedimientos indicados en sus respectivas documentaciones. Asimismo, se realizarán las pruebas de ejecución que estas indiquen.

Por otro lado, para la integración de estos servicios para la herramienta web se desarrollará una aplicación middleware que integre las funcionalidades requeridas para la curación de proteínas repetidas. Esta se realizará utilizando el *framework* Flask del lenguaje Python que permite desarrollar un API REST.

Para el desarrollo del servicio web integrado para la curación de proteínas repetidas se elaborará un diagrama de ejecución en el que se encontrará el detalle de funciones que se integrarán por cada servicio desplegado. Asimismo, se definirán casos de prueba que el servicio deberá cumplir. Para el diseño de la interfaz, se utilizará la plataforma Figma. Para el desarrollo de la interfaz, se podrá trabajar con el *framework* Vue del lenguaje JavaScript. Para la programación en estos *frameworks*, se trabajará con Visual Studio Code. Asimismo, se tomará en consideración los lineamientos de usabilidad para servicios web bioinformáticos con la finalidad de asegurar la facilidad de uso del servicio (Bezerra Brandao et al., 2021).

Finalmente, el servicio será evaluado cualitativamente mediante juicio experto en bioinformática, en donde la persona experta llenará un acta donde pondrá su veredicto y sus observaciones.

# Capítulo 2. Marco Conceptual

## 2.1 Introducción

El objetivo de este marco conceptual es presentar el sustento basado en literatura que contextualiza la temática de este proyecto de investigación. Este marco ha sido basado en los trabajos desarrollados en el grupo de investigación de Bioinformática de la PUCP (Bezerra Brandao et al., 2021; Muroya Tokushima & Hirsh, 2022; Palomino & Hirsh, 2021; Pedraza & Hirsh, 2019; Tenorio Ku & Hirsh, 2021; Urbina & Hirsh, 2021). En este capítulo se presentan los conceptos requeridos para comprender la presente investigación y el problema central que es la falta de una herramienta integrada para la curación de proteínas repetidas que considere la predicción de clasificación, estructura terciaria e identificación de unidades de repetición.

## 2.2 Desarrollo del marco

El presente proyecto de investigación está basado en la **Bioinformática**. Esta es la aplicación de tecnologías computacionales y la estadística a la gestión y análisis de datos biológicos como lo son las secuencias de ADN y de proteínas (Can, 2014). Las **proteínas** son las macromoléculas biológicas más abundantes y se encuentran en todas las células (Lodish et al., 2006). Dependiendo de la función, las proteínas pueden ser consideradas como estructurales, de transporte, reguladoras o receptoras (Lodish et al., 2006). Una característica particular es que las células pueden producir proteínas con diferentes propiedades y funciones al unir los mismos 20 aminoácidos en muchas combinaciones y secuencias diferentes (Nelson et al., 2008). Los **aminoácidos** son moléculas orgánicas que contienen un grupo amino ($NH_2$) en uno de los extremos de la molécula y un grupo ácido carboxílico (COOH) en el otro externo (Nelson et al., 2008). Como se mencionó previamente, los aminoácidos son las unidades que forman las proteínas y tanto estos como sus derivados participan en funciones celulares tan diversas como la transmisión nerviosa y la biosíntesis (Nelson et al., 2008).

Cuando se analiza una proteína se suele hacer hincapié en su estructura que puede ser descrita en varios niveles de complejidad (Nelson et al., 2008). La **estructura primaria**, también llamada **secuencia**, es una descripción de todos los enlaces covalentes, principalmente enlaces peptídicos y enlaces disulfuro, que unen residuos de aminoácidos en una cadena polipeptídica (Nelson et al., 2008). En otras palabras, el aspecto más importante de la estructura primaria es la secuencia de residuos de aminoácidos (Nelson et al., 2008). Desde el punto de vista de Bioinformática, esta secuencia de proteína hace

referencia a una cadena simple de letras que denota su secuencia de aminoácidos (Nelson et al., 2008). Por otro lado, la **estructura terciaria** de una proteína, también llamada **estructura 3D**, es la forma tridimensional final de una proteína que necesita adoptar para funcionar correctamente (Center for BioMolecular Modeling, 2022). Las interacciones y enlaces de las cadenas laterales dentro de una proteína particular determinan esta estructura (Moss, 2009). La forma tridimensional se define por sus coordenadas atómicas, las cuales pueden referirse a un dominio de proteína o a la estructura terciaria completa (Moss, 2009).

Un tipo de proteína que, durante los últimos años, se ha convertido en un tema de estudio e investigación muy frecuente son las **proteínas repetidas** (Brunette et al., 2015). Estas tienen una fuerte presencia en la naturaleza y cuentan con un rol crítico en procesos biológicos como la comunicación celular y el reconocimiento molecular (Brunette et al., 2015). Las proteínas repetidas se caracterizan por tener repetición de aminoácidos en su secuencia o, en su defecto, una estructura 3D repetitiva (Marcotte et al., 1999). Además, se estima que tienen una ocurrencia de una en cada tres proteínas humanas (Jorda & Kajava, 2010), por lo que tienen una gran relevancia en la salud (Jorda & Kajava, 2010) y son útiles para aplicaciones de ingeniería con proteínas (Mutter et al., 2008). En las proteínas repetidas, se le considera **unidad de repetición** al bloque mínimo que se repite más de una vez en una secuencia, ya sea en una forma idéntica o muy similar (Heringa, 1998). Las unidades repetitivas de proteínas son considerablemente diversas, desde la repetición de un solo aminoácidos hasta dominios de 100 o más residuos (Heringa, 1998).

La clasificación y predicción de unidades de repetición es una tarea complicada que actualmente se realiza mediante la **curación** manual de expertos (Hirsh et al., 2016). Este proceso no tiene una definición formal, pero puede describirse como el esfuerzo realizado para la clasificación, identificación de regiones junto a unidades de repetición y la anotación de funciones, clases estructurales y familia de una proteína (Hirsh et al., 2016, 2017; Muroya Tokushima & Hirsh, 2022; Palomino & Hirsh, 2021; Pedraza & Hirsh, 2019; Tenorio Ku & Hirsh, 2021; Urbina & Hirsh, 2021). Este proceso permite estructurar y ampliar la información conocida respecto a las proteínas repetidas para diversas aplicaciones relacionadas a la biotecnología (Hirsh et al., 2016).

# Capítulo 3.Estado del Arte

## 3.1 Introducción

Para el desarrollo de una herramienta que permita realizar la predicción de la estructura de una proteína repetida y sus unidades de repetición es necesario revisar las investigaciones y herramientas que se han llevado a cabo en los últimos años relacionadas a esta temática. Por lograrlo, se realizará una revisión de literatura empírica en bases de datos indizadas.

## 3.2 Objetivos de revisión

Las proteínas repetidas cuentan con un rol fundamental en la naturaleza (Kajava, 2012). Actualmente, se encuentran diversos estudios en este tipo de proteínas dada la relevancia que tiene en diversos procesos biológicos (Di Domenico et al., 2014). Es por ello que en esta revisión se busca conocer sobre los métodos y herramientas para la identificación de proteínas repetidas. Específicamente, los que están basados en estructura terciaria. Debido a esto, es importante investigar los métodos para la predicción de la estructura de una proteína en base a su secuencia. Finalmente, se requiere indagar respecto a los métodos de curación de unidades de repetición de proteínas repetidas.

## 3.3 Preguntas de revisión

Para definir las preguntas de investigación, se utiliza el método PICo (Lockwood et al., 2015). Esto consiste en identificar la población (P, del inglés *population*), el fenómeno de interés de la población (I, del inglés *interest*) y el contexto (Co, del inglés *context*) a investigar (Lockwood et al., 2015). Cada elemento puede ser visualizado en la tabla 4.

*Tabla 4 Método PICo (Lockwood et al., 2015) para el planteamiento de preguntas de investigación*

| | |
|---|---|
| **Población** | Proteínas repetidas |
| **Interés** | Identificación, predicción de estructura, curación de unidades de repetición |
| **Contexto** | Bioinformática, ciencias de la computación |

A continuación, se presentan las preguntas formuladas para esta revisión de literatura.

- P1. ¿En qué consisten y cuáles son los métodos y herramientas para la identificación de proteínas repetidas en base a su estructura?
- P2. ¿En qué consisten y cuáles son los métodos para la predicción de la estructura de una proteína según su secuencia?

- P3. ¿En qué consiste y cuáles son los métodos o herramientas para la curación de unidades de repetición de proteínas repetidas?

## 3.4 Estrategia de búsqueda

A continuación, se presenta el plan para la búsqueda de literatura primaria. Esta será esencial para responder a las preguntas planteadas para esta revisión.

### 3.4.1 Motores de búsqueda a usar

Para esta revisión de literatura se trabajarán con dos bases de datos indizadas. En primer lugar, se trabajará con Scopus. Esta es una base de datos de artículos STM (ciencia, tecnología y medicina por sus siglas en inglés) que facilita a los investigadores la búsqueda de publicaciones (Burnham, 2006). Además, se utilizará PubMed. Este también es un servicio de información desarrollado por el Centro Nacional para la Información Biotecnológica de Estados Unidos (NCBI por sus siglas en inglés) que cubre los temas de biomedicina y salud (Cañedo Andalia et al., 2015). Ya que las publicaciones requeridas se centran en bioinformática, ambas bases de datos son seleccionadas.

Además, se tiene conocimiento de que este tema ha sido trabajado previamente en proyectos de investigación del grupo de investigación de Bioinformática[2]. Estos han sido publicados como documentos de tesis en el repositorio institucional de la universidad y se incluyen en esta revisión.

### 3.4.2 Cadenas de búsqueda a usar

Para la revisión en las bases de datos indizadas, se ha elaborado una cadena de búsqueda por cada pregunta. Dado que las sintaxis de búsqueda avanzada de ambos motores difieren y tienen funcionalidades diferentes, las cadenas de cada preguntan han sido adaptadas para cada base de datos.

Para el caso de las tesis, se presentará en las siguientes secciones una selección manual de trabajos de investigación. Esta será validada por la coordinadora del grupo de investigación de Bioinformática.

Para la conformación de las cadenas de búsqueda en bases de datos indizadas para cada pregunta se siguió el proceso que se describe a continuación.

[2] Se puede encontrar mayor información relacionada al grupo en Bioinformatica.org

- SC1. En primer lugar, para cada pregunta, se determinaron las palabras claves a incluir en la búsqueda. Estas requirieron estar relacionadas con lo identificado en el método PICo (Lockwood et al., 2015) y la pregunta en sí.
    - SC1_P1. Para la revisión de métodos y herramientas para la identificación de proteínas repetidas en base a su estructura se propone la siguiente subcadena.

      "Repeat Protein*" AND ("Identification" OR "Prediction") AND ("Method" OR "Tool" OR "Service") AND "Structure"
    - SC1_P2. Para la revisión de métodos para la predicción de la estructura de una proteína en base a su secuencia se propone la siguiente subcadena.

      "Protein*" AND (("Tertiary Structure Prediction" OR "3D Structure Prediction")) AND ("Primary Structure" OR "Sequence") AND "Method"
    - SC1_P3. Para la revisión de curación de unidades de repetición en proteínas repetidas se propone la siguiente subcadena.

      "Protein*" AND ("Repeat* unit*" OR "Repeat* element*") AND ("Curation" OR "Identification" OR "Annotation")
- SC2. Luego, se definición una subcadena de exclusión. Se procuró excluir las investigaciones relacionadas al ADN, ARN. Asimismo, se excluyeron las investigaciones relacionadas a la estructura secundaria. Por último, se excluyeron investigaciones experimentales. Todos estos temas quedan fuera del alcance de este trabajo de investigación y, por ende, de esta revisión de literatura.

  NOT "Secondary Structure" AND NOT ("DNA" OR "cDNA" OR "Gene" OR "RNA") AND NOT ("Wet lab" OR "Experiment" OR "CASP*" OR "Mass Spectrometry"))
- SC3. Asimismo, se limitó la búsqueda a las áreas de Ciencias de la Computación, Bioquímica, Genética y Biología Molecular.

  (LIMIT-TO (SUBJAREA,"BIOC") OR LIMIT-TO (SUBJAREA,"COMP"))

Finalmente, se combinaron las tres subcadenas, por cada pregunta, y se obtuvieron las cadenas de búsqueda a utilizar. A continuación, se presenta cada una de estas cadenas adaptadas a los motores de búsqueda Scopus y PubMed.

#### 3.4.2.1 Cadenas de búsqueda para ¿En qué consisten y cuáles son los métodos y herramientas para la identificación de proteínas repetidas en base a su estructura?

Para la búsqueda en Scopus se propone la siguiente cadena.

TITLE-ABS-KEY("Repeat Protein*" AND ("Identification" OR "Prediction") AND ("Method" OR "Tool" OR "Service") AND "Structure" AND NOT “Secondary Structure” AND NOT ("DNA" OR “cDNA” OR "Gene" OR “RNA”) AND NOT ("Wet lab" OR "Experiment" OR "CASP*" OR "Mass Spectrometry")) AND (LIMIT-TO(SUBJAREA,"BIOC") OR LIMIT-TO (SUBJAREA,"COMP"))

Para la búsqueda en PubMed se propone la siguiente cadena.

"Repeat Protein*"[Title/Abstract]) AND ("Identification"[Title/Abstract] OR "Prediction"[Title/Abstract]) AND ("Method"[Title/Abstract] OR "Tool"[Title/Abstract] OR "Service"[Title/Abstract]) AND "Structure"[Title/Abstract] NOT (“Secondary Structure”) NOT ("DNA") NOT ("cDNA") NOT ("Gene") NOT ("RNA") NOT ("Wet Lab") NOT ("Experiment") NOT ("CASP*") NOT ("Mass Spectrometry")

#### 3.4.2.2 Cadenas de búsqueda para ¿En qué consisten y cuáles son los métodos para la predicción de la estructura de una proteína según su secuencia?

Para la búsqueda en Scopus se propone la siguiente cadena.

TITLE-ABS-KEY("Protein*" AND (("Tertiary Structure Prediction" OR "3D Structure Prediction")) AND ("Primary Structure" OR "Sequence") AND "Method" AND NOT "Secondary Structure" AND NOT ("DNA" OR "cDNA" OR "Gene" OR "RNA") AND NOT ("Wet lab" OR "Experiment" OR "CASP*" OR "Mass Spectrometry")) AND (LIMIT-TO (SUBJAREA,"BIOC") OR LIMIT-TO (SUBJAREA,"COMP"))

Para la búsqueda en PubMed se propone la siguiente cadena.

"Protein*"[Title/Abstract] AND ("Tertiary Structure Prediction"[Title/Abstract] OR "3D Structure Prediction"[Title/Abstract]) AND ("Primary Structure"[Title/Abstract] OR "Sequence"[Title/Abstract]) AND "Method"[Title/Abstract] NOT ("Secondary Structure") NOT ("DNA") NOT ("cDNA") NOT ("Gene") NOT ("RNA") NOT ("Wet Lab") NOT ("Experiment") NOT ("CASP*") NOT ("Mass Spectrometry")

#### 3.4.2.3 Cadenas de búsqueda para ¿En qué consiste y cuáles son los métodos o herramientas para la curación de unidades de repetición de proteínas repetidas?

Para la búsqueda en Scopus se propone la siguiente cadena.

```
TITLE-ABS-KEY("Protein*" AND ("Repeat* unit*" OR "Repeat* element*") AND
("Curation" OR "Identification" OR "Annotation") AND NOT "Secondary Structure”
AND NOT ("DNA" OR "cDNA" OR "Gene" OR "RNA") AND NOT ("Wet lab" OR "Experiment"
OR "CASP*" OR "Mass Spectrometry")) AND (LIMIT-TO (SUBJAREA,"BIOC") OR LIMIT-TO
(SUBJAREA,"COMP"))
```

Para la búsqueda en PubMed se propone la siguiente cadena.

```
"Protein*"[Title/Abstract] AND ("Repeat* unit*"[Title/Abstract] OR "Repeat*
element*"[Title/Abstract]) AND ("Curation"[Title/Abstract] OR
"Identification"[Title/Abstract] OR "Annotation"[Title/Abstract] "Annotation")
NOT "Secondary Structure" NOT ("DNA") NOT ("cDNA") NOT ("Gene") NOT ("RNA") NOT
("Wet Lab") NOT ("Experiment") NOT ("CASP*") NOT ("Mass Spectrometry")
```

### 3.4.3 Documentos encontrados

Con cada cadena, se realiza la búsqueda de literatura en las bases de datos indizadas seleccionadas. A continuación, la tabla 5 presenta el total de publicaciones encontradas por cada pregunta, excluyendo las repeticiones.

*Tabla 5 Resumen numérico de resultados por pregunta en cada motor de búsqueda*

| Motor de búsqueda | P1 | P2 | P3 |
|---|---|---|---|
| Scopus | 11 | 61 | 25 |
| PubMed | 6 | 13 | 2 |
| Total | 17 | 74 | 27 |

Asimismo, se encontraron 3 publicaciones que resultaron de las búsquedas para la pregunta 1 y la pregunta 3. Por ello, se tiene un total de 115 publicaciones iniciales para esta revisión.

### 3.4.4 Criterios de exclusión/inclusión

Para la revisión de las publicaciones obtenidas de las búsquedas en los motores seleccionados, es necesario identificar cuáles de los resultados son estrictamente relevantes para la revisión. Por este motivo, se proponen los siguientes criterios de inclusión y exclusión.

#### 3.4.4.1 Criterios de exclusión

- CE1. Excluir los resultados que no sean artículos (*Article*) o actas de conferencias (*Conference Paper*). Se ha identificado que solo este tipo de publicaciones

permitirán responder de manera directa las preguntas planteadas debido a que abordan específicamente los temas relacionados al objetivo.

- CE2. Excluir los artículos en idiomas diferentes al español, portugués e inglés. Considerar los resultados en idiomas diferentes a los mencionados representará una dificultad para su análisis puesto que se cuenta con una barrera lingüística y los traductores en línea no son cien por ciento fiables, especialmente en términos académicos.

#### 3.4.4.2 Criterios de inclusión

- CI1. Incluir los resultados cuyo año de publicación se encuentre entre los últimos seis años (2016-2022). Esto debido a que se desea revisar los avances actuales[3].
- CI2. Incluir los resultados que se enfoquen principalmente en métodos o herramientas para la clasificación de proteínas repetidas, predicción de la estructura de una proteína en base a su secuencia y/o la curación para la identificación de unidades de repetición en proteínas repetidas. Entre los resultados se han encontrado publicaciones que, a pesar de mencionar temas relacionados a las preguntas, no desarrollan la información que se espera usar.

### 3.4.5 Estudios primarios

Luego de aplicar los criterios de inclusión y exclusión, se encontrarán las publicaciones que no son excluidas por los criterios y que cumplen con todos los criterios de inclusión. Estas serán relevantes para responder a las tres preguntas que se plantearon para esta revisión. La tabla 6 detalla el resumen de publicaciones por criterios de exclusión.

*Tabla 6 Resumen de resultados por criterios de exclusión*

| Criterios de exclusión | Número de resultados |
|---|---|
| CE1. Excluir los resultados que no sean artículos (*Article*) o actas de conferencias (*Conference Paper*). | 8 |
| CE2. Excluir los artículos en idiomas diferentes al español, portugués e inglés. | 0 |
| Resultados que no son excluidos | 107 |

Según lo descrito anteriormente, los documentos que no han sido excluidos que serán útiles para absolver las interrogantes planteadas deberán cumplir con los dos criterios de

[3] Por lo general, para revisiones del estado del arte, la antigüedad de las publicaciones a revisar no debería exceder los 6 años.

inclusión. La tabla 7 presenta el resumen de publicaciones según criterio de inclusión, resumiendo cuántos podrán ser considerados artículos relevantes.

*Tabla 7 Resumen de resultados por criterios de inclusión*

| Criterios de inclusión | Número de resultados |
|---|---|
| CI1. Incluir los resultados cuyo año de publicación se encuentre entre los últimos seis años (2016-2022). | 26 |
| CI2. Incluir los resultados que se enfoquen principalmente en métodos u herramientas para la clasificación de proteínas repetidas, predicción de la estructura de una proteína en base a su secuencia y/o la curación para la identificación de unidades de repetición en proteínas repetidas. | 11 |
| Publicaciones no excluidas que cumplen con todos los criterios de inclusión y que son consideradas relevantes. | 11 |

Finalmente, se concluye que solo 11 de los resultados cumplen todos los criterios de inclusión y no han sido excluidos. Estos son los que se considerarán como los artículos relevantes de la revisión. A continuación, se presenta la lista de las publicaciones a utilizar, ordenados alfabéticamente por autor, para cada pregunta.

#### 3.4.5.1 Estudios primarios para ¿En qué consisten y cuáles son los métodos y herramientas para la identificación de proteínas repetidas en base a su estructura?

Chakrabarty, B., & Parekh, N. (2020). PRIGSA2: Improved version of protein repeat identification by graph spectral analysis. *Journal of Biosciences*, *45*(1), 95. https://doi.org/10.1007/s12038-020-00058-x

Espada, R., Parra, R. G., Mora, T., Walczak, A. M., & Ferreiro, D. U. (2017). Inferring repeat-protein energetics from evolutionary information. *PLOS Computational Biology*, *13*(6), e1005584. https://doi.org/10.1371/journal.pcbi.1005584

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

Pagès, G., & Grudinin, S. (2019). DeepSymmetry: Using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. *Bioinformatics*, *35*(24), 5113-5120. https://doi.org/10.1093/bioinformatics/btz454

#### 3.4.5.2 Estudios primarios para ¿En qué consisten y cuáles son los métodos para la predicción de la estructura de una proteína según su secuencia?

Chen, C., Wu, H., & Bian, K. (2017). β-Barrel Transmembrane Protein Predicting Using Support Vector Machine. En D.-S. Huang, A. Hussain, K. Han, & M. M. Gromiha (Eds.), *Intelligent Computing Methodologies* (Vol. 10363, pp. 360-368). Springer International Publishing. https://doi.org/10.1007/978-3-319-63315-2_31

Fefelova, I., Fefelov, A., Lytvynenko, V., Dzierżak, R., Lurie, I., Savina, N., Voronenko, M., & Vyshemyrska, S. (2020). Protein Tertiary Structure Prediction with Hybrid Clonal Selection and Differential Evolution Algorithms. En V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya, & S. Radetskaya (Eds.), *Lecture Notes in Computational Intelligence and Decision Making* (Vol. 1020, pp. 673-688). Springer International Publishing. https://doi.org/10.1007/978-3-030-26474-1_47

Palopoli, N., Monzon, A. M., Parisi, G., & Fornasari, M. S. (2016). Addressing the Role of Conformational Diversity in Protein Structure Prediction. *PLOS ONE*, *11*(5), e0154923. https://doi.org/10.1371/journal.pone.0154923

Shi, H., & Zhang, X. (2020). Component-Based Design and Assembly of Heuristic Multiple Sequence Alignment Algorithms. *Frontiers in Genetics*, *11*, 105. https://doi.org/10.3389/fgene.2020.00105

Takahashi, T., Chikenji, G., & Tokita, K. (2021). Lattice protein design using Bayesian learning. *Physical Review E*, *104*(1), 014404. https://doi.org/10.1103/PhysRevE.104.014404

#### 3.4.5.3 Estudios primarios para ¿En qué consiste y cuáles son los métodos o herramientas para la curación de unidades de repetición de proteínas repetidas?

Bliven, S. E., Lafita, A., Rose, P. W., Capitani, G., Prlić, A., & Bourne, P. E. (2019). Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm. *PLOS Computational Biology*, *15*(4), e1006842. https://doi.org/10.1371/journal.pcbi.1006842

Chakrabarty, B., & Parekh, N. (2022). DBSTRIPS: Database of structural repeats in proteins. *Protein Science*, *31*(1), 23-36. https://doi.org/10.1002/pro.4052

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

#### 3.4.5.4 Trabajos de investigación seleccionados del grupo de investigación de Bioinformática

A continuación, se presenta la lista de trabajos de investigación desarrollados dentro del marco de trabajo del grupo de investigación Bioinformática que han sido seleccionados para este proyecto de tesis. Esta selección ha sido validada por la coordinadora del grupo[4].

Bezerra Brandao, M., Hirsh, L., & Pow Sang, J. (2021). *Usabilidad en servicios web bioinformáticos* [Pontificia Universidad Católica del Perú]. http://hdl.handle.net/20.500.12404/19477

Hirsh, L., Bernardi, P., Tosatto, S. S. E., & Piovesan, D. (2017). *Solving the Structural Modeling Problems for Tandem Repeat Proteins* [Università Degli Studi Di Padova]. http://hdl.handle.net/11577/3424915

[4] El acta de conformidad de esta validación puede ser revisada en el siguiente enlace: https://drive.google.com/file/d/15ohatXizRuz4uBAcfMZE-bSDi-xx69Ku/view?usp=sharing

Muroya Tokushima, L. F., & Hirsh, L. (2022). *Identificación y clasificación automática de repeticiones en estructuras de proteínas repetidas* [Pontificia Universidad Católica del Perú]. http://hdl.handle.net/20.500.12404/21423

Palomino, S., & Hirsh, L. (2021). *Desarrollo de una herramienta para la predicción de estructuras terciarias de proteínas repetidas a partir de su estructura primaria*. Pontificia Universidad Católica del Perú.

Pedraza, K., & Hirsh, L. (2019). *Optimización de método para la clasificación de proteínas repetidas e identificación de unidades de repetición mediante el uso de perfiles de Modelos Ocultos de Markov*. Pontificia Universidad Católica del Perú.

Tenorio Ku, L. G., & Hirsh, L. (2021). *Clasificación de proteínas repetidas basado en su información estructural utilizando aprendizaje de máquina* [Pontificia Universidad Católica del Perú]. http://hdl.handle.net/20.500.12404/18199

Urbina, V., & Hirsh, L. (2021). *Herramienta para la curación de familias de proteínas repetidas*. Pontificia Universidad Católica del Perú.

### 3.5 Formulario de extracción de datos

Para poder obtener los datos que serán más útiles para absolver las preguntas propuestas en la revisión se utilizará un formulario de extracción. Se adjunta a este documento, mediante el anexo A “Formulario de extracción”, una hoja de cálculo con la estructura e información obtenida de los artículos relevantes. Este formulario será utilizado para obtener la información de los estudios primarios obtenidos mediante la búsqueda en bases de datos indizadas. A continuación, en la tabla 8, se detalla el nombre, descripción y pregunta relacionada de cada uno de los campos del formulario de extracción diseñado.

*Tabla 8 Descripción del formulario de extracción*

| Campo | Descripción | Pregunta |
|---|---|---|
| ID | E[Número] por ejemplo, E04 | General |
| P1 | Indica si la publicación viene de la búsqueda de la P1 | P1 |
| P2 | Indica si la publicación viene de la búsqueda de la P2 | P2 |
| P3 | Indica si la publicación viene de la búsqueda de la P3 | P3 |
| Autores | Autor o autores de la publicación | General |
| Título | Título de la publicación | General |
| Año | Año de publicación | General |
| Fuente | Nombre de la fuente de publicación | General |
| nº Citas | Número de veces que ha sido citado | General |
| DOI | Identificador DOI | General |
| Resumen | También llamado *Abstract* | General |
| Tipo de documento | *Article* o *Conference Paper* | General |
| Base de datos | Scopus o PubMed | General |
| Identificación de proteínas repetidas | Métodos y herramientas para la predicción de proteínas repetidas en base a su estructura | P1 |
| Predicción de estructura de proteína | Métodos para la predicción de la estructura de una proteína según su secuencia | P2 |
| Curación de unidades de repetición | Métodos y herramientas de curación de unidades de repetición de proteínas repetidas | P3 |

## 3.6 Resultados de la revisión

En esta sección, se presentan las respuestas elaboradas para cada pregunta utilizando la información de la literatura seleccionada por medio del formulario de extracción (anexo A). Se incluyen los proyectos de tesis seleccionados del grupo de investigación de Bioinformática.

### 3.6.1 Respuesta a pregunta ¿En qué consisten y cuáles son los métodos y herramientas para la identificación de proteínas repetidas en base a su estructura?

Las proteínas repetidas están compuestas por unidades de repetición de patrones estructurales similares, compuestas por aminoácidos, entre 20 y 40 diferentes en la naturaleza (Espada et al., 2017). Estas son el objetivo de estudio principal para el diseño de proteínas, con muchos resultados exitosos en diversas topologías (Brunette et al., 2015; Rowling et al., 2015; Urvoas et al., 2010). En los estudios primarios seleccionados se han encontrado algunos métodos y herramientas para la identificación de proteínas repetidas (Chakrabarty & Parekh, 2020; Hirsh et al., 2016).

En primer lugar, se encontró el algoritmo PRIGSA (Chakrabarty & Parekh, 2020). Este es un método generalizado para la identificación de repeticiones estructurales usando el perfil del eigenvector de centralidad ($A_{levc}$) y la arquitectura de la estructura secundaria de la proteína (Chakrabarty & Parekh, 2020). Este algoritmo identifica miembros de familias de proteínas repetidas conocidas comparando el perfil del $A_{levc}$ y la arquitectura de la estructura secundaria con los perfiles de familias conocidas, computarizadas previamente (Chakrabarty & Parekh, 2020). Asimismo, en su última actualización, PRIGSA2 captura nuevos patrones de repetición para identificar proteínas repetidas que aún no han sido clasificadas (Chakrabarty & Parekh, 2020). Este método está enfocado principalmente en la detección de las clases III y IV de proteínas repetidas (Chakrabarty & Parekh, 2020; Kajava, 2012). Esta herramienta consiste de tres módulos principales. El primero, se encarga de la construcción de la red de procesamiento, el segundo de la identificación de repeticiones y el último es un módulo dedicado al procesamiento posterior (Chakrabarty & Parekh, 2020).

Asimismo, dentro de la literatura primaria se encuentra ReUPred, un predictor para la clasificación de proteínas repetidas (Hirsh et al., 2016). Este método está basado en la clasificación RepeatsDB y solo considera proteínas repetidas solenoides (clases III.1 al III.3) ya que es considerada la clase de proteínas más abundante en la naturaleza (Hirsh et al., 2016; Kajava, 2001, 2012). El algoritmo analiza la composición de la estructura de la proteína mediante una estrategia del tipo “divide y vencerás” (Hirsh et al., 2016). Esta herramienta busca imitar la evolución para realizar la predicción de la clasificación (Hirsh et al., 2016). Esto se fundamenta en que se ha demostrado que las unidades solenoides

evolucionan de una unidad representativa a múltiples copias mediante la duplicación de repeticiones (Björklund et al., 2006; Hirsh et al., 2016).

Por otro lado, se encontraron métodos relacionados a la identificación de distribución de energía y repeticiones de estructura que están directamente relacionadas con la identificación de proteínas repetidas (Espada et al., 2017; Pagès & Grudinin, 2019). Por ejemplo, DeepSymmetry es un método basado en redes convolucionales 3D que detecta repeticiones estructurales en proteínas y su mapa de densidad (Pagès & Grudinin, 2019). Este algoritmo está diseñado para identificar repeticiones proteicas, proteínas con simetrías internas y las simetrías en los mapas de densidad, junto con su orden y ejes de simetría (Pagès & Grudinin, 2019). Para estos últimos, el algoritmo se basa en el aprendizaje de mapeo Veronese de seis dimensiones de vectores 3D (Pagès & Grudinin, 2019). En este modelo, las estructuras de proteínas son representadas por el mapa de densidad de electrones (Pagès & Grudinin, 2019).

De la misma manera, se encontraron menciones a métodos para analizar mutaciones correlacionadas entre familias de proteínas como mfDCA (Morcos et al., 2011), plmDCa (Ekeberg et al., 2013, 2014) y Gremlin (Balakrishnan et al., 2011). La hipótesis principal en estos métodos es que los cambios bioquímicos producidos en un punto de mutación deben estar compensados por otras mutaciones, a través de las líneas de evolución, para mantener la viabilidad o función de la proteína (Espada et al., 2017). Se cree que las proteínas repetidas evolucionan mediante la duplicación y reordenamiento de repeticiones, lo que resulta en una simetría inherente en la que el análisis de la secuencia suele perder precisión (Espada et al., 2017). Por todo esto, se desarrolló un modelo estadístico para considerar los detalles de la distribución de energía en familias de proteínas repetidas usando únicamente secuencias de aminoácidos, tomando como eje central a las familias de proteínas repetidas ANK (*ankyrin*), LRR (*leucine-rich*) y TPR (*tetratricopeptide-like*) (Espada et al., 2017).

Dentro de los proyectos realizados en el grupo de investigación de Bioinformática, se encuentra un método de predicción de identificación y clasificación de repeticiones en estructuras de proteínas repetidas basado en aprendizaje automático (Muroya Tokushima & Hirsh, 2022). Este tiene un enfoque de conversión de la información estructural en un mapa de volumen electrónico para el filtro de repeticiones y en una imagen 2D para la clasificación (Muroya Tokushima & Hirsh, 2022). Esta predicción se centra en la clase IV (IV.1, IV.2 y IV.4) y se alimenta de las estructuras disponibles en el banco de datos de

proteínas PDB (Berman et al., 2000) y las estructuras de regiones clase IV encontradas en RepeatsDB (Di Domenico et al., 2014) (Muroya Tokushima & Hirsh, 2022). Este método obtuvo un 80% de sensibilidad en un universo potencial de aproximadamente 130 mil repeticiones con una precisión de 95% sobre el conjunto de pruebas (Muroya Tokushima & Hirsh, 2022).

Asimismo, en otro proyecto se desarrolló un algoritmo para la clasificación de proteínas repetidas basado en su información estructural primaria (Tenorio Ku & Hirsh, 2021). Su principal objetivo es detectar la presencia de regiones repetidas dentro de una cadena protéica con la finalidad de evitar procesar información irrelevante en métodos más complejos como ReUPred (Hirsh et al., 2017) que requieren de muchos recursos (Tenorio Ku & Hirsh, 2021). Si bien en este proyecto no se considera la estructura 3D, resulta útil considerar el desarrollo de este algoritmo ya que existen métodos para la predicción de estructuras terciarias en proteínas (Palomino & Hirsh, 2021).

Se concluye que existen algunos estudios para la clasificación de proteínas repetidas en base a la estructura (Chakrabarty & Parekh, 2020; Hirsh et al., 2016; Muroya Tokushima & Hirsh, 2022). Asimismo, estos están basados en aprendizaje de máquina, biofísica, bioquímica, modelos estadísticos y algoritmos computacionales (Chakrabarty & Parekh, 2020; Espada et al., 2017; Hirsh et al., 2016; Muroya Tokushima & Hirsh, 2022; Pagès & Grudinin, 2019). Finalmente, muchos de estos algoritmos se encuentran desplegados en un servicio web (Chakrabarty & Parekh, 2020; Hirsh et al., 2016; Muroya Tokushima & Hirsh, 2022; Pagès & Grudinin, 2019).

### 3.6.2 Respuesta a pregunta ¿En qué consisten y cuáles son los métodos para la predicción de la estructura de una proteína según su secuencia?

La predicción de la estructura 3D de una proteína según su secuencia de aminoácidos es conocida como un problema de gran interés en la ciencia (Al-Lazikani et al., 2001). Por ello, para el presente proyecto de investigación es importante revisar los métodos para realizar esta predicción que hayan sido desarrollados en los últimos años. Se encontraron publicaciones que proponen algoritmos computacionales (Fefelova et al., 2020) y que desarrollan mejoras a métodos existentes o describen estudios muy relacionados con la predicción de estructura terciaria de una proteína (Chen et al., 2017; de Lima Salgado et al., 2017; Shi & Zhang, 2020; Takahashi et al., 2021)

Principalmente, se encontró un algoritmo evolutivo basado en métodos especiales de codificación y decodificación de unidades (Fefelova et al., 2020). Con la finalidad de

aumentar la velocidad de los cálculos computacionales, se generó un híbrido de funciones existentes junto con una función propia de afinidad que pudo reducir la cantidad de confirmaciones incorrectas (Fefelova et al., 2020).

Por otro lado, otros autores mencionan que, en esta predicción de estructuras 3D, usualmente se ignora que un conjunto de conformadores en equilibrio se encuentran en el estado primitivo de las proteínas (de Lima Salgado et al., 2017). Por ello, se realizó una recolección de modelos de acceso abierto y sus estructuras finales correspondientes que hayan sido resueltas experimentalmente (de Lima Salgado et al., 2017). Con ello, se pudo estudiar como estos describen la diversidad conformacional de una proteína (de Lima Salgado et al., 2017). Como resultado, se muestra que un 70% de las proteínas del conjunto son estructuralmente cercanas a diferentes conformadores del mismo modelo 3D curado experimentalmente. Con ello, se concluyó que el modelamiento de la estructura de proteínas resulta útil para la identificación de miembros del ensamble nativo (de Lima Salgado et al., 2017). Además, se resalta la importancia de considerar la diversidad conformacional en la evaluación de estructuras 3D de proteínas (de Lima Salgado et al., 2017).

Asimismo, se encontró que el alineamiento múltiple de secuencias es usado para la predicción de estructuras secundarias y terciarías de proteínas (Shi & Zhang, 2020). En la actualidad, los algoritmos de alineamiento múltiple de secuencias son complejos, redundantes y difíciles de entender, lo que resulta en que los investigadores no puedan seleccionar el algoritmo apropiado para evitar errores computacionales (Shi & Zhang, 2020). Mediante un profundo análisis respecto a los algoritmos heurísticos de alineamiento múltiple de secuencias (HMSAA, por sus siglas en inglés), se ha logrado desarrollar un modelo interactivo orientado a funcionalidades de HMSAA (Shi & Zhang, 2020). Este está acorde al método de programación generativo (Shi & Zhang, 2020).

Del mismo modo, se conoce que la predicción de la clasificación de las secciones transmembranas *β-barrel* de una proteína de acuerdo a la secuencia de aminoácidos resulta de suma importancia para el modelamiento de la estructura tridimensional y el análisis de funcionalidades (Chen et al., 2017). Por ello, se encontró el desarrollo de un predictor de estas secciones transmembranas de proteínas basado en *Support Vector Machine* (SVM) (Chen et al., 2017). Este puede proveer de mejoras válidas para la predicción de estructuras 3D de proteínas transmembranas y su respectivo análisis de funciones, puesto que cuenta con una precisión de 88.36% (Chen et al., 2017),

Finalmente, se encontró que el diseño de proteínas es el procedimiento inverso de predicción de la estructura 3D (Takahashi et al., 2021). Así, se puede inferir la relación de la estructura terciaria y la secuencia de aminoácidos (Takahashi et al., 2021). Para esto, se requieren dos ciclos de procesamiento (Takahashi et al., 2021). Un primer ciclo para detectar los cambios de la secuencia de aminoácidos y otra secuencia para la búsqueda conformacional exhaustiva para cada secuencia de aminoácidos (Takahashi et al., 2021).

Dentro de los proyectos realizados en el grupo de investigación de Bioinformática, se desarrolló un algoritmo capaz de explotar los datos existentes de estructuras primarias de proteínas en general para predecir estructuras terciarias de proteínas repetidas (Palomino & Hirsh, 2021). Para esto, se adaptó el algoritmo DMPfold (Greener et al., 2019) el cual fue implementado bajo una perspectiva general aplicable a todas las proteínas existentes (Palomino & Hirsh, 2021). El proceso de adaptación del algoritmo se fundamenta primordialmente en la afinidad de las funciones y características entre las proteínas repetidas que comparten la misma familia (Palomino & Hirsh, 2021).

Con todo lo presentado, se concluye que se han desarrollado algoritmos para la predicción de estructura terciaria según la secuencia de aminoácidos de una proteína (Fefelova et al., 2020; Palomino & Hirsh, 2021). Estos son algoritmos computacionales basados en modelos de aprendizaje de máquina (Fefelova et al., 2020; Palomino & Hirsh, 2021). Asimismo, se han realizado investigaciones relacionadas a esta predicción con enfoque al alineamiento múltiple de secuencias, la predicción de clasificación de secciones transmembrana, la diversidad conformacional y el diseño de proteínas (Chen et al., 2017; de Lima Salgado et al., 2017; Shi & Zhang, 2020; Takahashi et al., 2021).

### 3.6.3 Respuesta a pregunta ¿En qué consiste y cuáles son los métodos o herramientas para la curación de unidades de repetición de proteínas repetidas?

El creciente interés en las proteínas repetidas se debe a los pliegues estructurales, la alta conservación evolutiva y un repertorio de funciones de este tipo de proteínas (Chakrabarty & Parekh, 2022). La curación de unidades de repetición de proteínas repetidas no se encuentra definida explícitamente en las publicaciones revisadas. Sin embargo, se puede describir como la anotación manual que se realiza respecto a la clasificación, función y evolución de estas unidades de repetición (Bliven et al., 2019; Chakrabarty & Parekh, 2022; Hirsh et al., 2016). A continuación, se describen los resultados de investigaciones relacionadas a la curación de unidades de repetición de proteínas repetidas.

Se encontró la herramienta CE-Symm, utilizada para la detección sistemática de simetrías internas y repeticiones estructurales en proteínas (Bliven et al., 2019). Esta permite reportar los tipos de simetrías internas, identificar la menor unidad de repetición, describir el arreglo de repeticiones mediante operaciones transformacionales y sus ejes de simetría (Bliven et al., 2019). Asimismo, permite comparar la similitud de todas las repeticiones en el nivel residual (Bliven et al., 2019). Se desarrolló CE-Symm 2.0, para incorporar la curación de unidades de repetición mediante la clasificación, anotación de funciones y el análisis evolutivo (Bliven et al., 2019).

De la misma manera, se encontró la base de datos de repeticiones estructurales en proteínas DbStRiPs (Chakrabarty & Parekh, 2022). Mediante una curación manual de la clasificación Kajava de proteínas (Kajava, 2001, 2012), se desarrolló una base de datos de repeticiones estructurales de proteínas usando una conexión basada en redes convolucionales (Chakrabarty & Parekh, 2022). Una característica importante de esta base de datos es la disponibilidad de la información existente de familias de secuencias repetidas que se encuentran mediante una búsqueda en Pfam (El-Gebali et al., 2019), una base de datos de familias de proteínas (Chakrabarty & Parekh, 2022). Adicionalmente, se realizó un análisis de registros disponibles en PDB (Base de Datos de Proteínas, por sus siglas en inglés) (Berman et al., 2000), con 16472 anotaciones de repeticiones en 15141 cadenas de proteínas (Chakrabarty & Parekh, 2022). Además, se encontré una familia de proteínas repetidas sin clasificar, a la que se le llamó *left-handed beta helix,* y 33 clústeres de proteínas repetidas (Chakrabarty & Parekh, 2022).

Asimismo, con el desarrollo de ReUPred (Hirsh et al., 2016) se describió que la anotación de proteínas repetidas puede realizarse mediante búsquedas estructurales basadas en fragmentos de estructuras a modo de plantillas (Hirsh et al., 2016). En este trabajo también se demostró que esta anotación puede realizarse a gran escala sobre entradas de RepeatsDB (Di Domenico et al., 2014) sin caracterizar, descubriendo nuevos escenarios para el análisis del universo de proteínas repetidas.

Dentro de los proyectos realizados en el grupo de investigación de Bioinformática de la PUCP se encuentra una optimización para la clasificación de proteínas repetidas e identificación de unidades de repetición mediante el uso de perfiles de Modelos Ocultos de Markov (Pedraza & Hirsh, 2019) que es utilizada en una herramienta para la curación de familias de proteínas repetidas (Urbina & Hirsh, 2021).

El algoritmo optimizado se trata de ReUPred (Hirsh et al., 2016, 2017), al que se le incorporó un modelo clasificador de modelos ocultos de Markov de manera que inicialmente realice una clasificación de la secuencia que recibe (Pedraza & Hirsh, 2019). Este considera 86 familias de proteínas y cuenta con una precisión de 64% y una sensibilidad de 72%, en promedio (Pedraza & Hirsh, 2019).

El algoritmo optimizado fue incorporado una herramienta que proporciona funcionalidades básicas, textuales y gráficas para realizar más eficientemente el proceso de curación de familias de proteínas repetidas (Urbina & Hirsh, 2021). Con esta herramienta se puede controlar el tiempo invertido por los usuarios durante el proceso de curación pues se reduce el margen de error y se automatizan tareas que suelen ser manuales en el proceso clásico de curación (Urbina & Hirsh, 2021). Además, toma en consideración lineamientos de usabilidad en servicios web bioinformáticos (Bezerra Brandao et al., 2021) para mejorar la facilidad de uso y entendimiento de la herramienta (Urbina & Hirsh, 2021).

Estos proyectos, a pesar de estar orientados a familias de proteínas repetidas, son útiles para usar como base en el proceso de curación de unidades de repetición. Por esto y todo lo escrito anteriormente, se concluye que existen investigaciones que desarrollan herramientas que facilitan tareas de curación de unidades de repetición en proteínas como CE-Symm 2.0, DbStRiPs y ReUPred (Bliven et al., 2019; Chakrabarty & Parekh, 2022; Hirsh et al., 2016).

## 3.7 Conclusiones

Se observa que existe interés en estudiar proteínas repetidas debido a sus peculiares características y múltiples aplicaciones en biotecnología (Chakrabarty & Parekh, 2022). Una actividad necesaria para continuar incrementando el conocimiento existente sobre este tipo de proteínas es la curación de unidades de repetición (Hirsh et al., 2016, 2017).

La curación de proteínas repetidas requiere de diversos métodos y herramientas que puedan facilitar las tareas que permitan determinar aspectos específicos de cada proteína (Hirsh et al., 2017). Por ejemplo, se requieren servicios que permiten predecir la clasificación de una proteína repetida (Muroya Tokushima & Hirsh, 2022; Tenorio Ku & Hirsh, 2021). De la misma manera, se necesitan servicios que permitan obtener la estructura tridimensional de una proteína (Berman et al., 2000) o predecirla, en caso no sea conocida (Palomino & Hirsh, 2021). Asimismo, se precisa contar con métodos de aproximación para la identificación de unidades de repetición (Hirsh et al., 2016; Pedraza

& Hirsh, 2019) con la finalidad de que el curador pueda discernir unidades y regiones de repetición y realizar anotación respecto a la familia y funciones de una proteína repetida (Hirsh et al., 2017).

En esta revisión de literatura primaria, se han encontrado diversas herramientas que permiten realizar estas tareas (Chakrabarty & Parekh, 2020, 2022; Fefelova et al., 2020; Hirsh et al., 2016; Palopoli et al., 2016). Asimismo, se han observado diversos resultados del grupo de investigación de Bioinformática de la PUCP que buscan solucionar estas necesidades (Hirsh et al., 2017; Muroya Tokushima & Hirsh, 2022; Palomino & Hirsh, 2021; Pedraza & Hirsh, 2019; Tenorio Ku & Hirsh, 2021). Sin embargo, no se ha encontrado una propuesta de solución que integre todas estas funcionalidades para la curación de proteínas repetidas.

Por ello, se concluye que resulta fundamental realizar un proyecto que busque desarrollar un servicio que integre las necesidades de los investigadores para la curación de proteínas repetidas, tal como se realizó con la curación de familias de proteínas repetidas (Urbina & Hirsh, 2021). Para esto, es importante considerar lineamientos de usabilidad en servicios web bioinformáticos para asegurar la facilidad de uso en el desarrollo de una herramienta que busque solucionar esta problemática (Bezerra Brandao et al., 2021).

# Capítulo 4.Integración de servicios de predicción de clasificación de proteínas repetidas

## 4.1 Introducción

En este capítulo se presentan los resultados alcanzados para el primer objetivo específico de este proyecto de tesis. Como primer problema a atacar en este trabajo de investigación se encuentra la falta de integración de servicios de predicción de clasificación de proteínas repetidas en el proceso de curación. Por ello, se ha plantea que el primer objetivo específico a lograr sea integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura primaria y/o terciaria.

## 4.2 Resultados alcanzados

Con el fin de lograr el primer objetivo específico de este proyecto se plantearon una serie de resultados esperados a conseguir mediante una serie de actividades. Esta se encuentra detallada en el anexo B del presento documente titulado Plan de proyecto. A continuación, se procede a describir cada uno de estos resultados.

### 4.2.1 Descripción de un servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria

El primer resultado alcanzado en este proyecto de tesis es la revisión de un servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria. Para esto, se decidió trabajar con el servicio TRNET-lite (Tenorio Ku & Hirsh, 2021) desarrollando dentro del marco de trabajo del grupo de investigación en Bioinformática.

Esta aplicación considera todas las clases y subclases de proteínas repetidas a excepción de las clases I y II debido a falta de información (Tenorio Ku & Hirsh, 2021). Asimismo, trabaja con la información estructural disponible dentro del servicio RCSB PDB (Berman et al., 2000). En este, se puede obtener la estructura de una proteína en un archivo de formato PDB, el cual contiene principalmente anotaciones acerca de los métodos que se utilizaron para obtener los datos moleculares e información de cada molécula revisada de manera organizada (Berman et al., 2000; Tenorio Ku & Hirsh, 2021). Además, incluye las coordenadas de cada átomo dentro de su estructura tridimensional, la información que resulta más relevante para el servicio TRNET-lite (Berman et al., 2000; Tenorio Ku & Hirsh, 2021).

Para el procesamiento, se definieron representaciones de datos que reduce el número de dimensiones considerando 5 grupos de aminoácidos (Tenorio Ku & Hirsh, 2021). Luego, se desarrolló un modelo de red neuronal que pueda identificar los tipos de proteínas repetidas de las 26 existentes a partir de las representaciones definidas (Tenorio Ku & Hirsh, 2021). Este modelo se enfocó en la simplificación del problema en diversas clasificaciones binarias ordenadas por relevancia (Tenorio Ku & Hirsh, 2021). Bajo el framework Tensorflow, se realizaron 3 propuestas de las cuales tuvieron una validación cruzada utilizando las clases III.3, IV.1, IV.2 y IV.4 (Tenorio Ku & Hirsh, 2021). Con una media mayor a 98.6% de métrica AUC (Área bajo la curva ROC, por sus siglas en inglés)

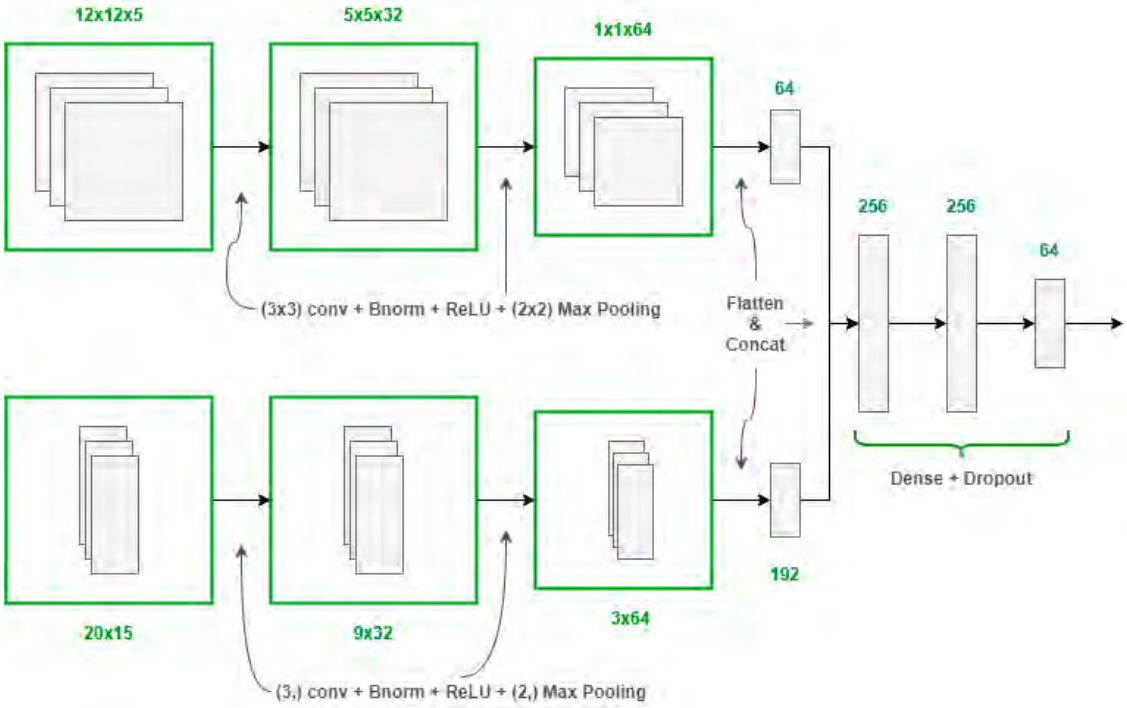



*Figura 2 "arq7": Arquitectura planteada que utiliza que utiliza el diagrama de Ramachandran y los histogramas de distancia (Tenorio Ku & Hirsh, 2021).*

en cada clase, se seleccionó "arq7", una arquitectura que utiliza el diagrama de Ramachandran y los histogramas de distancia (Tenorio Ku & Hirsh, 2021). Esta puede ser observada en la figura 2.

Con el modelo desarrollado, se preparó un servicio web que permite obtener la probabilidad de clasificación de una proteína en los grupos III, IV y V y sus subclases (Tenorio Ku & Hirsh, 2021). A continuación, se presenta el modelo de datos del servicio TRNET-lite. En este, se detallan los datos de entrada y de salida del servicio.

```
MODELO DE DATOS TRNET-lite {
    ENTRADA {
        "idPDB":String
            /*Es una cadena de cuatro caracteres que es utilizado como
            identificador de una proteína en la base de datos RCSB PDB.
            Con este identificador, se obtiene el archivo que contiene
            la estructura de la proteína para realizar la predicción. */
    }

    SALIDA {
        "chainList":List
            /*Es una lista con los valores de predicción de clase para
            cada cadena de la proteína evaluada*/
            [
                { "chainID":Char
                    /*Carácter identificador de la cadena
                    evaluada*/
                "III_1":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase III.1*/
                "III_2":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase III.2*/
                "III_3":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase III.3*/
                "III_4":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase III.4*/
                "III_5":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase III.5*/
                "III_6":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase III.6*/
                "IV_1":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase IV.1*/
                "IV_2":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase IV.2*/
                "IV_3":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase IV.1*/
                "IV_4":Double
                    /*Valor de probabilidad de clasificación de la
                    cadena para la clase IV.4*/
```

```
                    "IV_5":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase IV.5*/
                    "IV_6":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase IV.6*/
                    "IV_7":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase IV.7*/
                    "IV_8":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase IV.8*/
                    "IV_9":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase IV.9*/
                    "IV_10":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase IV.10*/
                    "V_1":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase V.1*/
                    "V_2":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase V.2*/
                    "V_3":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase V.3*/
                    "V_4":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase V.4*/
                    "V_5":Double
                        /*Valor de probabilidad de clasificación de la
                        cadena para la clase V.5*/
                    }
                    ...
                ]
        }
}
```

Como se observa en el modelo de datos, el servicio recibe el identificador PDB de la proteína a evaluar. Con este, obtiene el archivo con la estructura y lo procesa por cada cadena. Con ello, para cada cadena, se obtienen los porcentajes de probabilidad de clasificación para cada clase y subclase de proteína repetida. Así, el investigador puede observar y decidir de que manera trabajar la curación de unidades de repetición, dependiendo de la clase.

Este resultado fue verificado mediante la evaluación cualitativa de juicio experto. Se elaboró un acta de aceptación con una experta en el área de Bioinformática la cual puede ser revisada en el anexo D.

### 4.2.2 Integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria

El segundo resultado alcanzado de este proyecto es el despliegue del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria y su integración. Para esto, se realizó la comunicación con los respectivos autores para obtener el código fuente del servicio. Se encontró que la arquitectura en la que el servicio fue desplegado era de tipo *serverless* mediante Amazon Web Services. No obstante, el modelo de aprendizaje de máquina para la predicción de clase y subclase de la proteína se encontraba en un Docker mientras que las funciones *serverless* eran utilizadas para manejar las solicitudes de procesamiento. En la figura 3 se puede observar la arquitectura que se desarrolló para el servicio.

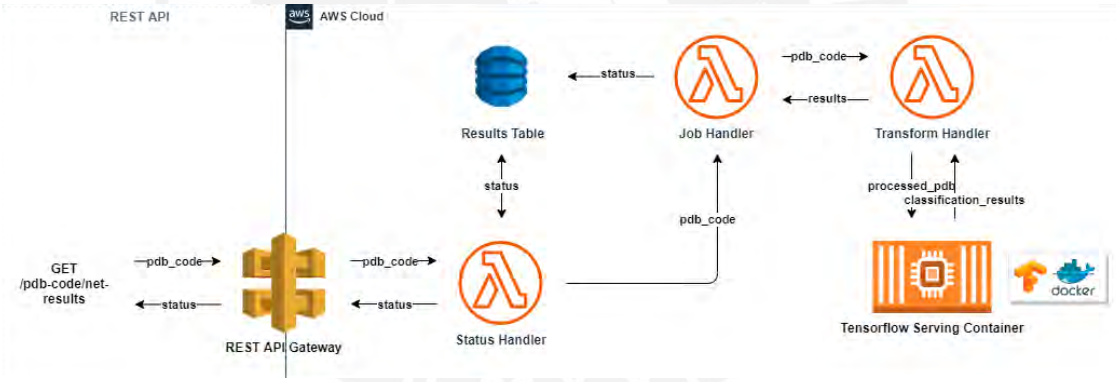



*Figura 3 Arquitectura serverless desarrollada para el servcio TRNET-lite (Tenorio Ku & Hirsh, 2021)*

Por ello, se realizaron los siguientes ajustes al código fuente antes de su despliegue. Un punto importante a modificar del servicio fueron los datos de entrada. El servicio se desarrolló para trabajar únicamente con la base de datos de estructuras de proteínas PDB (Berman et al., 2000). Por ello, el algoritmo utilizaba un identificador de esta base de datos para obtener la información de la estructura solicitada y realizar la predicción de clase y subclase de proteína. Debido a las necesidades de este proyecto, se modificó el algoritmo para que pueda funcionar recibiendo una estructura de proteína en formato CIF (*Crystallographic Information File*, por sus siglas). Esto se realizó debido a que se plantea integrar el servicio de predicción de estructura tridimensional de proteína en base a su secuencia con este servicio. Por otro lado, de modificó la arquitectura *serverless* del

servicio a una arquitectura orientada a servicios. Para esto, se utilizó el modelo entrenado exportado del proyecto original. Con este, aplicando las funciones de preprocesamiento de archivos de entrada, se creo un servicio que permite recibir un archivo de estructura tridimensional en formato CIF para obtener las probabilidades de clasificación de cada subclase de proteína repetida. Además, se tomó en consideración reutilizar la carga del modelo para procesar múltiples estructuras.

Como caso de prueba, se tiene como dato de entrada un archivo en formato CIF. Como salida, se obtiene un porcentaje de probabilidad (entre 0 y 1) para las clases III.1, III.2, III.3, III.4, III.5, III.6, IV.1, IV.2, IV.3, IV.4, IV.5, IV.6, IV.7, IV.8, IV.9, IV.10, V.1, V.2, V.3, V.4 y V.5.

Finalmente, una vez desplegado el servicio web TRNET-lite (Tenorio Ku & Hirsh, 2021) modificado, se procedió a ejecutar los casos de prueba, obteniendo un 100% de éxito. Esto cumple con el indicador objetivamente verificable esperado para el resultado. El código fuente del servicio desarrollado puede obtenerse en el Anexo F del presente documento.

### 4.3 Discusión

Para poder resolver la falta de una herramienta integrada para la curación de proteínas repetidas que considere las actividades más relevantes de esta tarea, primero fue necesario identificar, describir, desplegar e integrar servicios de predicción de clase y subclase de proteínas repetidas.

Para esto, se trabajó con el servicio TRNET-lite (Tenorio Ku & Hirsh, 2021). Este servicio permite predecir la clase y subclase de proteína según la estructura tridimensional. Se realizó la descripción del servicio, así como el modelo de datos que utiliza.

Para que el servicio esté adecuadamente integrado al objetivo general del proyecto, se realizó una modificación para permitir que el procesamiento se realice con una estructura terciaria de proteína en formato CIF. Con ello, el servicio fue desplegado para ser utilizado en el proyecto.

Con los resultados obtenidos, que cumplen con los indicadores objetivamente verificables, se afirma que se logró integrar un servicio de predicción de clasificación de proteínas repetidas en base a su estructura terciaria.

# Capítulo 5. Incorporación de servicios para predicción de la estructura de una proteína

## 5.1 Introducción

En este capítulo se presentan los resultados alcanzados para el segundo objetivo específico de este proyecto de tesis. Como segundo problema a atacar en este trabajo de investigación se encuentra la falta de integración de servicios de predicción de estructuras de proteínas en el proceso de curación de unidades de repetición. Por ello, se ha plantea que el segundo objetivo específico a lograr sea integrar servicios de predicción de estructura de proteínas en base a su secuencia.

## 5.2 Resultados alcanzados

Con el fin de lograr el segundo objetivo específico de este proyecto se plantearon una serie de resultados esperados a conseguir mediante una serie de actividades. Esta se encuentra detallada en el anexo B del presento documente titulado Plan de proyecto. A continuación, se procede a describir cada uno de estos resultados.

### 5.2.1 Descripción de un servicio de predicción de estructura de proteína en base a su secuencia

El tercer resultado alcanzado en este proyecto de tesis es la revisión de un servicio de predicción de estructura de proteína basada en su secuencia. Para esto, primero se decidió trabajar con el servicio DeepReSPred (Palomino & Hirsh, 2021) desarrollando dentro del marco de trabajo del grupo de investigación en Bioinformática.

Para este servicio, se realizó una adaptación del algoritmo DMPfold (Greener et al., 2019) considerando una característica particular de las proteínas repetidas, los patrones de repetición dentro de una misma familia (Kajava, 2012; Palomino & Hirsh, 2021). Este algoritmos es una continuación del método de predicción de contacto conocido como DeepMetaPSICOV (Greener et al., 2019; Palomino & Hirsh, 2021). Se enfoca principalmente en la predicción de límites de distancia interatómica, ángulos de torsión y enlaces de hidrógeno de la cadena principal a través de redes neuronales profundas (Greener et al., 2019; Palomino & Hirsh, 2021).

No obstante, se han observado recientes avances relevantes en el uso de modelos de redes neuronales para la predicción de la estructura tridimensional de una proteína en base a su secuencia. Un gran ejemplo de esto es el algoritmo AlphaFold (Jumper et al.,

2021; Tunyasuvunakool et al., 2021; Varadi et al., 2022). Este algoritmo permite realizar predicciones más precisas que las realizadas por DeepResPred sin la necesidad de contar con una familia de proteínas conocidas para evaluar los resultados. Además, actualmente se cuenta con la cuarta versión de estructuras predichas en base a toda la base de conocimiento de secuencias de proteína UniProt (The UniProt Consortium, 2019; Varadi et al., 2022).

Por este motivo, se decide trabajar con la base de datos de estructuras terciarias de AlphaFold (Varadi et al., 2022). A continuación, se describe de datos para utilizar este servicio. En este se detallan los datos de entrada y de salida del servicio.

```
MODELO DE DATOS AlphaFold Database {
      ENTRADA {
            "UniProtID":String
                  /*Es una cadena de caracteres que representa el
                  identificador en la base de conocimientos UniProt (The
                  UniProt Consortium, 2019). */
      }

      SALIDA {
            "predictedStructurePDBFile":File
                  /*Es el archivo de estructura tridimensional predicha por el
                  modelo AlphaFold en formato .pdb*/
            "predictedStructureCIFFile":File
                  /*Es el archivo de estructura tridimensional predicha por el
                  modelo AlphaFold en formato .cif*/
            "predictedErrorsFile":File
                  /*Es un archivo que contiene el error de predicción para
                  cada parte de la estructura predicha de proteína*/
      }
}
```

Como se pude observar, el componente solo requiere del identificador UniProt para obtener los archivos estructurales predichos en formato .pdb y .cif. Este resultado fue verificado mediante la evaluación cualitativa de juicio experto. Se elaboró un acta de aceptación con una experta en el área de Bioinformática la cual puede ser revisada en el anexo D.

### 5.2.2 Integración del servicio de predicción de estructura de proteína en base a su secuencia

El cuarto resultado alcanzado de este proyecto es el despliegue del servicio de predicción de estructura de proteína en base a su secuencia y su integración. Para esto, se realizó la

comunicación con los respectivos autores para obtener el código fuente del servicio. Se encontró que la arquitectura en la que el servicio fue desplegado era orientado a servicios mediante Amazon Web Services. Asimismo, se encontró que se utilizaba un servicio de almacenamiento para los elementos estáticos de la interfaz y un base de datos para almacenar las solicitudes y resultados de cada procesamiento. En la figura 4 se puede observar la arquitectura que se desarrolló para el servicio.

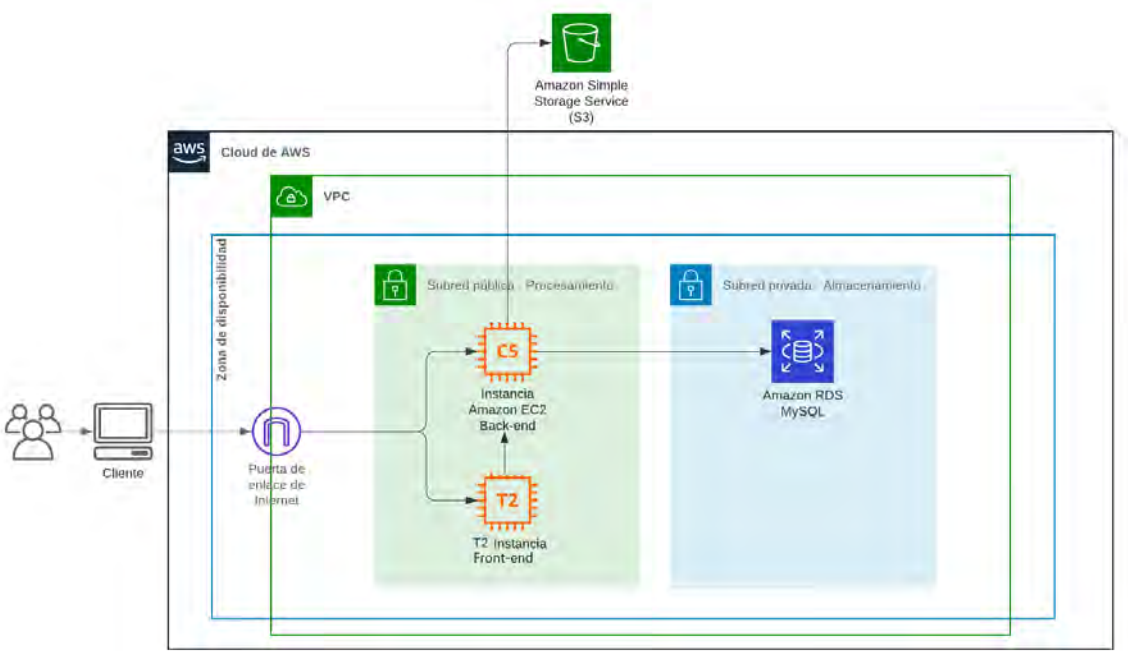



*Figura 4 Diagrama de arquitectura en la nube Amazon Web Services de DeepReSPred (Palomino & Hirsh, 2021)*

No obstante, debido a los recientes avances en el uso de redes neuronales para la predicción de estructura tridimensional de proteína en base a su secuencia, se procedió a crear un servicio que permita acceder a los archivos de la base de datos AlphaFold (Varadi et al., 2022). Esto permitirá que el servicio integrado pueda trabajar tanto con estructuras conocidas de PDB (Berman et al., 2000) como estructuras predichas en base a las secuencias de UniProt (The UniProt Consortium, 2019).

Para esto, se desarrolló un componente que se conecta a ambas bases de datos y permite descargar los archivos de estructuras terciarias de una proteína en base a su identificador. Si es una proteína con estructura conocida, el componente se conecta a PDB según su *accession identifier*. Si se busca una predicción de estructura de proteína, el componente se conecta a la base de datos de AlphaFold según el identificador UniProt.

Como caso de prueba, se tiene como dato de entrada un identificador PDB o UniProt. Como salida, se obtienen los archivos de estructura terciaria en formado PDB y CIF. En el caso de predicción, también se obtiene el archivo de errores.

Finalmente, una vez desplegado el servicio web para la conexión con las bases de datos PDB y AlphaFold (Berman et al., 2000; Varadi et al., 2022), se procedió a ejecutar los casos de prueba, obteniendo un 100% de éxito. Esto cumple con el indicador objetivamente verificable esperado para el resultado. El código fuente del componente desarrollado puede obtenerse en el Anexo F del presente documento.

## 5.3 Discusión

Para poder resolver la falta de una herramienta integrada para la curación de proteínas repetidas que considere las actividades más relevantes de esta tarea, también fue necesario identificar, describir, desplegar e integrar servicios de predicción estructura terciaria de una proteína en base a su secuencia.

Para esto, se primero se trabajó con el servicio DeepReSPred (Palomino & Hirsh, 2021). Este servicio permite realizar una predicción de estructura terciaria mediante la secuencia de una proteína. Se realizó la descripción del servicio, así como el modelo de datos que utiliza.

Por otro lado, según lo recientes avances en el uso de modelos de redes neuronales para la predicción de información estructural de proteínas (Jumper et al., 2021; Tunyasuvunakool et al., 2021; Varadi et al., 2022), se decidió implementar un componente que permita acceder a las bases de conocimiento correspondientes.

Para que el servicio esté integrado al objetivo general del proyecto, se desarrolló el componente que permite la conexión con las bases de datos PDB (Berman et al., 2000) y AlphaFold (Varadi et al., 2022) según su PDB id o UniProt id, respectivamente. Con ello, el servicio fue desplegado para ser utilizado en el proyecto.

Con los resultados obtenidos, que cumplen con los indicadores objetivamente verificables, se afirma que se logró incorporar un servicio de predicción de estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación.

# Capítulo 6.Servicio web integrado para la curación de proteínas repetidas

## 6.1 Introducción

En este capítulo se presentan los resultados alcanzados para el último objetivo específico de este proyecto de tesis. Como tercer problema a atacar en este trabajo de investigación se encuentra la falta de integración de servicios de identificación de unidades de repetición en el proceso de curación de proteínas repetidas. Así, se podrá resolver la falta de una herramienta integrada para la curación de proteínas repetidas que considere la predicción y clasificación como predicción de la estructura terciaria de una proteína. Por ello, se ha plantea que el tercer objetivo específico a lograr sea Implementar un servicio web para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición.

## 6.2 Resultados alcanzados

Con el fin de lograr el tercer objetivo específico de este proyecto se plantearon una serie de resultados esperados a conseguir mediante una serie de actividades. Esta se encuentra detallada en el anexo B del presento documente titulado Plan de proyecto. A continuación, se procede a describir cada uno de estos resultados.

### 6.2.1 Descripción de un servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición

El quinto resultado alcanzado en este proyecto de tesis es la revisión de un servicio de predicción de unidades de repetición de una proteína. Para esto, se decidió trabajar con el servicio RepeatsDB-lite (Hirsh et al., 2018) desarrollando dentro del marco de trabajo del grupo de investigación en Bioinformática.

Este servicio consiste en la predicción de elementos estructurales repetitivos y unidades de repetición en proteínas (Hirsh et al., 2018). Este servicio extiende la predicción de unidades de repetición a todas las clases y mejora el rendimiento de la predicción en términos de procesamiento computacional y precisión frente a otros métodos (Hirsh et al., 2018). Asimismo, tiene una precisión mayor al 95% para estructuras solenoides (Hirsh et al., 2018).

El algoritmo utilizado en RepeatsDB-lite es una evolución del método ReUPred (Hirsh et al., 2016, 2017, 2018). Utiliza una búsqueda estructural iterativa en una librería de

unidades de repetición para encontrar elementos repetitivos en estructuras de proteínas (Hirsh et al., 2018). Este método requiere de la estructura a procesar y el conjunto de unidades de repetición, que representa el espacio conformacional y la diversidad de repeticiones validadas por la comunidad científica (Hirsh et al., 2018). El algoritmo aprovecha la librería alineándola con la estructura objetivo utilizando un enfoque basado en los algoritmos “divide y vencerás” (Hirsh et al., 2018).

Una vez el algoritmo haya definido la mejor unidad en base a la similitud estructural, a la que se le llamará unidad maestra, se fija y el algoritmo se divide (Hirsh et al., 2018). Se procede a crear dos entradas nuevas, que corresponden a los fragmentos terminales N- y C- de la unidad predicha y se realizan dos nuevos ciclos de búsqueda estructural (Hirsh et al., 2018). Las búsquedas se realizan en una nueva librería de unidades creadas durante el procesamiento en base a la unidad maestra y todas las nuevas predicciones de unidades son incluidas en las búsquedas de los siguientes ciclos (Hirsh et al., 2018). Mediante este enfoque, la región de repetición se expande hasta que los nuevos fragmentos a procesar son demasiado cortos, por lo que se termina el procesamiento (Hirsh et al., 2018).

Las unidades predichas se recolectan y se evalúan en conjunto (Hirsh et al., 2018). En esta fase, los fragmentos incluidos en la región que fueron derivados de la unidad estructural original son anotados como inserciones (Hirsh et al., 2018). Al final, si la región tiene menos de tres unidades, la siguiente posible unidad de la librería es usada como unidad maestra por un máximo de cuatro repeticiones del algoritmo (Hirsh et al., 2018).

Comparado con ReUPred, el nuevo algoritmo descarta alineamientos estructurales en las que las unidades limites interfieren con los elementos de la estructura secundaria (Hirsh et al., 2018). Asimismo, el algoritmo ahora puede detectar múltiples regiones en la misma cadena (Hirsh et al., 2018). El tiempo de ejecución del algoritmo para una única cadena es de unos minutos pero depende de la clase de la unidad maestra (Hirsh et al., 2018).

Con el algoritmo modificado, se desarrolló una interfaz para la identificación de unidades de repetición en una proteína. En la figura 5 se puede apreciar una captura de la interfaz de RepeatsDB-lite y sus principales elementos.

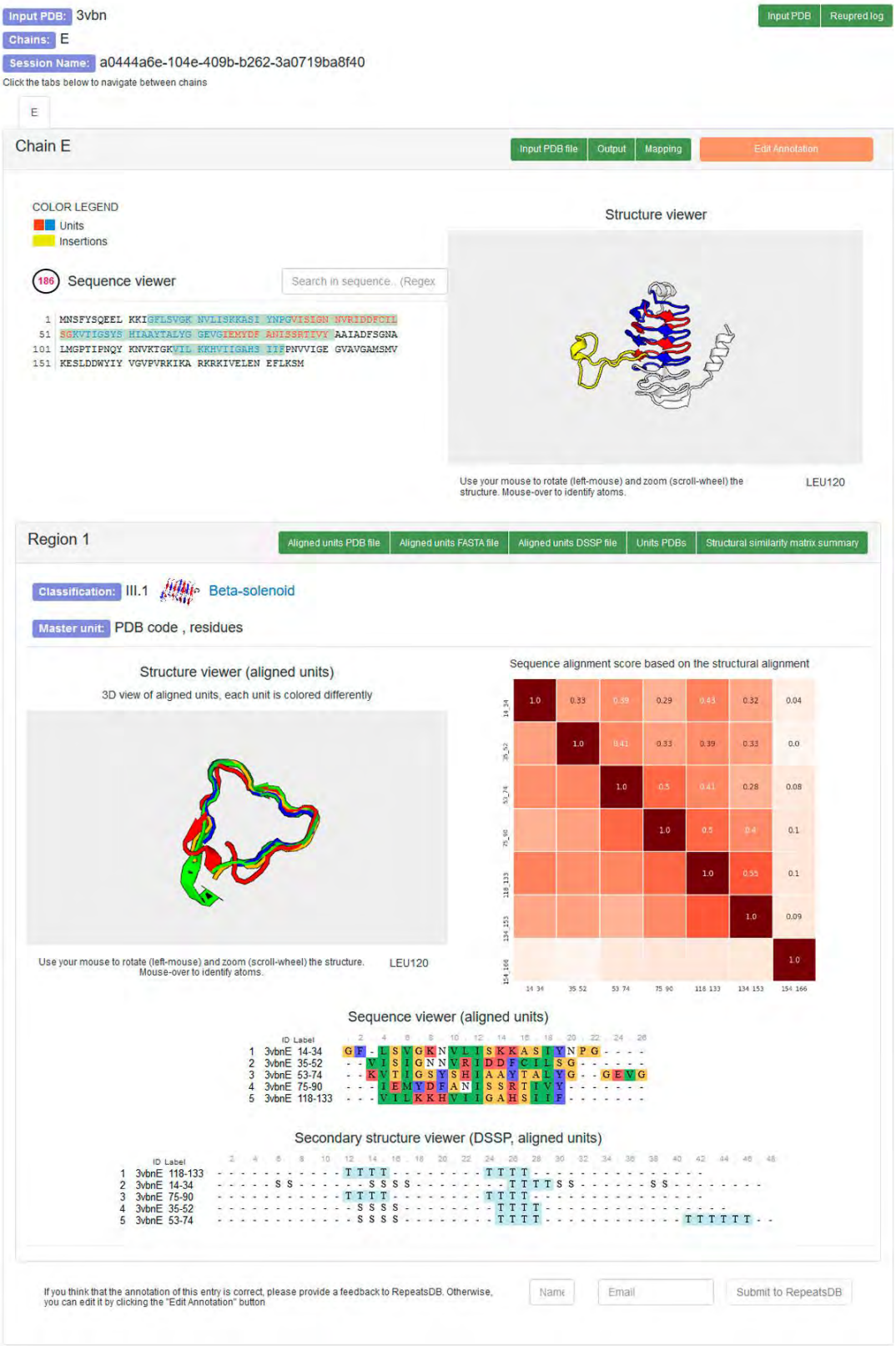



*Figura 5 Interfaz e RepeatsDB-lite. La cabecera de la interfaz presenta un resumen sobre el procesamiento. Se muestran pestañas para la navegación entre cadenas. En cada pestaña se muestra información general sobre la cadena y un resultado por cada región (Hirsh et al., 2018)*

A continuación, se describe el modelo de datos para del servicio RepeatsDB-lite. En este, se detallan los datos de entrada y de salida del servicio.

```
MODELO DE DATOS RepeatsDB-lite {
    ENTRADA {
        "pdbID":String
            /*Es una cadena de cuatro caracteres que es utilizado como
            identificador de una proteína en la base de datos RCSB PDB.
            Con este identificador, se obtiene el archivo que contiene
            la estructura de la proteína para realizar la predicción. */
        "chainID":Char
            /*Es un carácter que identifica a una cadena de la
            estructura a procesar. Este valor es enviado en caso solo se
            quiera procesar una cadena en particular */
        "pdbFile":File
            /*Es un archivo en formato .pdb que contiene una estructura
            de proteína para que sea procesada. Puede cargarse un
            archivo o utilizar el identificador PDB para el
            procesamiento. */
        "allChains":Bool
            /*Es un valor booleano que indica si se procesarán todas las
            cadenas de la estructura de la proteína ingresada o si solo
            se procesará una en particular. En el segundo caso, es
            necesario indicar el identificador de cadena. */
    }
    SALIDA {
        "inputPDB":String
            /*Es una cadena de cuatro caracteres que se ingresó como
            identificador de una proteína en la base de datos RCSB PDB.
            Con este identificador, se obtiene el archivo que contiene
            la estructura de la proteína para realizar la predicción. */
        "chains":List
            /*Es la lista de las cadenas procesadas y los resultados
            obtenidos para cada una de estas. */
            [
                { "chainID":Char
                    /*Es el identificador de la cadena procesada.
                    */
                "inputPDBChainFile":File
                    /*Es el archivo en formato .pdb con la
                    estructura de la cadena procesada. */
                "outputChainfile":File
                    /*Es el archivo en formato .db con las
                    anotaciones de las regiones de las unidades de
                    repetición identificadas. */
                "mappingChainFile":File
```

/*Es el archivo en formato .mapping que contiene la concordancia de posiciones de los aminoácidos entre la secuencia y estructura*/

**“regions”:***List*

/*Es la lista de resultados por regiones de repetición identificados en la cadena de la estructura de la proteína. */

[

{ **“classNumber”:***String*

/*Es el número de clase y subclase identificado en la región. */

**“className”:***String*

/*Es el nombre de clase y subclase identificado en la región. */

**“masterUnitId”:***String*

/*Es el identificador en PDB de la unidad maestra identificada para la región. */

**“masterUnitBeg”:***Integer*

/*Es la posición inicial en la secuencia de residuos donde inicia la unidad de repetición maestra. */

**“masterUnitEnd”:***Integer*

/*Es la posición final en la secuencia de residuos donde inicia la unidad de repetición maestra. */

**“alignedUnitsPDB”:***File*

/*Es el archivo en formato .pdb con la alineación de unidades de repetición en la estructura. */

**“alignedUnitsFASTA”:***File*

/*Es el archivo en formato .fasta con la alineación de unidades de repetición en la secuencia. */

**“alignedUnitsDSSP”:***File*

/*Es el archivo .dssp que contiene la representación de la estructura secundaria de cada unidad dentro de la región de repetición/

**“structuralSimilarityMatrix”:***File*

/*Es el archivo en formato .txt que contiene la matriz de similitud estructural de la región analizada para la cadena. */

}

...

```
                        ]
                    }
                    ...
                ]
        }
}
```


Este resultado fue verificado mediante la evaluación cualitativa de juicio experto. Se elaboró un acta de aceptación con una experta en el área de Bioinformática la cual puede ser revisada en el anexo D.

### 6.2.2 Integración del servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición

Tal como fue descrito el servicio RepeatsDB-lite (Hirsh et al., 2018), el algoritmo está diseñado para buscar una unidad de repetición maestra entre más de 25000 unidades. En sus inicios, esta muestra de unidades era más pequeña, por lo que la ejecución del algoritmo no tomaba mucho tiempo de procesamiento (Hirsh et al., 2018). No obstante, con la cantidad de unidades a revisar actualmente, se requiere realizar un ajuste al performance del algoritmo.

Para lograr esta mejora, se utiliza la predicción de clase y subclase de proteína repetida para comenzar la ejecución del algoritmo en las muestras más relevantes. Para esto, la búsqueda de unidad maestra se inicia en las muestras de unidades de repetición que pertenezcan a la clase y subclase obtenida previamente. Con este cambio, se logró mejorar el tiempo de procesamiento del algoritmo.

Finalmente, una vez desplegado el servicio web RepeatsDB-lite (Hirsh et al., 2018) modificado, se procedió a ejecutar los casos de prueba, obteniendo un 100% de éxito. Esto cumple con el indicador objetivamente verificable esperado para el resultado. El código fuente del servicio desarrollado puede obtenerse en el Anexo F del presente documento.

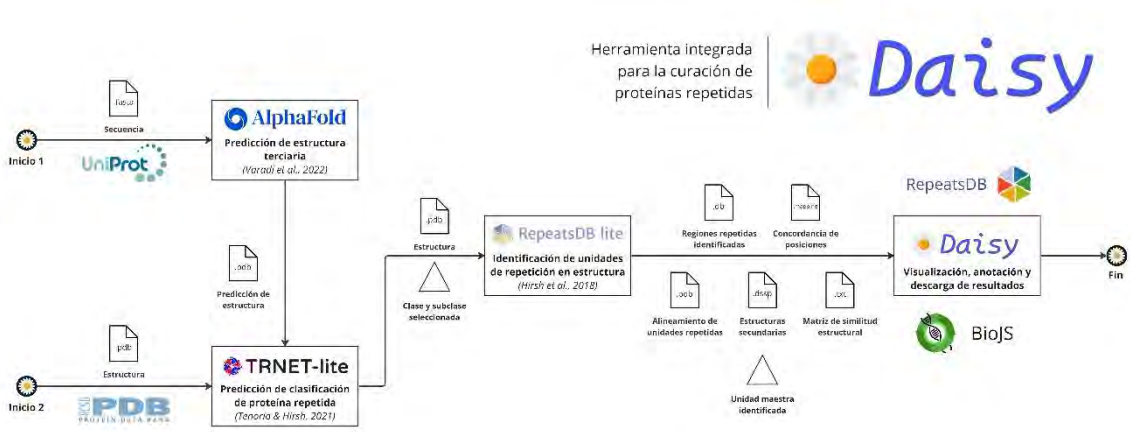
### 6.2.3 Daisy: Servicio web integrado para la curación de proteínas repetidas

El séptimo resultado alcanzado en este proyecto de tesis es desarrollo de un servicio web integrado para la curación de proteínas repetidas, considerando la predicción de clasificación y estructura terciaria, siguiendo lineamientos de usabilidad para servicios web bioinformáticos.

Este resultado integra todos los resultados anteriores y resuelve la problemática central de este proyecto de tesis. Por su extensión, se describirán los pasos y medios de verificación desarrollados en las siguientes subsecciones.

#### 6.2.3.1 Diagrama de flujo del proceso del servicio web integrado

Para el desarrollo del servicio integrado para la curación de proteínas repetidas se




definición un diagrama de flujo de información a través de los diversos servicios seleccionados y desplegados. A continuación, la figura 6 muestra el diagrama de flujo del proceso del servicio web integrado.

*Figura 6 Diagrama de flujo del proceso del servicio web integrado para la curación de proteínas repetidas*

El proceso del servicio integrado propuesto tiene dos inicios posibles. Si la proteína a evaluar ha sido trabajada previamente por la comunidad científica y presenta un registro validado de su estructura terciaria, se comenzará el proceso mediante la predicción de clasificación de proteína repetida. Por otro lado, cuando se requiere curar una proteína la cuál no cuenta con una estructura terciaria previamente validada por la comunidad científica. Para este caso, se trabajará con el identificador UniProt para utilizar la predicción de estructura tridimensional en la base de datos de AlphaFold (Varadi et al., 2022). Con esto, se obtendrá la estructura terciaría en formato PDB y CIF con el que se seguirá el proceso.

Continuando con el proceso, con la estructura terciaria de la proteína a evaluar, sea obtenida del Protein Data Bank (Berman et al., 2000) o haya sido generada mediante la el modelo AlphaFold (Jumper et al., 2021) previamente descrito, se realizará la predicción de clase y subclase mediante el modelo de aprendizaje de máquina del servicio integrado TRNET-lite (Tenorio Ku & Hirsh, 2021). Se obtendrá como resultado las probabilidades de

clasificación para las clases III, IV y V y sus subclases. Con estos valores, el usuario podrá determinar con que clasificación específica continuará con el proceso de curación.

La clase y subclase seleccionada permitirán ejecutar la optimización del algoritmo RepeatsDB-lite (Hirsh et al., 2018) el cual procesará la estructura evaluada para identificar sus unidades de repetición. Se obtendrán las regiones repetidas identificadas, la concordancia de posiciones entre las secuencias y las estructuras, la unidad maestra de repetición seleccionada, el alineamiento de las unidades de repetición, las estructuras secundarias de las regiones repetidas y una matriz de similitud estructural.

Luego de todo el procesamiento, el investigador podrá visualizar los resultados mediante la librería BioJS (Gómez et al., 2013). En está, podrá visualizar la estructura tridimensional de la proteína, así como el alineamiento estructural de las unidades repetidas identificadas. Además, se podrá observar la secuencia, el alineamiento de secuencia de las unidades repetidas, la matriz de similitud estructural y las estructuras secundarias de las unidades identificadas.

El usuario podrá realizar anotaciones y afinar la predicción de regiones de repetición de la proteína analizada. Asimismo, podrá descargar todos los resultados. Finalmente, el usuario podrá presentar su curación a RepeatsDB (Di Domenico et al., 2014) para que se actualice el registro de la proteína repetida identificada y evaluada.

#### 6.2.3.2 Lista de requerimientos funcionales y no funcionales del servicio web integrado

Para el diseño de este servicio web integrado para la curación de proteínas repetidas, se plantearon requerimientos funcionales y no funcionales. En la tabla x, se detalla cada uno de estos requerimientos.

*Tabla 9 Catálogo de requerimientos funcionales y no funcionales del servicio web integrado para la curación de proteínas repetidas.*

| # | Requerimiento | Tipo |
|---|---|---|
| 1 | El servicio podrá recibir una identificador UniProt para obtener la predicción de estructura desde la base de datos AlphaFold. | Funcional |
| 2 | El servicio obtendrá el resultado de la predicción de estructura terciaria en función a una estructura primaria en formato PDB y CIF. | Funcional |
| 3 | El servicio podrá recibir un identificador PDB, un archivo o texto en formato CIF que contenga la estructura de una proteína para predecir la clase y subclase de repetición utilizando el modelo | Funcional |

| | | |
|---|---|---|
| | del servicio TRNET-lite, considerando las clases III, IV y V. | |
| 4 | El servicio podrá procesar la estructura terciaria predicha para la predicción de clase y subclase de proteína repetida utilizando el modelo del servicio TRNET-lite, considerando las clases III, IV y V. | Funcional |
| 5 | El servicio utilizará la estructura terciaria procesada junto a la clase y subclase predicha seleccionada para la identificación de unidades de repetición en la estructura utilizando un algoritmo basado en el servicio RepeatsDB-lite. | Funcional |
| 6 | El servicio obtendrá las regiones repetidas identificadas en formato DB, la concordancia de posiciones entre las secuencias y las estructuras en formato MAPPING, la unidad maestra de repetición seleccionada, el alineamiento de las unidades de repetición en formato PDB, las estructuras secundarias de las regiones repetidas en formato DSSP y una matriz de similitud estructural de una estructura terciaria de proteína repetida. | Funcional |
| 7 | El servicio permitirá visualizar la estructura tridimensional de la proteína, así como el alineamiento estructural de las unidades repetidas identificadas. Además, se podrá observar la secuencia, el alineamiento de secuencia de las unidades repetidas, la matriz de similitud estructural y las estructuras secundarias de las unidades identificadas. | Funcional |
| 8 | El servicio permitirá realizar anotaciones y afinar la predicción de regiones de repetición de la proteína analizada. | Funcional |
| 9 | El servicio permitirá descargar los resultados obtenidos durante el procesamiento. | Funcional |
| 10 | El servicio generará un registro de procesamiento para que el usuario pueda revisar el estado y resultados de su solicitud. | Funcional |
| 11 | El servicio notificará al usuario por correo electrónico cuando una etapa del procesamiento haya finalizado. | Funcional |
| 12 | El servicio permitirá que el usuario envíe sus anotaciones curadas de la proteína repetida analizada al servicio RepeatsDB | Funcional |
| 13 | El servicio seguirá los lineamientos de usabilidad para servicios web bioinformáticos para el diseño de interfaz | No funcional |
| 14 | El servicio deberá obtener un resultado mayor a 55 puntos en una evaluación de usabilidad utilizando el cuestionario SUS-BWS | No funcional |
| 15 | El componente *backend* del servicio será desarrollado en el lenguaje Python | No funcional |
| 16 | El componente *frontend* del servicio será desarrollado en el lenguaje Javascript | No funcional |

#### 6.2.3.3 Mockups navegables con el diseño de la interfaz de usuario del servicio web integrado

El diseño de la interfaz del servicio web integrado para la curación de proteínas repetidas se realizó en la plataforma Figma (Design, 2017), siguiendo lineamientos de usabilidad para servicios web bioinformáticos (Bezerra Brandao et al., 2021). El proyecto se puede revisar en el anexo E del presento documento. A continuación, las figuras del 7 al 13 presentan las pantallas más relevantes diseñadas para el servicio.

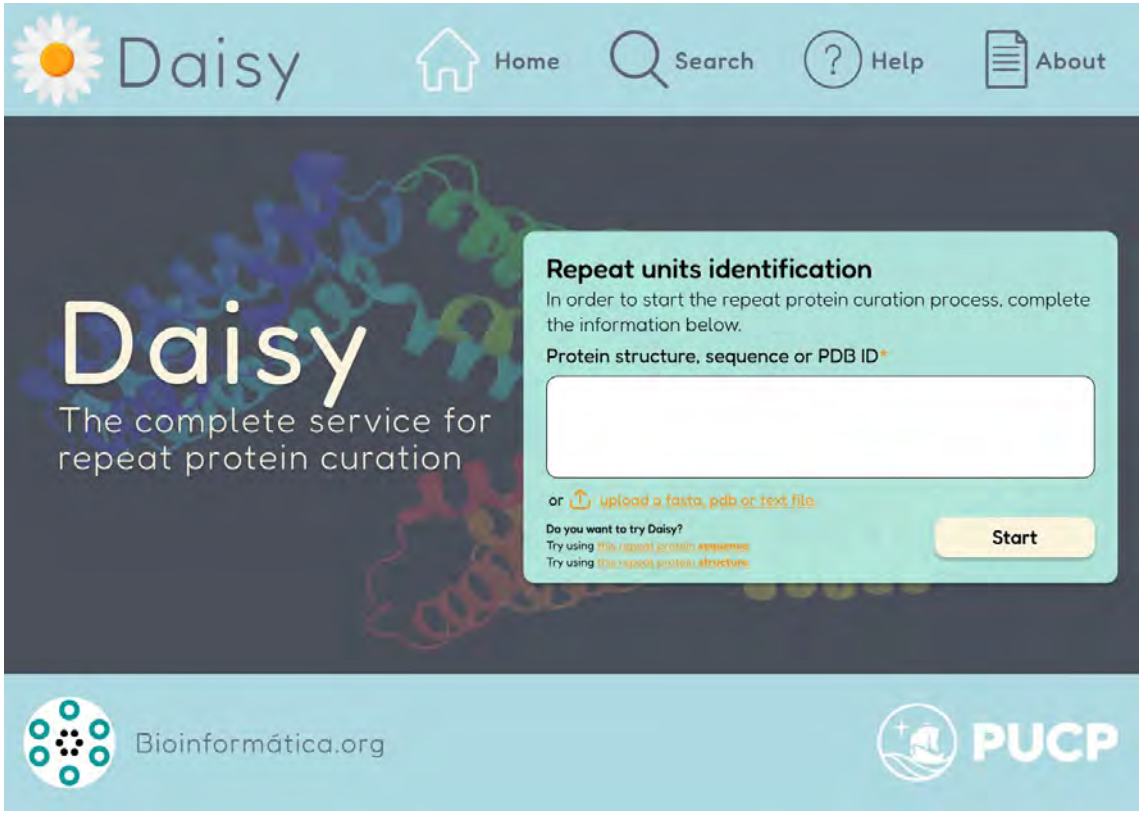



*Figura 7 Diseño de la página principal para el servicio web integrado para la curación de proteínas repetidas*

En la figura 7 se observa la página principal del servicio. En esta, el usuario podrá ingresar una estructura o secuencia de proteína que desee analizar. Asimismo, podrá cargar esta información como un archivo. Es importante mencionar que las otras secciones de la página del servicio pueden ser accedidas desde la cabecera.

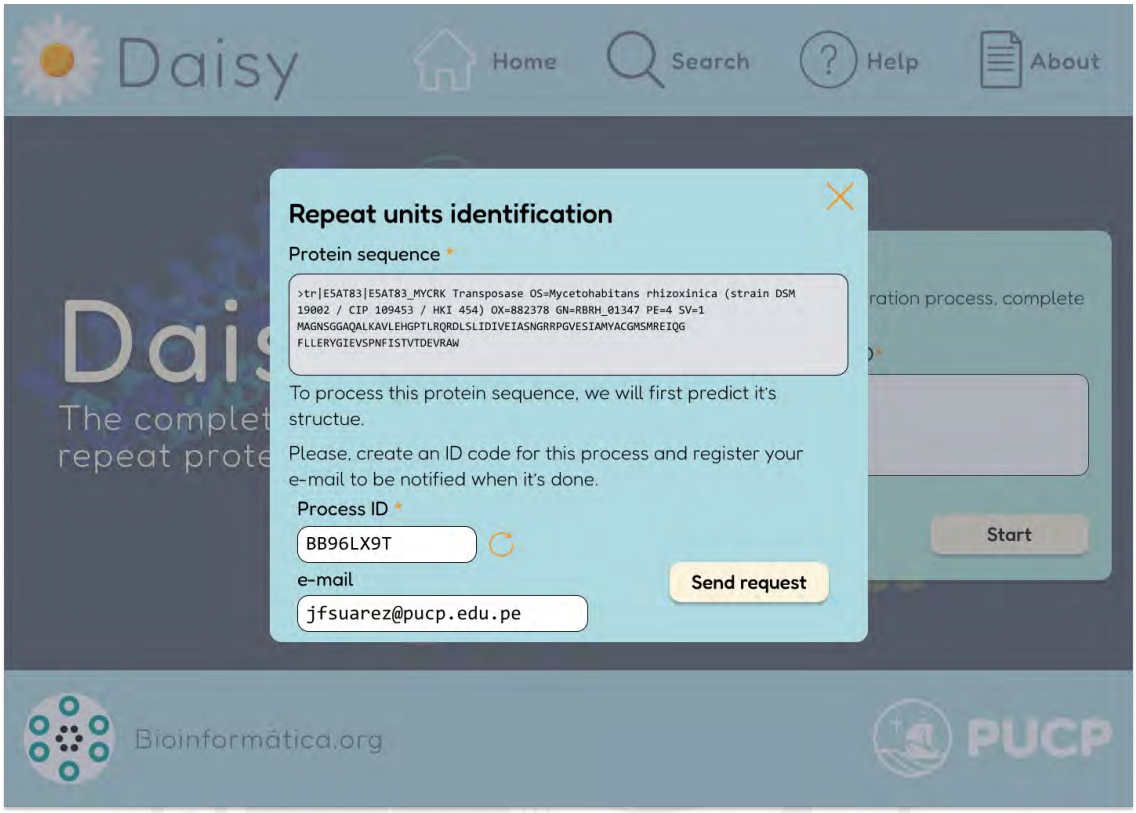



*Figura 8 Diseño de la ventana de registro de proceso para una secuencia de proteína*

En la figura 8 se puede observar el registro del proceso cuando el usuario decide ingresar la secuencia de una proteína. Para esto se le informa que el primer paso que se realizará en el servicio es obtener una predicción de la estructura. Para esto, antes de iniciar, el usuario debe crear un identificador de proceso o utilizar el autogenerado. Además, puede registrar su correo electrónico para que sea notificado de los resultados del proceso.

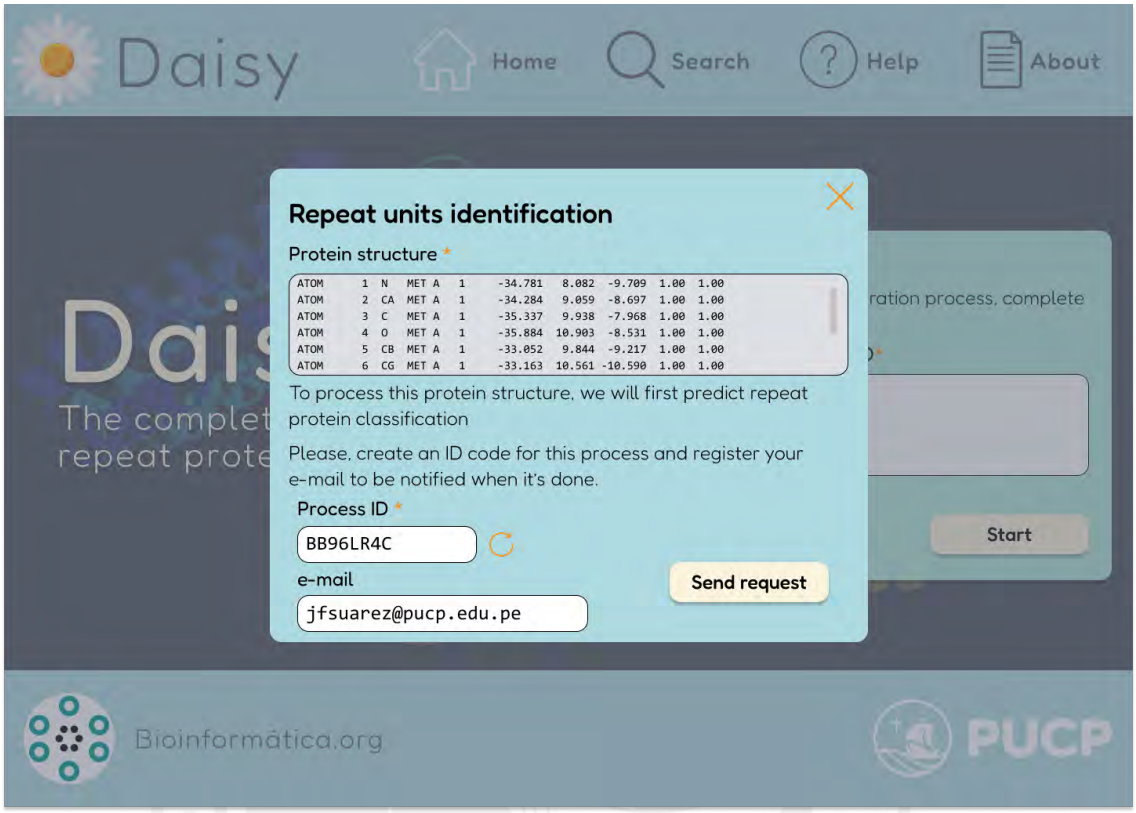



*Figura 9 Diseño de la ventana de registro de proceso de una estructura de proteína*

En la figura 9 se puede observar el registro del proceso cuando el usuario decide ingresar la estructura de una proteína. Para esto se le informa que el primer paso que se realizará en el servicio es predecir la clase y subclase de proteína repetida. Para esto, antes de iniciar, el usuario debe crear un identificador de proceso o utilizar el autogenerado. Además, puede registrar su correo electrónico para que sea notificado de los resultados del proceso.

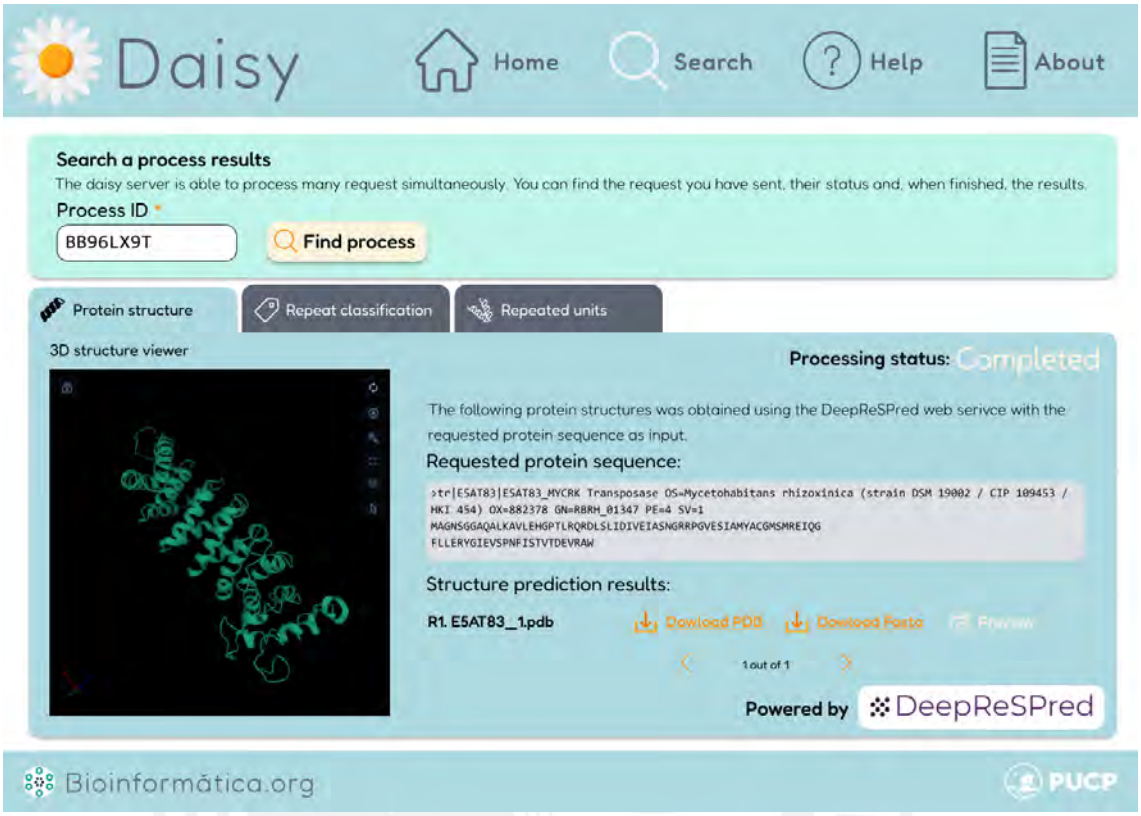



*Figura 10 Diseño de la ventana de resultados de predicción de estructura de proteína en base a su secuencia*

En la figura 10 se puede observar la visualización de resultados de predicción de estructura tridimensional de proteína repetida en base a su secuencia. En la parte superior se observa la opción de búsqueda de resultados según el identificador de proceso. Asimismo, estos resultados son los obtenidos mediante la base de datos AlphaFold (Varadi et al., 2022).

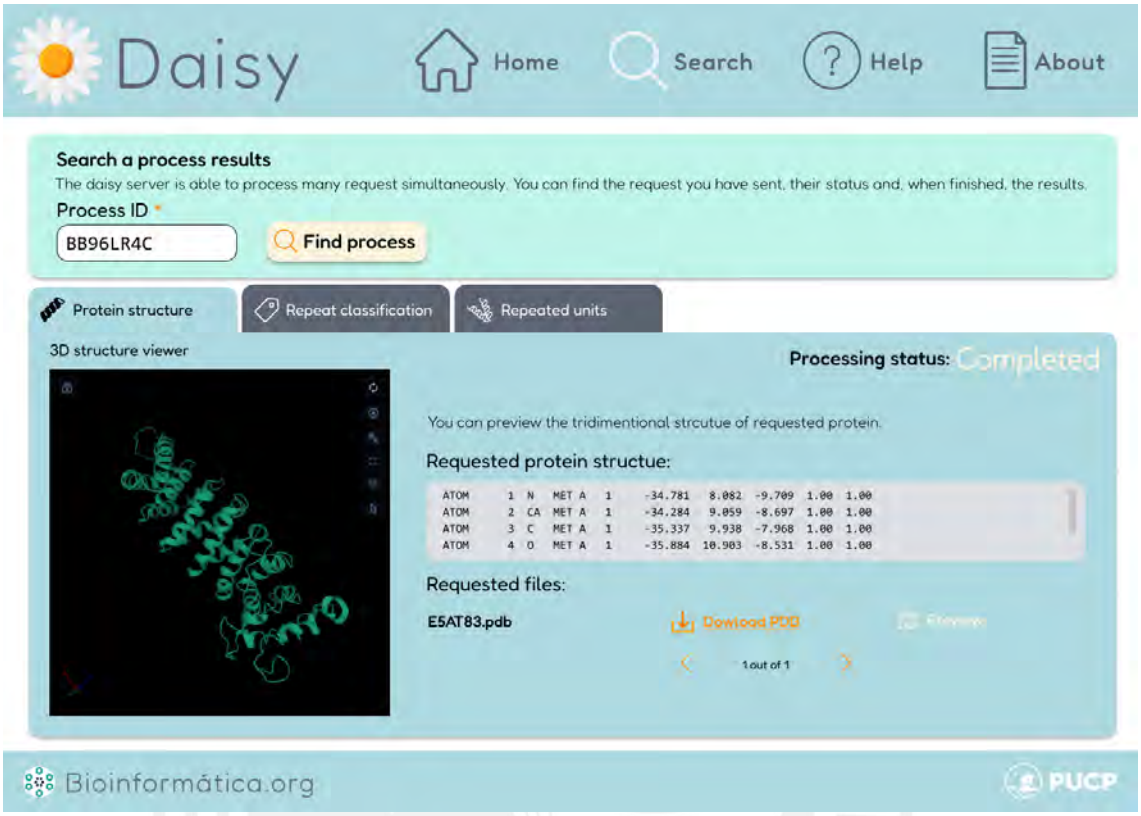



*Figura 11 Diseño de la ventana de visualización de estructura de proteína ingresada*

En la figura 11 se observa el diseño realizado para visualizar la estructura de proteína ingresada antes del procesamiento del flujo. Para esta visualización, se utiliza la librería PyMol (Schrödinger, LLC, 2015).

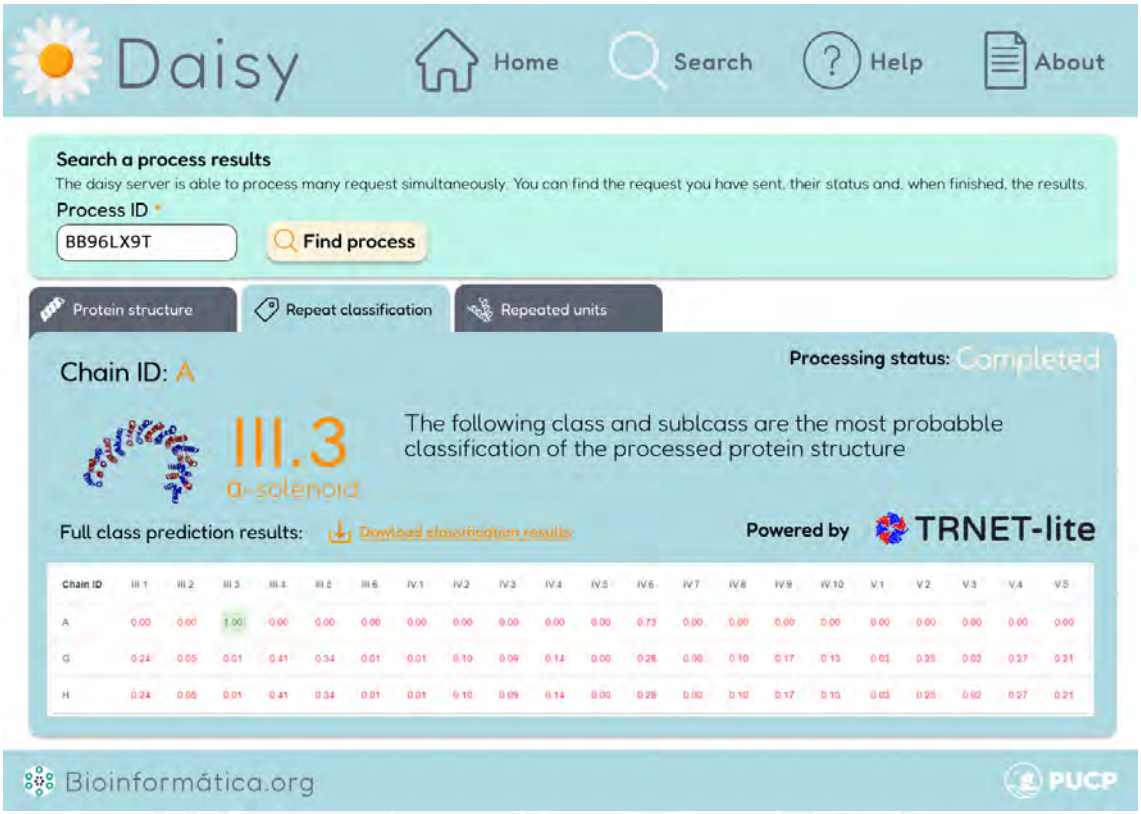



*Figura 12 Diseño de la ventana de resultados de predicción de clase y subclase de proteína repetida.*

En la figura 12 se observa el diseño realizado para la pantalla que muestra los resultados de predicción de clase y subclase de proteína repetida. Estos resultados son obtenidos mediante el servicio TRNET-lite (Tenorio Ku & Hirsh, 2021). Asimismo, se utiliza la predicción de estructura terciaria o la estructura ingresada para el procesamiento de secuencia o estructura, respectivamente. En el resultado, las subclases las cuales tengan una probabilidad de clasificación mayor a la del umbral se resaltan en verde.

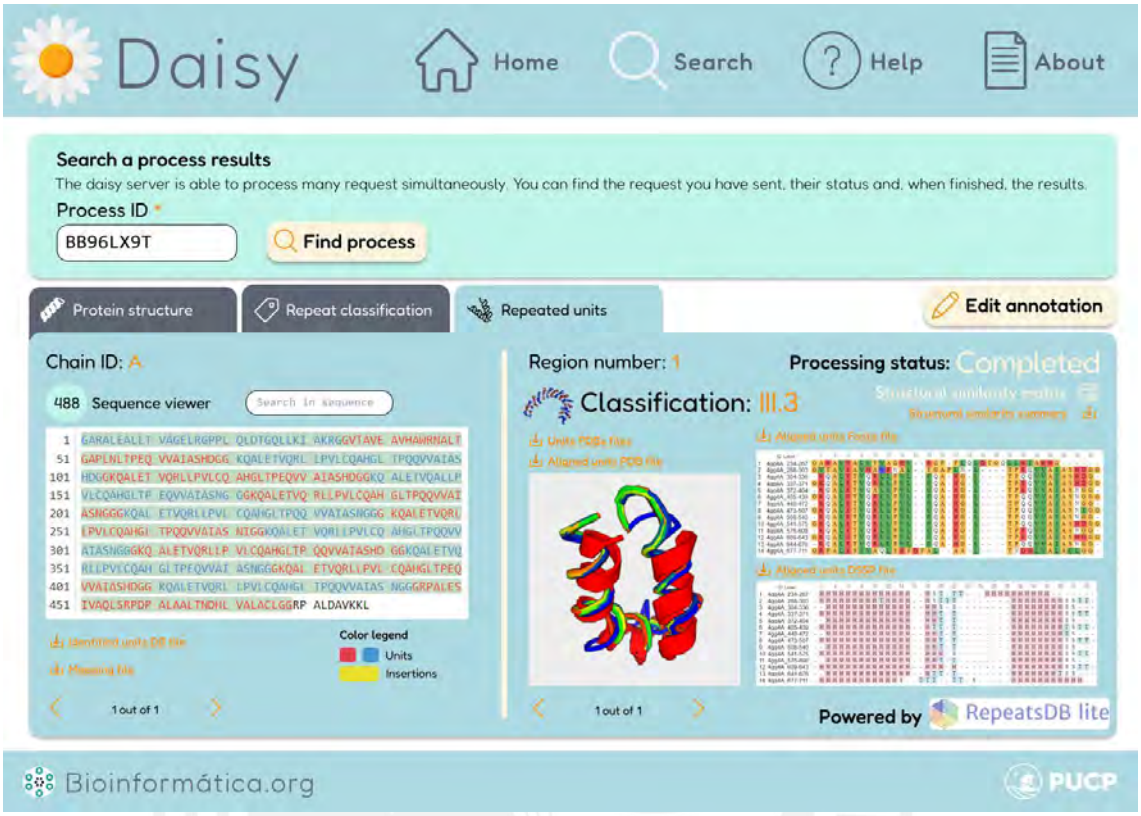



*Figura 13 Diseño de la ventana de resultados de predicción de unidades de repetición y edición de anotaciones*

En la figura 13 se observa el diseño realizado para la visualización de los resultados de predicción de unidades de repetición. Por cada cadena de la proteína, se muestran las regiones de repetición. Asimismo, se detalla la alineación de secuencia y estructura. Asimismo, se muestra la clase y subclase identificada para la unidad de repetición. Estos resultados se obtienen mediante el servicio RepeatsDB lite (Hirsh et al., 2018).

#### 6.2.3.4 Integración, desarrollo y ejecución de pruebas del servicio web integrado para la curación de proteínas repetidas

Con la definición de requerimientos y el diseño de interfaces, se desarrolló el servicio web integrado para la curación de proteínas repetidas Daisy. Este considera la predicción de estructura terciaria de una proteína, la predicción de clase y subclase de proteína repetida y la predicción de unidades de repetición.

Se desarrolló un *middleware* que orquesta las llamadas a los servicios desplegados. Se interconectó el componente de descarga de archivos de estructura desde PDB y AlphaFold (Berman et al., 2000; Varadi et al., 2022), el servicio modificado TRNET-lite (Tenorio Ku & Hirsh, 2021) y el servicio modificado RepeatsDB-lite (Hirsh et al., 2018).

Este *middleware* se desarrolló en Python, considerando el *framework* Flask. Asimismo, se desarrollará una interfaz gráfica web que sigue estándares de usabilidad en servicios web bioinformáticos (Bezerra Brandao et al., 2021). Esta interfaz se deberá construir en JavaScript considerando el *framework* Vue. La figura 14 muestra el diagrama de arquitectura del servicio web integrado desarrollado Daisy. Se utilizó una nube virtual privada en AWS *(Amazon Web Services*, por sus siglas en inglés). Asimismo, el servicio web está disponible a través de la página web del grupo de trabajo en Bioinformática de la Pontificia Universidad Católica del Perú[5].

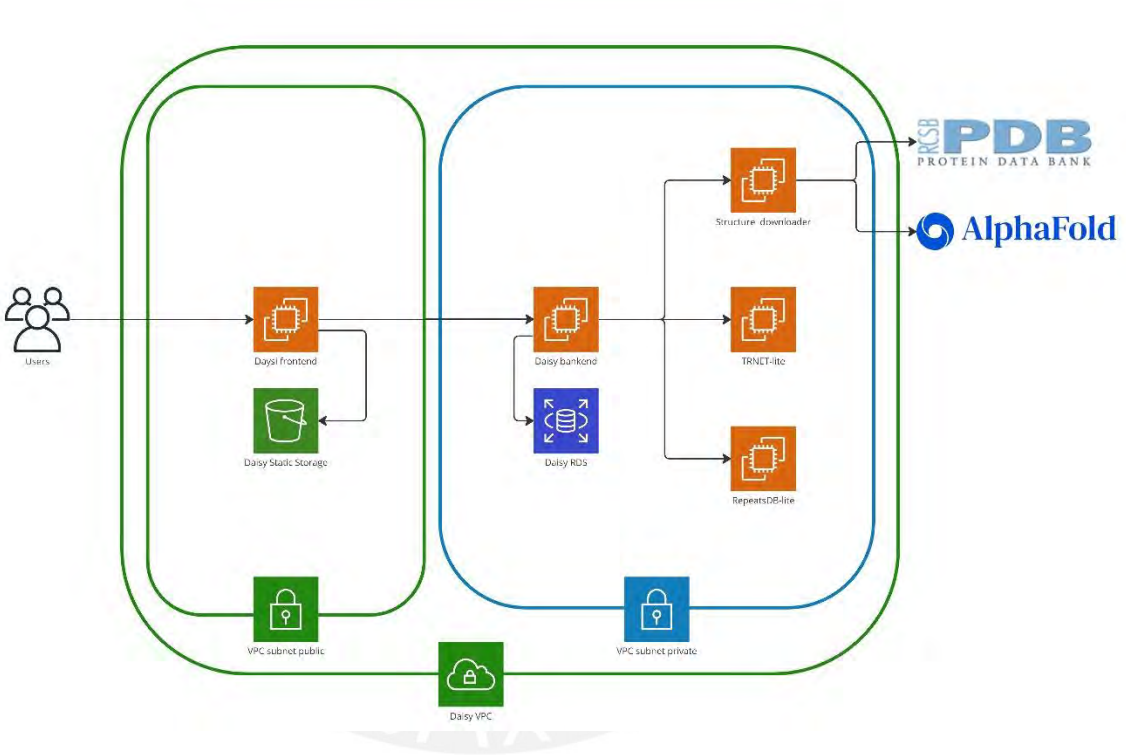



*Figura 14 Diagrama de arquitectura del servicio web integrado Daisy*

Se definieron casos de prueba en base a los requerimientos del servicio. Se procedió a ejecutar los casos de prueba, obteniendo un 100% de éxito. Asimismo, se obtuvo la aprobación general en una revisión cualitativa mediante juicio experto. Esto cumple con el indicador objetivamente verificable esperado para el resultado. El código fuente del servicio desarrollado puede obtenerse en el Anexo F del presente documento. Asimismo, la documentación del servicio API desarrollado puede revisarse en el Anexo G.

[5] Esta página puede ser accedida mediante el enlace https://bioinformatica.org/

## 6.3 Discusión

Para poder resolver la falta de una herramienta integrada para la curación de proteínas repetidas que considere las actividades más relevantes de esta tarea, también fue necesario identificar, describir, desplegar e integrar servicios de predicción de unidades de repetición en base a la estructura tridimensional de una proteína.

Para esto, se trabajó con el servicio RepeatsDB-lite (Hirsh et al., 2018). Este servicio permite realizar una predicción de unidades en una cadena de la estructura de una proteína. Se realizó la descripción del servicio, así como el modelo de datos que utiliza.

Para que el servicio esté adecuadamente integrado al objetivo general del proyecto, se realizó una modificación para permitir que el procesamiento consuma menos tiempo de procesamiento. Esto se requiere debido a que el algoritmo revisa 25000 muestras de unidades de repetición para la identificación de la unidad maestra. Para esto, el algoritmo toma como dato de entrada la clase y subclase de proteína repetida para realizar la búsqueda únicamente en las muestras de unidades relevantes. Gracias a esto, el algoritmo mejora notablemente respecto al tiempo de ejecución.

Asimismo, se analizó, diseñó y desarrolló un servicio web integrado para la curación de proteínas repetidas. Primero, se planteó el diagrama de flujo de ejecución para el proceso de curación de proteína repetida, considerando los servicios revisados, desplegados e integrados. Con ello, se elaboró una lista de requerimientos funcionales y funcionales. Luego de ello, se procedió a realizar un diseño de interfaces gráficas considerando lineamientos de usabilidad en servicios web bioinformáticos (Bezerra Brandao et al., 2021). También se diseñó el diagrama de despliegue incluyendo todos los servicios integrados previamente y un *middleware* que orquesta la ejecución del proceso de curación. Finalmente, se procedió a desarrollar el servicio con el *framework* definido en los requerimientos. Este servicio fue probado y desplegado en la nube de AWS.

Con los resultados obtenidos, que cumplen con los indicadores objetivamente verificables, se afirma que se logró implementar un servicio web para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición.

# Capítulo 7.Conclusiones y trabajos futuros

## 7.1 Conclusiones

Con la finalidad de resolver la falta de una herramienta integrada para la curación de proteínas repetidas que considere la predicción de clasificación, estructura terciaria e identificación de unidades de repetición, se planteó desarrollar una herramienta que permita realizar este proceso de curación con las funcionalidades detalladas. Para resolver este problema, se plantearon tres objetivos específicos.

El primer objetivo específico de este proyecto de tesis es integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura terciaria. Para esto, se elaboró la descripción y modelo de datos del servicio TRNET-lite (Tenorio Ku & Hirsh, 2021). Este servicio permite la predicción de clase y subclase de una proteína repetida en base a su estructura. Una modificación realizada al servicio fue cambiar los datos de ingreso de un identificador de la base de datos PDB (Berman et al., 2000) a un archivo de estructura de proteína en formato PDB. Esto se realizó debido a que el servicio debe funcionar para predicción de estructuras tridimensionales de proteínas. Finalmente, el servicio modificado fue desplegado para ser integrado con la herramienta. Por lo descrito, se afirma que se logró el primer objetivo específico de este proyecto de tesis.

Luego de ello, el segundo objetivo específico trabajo fue incorporar servicios de predicción de estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación. Para esto, se elaboró la descripción y modelo de datos del servicio DeepReSPred (Palomino & Hirsh, 2021). Este servicio permite la predicción de estructura terciaria de una proteína en base a su secuencia. Una modificación realizada al servicio fue hacer que se mantenga la información ingresada en la primera línea del formato FASTA de la secuencia de proteína en la predicción. Asimismo, considerando los recientes avances en el uso de redes neuronales para la predicción de la estructura tridimensional de proteína en base a su secuencia se consideró integrarlo al servicio (Jumper et al., 2021; Tunyasuvunakool et al., 2021). Por ello, se desarrolló un componente para obtener los archivos de estructura desde las bases de dato PDB y AlphaFold (Berman et al., 2000; Varadi et al., 2022). Por lo descrito, se afirma que se logró el segundo objetivo específico de este proyecto de tesis.

Como último objetivo específico a lograr, se tenía implementar un servicio web para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición. Para

esto, se elaboró la descripción y modelo de datos del servicio RepeatsDB-lite (Hirsh et al., 2018). Este servicio permite la predicción de unidades de repetición en una cadena de estructura tridimensional de una proteína. Una modificación realizada al servicio fue modificar el flujo de ejecución del algoritmo. Para esto, en lugar de revisar las 25000 muestras de unidades de repetición para hallar una unidad maestra, se acortó la búsqueda tomando como referencia una clase y subclase de proteína repetida. Esto permitió que el servicio tome menos tiempo de ejecución durante su procesamiento. Finalmente, el servicio modificado fue desplegado para ser integrado con la herramienta. Luego de ello, se procedió con el diseño y desarrollo del servicio web integrado para la curación de proteínas repetidas Daisy. Para esto, se diseño el diagrama de flujo del proceso de curación de proteína repetida tomando en consideración los servicios modificados desplegados. Con esto, se procedió a la definición de requisitos funcionales y no funcionales del servicio web. Luego se procedió a realizar el diseño de interfaces gráficas considerando lineamientos de usabilidad para servicios web bioinformáticos (Bezerra Brandao et al., 2021). Asimismo, se elaboró un diagrama de arquitectura considerando los servicios de AWS (*Amazon Web Services*, por sus siglas en inglés) para la herramienta. En este, se detalla un *middleware* que se encarga de orquestar el procesamiento integrando los servicios desplegados. Por último, se procedió con el desarrollo, despliegue y evaluación del servicio web para la curación de proteínas repetidas, Daisy. Por lo descrito, se afirma que se logró el último objetivo específico de este proyecto de tesis.

Ya que se han cumplido todos los objetivos específicos de este proyecto de tesis, se puede concluir que se logró conseguir el objetivo general. Mediante la integración de servicios para la predicción de clasificación, estructura terciaria e identificación de unidades de repetición de una proteína y el desarrollo de un servicio web para la curación de proteínas repetidas que consideren estos aspectos, se logró concluir satisfactoriamente este proyecto de investigación

## 7.2 Trabajos futuros

En este proyecto se ha desarrollado un servicio web integrado para la curación de proteínas repetidas. Dentro de este servicio, se ha considerado el uso de una predicción de la estructura tridimensional de una proteína en base a su secuencia. Como trabajo futuro, se tiene modificar el servicio que realiza esta tarea por medio de generación de resultados usando el algoritmo AlphaFold (Jumper et al., 2021). Este es un servicio de

mayor precisión para la predicción de estructura tridimensional de una proteína en base a su secuencia, pero que tiene un costo computacional mayor. Este servicio utiliza un algoritmo que permite realizar predicción que tienen precisión atómica incluso en casos donde no hay una estructura similar conocida (Jumper et al., 2021). Este es una cualidad que no posee el servicio DeepReSPred (Palomino & Hirsh, 2021).

Por otro lado, se planea desarrollar la interfaz gráfica web del servicio y la función de notificación por correo electrónico para las solicitudes de los usuarios. Además, se desea implementar la posibilidad de procesar archivos en formato FASTA, PDB y/o CIF. Esto permitirá que el usuario pueda obtener los resultados específicos que desea, considerando también la selección manual de clase y subclase de proteína repetida para la identificación de unidades y regiones de repetición.

Asimismo, se plantea integrar una funcionalidad que permita registrar la curación de la proteína repetida a una base de datos como RepeatsDB (Di Domenico et al., 2014). Esta es una base de datos con anotaciones de estructuras de proteínas repetidas (Di Domenico et al., 2014). Asimismo, esta base de datos es un esfuerzo continuo para sistematizar la clasificación y anotación estructural en unidades de repetición de proteínas de manera consistente (Di Domenico et al., 2014). Así, se podrá compartir los resultados del proceso de curación y las anotaciones estructurales que realicen los investigadores que utilicen el servicio web integrado desarrollado en este proyecto.

## Referencias


Al-Lazikani, B., Jung, J., Xiang, Z., & Honig, B. (2001). Protein structure prediction. *Current opinion in chemical biology*, *5*(1), 51–56.

Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I., & Langmead, C. J. (2011). Learning generative models for protein fold families: Generative Models for Protein Fold Families. *Proteins: Structure, Function, and Bioinformatics*, *79*(4), 1061–1078. https://doi.org/10.1002/prot.22934

Baumer, D., Bischofberger, W., Lichter, H., & Zullighoven, H. (1996). User interface prototyping-concepts, tools, and experience. *Proceedings of IEEE 18th International Conference on Software Engineering*, 532–541. https://doi.org/10.1109/ICSE.1996.493447

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bezerra Brandao, M., Hirsh, L., & Pow Sang, J. (2021). *Usabilidad en servicios web bioinformáticos* [Pontificia Universidad Católica del Perú]. http://hdl.handle.net/20.500.12404/19477

Björklund, Å. K., Ekman, D., & Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS computational biology*, *2*(8), e114.

Bliven, S. E., Lafita, A., Rose, P. W., Capitani, G., Prlić, A., & Bourne, P. E. (2019). Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm. *PLOS Computational Biology*, *15*(4), e1006842. https://doi.org/10.1371/journal.pcbi.1006842

Brunette, T., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., & Baker, D. (2015). Exploring the repeat protein universe through computational protein design. *Nature*, *528*(7583), 580–584. https://doi.org/10.1038/nature16162

Burnham, J. F. (2006). Scopus database: A review. *Biomedical digital libraries*, *3*(1), 1.

Can, T. (2014). Introduction to Bioinformatics. En M. Yousef & J. Allmer (Eds.), *MiRNomics: MicroRNA Biology and Computational Analysis* (Vol. 1107, pp. 51–71). Humana Press. https://doi.org/10.1007/978-1-62703-748-8_4

Cañedo Andalia, R., Nodarse Rodríguez, M., & Mulet, N. L. (2015). Similitudes y diferencias entre PubMed, Embase y Scopus. *Revista Cubana de Información en Ciencias de la Salud (ACIMED)*, *26*(1), 84–91.

Center for BioMolecular Modeling. (2022). *Tertiary Structure: The Overall 3-Dimensional Shape of a Protein*. Tertiary Structure Protein Structure Tutorials. https://cbm.msoe.edu/teachingResources/proteinStructure/tertiary.html

Chakrabarty, B., & Parekh, N. (2020). PRIGSA2: Improved version of protein repeat identification by graph spectral analysis. *Journal of Biosciences*, *45*(1), 95. https://doi.org/10.1007/s12038-020-00058-x

Chakrabarty, B., & Parekh, N. (2022). DBSTRIPS: Database of structural repeats in proteins. *Protein Science*, *31*(1), 23–36. https://doi.org/10.1002/pro.4052

Chen, C., Wu, H., & Bian, K. (2017). β-Barrel Transmembrane Protein Predicting Using Support Vector Machine. En D.-S. Huang, A. Hussain, K. Han, & M. M. Gromiha (Eds.), *Intelligent Computing Methodologies* (Vol. 10363, pp. 360–368). Springer International Publishing. https://doi.org/10.1007/978-3-319-63315-2_31

de Lima Salgado, A., Agostini do Amaral, L., Fortes, R. P. M., Chagas, M. H. N., & Joyce, G. (2017). Addressing mobile usability and elderly users: Validating contextualized heuristics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10288 LNCS*, 379–394. https://doi.org/10.1007/978-3-319-58634-2_28

Design, F. (2017). Figma: The collaborative interface design tool. *Retrieved September*, *17*, 2017.

Di Domenico, T., Potenza, E., Walsh, I., Gonzalo Parra, R., Giollo, M., Minervini, G., Piovesan, D., Ihsan, A., Ferrari, C., Kajava, A. V., & Tosatto, S. C. E. (2014). RepeatsDB: A database of tandem repeat protein structures. *Nucleic Acids Research*, *42*(D1), D352–D357. https://doi.org/10.1093/nar/gkt1175

Dudding, C. C. (2009). Digital Videoconferencing: Applications Across the Disciplines. *Communication Disorders Quarterly*, *30*(3), 178–182. https://doi.org/10.1177/1525740108327449

Ekeberg, M., Hartonen, T., & Aurell, E. (2014). Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics*, *276*, 341–356. https://doi.org/10.1016/j.jcp.2014.07.024

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., & Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, *87*(1), 012707. https://doi.org/10.1103/PhysRevE.87.012707

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein

families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Espada, R., Parra, R. G., Mora, T., Walczak, A. M., & Ferreiro, D. U. (2017). Inferring repeat-protein energetics from evolutionary information. *PLOS Computational Biology*, *13*(6), e1005584. https://doi.org/10.1371/journal.pcbi.1005584

Fefelova, I., Fefelov, A., Lytvynenko, V., Dzierżak, R., Lurie, I., Savina, N., Voronenko, M., & Vyshemyrska, S. (2020). Protein Tertiary Structure Prediction with Hybrid Clonal Selection and Differential Evolution Algorithms. En V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya, & S. Radetskaya (Eds.), *Lecture Notes in Computational Intelligence and Decision Making* (Vol. 1020, pp. 673–688). Springer International Publishing. https://doi.org/10.1007/978-3-030-26474-1_47

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, *20*(6), 1981–1996. https://doi.org/10.1093/bib/bby063

Gómez, J., García, L. J., Salazar, G. A., Villaveces, J., Gore, S., García, A., Martín, M. J., Launay, G., Alcántara, R., Del-Toro, N., & others. (2013). BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics*, *29*(8), 1103–1104.

Goncearenco, A., & Berezovsky, I. N. (2015). Protein function from its emergence to diversity in contemporary proteins. *Physical Biology*, *12*(4), 045002. https://doi.org/10.1088/1478-3975/12/4/045002

Greener, J. G., Kandathil, S. M., & Jones, D. T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural

constraints. *Nature Communications*, *10*(1), 3977. https://doi.org/10.1038/s41467-019-11994-0

Heringa, J. (1998). Detection of internal repeats: How common are they? *Current Opinion in Structural Biology*, *8*(3), 338–345. https://doi.org/10.1016/S0959-440X(98)80068-7

Hirsh, L., Bernardi, P., Tosatto, S. S. E., & Piovesan, D. (2017). *Solving the Structural Modeling Problems for Tandem Repeat Proteins* [Università Degli Studi Di Padova]. http://hdl.handle.net/11577/3424915

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

Jain, M. N., & Karnad, V. (2017). *Online Forms for Data Collection and its Viability in Fashion and Consumer Buying Behavior Survey–A Case Study*.

Jorda, J., & Kajava, A. V. (2010). Protein Homorepeats. En *Advances in Protein Chemistry and Structural Biology* (Vol. 79, pp. 59–88). Elsevier. https://doi.org/10.1016/S1876-1623(10)79002-7

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with

AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kajava, A. V. (2001). Review: Proteins with Repeated Sequence—Structural Prediction and Modeling. *Journal of Structural Biology*, *134*(2–3), 132–144. https://doi.org/10.1006/jsbi.2000.4328

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

Lockwood, C., Munn, Z., & Porritt, K. (2015). Qualitative research synthesis: Methodological guidance for systematic reviewers utilizing meta-aggregation. *International Journal of Evidence-Based Healthcare*, *13*(3), 179–187. https://doi.org/10.1097/XEB.0000000000000062

Lodish, H. F., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2006). *Molecular cell biology* (Vol. 4). Citeseer.

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

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., & Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, *108*(49). https://doi.org/10.1073/pnas.1111471108

Moss, G. P. (2009). *Basic Terminology of Stereochemistry: (IUPAC Recommendations 1996)* [Data set]. De Gruyter. https://doi.org/10.1515/iupac.68.3330

Muroya Tokushima, L. F., & Hirsh, L. (2022). *Identificación y clasificación automática de repeticiones en estructuras de proteínas repetidas* [Pontificia Universidad Católica del Perú]. http://hdl.handle.net/20.500.12404/21423

Mutter, S., Pfahringer, B., & Holmes, G. (2008). Propositionalisation of Profile Hidden Markov Models for Biological Sequence Analysis. En W. Wobcke & M. Zhang (Eds.), *AI 2008: Advances in Artificial Intelligence* (Vol. 5360, pp. 278–288). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-89378-3_27

Nelson, D. L., Lehninger, A. L., & Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

Pagès, G., & Grudinin, S. (2019). DeepSymmetry: Using 3D convolutional networks for identification of tandem repeats and internal symmetries in protein structures. *Bioinformatics*, *35*(24), 5113–5120. https://doi.org/10.1093/bioinformatics/btz454

Palomino, S., & Hirsh, L. (2021). *Desarrollo de una herramienta para la predicción de estructuras terciarias de proteínas repetidas a partir de su estructura primaria*. Pontificia Universidad Católica del Perú.

Palopoli, N., Monzon, A. M., Parisi, G., & Fornasari, M. S. (2016). Addressing the Role of Conformational Diversity in Protein Structure Prediction. *PLOS ONE*, *11*(5), e0154923. https://doi.org/10.1371/journal.pone.0154923

Pedraza, K., & Hirsh, L. (2019). *Optimización de método para la clasificación de proteínas repetidas e identificación de unidades de repetición mediante el uso de perfiles de Modelos Ocultos de Markov*. Pontificia Universidad Católica del Perú.

Protein Data Bank. (2022). *Atomic Coordinate Entry Format Version 3.3*. Wwpdb.org. http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html

Rowling, P. J. E., Sivertsson, E. M., Perez-Riba, A., Main, E. R. G., & Itzhaki, L. S. (2015). Dissecting and reprogramming the folding and assembly of tandem-repeat proteins. *Biochemical Society Transactions*, *43*(5), 881–888. https://doi.org/10.1042/BST20150099

Schrödinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.

Shi, H., & Zhang, X. (2020). Component-Based Design and Assembly of Heuristic Multiple Sequence Alignment Algorithms. *Frontiers in Genetics*, *11*, 105. https://doi.org/10.3389/fgene.2020.00105

Takahashi, T., Chikenji, G., & Tokita, K. (2021). Lattice protein design using Bayesian learning. *Physical Review E*, *104*(1), 014404. https://doi.org/10.1103/PhysRevE.104.014404

Tenorio Ku, L. G., & Hirsh, L. (2021). *Clasificación de proteínas repetidas basado en su información estructural utilizando aprendizaje de máquina* [Pontificia Universidad Católica del Perú]. http://hdl.handle.net/20.500.12404/18199

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Urbina, V., & Hirsh, L. (2021). *Herramienta para la curación de familias de proteínas repetidas*. Pontificia Universidad Católica del Perú.

Urvoas, A., Guellouz, A., Valerio-Lepiniec, M., Graille, M., Durand, D., Desravines, D. C., van Tilbeurgh, H., Desmadril, M., & Minard, P. (2010). Design, Production and Molecular Structure of a New Family of Artificial Alpha-helicoidal Repeat Proteins (αRep) Based on Thermostable HEAT-like Repeats. *Journal of Molecular Biology*, *404*(2), 307–327. https://doi.org/10.1016/j.jmb.2010.09.048

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., & others. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439–D444.

Yang Zhang Lab. (2022). *What is FASTA format?* FASTA Format. https://zhanggroup.org/FASTA

# Anexos

A continuación, se detallan los diferentes anexos a este documento.

## Anexo A: Formulario de extracción

Este anexo contiene el formulario de extracción utilizado para el Capítulo 3 - Estado del arte se encuentra en la hoja de cálculo nombrada como *20161811_ManuelBezerraBrandao_LaylaHirsh _Anexo_A.xlsx* y se puede acceder a través de la nube por medio del siguiente enlace:

https://docs.google.com/spreadsheets/d/1aahVGTREvCLnW2n_SxtEV6mZYb20PoWc/edit?usp=sharing&ouid=115633799306253091037&rtpof=true&sd=true

# Anexo B: Plan de Proyecto

En este anexo se muestran las diferentes partes de la planificación para la ejecución del presente proyecto de tesis.

## 1. Justificación

Actualmente, dentro de la de la Bioinformática Estructural, existe un creciente interés en las proteínas repetidas dado que tienen diversas aplicaciones en la ingeniería, en el estudio de enfermedades humanas y la diversidad estructural (Goncearenco & Berezovsky, 2015; Kajava, 2012). Se han desarrollado múltiples herramientas para el estudio de secuencias, estructuras, funciones y familias de proteínas repetidas y sus familias (Berman et al., 2000; Di Domenico et al., 2014; El-Gebali et al., 2019; The UniProt Consortium, 2019).

Sin embargo, no existe alguna publicación en donde se haya implementado una herramienta que facilite la tarea más complicada y manual en el estudio de proteínas repetidas, la curación(Hirsh et al., 2017; Urbina & Hirsh, 2021).

Es por ello que, el presente proyecto de investigación tiene como objetivo cubrir la falta de una herramienta integrada para la curación de proteínas repetidas que considere la predicción de clasificación, estructura terciaria e identificación de unidades de repetición. Para ello, se consideran servicios desarrollados con el enfoque de curación para atender esta necesidad (Bezerra Brandao et al., 2021; Hirsh et al., 2017; Muroya Tokushima & Hirsh, 2022; Palomino & Hirsh, 2021; Pedraza & Hirsh, 2019; Tenorio Ku & Hirsh, 2021; Urbina & Hirsh, 2021).

## 2. Viabilidad

Para este proyecto, se requerirá integrar diversos servicios existentes y diseñar un servicio web para la curación de proteínas repetidas. Las aplicaciones seleccionadas para la integración han sido desarrolladas como parte de proyectos de investigación del grupo de investigación de Bioinformática[6] y se tiene comunicación directa con cada uno de sus desarrolladores (Bezerra Brandao et al., 2021; Hirsh et al., 2017; Muroya Tokushima & Hirsh, 2022; Pedraza & Hirsh, 2019; Tenorio Ku & Hirsh, 2021; Urbina & Hirsh, 2021). A nivel ténico, se cuenta con la formación necesaria para realizar el diseño y desarrollo de

[6] Página web del grupo de investigación: bioinformatica.org

un servicio web, la cual será complementada con un curso virtual de desarrollo *frontend* en el *framework* Vue de Javascript en la plataforma Udemy[7].

Por todo esto, se afirma que el proyecto propuesto es viable. Además, todos los servicios a utilizar son de acceso abierto y se contará con la asesoría de sus creadores en caso de requerirla.

## 3. Alcance

Para este proyecto se han planteado una serie de objetivos específicos. El primero es integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura primaria y/o terciaria. Para ello se debe realizar el despliegue e integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria (Tenorio Ku & Hirsh, 2021).

El segundo objetivo es incorporar servicios de predicción de la estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación. Para ello se debe realizar el despliegue e integración del servicio de predicción de estructura de una proteína basada en su secuencia (Palomino & Hirsh, 2021).

Finalmente, el tercer y último objetivo es implementar un servicio web integrado para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición. Para ello se debe realizar el despliegue e integración del servicio para la curación la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición (Hirsh et al., 2016, 2017; Pedraza & Hirsh, 2019). Asimismo, se requiere desarrollar el servicio web para predecir la estructura de una proteína repetida e identificar sus unidades de repetición siguiendo los lineamientos de usabilidad para servicios web bioinformáticos (Bezerra Brandao et al., 2021; Urbina & Hirsh, 2021).

Para el desarrollo de este servicio se contempla la elaboración de un diagrama de ejecución, *mockups* navegables y una evaluación de usabilidad utilizando el cuestionario SUS-BWS (Bezerra Brandao et al., 2021). Por otro lado, el mantenimiento o modificación significativa de los servicios seleccionados está excluido del alcance de este proyecto.

## 4. Limitaciones

Se requerirá trabajar con servicios desarrollados dentro del grupo de investigación de Bioinformática. Estos han sido compartidos junto a su documentación desarrollada mas

[7] Página web del curso adquirido: https://www.udemy.com/course/vuejs-fh/

no se asegura el correcto funcionamiento de estos. Por otro lado, para las pruebas de usabilidad del servicio integrado a desarrollar en este proyecto, se tiene una limitación respecto a la disponibilidad de tiempo de los participantes usuarios de servicios web bioinformáticos.

## 5. Identificación de los riesgos del proyecto

A continuación, se presenta una lista conteniendo los riesgos de este proyecto, presentando síntomas, probabilidad, impacto y severidad. Adicionalmente se presentan la mitigación y contingencia planeada para estos riesgos.

RG1. Alguno de los servicios que han sido seleccionados no cumplen sus funciones correctamente una vez que son desplegados.

Síntomas:

- Las salidas del servicio son incorrectas o incoherentes a los resultados esperados.
- El servicio no arroja salidas y se cae a pesar de seguir las instrucciones de la documentación.

Probabilidad: 0.30 (Baja)

Impacto: 0.80 (Muy alto)

Severidad: 0.24 (Alta)

Mitigación:

- Comunicarse con los desarrolladores del servicio en caso haya algún error que no pueda ser solucionado con la documentación.
- Realizar un mantenimiento al servicio.

Contingencia:

- Elegir un nuevo servicio que cumpla con la misma función en caso el mantenimiento requerido sea excesivo.

RG2. La disponibilidad de los usuarios que participarán en la evaluación de usabilidad no concuerda con la planificación del proyecto.

Síntomas:

- Dificultad para contactar con alguno de los usuarios.
- Demora en el tiempo de respuesta de alguno de los usuarios.

Probabilidad: 0.50 (Moderada)

Impacto: 0.10 (Bajo)

Severidad: 0.05 (Baja)

Mitigación:

- Coordinar previamente con los usuarios participantes.
- Contar con una referencia de posibles usuarios alternativos a participar en la evaluación de usabilidad.

Contingencia:

- Coordinar con uno de los usuarios alternativos a participar en la evaluación de usabilidad.

Como parte de la identificación de riesgos del proyecto, se presentan las respectivas leyendas de los valores utilizados para la probabilidad, impacto y severidad de los riesgos en las tablas 10, 11 y 12, respectivamente

*Tabla 10 Anexo B: Leyenda de probabilidad*

| Probabilidad | Valor |
|---|---|
| Muy baja | 0.10 |
| Baja | 0.30 |
| Moderada | 0.50 |
| Alta | 0.70 |
| Muy alta | 0.90 |

*Tabla 11 Anexo B: Leyenda de impacto*

| Impacto | Valor |
|---|---|
| Muy bajo | 0.05 |
| Bajo | 0.10 |
| Moderado | 0.20 |
| Alto | 0.40 |
| Muy alto | 0.80 |

*Tabla 12 Anexo B: Leyenda de severidad*

| Severidad | Valor |
|---|---|
| Baja | ≤0.05 |
| Media | >0.0.5 ∧ <0.14 |
| Alta | ≥0.14 |

## 6. Estructura de descomposición del trabajo (EDT)

A continuación, se presenta la estructura de descomposición del trabajo mediante una

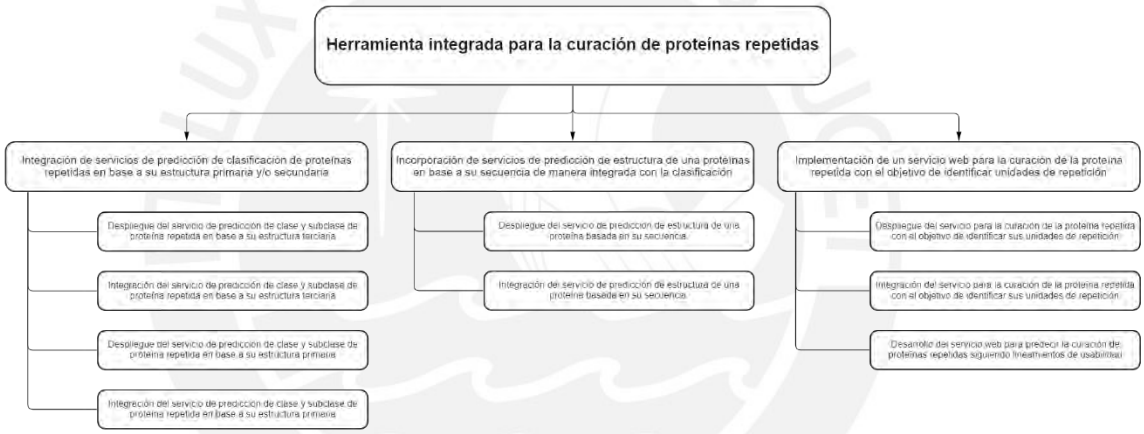



*Figura 15 Anexo B: Estructura de descomposición del trabajo del proyecto: Parte superior, título del proyecto. Zona media, objetivos del trabajo. Zona inferior, resultados esperados y medios de verificación.*

representación esquemática para el proyecto de tesis en la figura 15.

## 7. Lista de tareas

En esta sección del documento se presenta la lista de tareas de la planificación del proyecto de tesis, detallando la duración estimada, esfuerzo asociado y costo estimado. El resumen total de esfuerzos y costos estimados de esta investigación se encuentran en la sección *Costeo del proyecto*. Toda esta información se muestra en la tabla 13.

*Tabla 13 Lista de actividades del proyecto de tesis con duración, esfuerzo y costos estimados*

| ID | Actividad | Duración estimada | Esfuerzo asociado | Costo estimado |
|---|---|---|---|---|
| **Herramienta integrada para la curación de proteínas repetidas** | | **20 semanas** | **420 horas-persona** | **S/.25200.00** |
| **1** | **Integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura primaria y/o terciaria.** | **28 días** | **84 horas-persona** | **S/.5040.00** |
| 1.1 | Despliegue del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria. | 7 días | 21 horas-persona | S/.1260.00 |
| 1.1.1 | Elaborar el informe de despliegue del servicio | 4 días | 12 horas-persona | S/.720.00 |
| 1.1.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 3 días | 9 horas-persona | S/.540.00 |
| 1.2 | Integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria. | 7 días | 21 horas-persona | S/.1260.00 |
| 1.2.1 | Elaborar informe de integración del servicio | 3 días | 9 horas-persona | S/.540.00 |
| 1.2.2 | Elaborar propuesta de casos de prueba del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |
| 1.2.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |
| 1.3 | Despliegue del servicio de predicción de clase y subclase de | 7 días | 21 horas-persona | S/.1260.00 |

| | | | | |
|---|---|---|---|---|
| | proteína repetida en base a su estructura primaria. | | | |
| 1.3.1 | Elaborar el informe de despliegue del servicio | 4 días | 12 horas-persona | S/.720.00 |
| 1.3.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 3 días | 9 horas-persona | S/.540.00 |
| 1.4 | Integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura primaria. | 7 días | 21 horas-persona | S/.1260.00 |
| 1.4.1 | Elaborar informe de integración del servicio | 3 días | 9 horas-persona | S/.540.00 |
| 1.4.2 | Elaborar propuesta de casos de prueba del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |
| 1.4.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |
| **2** | **Predecir la estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación.** | **14 días** | **42 horas-persona** | **S/.2520.00** |
| 2.1 | Despliegue del servicio de predicción de estructura de una proteína basada en su secuencia. | 7 días | 21 horas-persona | S/.1260.00 |
| 2.1.1 | Elaborar el informe de despliegue del servicio | 4 días | 12 horas-persona | S/.720.00 |
| 2.1.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 3 días | 9 horas-persona | S/.540.00 |

| | | | | |
|---|---|---|---|---|
| 2.2 | Integración del servicio de predicción de estructura de una proteína basada en su secuencia. | 7 días | 21 horas-persona | S/.1260.00 |
| 2.2.1 | Elaborar informe de integración del servicio | 3 días | 9 horas-persona | S/.540.00 |
| 2.2.2 | Elaborar propuesta de casos de prueba del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |
| 2.2.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |
| **3** | **Implementar un servicio web integrado para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición.** | **98 días** | **294 horas-persona** | **S/.17640.00** |
| 3.1 | Despliegue del servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición. | 7 días | 21 horas-persona | S/.1260.00 |
| 3.1.1 | Elaborar el informe de despliegue del servicio | 4 días | 12 horas-persona | S/.720.00 |
| 3.1.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 3 días | 9 horas-persona | S/.540.00 |
| 3.2 | Integración del servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición. | 7 días | 21 horas-persona | S/.1260.00 |
| 3.2.1 | Elaborar informe de integración del servicio | 3 días | 9 horas-persona | S/.540.00 |
| 3.2.2 | Elaborar propuesta de casos de prueba del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |

| | | | | |
|---|---|---|---|---|
| 3.2.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 2 días | 6 horas-persona | S/.360.00 |
| 3.3 | Desarrollo del servicio web para predecir la estructura de una proteína repetida e identificar sus unidades de repetición siguiendo los lineamientos de usabilidad para servicios web bioinformáticos. | 84 días | 252 horas-persona | S/.15120.00 |
| 3.3.1 | Elaborar el diagrama de ejecución del proceso del servicio integrado | 7 días | 21 horas-persona | S/.1260.00 |
| 3.3.2 | Diseñar mockups navegables para las interfaces del servicio integrado | 14 días | 42 horas-persona | S/.2520.00 |
| 3.3.3 | Desarrollar el servicio integrado | 42 días | 126 horas-persona | S/.7560.00 |
| 3.3.4 | Elaborar propuesta de casos de prueba del servicio integrado desarrollado | 7 días | 21 horas-persona | S/.1260.00 |
| 3.3.5 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 7 días | 21 horas-persona | S/.1260.00 |
| 3.3.6 | Elaborar el informe de adaptación de lineamientos y evaluación de usabilidad. | 7 días | 21 horas-persona | S/.1260.00 |

## 8. Cronograma del proyecto

A continuación, se detalla el cronograma del proyecto de tesis, especificando la fecha de inicio y la fecha de fin de cada actividad en la tabla 14. Adicionalmente, en el anexo C al documento se encuentra una hoja de cálculo detallando de manera visual la programación de cada actividad.

*Tabla 14 Cronograma del proyecto de tesis*

| ID | Actividad | Inicio | Fin |
|---|---|---|---|
| **Herramienta integrada para la curación de proteínas repetidas** | | **1-Ago** | **18-Dic** |
| **1** | **Integrar servicios de predicción de clasificación de proteínas repetidas en base a su estructura primaria y/o terciaria.** | **1-Ago** | **28-Ago** |
| 1.1 | Despliegue del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria. | 1-Ago | 7-Ago |
| 1.1.1 | Elaborar el informe de despliegue del servicio | 1-Ago | 4-Ago |
| 1.1.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 5-Ago | 7-Ago |
| 1.2 | Integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria. | 8-Ago | 14-Ago |
| 1.2.1 | Elaborar informe de integración del servicio | 8-Ago | 10-Ago |
| 1.2.2 | Elaborar propuesta de casos de prueba del servicio integrado | 11-Ago | 12-Ago |
| 1.2.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 13-Ago | 14-Ago |

| | | | |
|---|---|---|---|
| 1.3 | Despliegue del servicio de predicción de clase y subclase de proteína repetida en base a su estructura primaria. | 15-Ago | 21-Ago |
| 1.3.1 | Elaborar el informe de despliegue del servicio | 15-Ago | 18-Ago |
| 1.3.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 19-Ago | 21-Ago |
| 1.4 | Integración del servicio de predicción de clase y subclase de proteína repetida en base a su estructura primaria. | 22-Ago | 28-Ago |
| 1.4.1 | Elaborar informe de integración del servicio | 22-Ago | 24-Ago |
| 1.4.2 | Elaborar propuesta de casos de prueba del servicio integrado | 25-Ago | 26-Ago |
| 1.4.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 27-Ago | 28-Ago |
| **2** | **Predecir la estructura de una proteína basada en su secuencia de manera integrada con la predicción de clasificación.** | **29-Ago** | **11-Set** |
| 2.1 | Despliegue del servicio de predicción de estructura de una proteína basada en su secuencia. | 29-Ago | 4-Set |
| 2.1.1 | Elaborar el informe de despliegue del servicio | 29-Ago | 1-Set |
| 2.1.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 2-Set | 4-Set |
| 2.2 | Integración del servicio de predicción de estructura de una proteína basada en su secuencia. | 5-Set | 11-Set |

| | | | |
|---|---|---|---|
| 2.2.1 | Elaborar informe de integración del servicio | 5-Set | 7-Set |
| 2.2.2 | Elaborar propuesta de casos de prueba del servicio integrado | 8-Set | 9-Set |
| 2.2.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 10-Set | 11-Set |
| **3** | **Implementar un servicio web integrado para la curación de proteínas repetidas con el objetivo de identificar unidades de repetición.** | **12-Set** | **18-Dic** |
| 3.1 | Despliegue del servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición. | 12-Set | 18-Set |
| 3.1.1 | Elaborar el informe de despliegue del servicio | 12-Set | 15-Set |
| 3.1.2 | Elaborar el informe de ejecución de pruebas de la documentación del servicio | 16-Set | 18-Set |
| 3.2 | Integración del servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición. | 19-Set | 25-Set |
| 3.2.1 | Elaborar informe de integración del servicio | 19-Set | 21-Set |
| 3.2.2 | Elaborar propuesta de casos de prueba del servicio integrado | 22-Set | 23-Set |
| 3.2.3 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 24-Set | 25-Set |

| | | | |
|---|---|---|---|
| 3.3 | Desarrollo del servicio web para predecir la estructura de una proteína repetida e identificar sus unidades de repetición siguiendo los lineamientos de usabilidad para servicios web bioinformáticos. | 26-Set | 18-Dic |
| 3.3.1 | Elaborar el diagrama de ejecución del proceso del servicio integrado | 26-Set | 2-Oct |
| 3.3.2 | Diseñar *mockups* navegables para las interfaces del servicio integrado | 3-Oct | 16-Oct |
| 3.3.3 | Desarrollar el servicio integrado | 17-Oct | 27-Nov |
| 3.3.4 | Elaborar propuesta de casos de prueba del servicio integrado desarrollado | 28-Nov | 4-Dic |
| 3.3.5 | Redactar informe de ejecución de casos de prueba propuestos del servicio integrado | 5-Dic | 11-Dic |
| 3.3.6 | Elaborar el informe de adaptación de lineamientos y evaluación de usabilidad. | 12-Dic | 18-Dic |

## 9. Lista de recursos

En esta sección del documento se presentan todos los recursos requeridos para la ejecución de este proyecto.

### 9.1. Personas involucradas y necesidades de capacitación

A continuación, se describen todas las personas que están involucradas directamente en este proyecto de investigación.

- Asesora de tesis: Layla Hirsh Martinez, Ingeniera Informática. Cuenta con una maestría en Ciencias de la Computación y un doctorado en Biociencia y Biotecnología. Profesora asociada de la Pontificia Universidad Católica del Perú en el Departamento de Ingeniería. Actual Directora de Estudios de Generales Ciencias. Colabora del Instituto Europeo de Bioinformática (EMBL-EBI).

- Usuarios de servicios web bioinformáticos: Investigadores con los cuales se realizará la evaluación de usabilidad sobre el servicio integrado para la curación de proteínas repetidas
- Tesista: Estudiante de posgrado que desarrolla el proyecto de tesis. Cumple con la formación requerida para este proyecto. No obstante, se adquiere una capacitación en desarrollo *frontend*.

### 9.2. Materiales requeridos para el proyecto

Para el desarrollo de este proyecto de investigación, no se requerirán materiales.

### 9.3. Estándares utilizados en el proyecto

En este proyecto de tesis se tendrá que diseñar la interfaz para una herramienta que soportará el proceso de curación de proteínas repetidas. Para esta tarea, se trabajarán con lineamientos de usabilidad para servicios web bioinformáticos (Bezerra Brandao et al., 2021). Esto permitirá que la aplicación desarrollada presente facilidad de uso para el público objetivo.

### 9.4. Equipamiento requerido

Para este proyecto se necesitarán espacios en la nube. Se trabajará con Amazon Web Services (AWS), en donde se diseñará la arquitectura del servicio, considerando un ambiente para el desarrollo y otro para las pruebas y despliegue.

### 9.5. Herramientas requeridas

A continuación, se describen las herramientas requeridas para el desarrollo de este proyecto de tesis.

- Formularios en línea. Estos son cuestionarios en los que un público objetivo puede enviar datos por medio de la web (Jain & Karnad, 2017). Serán utilizados para la evaluación de usabilidad del servicio desarrollado.
- Herramientas para la elaboración de *mockups* navegables de la interfaz gráfica de usuario. La elaboración de prototipos de interfaces gráficas de usuario es un método de desarrollo utilizado para mejorar el planteamiento y ejecución de proyectos de software con fines experimentales (Baumer et al., 1996). Para este proyecto se utilizará Figma.
- Entornos de desarrollo Integrado (IDE). Son interfaces que facilitan la programación de una aplicación. Para el desarrollo en el *framework* Vue de

JavaScript y el *framework* Flask de Python se trabajará en Visual Studio Code con *plugins* para los *frameworks* respectivos.

- Herramientas de videoconferencia. Estas se utilizarán para compartir en tiempo real audio video y, en algunos casos, pantalla entre dos o más puntos del mundo (Dudding, 2009). Para este proyecto se utilizará la plataforma Zoom.

## 10. Costeo del Proyecto

A continuación, se presenta la especificación de la estimación de los costos del proyecto para establecer el presupuesto en la tabla 15.

*Tabla 15 Anexo B: Costeo del proyecto*

| Ítem | Descripción | Unidad | Cantidad | Valor Unidad (S/.) | Monto Parcial (S/.) | Monto Total (S/.) |
|---|---|---|---|---|---|---|
| **0** | **Costo total del proyecto** | **---** | **---** | **---** | **---** | **52,600** |
| **1.** | **Participantes del proyecto** | **---** | **---** | **---** | **---** | **33,200** |
| 1.1 | Asesora de tesis | Horas | 40[8] | 150 | 6,000 | |
| 1.2 | Usuarios de servicios web bioinformáticos | Horas | 5 | 400 | 2,000 | |
| 1.3 | Tesista | Horas | 420 | 60[9] | 25,200 | |
| **2.** | **Bienes, equipos y servicios** | **---** | **---** | **---** | **---** | **18,650** |
| 2.1 | Laptop MSI Pulse GL66 | Equipo | 1 | 9,500 | 9,500 | |
| 2.2 | Servicio de conexión a internet | Meses | 6 | 330 | 1,980 | |
| 2.3 | Nube AWS | Meses | 6 | 1,200 | 7,200 | |
| **3.** | **Licencias de software** | **---** | **---** | **---** | **---** | **750** |
| 3.1 | Google Forms | Meses | 6 | 20 | 120 | |
| 3.2 | Figma | Meses | 6 | 50 | 300 | |
| 3.3 | Zoom | Meses | 6 | 55 | 330 | |
| 3.4 | Visual Studio Code | Meses | 6 | 0 | 0 | |

[8] Considerando 2 horas semanales de asesoría para el proyecto

[9] Considerando la retribución de un Jefe de Práctica en la Pontificia Universidad Católica del Perú

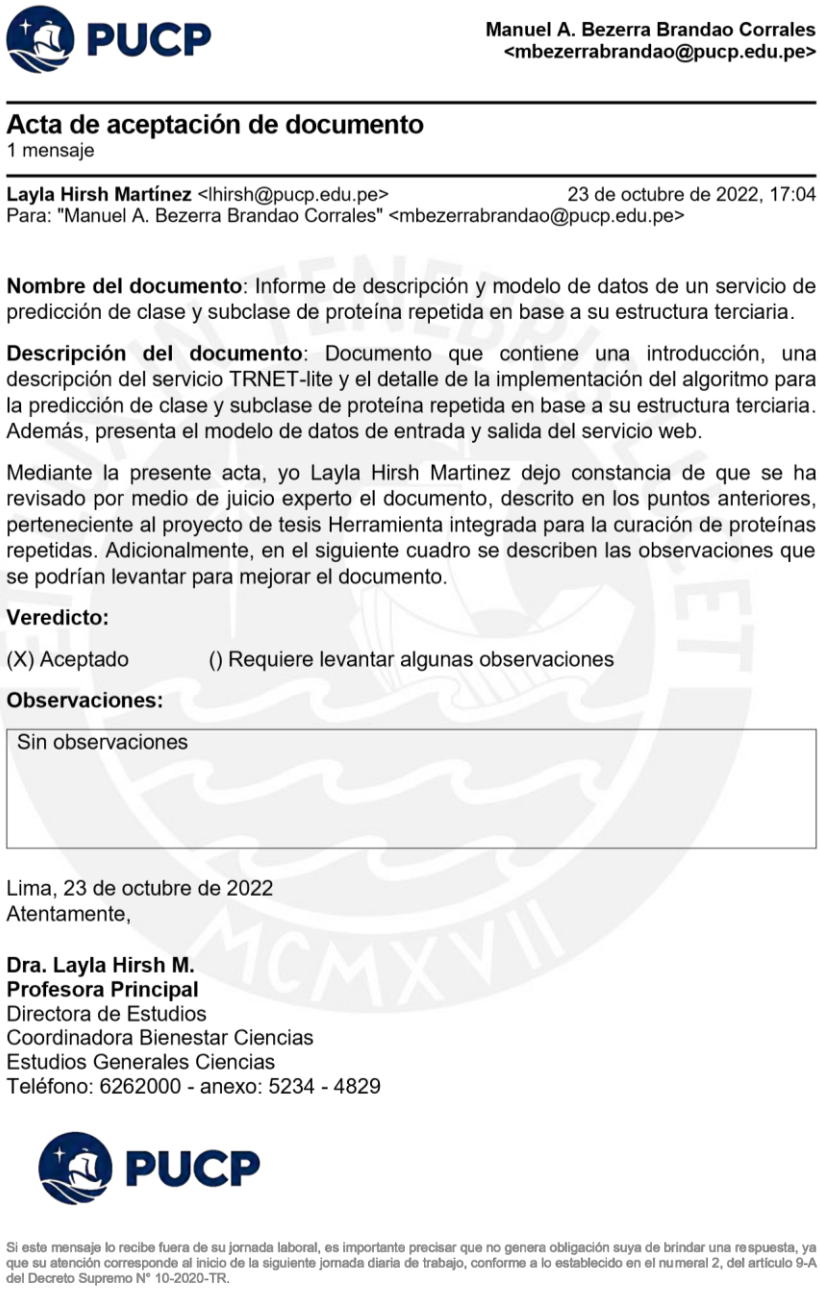
## Anexo C: Cronograma de proyecto

Este anexo contiene el cronograma del proyecto con detalle visual mencionado en el Anexo B y se encuentra en la hoja de cálculo nombrada como *20161811_ManuelBezerraBrandao_LaylaHirsh _Anexo_C.xlsx* y se puede acceder a través de la nube por medio del siguiente enlace: https://docs.google.com/spreadsheets/d/1lWqu-fnVbQpmsyU5fpVTg9m1UiT2pAP_/edit?usp=sharing&ouid=115633799306253091037&rtpof=true&sd=true

## Anexo D: Actas de revisión mediante juicio experto

En este anexo se presentan las actas de revisión y validación de resultados mediante juicio experto en Bioinformática. Estas pertenecen a los resultados esperados que cuentan con un indicador objetivamente verificable relacionada a la aceptación mediante juicio experto.

1. **Acta de aceptación del informe de descripción y modelo de datos de un servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria**



PUCP

**Manuel A. Bezerra Brandao Corrales**
**<mbezerrabrandao@pucp.edu.pe>**

**Acta de aceptación de documento**
1 mensaje

**Layla Hirsh Martínez** <lhirsh@pucp.edu.pe> 23 de octubre de 2022, 17:04
Para: "Manuel A. Bezerra Brandao Corrales" <mbezerrabrandao@pucp.edu.pe>

**Nombre del documento**: Informe de descripción y modelo de datos de un servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria.

**Descripción del documento**: Documento que contiene una introducción, una descripción del servicio TRNET-lite y el detalle de la implementación del algoritmo para la predicción de clase y subclase de proteína repetida en base a su estructura terciaria. Además, presenta el modelo de datos de entrada y salida del servicio web.

Mediante la presente acta, yo Layla Hirsh Martinez dejo constancia de que se ha revisado por medio de juicio experto el documento, descrito en los puntos anteriores, perteneciente al proyecto de tesis Herramienta integrada para la curación de proteínas repetidas. Adicionalmente, en el siguiente cuadro se describen las observaciones que se podrían levantar para mejorar el documento.

**Veredicto:**

(X) Aceptado () Requiere levantar algunas observaciones

**Observaciones:**

| |
|---|
| Sin observaciones |

Lima, 23 de octubre de 2022
Atentamente,

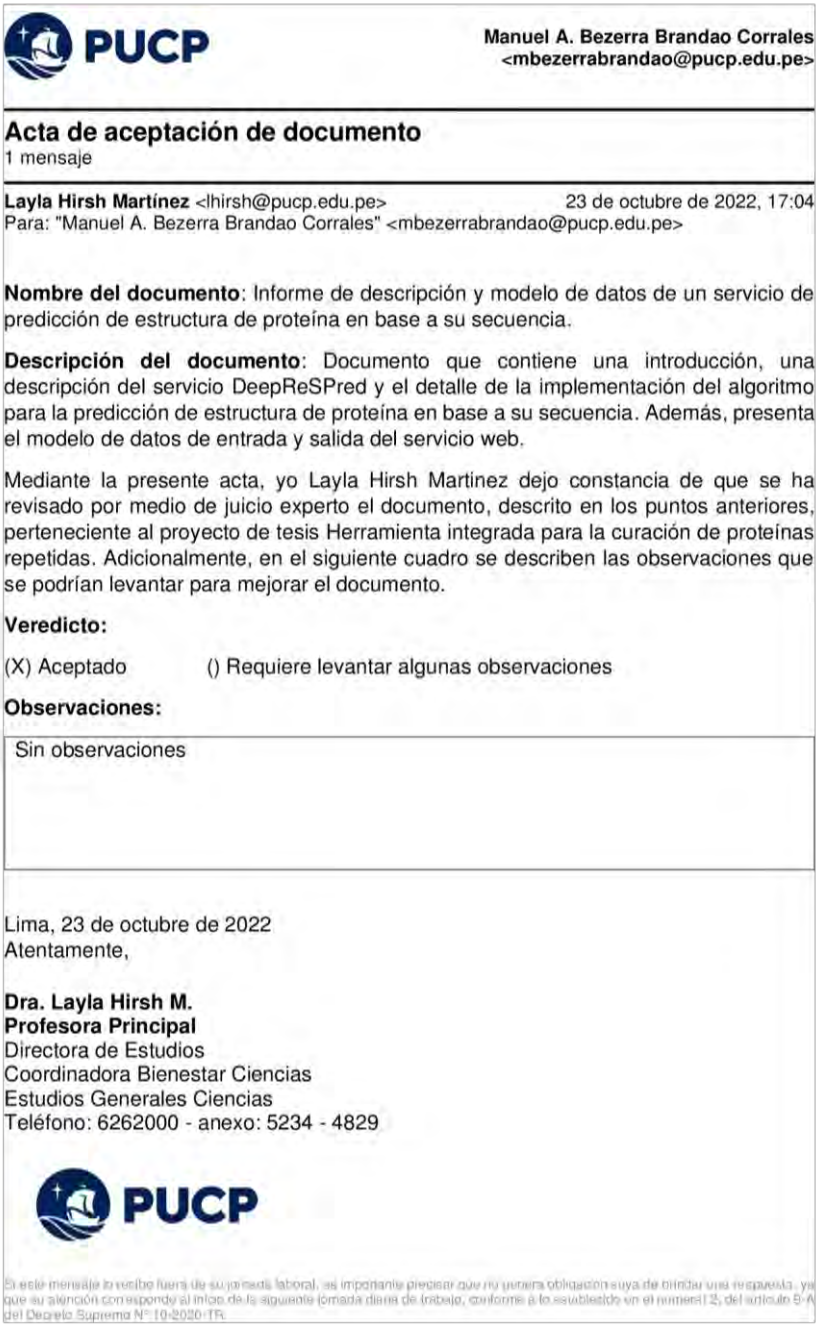
**Dra. Layla Hirsh M.**
**Profesora Principal**
Directora de Estudios
Coordinadora Bienestar Ciencias
Estudios Generales Ciencias
Teléfono: 6262000 - anexo: 5234 - 4829

PUCP

Si este mensaje lo recibe fuera de su jornada laboral, es importante precisar que no genera obligación suya de brindar una respuesta, ya que su atención corresponde al inicio de la siguiente jornada diaria de trabajo, conforme a lo establecido en el numeral 2, del artículo 9-A del Decreto Supremo N° 10-2020-TR.

*Figura 16 Anexo D: Acta de aceptación del informe de descripción y modelo de datos de un servicio de predicción de clase y subclase de proteína repetida en base a su estructura terciaria*

2. **Acta de aceptación del informe de descripción y modelo de datos de un servicio de predicción de estructura de proteína en base a su secuencia**



PUCP

Manuel A. Bezerra Brandao Corrales
<mbezerrabrandao@pucp.edu.pe>

**Acta de aceptación de documento**
1 mensaje

**Layla Hirsh Martínez** <lhirsh@pucp.edu.pe> 23 de octubre de 2022, 17:04
Para: "Manuel A. Bezerra Brandao Corrales" <mbezerrabrandao@pucp.edu.pe>

**Nombre del documento**: Informe de descripción y modelo de datos de un servicio de predicción de estructura de proteína en base a su secuencia.

**Descripción del documento**: Documento que contiene una introducción, una descripción del servicio DeepReSPred y el detalle de la implementación del algoritmo para la predicción de estructura de proteína en base a su secuencia. Además, presenta el modelo de datos de entrada y salida del servicio web.

Mediante la presente acta, yo Layla Hirsh Martinez dejo constancia de que se ha revisado por medio de juicio experto el documento, descrito en los puntos anteriores, perteneciente al proyecto de tesis Herramienta integrada para la curación de proteínas repetidas. Adicionalmente, en el siguiente cuadro se describen las observaciones que se podrían levantar para mejorar el documento.

**Veredicto:**

(X) Aceptado () Requiere levantar algunas observaciones

**Observaciones:**

Sin observaciones

Lima, 23 de octubre de 2022
Atentamente,

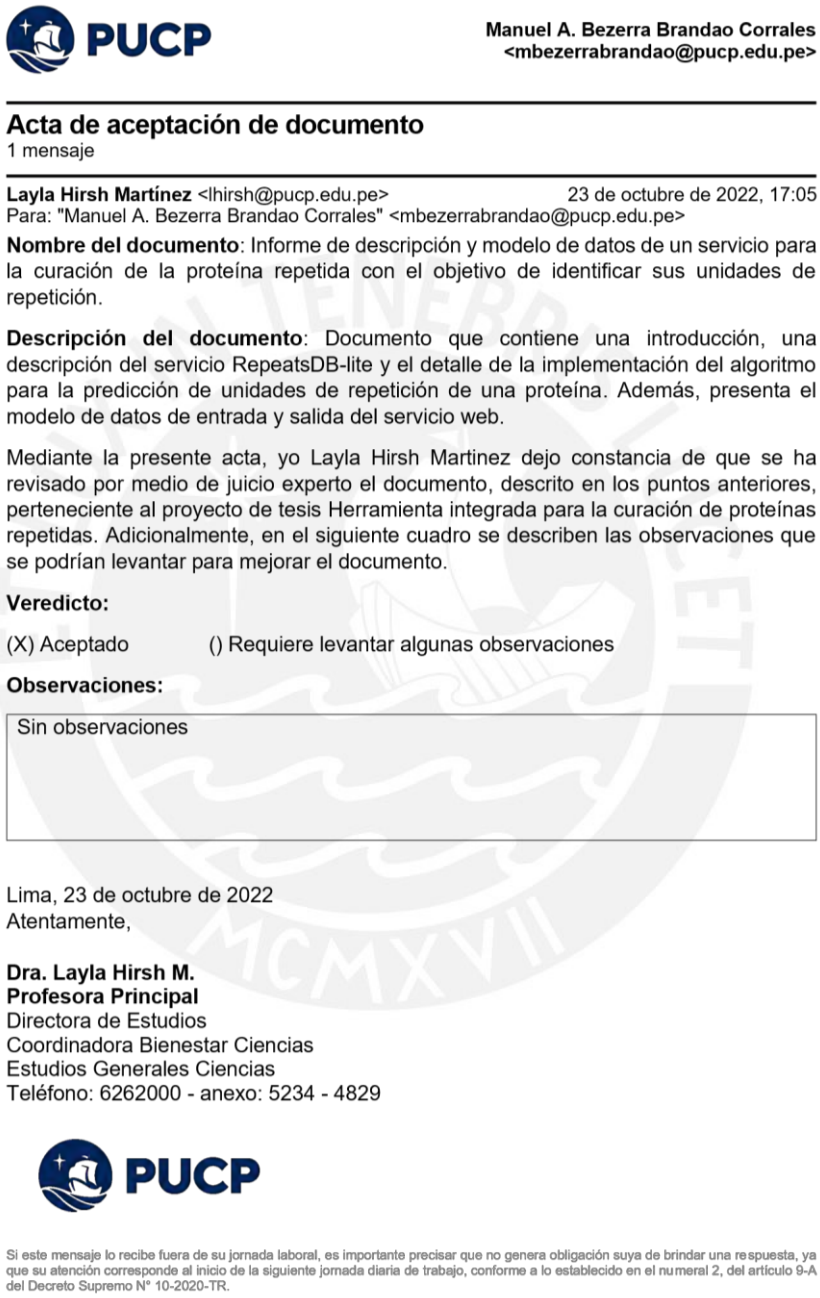
**Dra. Layla Hirsh M.**
**Profesora Principal**
Directora de Estudios
Coordinadora Bienestar Ciencias
Estudios Generales Ciencias
Teléfono: 6262000 - anexo: 5234 - 4829

PUCP

Si este mensaje lo recibe fuera de su jornada laboral, es importante precisar que no tendrá obligación suya de brindar una respuesta, ya que su atención corresponde al inicio de la siguiente jornada diaria de trabajo, conforme a lo establecido en el numeral 2, del artículo 9-A del Decreto Supremo N° 10-2020-TR.

*Figura 17 Anexo D: Acta de aceptación del informe de descripción y modelo de datos de un servicio de predicción de estructura de proteína en base a su secuencia*

3. **Acta de aceptación del informe de descripción y modelo de datos de un servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición.**



PUCP

**Manuel A. Bezerra Brandao Corrales**
**<mbezerrabrandao@pucp.edu.pe>**

**Acta de aceptación de documento**
1 mensaje

**Layla Hirsh Martínez** <lhirsh@pucp.edu.pe> 23 de octubre de 2022, 17:05
Para: "Manuel A. Bezerra Brandao Corrales" <mbezerrabrandao@pucp.edu.pe>

**Nombre del documento**: Informe de descripción y modelo de datos de un servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición.

**Descripción del documento**: Documento que contiene una introducción, una descripción del servicio RepeatsDB-lite y el detalle de la implementación del algoritmo para la predicción de unidades de repetición de una proteína. Además, presenta el modelo de datos de entrada y salida del servicio web.

Mediante la presente acta, yo Layla Hirsh Martinez dejo constancia de que se ha revisado por medio de juicio experto el documento, descrito en los puntos anteriores, perteneciente al proyecto de tesis Herramienta integrada para la curación de proteínas repetidas. Adicionalmente, en el siguiente cuadro se describen las observaciones que se podrían levantar para mejorar el documento.

**Veredicto:**

(X) Aceptado () Requiere levantar algunas observaciones

**Observaciones:**

Sin observaciones

Lima, 23 de octubre de 2022
Atentamente,

**Dra. Layla Hirsh M.**
**Profesora Principal**
Directora de Estudios
Coordinadora Bienestar Ciencias
Estudios Generales Ciencias
Teléfono: 6262000 - anexo: 5234 - 4829

PUCP

Si este mensaje lo recibe fuera de su jornada laboral, es importante precisar que no genera obligación suya de brindar una respuesta, ya que su atención corresponde al inicio de la siguiente jornada diaria de trabajo, conforme a lo establecido en el numeral 2, del artículo 9-A del Decreto Supremo N° 10-2020-TR.

*Figura 18 Anexo D: Acta de aceptación del informe de descripción y modelo de datos de un servicio para la curación de la proteína repetida con el objetivo de identificar sus unidades de repetición.*

## Anexo E: Mockups navegables del diseño de interfaz para el servicio web integrado para la curación de proteínas repetidas

En este anexo, se presenta el enlace para poder revisar los mockups navegables que muestran el diseño propuesto para la interfaz del servicio web integrado para la curación de proteínas repetidas.

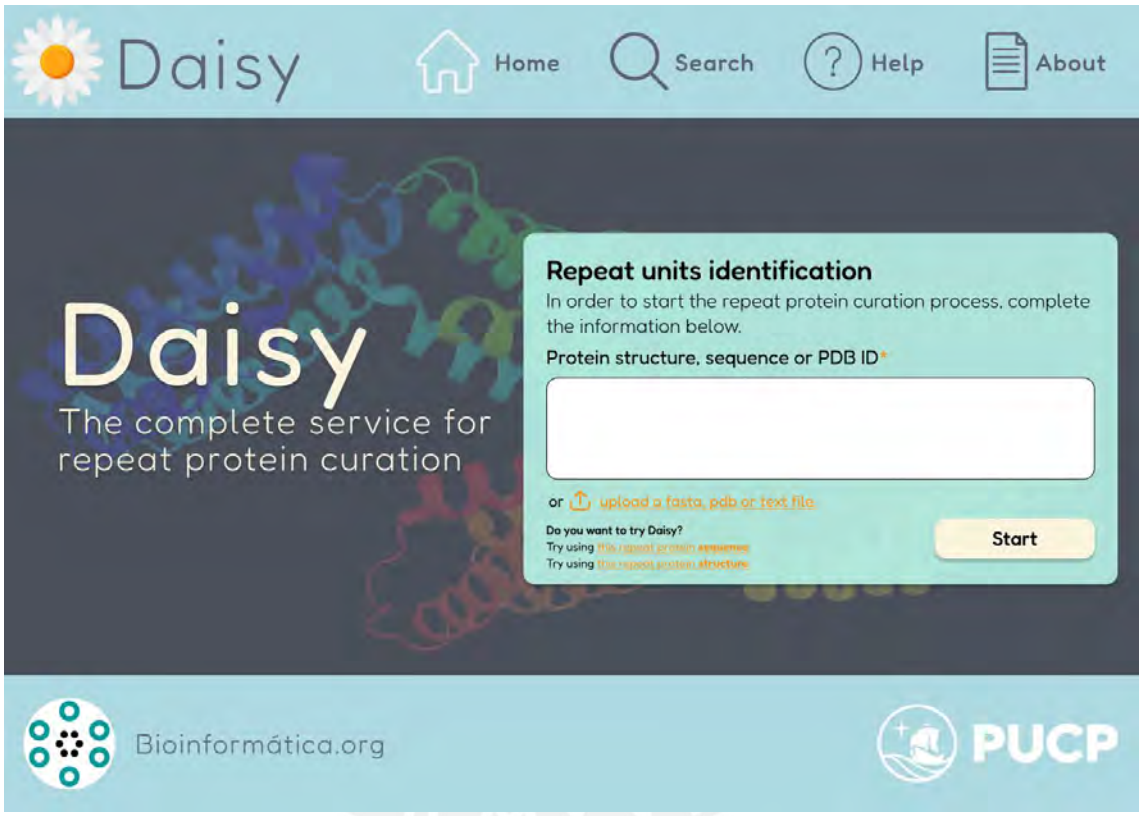



*Figura 19 Anexo E: Diseño de la pantalla principal del servicio web integrado para la curación de proteínas repetidas*

Enlace: https://www.figma.com/proto/vCzZqMIxCbOOA4o1thxszC/Mockups?page-id=0%3A1&node-id=1%3A5&viewport=-271%2C574%2C0.27&scaling=scale-down&starting-point-node-id=1%3A5

## Anexo F: Repositorios de código fuente de los componentes del servicio desarrollado

A continuación, se presentan los enlaces a los repositorios en GitLab de los componentes del servicio web integrado para la curación de proteínas repetidas. Estos son de acceso restringido. Para obtener acceso a alguno de estos componentes por favor escribir a mbezerrabrandao@pucp.edu.pe.

**Componente TRNET-lite:**

https://gitlab.com/daisy-repeatproteincurator/trnet-lite

**Componente Structures Downloader:**

https://gitlab.com/daisy-repeatproteincurator/alphafold-downloader

**Componente RepeatsDBLite 2.0:**

https://gitlab.com/daisy-repeatproteincurator/repeatsdblite2.0

**Componente Daisy back-end:**

https://gitlab.com/daisy-repeatproteincurator/daisy-backend

## Anexo G: Documentación del servicio API desarrollado para la herramienta integrada para la curación de proteínas repetidas

A continuación, se presenta la documentación del servicio API desarrollado para la herramienta integrada para la curación de proteínas repetidas. Este servicio ha sido desarrollado en e *framework* Flask de Python.

**Enviar un request [POST]**

URL: daisy-back.url/request

Body: {

“proteinID”:String //Es el identificador PDB o AlphaFold (Uniprot) de la proteína a procesar

“email”:String //Es el correo del usuario que hace la solicitud

}

Response: {

“isReady”: Boolean //Falso si recién será procesada la proteína, verdadero si ya se encuentran listos los resultados

“requestID”: Integer // ID en base de datos de la solicitud que el usuario debe tener para ver sus resultados.

“result”: Boolean //Verdadero si se insertó correctamente el request en base de datos

“proteinResult”: Protein Json //Solo llegará si isReady es verdadero. Ver documentación más adelante

}

**Consultar un request [GET]**

URL: daisy-back.url/request/<requestID>

Params:

- requestID: Integer. Es el identificador en base de datos de la solicitud del usuario

Response: {

“isReady”: Boolean //Falso si aún no termina el procesamiento de la proteína, verdadero si ya se encuentran listos los resultados

“proteinResult”: Protein Json //Solo llegará si isReady es verdadero. Ver documentación más adelante

}

### Clases de JSON

#### Protein JSON

Todos los resultados de la proteína procesadas se podrán leer desde el Protein Json. A continuación se detalla el contenido.

{

“id”: String //Identificador de la proteína (Sea PDB o UNIPROT)

“type”: String //Tipo de proteína (“PDB”, “AlphaFold” o “ERROR”), si es del tipo Error, mostrar que el identificador solicitado no es válido o que no fue posible procesar esta proteína.

“isRepeat”: Boolean //Verdadero si la proteína tiene al menos una cadena con regiones repetidas.

“isProcessed”: Boolean //Verdadero si la cadena fue procesada, debería ser siempre verdadero.

“chains”:[Chain Json] //Lista de Cadenas solo si isRepeat es verdadero. Ver documentación más adelante

}

#### Chain JSON

Todas las cadenas de una proteína estarán presentes en el Json, sean repetidas o no. Es importante mostrar la predicción de clases repetidas.

{

“name”: Char //Nombre de la cadena

“isRepeat”: Boolean //Verdadero si la cadena tiene regiones repetidas

“classPrediction”: {

“III_1”:Float //Probabilidad predicha para la clase y subclase

“III_2”:Float //Probabilidad predicha para la clase y subclase

“III_3”:Float //Probabilidad predicha para la clase y subclase

“III_4”:Float //Probabilidad predicha para la clase y subclase

“III_5”:Float //Probabilidad predicha para la clase y subclase

“III_6”:Float //Probabilidad predicha para la clase y subclase

“IV_1”:Float //Probabilidad predicha para la clase y subclase

“IV_2”:Float //Probabilidad predicha para la clase y subclase

“IV_3”:Float //Probabilidad predicha para la clase y subclase

“IV_4”:Float //Probabilidad predicha para la clase y subclase

“IV_5”:Float //Probabilidad predicha para la clase y subclase

“IV_6”:Float //Probabilidad predicha para la clase y subclase

“IV_7”:Float //Probabilidad predicha para la clase y subclase

“IV_8”:Float //Probabilidad predicha para la clase y subclase

“IV_9”:Float //Probabilidad predicha para la clase y subclase

“IV_10”:Float //Probabilidad predicha para la clase y subclase

“V_1”:Float //Probabilidad predicha para la clase y subclase

“V_2”:Float //Probabilidad predicha para la clase y subclase

“V_3”:Float //Probabilidad predicha para la clase y subclase

“V_4”:Float //Probabilidad predicha para la clase y subclase

“V_5”:Float //Probabilidad predicha para la clase y subclase

}

“regions”: [Region Json] //Lista de regiones repetidas de la cadena, solo si isRepeat es verdadero. Ver documentación más adelante

**Region JSON**

Todas las regiones identificadas por RepeatsDBLite para una cadena. Con toda la información se podrán descargar los archivos resultantes.

{

“repeatClass”:String //Clase de proteína repetida

“repeatSubclass”:String //Subclase de proteína repetida

“classRegionNumber”:Int //Número de región repetida para la clase y subclase. Cabe mencionar que puede haber n regiones de la misma clase. Asimismo, podría haber diferentes regiones de diferentes clases. Todas estarán contenidas una por una en la lista de la cadena.

**Obtener archivo PDB de proteína (PDB) [GET]**

URL: daisy-back.url/file/<proteinID>/pdb

Params:

- proteinID: String. Es el identificador de la proteína PDB

**Obtener archivo CIF de proteína (AlphaFold) [GET]**

URL: daisy-back.url/file/<proteinID>/cif

Params:

- proteinID: String. Es el identificador de la proteína Uniprot

**Obtener archivo PDB de cadena [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/pdb

Nota: Para este caso, usar la clase y subclase de la primera región de la cadena. Se puede tener una bandera de “First Region” para cada cadena para obtener los archivos generales de la candena.

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”
- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena

**Obtener archivo DB de cadena [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/db

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”
- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena

**Obtener archivo Mapping de cadena [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/mapping

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”
- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena

**Obtener archivo ZIP de unidades PDB de región [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/<regionNumber>/units

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”
- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena
- regionNumber: Integer. Es el número de la región para la clase y subclase de proteína repetida de la cadena

**Obtener archivo PDB de unidades alineadas de región [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/<regionNumber>/pdb

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”
- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena
- regionNumber: Integer. Es el número de la región para la clase y subclase de proteína repetida de la cadena

**Obtener archivo matriz de alineamiento de región [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/<regionNumber>/matrix

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”

- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena
- regionNumber: Integer. Es el número de la región para la clase y subclase de proteína repetida de la cadena

**Obtener archivo de alineamiento Fasta de región [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/<regionNumber>/afasta

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”
- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena
- regionNumber: Integer. Es el número de la región para la clase y subclase de proteína repetida de la cadena

**Obtener archivo de alineamiento DSSP de región [GET]**

URL: daisy-back.url/file/<proteinID>/<proteinType>/<chainName>/<repeatClass>/<repeatSubclass>/<regionNumber>/dssp

Params:

- proteinID: String. Es el identificador de la proteína Uniprot o PDB
- proteinType: String. Es el tipo de proteína “PDB” o “AlphaFold”
- chainName: Char. Es el nombre de la cadena
- repeatClass: String. Es la clase repetida de la cadena
- repeatSubclass: String. Es la subclase repetida de la cadena
- regionNumber: Integer. Es el número de la región para la clase y subclase de proteína repetida de la cadena