FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA CIVIL EN BIOINFORMÁTICA

PREDICCIÓN DE SITIOS DE UNIÓN AL ADN EN LA SUPERFAMILIA FUR MEDIANTE INTELIGENCIA ARTIFICIAL Y DESCRIPTORES MOLECULARES.

JESSICA FERNANDA LARA MUÑOZ

Profesor Tutor: Mauricio Arenas Salinas

Profesor Co-Tutor: José Antonio Reyes

Profesor Informante: Jans Alzate-Morales

Memoria para optar al título de Ingeniera Civil en Bioinformática

30 de septiembre, 2022

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su unidad de procesos técnicos certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Talca, 2022

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA CIVIL EN BIOINFORMÁTICA

PREDICCIÓN DE SITIOS DE UNIÓN AL ADN EN LA SUPERFAMILIA FUR MEDIANTE INTELIGENCIA ARTIFICIAL Y DESCRIPTORES MOLECULARES.

JESSICA FERNANDA LARA MUÑOZ

Mauricio Arenas Salinas : _____
Profesor Tutor

José Antonio Reyes : _____
Profesor Co-Tutor

Jans Alzate-Morales : _____
Profesor Informante

Memoria para optar al título de Ingeniera Civil en Bioinformática

30 de septiembre, 2022

Financiamiento

Esta Memoria de Título fue financiada mediante el Proyecto FONDECYT 11180665, con el título “Identification and characterization of the regulatory iron binding site from the acidithiobacillus ferrooxidans transcription factor Fur”.

ÍNDICE

ÍNDICE 4

ÍNDICE DE FIGURAS 6

INDICE DE TABLAS 7

RESUMEN 8

ABSTRACT 9

INTRODUCCIÓN 10

1. Metaloproteínas 10
1.1 Superfamilia Fur 10
1.2 Tipos de proteínas Fur 11
1.2.1 Ferric uptake regulator (Fur) 12
1.2.2 Zinc uptake regulator (Zur) 12
1.2.3 Peroxide stress sensing Regulator (PerR) 13
1.2.4 Manganese uptake regulator (Mur) 14
1.2.5 Niquel-responsive regulator (Nur) 14
1.3 Sitio de unión al ADN 15

2. Predicción de sitios de unión al ADN 19
2.1 Machine Learning 20
2.2 Métodos de predicción basados en machine learning 20

3. Descriptores Moleculares 22
3.1 Basados en secuencia aminoacídica 23
3.2 Basados en estructura tridimensional 24

HIPÓTESIS Y OBJETIVOS 26

1. Hipótesis 26

2. Objetivo general 26

3. Objetivos específicos 26

MATERIALES Y MÉTODOS 27

1. Construcción del set de datos 27
1.1 Descargar estructuras de proteínas unidas al ADN. 28
1.2 Eliminación de redundancia del set de datos 29
1.3 Selección de fragmentos con y sin capacidad de unión al ADN. 29

2. Caracterización del set de datos 30
2.1 Descriptores de secuencia aminoacídica 31
2.2 Descriptores de estructura tridimensional 32
2.3 Preprocesamiento y limpieza de datos 33

3. Desarrollo de los modelos de predicción 33
3.1 Selección de características 33

3.2 Modelos de clasificación SVM Y RF 34

RESULTADOS ***37***

1. Resultados objetivo específico 1: Construir un set de datos que contenga información de secuencia aminoacídica y/o de estructural de los fragmentos de unión al ADN en proteínas de la superfamilia Fur y otros factores de transcripción**.** **37**

1.1 Estructuras utilizadas en el estudio. **37**

2. Resultados objetivo específico 2: Caracterizar el set de datos con descriptores moleculares a nivel de secuencia aminoacídica y estructural**.** **40**

3. Resultados objetivo específico 3: Evaluar modelos Máquinas de Vectores de Soporte y Random Forest para su aplicación en la predicción de sitios de unión al ADN**.** **41**

3.1 Selección de características y aplicación de modelos. 41

3.1.1 Atributos de secuencia 42

3.1.2 Atributos de estructura 46

DISCUSIÓN ***50***

CONCLUSIONES ***54***

REFERENCIAS ***56***

ANEXOS ***64***

1. Anexo 1: Ranking de atributos de secuencia. **64**

2. Anexo 2: Selección de atributos de secuencia. **67**

3. Anexo 3: Ranking de atributos de estructura. **68**

4. Anexo 4: Selección de atributos de secuencia. **70**

ÍNDICE DE FIGURAS

Figura 1. Representación del modelo estructural de una proteína Fur dímerica. _____ *11*

Figura 2. Procesos celulares modulados por reguladores de absorción férricos. _____ *15*

Figura 3. Representación de la estructura MRS2-Fur (código PDB: 4RB2). _____ *18*

Figura 4. Número de complejos proteína-ADN liberadas en PDB cada año. _____ *19*

Figura 5. Número de estructuras proteicas liberadas en PDB cada año. _____ *25*

Figura 6. Diagrama de la metodología general que se aplicará en esta investigación. ____ *27*

Figura 7. Estructuras utilizadas en la investigación. _____ *37*

Figura 8. Alineamiento estructural de la zona de unión al ADN. _____ *39*

Figura 9. Comparación estructural de zona de unión. _____ *40*

Figura 10. Atributos de secuencia acumulados. _____ *43*

Figura 11. Rendimiento de modelos SVM y RF en atributos de secuencia. _____ *45*

Figura 12. Atributos de estructura acumulados. _____ *46*

Figura 13. Rendimiento de modelos SVM y RF en atributos estructurales. _____ *48*

INDICE DE TABLAS

Tabla 1. Matriz de confusión. _____ *35*

Tabla 2. Matriz de distancia p de aminoácidos. _____ *38*

Tabla 3. Resultado de los modelos SMV y RF. _____ *48*

RESUMEN

La Superfamilia Fur se compone de proteínas reguladoras de absorción de iones, estas poseen una similitud funcional y capacidad de unión al ADN, varían según el cofactor metálico que utilicen dentro de las cuales podemos encontrar del tipo Fur, Zur, Mur, Nur y PerR. El estudio del sitio de unión en proteínas de la Superfamilia Fur y factores de transcripción es un tema de gran interés debido a que este tipo de proteínas pueden ser potencialmente utilizadas en desarrollos y aplicaciones antimicrobianas. En este estudio se caracterizaron 63 sitios de unión al ADN de organismos bacterianos pertenecientes a Fur y diversos factores de transcripción, mediante descriptores moleculares de secuencia y estructura. Utilizando técnicas de inteligencia artificial se realizó la identificación de características relevantes para entrenar los modelos predictivos SVM y RF. Del análisis realizado se determinó que el mejor modelo entrenado con características de secuencia es SVM logrando un rendimiento de exactitud (Accuracy) de 84.6%, mientras que utilizando características estructurales el modelo RF es el que logra un mejor rendimiento con una exactitud (Accuracy) de 77.3%. Los resultados de la predicción indican que es posible desarrollar un modelo predictivo de zonas de unión al ADN en proteínas con una alta precisión utilizando solo la información de secuencia.

ABSTRACT

The Fur superfamily is composed of ion absorption regulatory proteins. These have functional similarities and DNA binding capacity. The diversity depends on the metal cofactor used, among which they can find Fur, Zur, Mur, Nur, and PerR types. The study of metal binding site proteins in the Fur superfamily and transcription factors is a very interesting topic because these types of proteins can potentially be used in antimicrobial development and its applications. In this investigation, 63 DNA-binding sites of bacterial organisms belonging to the Fur family and various transcription factors were characterized by molecular descriptors of sequence and structure. The use of intelligent artificial intelligence techniques made it possible to identify relevant characteristics for training SVM and RF predicting models. The performed analysis determined that the best predicting model using sequence characteristics is SVM, achieving a performance of 84.6% accuracy while using structural characteristics the RF model achieved the best performance with 77.3% accuracy. The results of the prediction indicate that it is possible to develop a DNA-binding site predictive model in proteins with high accuracy using only sequence information.

INTRODUCCIÓN

1. Metaloproteínas

Las metaloproteínas son aquellas que presentan un ion metálico como cofactor, son diversas y varían según el metal que utilicen. Estas proteínas son importantes para las funciones biológicas de las células, teniendo implicancia en reacciones químicas, funciones estructurales y reguladoras (Yan Zhang & Zheng, 2020). El cofactor presente en este tipo de proteínas ya sea iones metálicos, no metálicos o moléculas orgánicas es lo que le da una diversidad de funciones, siendo los cofactores metálicos los que confieren cambios conformacionales, estabilidad en su estructura y plegamiento, eventos de catálisis nucleofílica o transferencia de electrones (Akcapinar & Sezerman, 2017).

1.1 Superfamilia Fur

La superfamilia Fur se compone de proteínas no homólogas que poseen similitud funcional, siendo la proteína Fur, o ferric uptake regulator, la principal dentro de esta familia de cofactores de transcripción en procariotas. Éstas proteínas se encargan de regular la captación y concentración de iones metálicos dentro de la célula (Pohl et al., 2003). Los microorganismos necesitan de un mecanismo de regulación de iones metálicos debido a que la carencia o exceso de éstos es negativo para su fisiología. Las proteínas metaloreguladoras son las encargadas de mantener la estabilidad y homeostasis de la célula ante diferentes cofactores metálicos (Gilston et al., 2014).

El hierro es uno de los cofactores principales para la vida celular, ya que con este elemento se activan enzimas esenciales. Sin embargo, un exceso de hierro en la célula puede ser letal para su funcionamiento, es por esto que proteínas como la Fur que tienen afinidad por el hierro se encargan de controlar la absorción de éste metal (Lee & Helmann, 2007).

La familia de metaloproteínas reguladoras Fur, en general, presentan un motivo conservado, rico en histidinas, que se ubica entre la región del extremo (C-terminal) y que está encargado de la detección de metales, y el dominio de unión al ADN (Pinochet-Barros & Helmann, 2018). Diferencias en el sitio de unión a metales es lo que ha dado selectividad a los miembros de esta familia entre los que se encuentran afinidad por hierro (Fur), zinc (Zur), manganeso (Mur), níquel (Nur) y regulador de peróxido (Per).

1.2 Tipos de proteínas Fur

Las proteínas Fur son complejos proteicos que son factores de transcripción, ya que se unen a una región específica del ADN promoviendo o impidiendo la función de la ARN polimerasa. Están presentes en procariotas como dímeros o tetrámeros (Agriesti et al., 2014) y una de las proteínas más estudiadas dentro de esta familia de factores de transcripción es Fur (Figura 1).

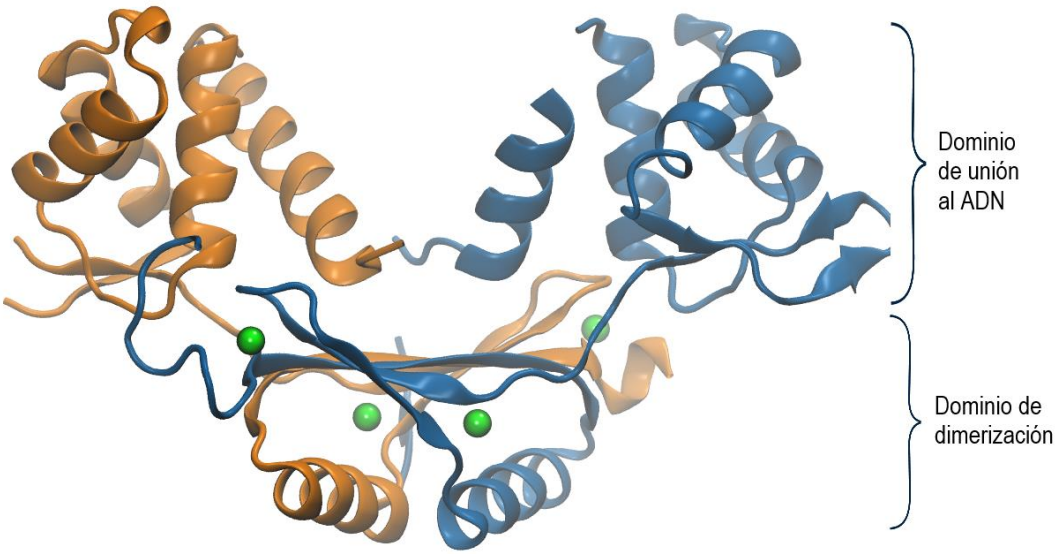

Figura 1. Representación del modelo estructural de una proteína Fur dímerica. Estructura cristalina de la proteína de código 4ETS, presente en el organismo *Campylobacter jejuni*. Los monomeros A y B se representan en café y azul, respectivamente. Los cofactores metálicos Zn^{2+} se encuentran en los sitios S1 (inferior) y S3 (superior), representados en esferas verde. Mientras que el sitio regulador S2 se encuentra vacío en esta representación estructural (Butcher et al., 2012).

1.2.1 Ferric uptake regulator (Fur)

El regulador global de la homeostasis del hierro (Fur), estructuralmente es un dímero que presenta en cada uno de sus monómeros un dominio de unión al ADN (N-terminal) y un dominio de dimerización (C-terminal). En cada monómero generalmente se encuentran dos sitios de unión a metales (Pohl et al., 2003), aunque también se han reportado tres sitios de unión a metales en organismos como *H. pylori* (Dian et al., 2011). El mecanismo de regulación consiste en que ante un exceso de hierro en la célula, Fur se une al metal Fe^{2+} y esto produce un cambio conformacional que causa la unión al ADN (Kaushik et al., 2016). Por el contrario, cuando existe una concentración menor de hierro, el metal Fe^{2+} se libera de Fur, el complejo se separa del ADN y permite la unión de la ARN polimerasa para producir transcripción (Kaushik et al., 2016).

En base a estudios y datos experimentales, las proteínas Fur se pueden clasificar en subclases con distintas funciones, una primera subclase de proteínas Fur se encargaría principalmente del control de homeostasis de hierro, tolerancia al ácido y protección contra el estrés oxidativo. Luego, una segunda subclase se vería implicada en la expresión de genes, como respuesta de las bacterias ante el estrés oxidativo, finalmente se encontraría la subclase de proteínas homólogas a Fur, como por ejemplo, la denominada proteína reguladora de la absorción de zinc (Zur), que se encuentran en organismos como *Eschericia Coli* (Pohl et al., 2003).

1.2.2 Zinc uptake regulator (Zur)

Zur es uno de los primeros homólogos de Fur identificados en *E. Coli*, y, al igual que Fur está implicado en la regulación de genes que desempeñan un papel importante no solo como regulador global, sino también en el almacenamiento y movilización de metales. Por ejemplo, tiene la capacidad de reprimir genes de proteínas ribosómicas homólogas (Panina et al., 2003). Zur contiene dos sitios de unión a metales (igual que PA-Fur, Pohl et al., 2003), y cuando contiene solo un Zn^{2+} en un sitio de unión a metal no logra

la unión al ADN, pero al existir un exceso de zinc y tener los dos sitios con iones metálicos se produce el cambio en la estructura y evita la unión de la ARN polimerasa (Lee & Helmann, 2007).

1.2.3 Peroxide stress sensing Regulator (PerR)

PerR es el principal regulador ante el estrés por peróxido, y la existencia de radicales hidroxilo o aniones de superóxido en las bacterias puede provocar muerte celular para combatir este efecto las bacterias han desarrollado mecanismos de protección mediante reguladores de transcripción (Makthal et al., 2013). PerR es un dímero con una topología similar a la estructura de la familia Fur, siendo el mecanismo que utiliza para detectar cambios ambientales (Fillat, 2014) una de las principales diferencias encontradas entre PerR y Fur. Estructuralmente PerR se diferencia en el brazo N-terminal, que contiene 11 aminoácidos, esta región juega un papel importante en la unión a metales y detección del peróxido, además no está presente en otras proteínas de la familia Fur (Makthal et al., 2013). En cuanto a los sitios de unión, PerR puede poseer dos iones metálicos en cada monómero, presentando en el sitio 1 el metal Zn^{2+} y el sitio 2, que está encargado de la detección de metales en medios abundantes en hierro, se puede unir a Fe^{2+} (Per: Zn, Fe) respondiendo al estrés oxidativo. Mientras tanto, en un exceso de Mn^{2+} se une a éste metal (Per: Zn, Mn), y este tipo de unión depende de un promotor particular (Pinochet-Barros & Helmann, 2018).

Existen organismos bacterianos que pueden poseer más de un homólogo de la familia Fur para su regulación como es el ejemplo de *Bacillus subtilis* que presenta a los parálogos Zur, Fur y Per, donde Zur se encarga de reprimir la transcripción de genes que codifican al sistema de transporte ABC de zinc o casete de unión ATP (Bsat et al., 1998). Se ha encontrado cooperatividad entre reguladores transcripcionales como Fur-Per, Zur-Per, Fur-Zur entre otros. La relación entre el metabolismo del hierro, regulación de genes y generación de radicales libres, conlleva a una cooperatividad

necesaria entre Fur y Per para responder ante el estrés oxidativo (Fillat, 2014).

1.2.4 Manganese uptake regulator (Mur)

Dentro de la familia de proteínas también encontramos las reguladoras de absorción de manganeso (Mur). Éstas se encuentran en el citoplasma y en condiciones de exceso de Mn^{2+}, reprimen la transcripción del operón sitABCD, pero no se ha observado represión en condiciones de exceso de Fe^{2+} (Bellini & Hemmings, 2006). Estructuralmente en Mur falta un sitio de unión de Zn^{2+}, lo que se modifica por la unión de dos iones Mn^{2+} por dímero (Lee & Helmann, 2007). Las reacciones que involucran manganeso o níquel son menores en comparación a hierro y zinc, sin embargo, estas son importantes ya que, el níquel está presente en varias enzimas (ureasas, deshidrogenasas, metil reductasa etc.) y es necesario para la oxidación de hidrógeno molecular.

1.2.5 Niquel-responsive regulator (Nur)

Un exceso de níquel puede provocar reacciones adversas, como la generación de especies reactivas de oxígeno o actuar como un potente carcinógeno en humanos (Ahn et al., 2006). Para mantener la homeóstasis de este metal existe el regulador de absorción a Níquel (Nur), el cual tiene una similitud de secuencia con otros miembros de la familia Fur, en un 48% con Fur del microorganismo *E. coli*. Por otro lado, tiene 4 de los 5 aminoácidos conservados en el sitio de unión a metales en comparación con PerR y el quinto se reemplaza por His (histidina), que sería el aminoácido que le confiere la selectividad por Ni^{2+} (Lee & Helmann, 2007). Esta proteína regula negativamente el grupo de genes transportadores de nikABCDE mediante la unión a las regiones promotoras en presencia de níquel, y a diferencia de otras proteínas Fur que se pueden activar mediante otros metales, ésta es altamente específica. Nur también participaría en la respuesta antioxidante en *S. coelicolor* a través de regulación negativa (Fillat, 2014).

Las proteínas que componen la superfamilia Fur son capaces de desempeñar distintos procesos celulares (Figura 2) y muchas veces establecen redes de transcripción cruzada entre parálogos (Fillat, 2014). Los estudios de genética, bioquímica molecular, bioinformática, cristalización de nuevas estructuras han permitido conocer más sobre los mecanismos de esta importante familia de reguladores que pueden ser la base para nuevos desarrollos y aplicaciones. Por ejemplo, debido a que no existen proteínas Fur homólogas en eucariontes, puede ser usado potencialmente como un agente antimicrobiano (Deng et al., 2015), o en el combate de enfermedades como la tularemia (Pérard et al., 2018).

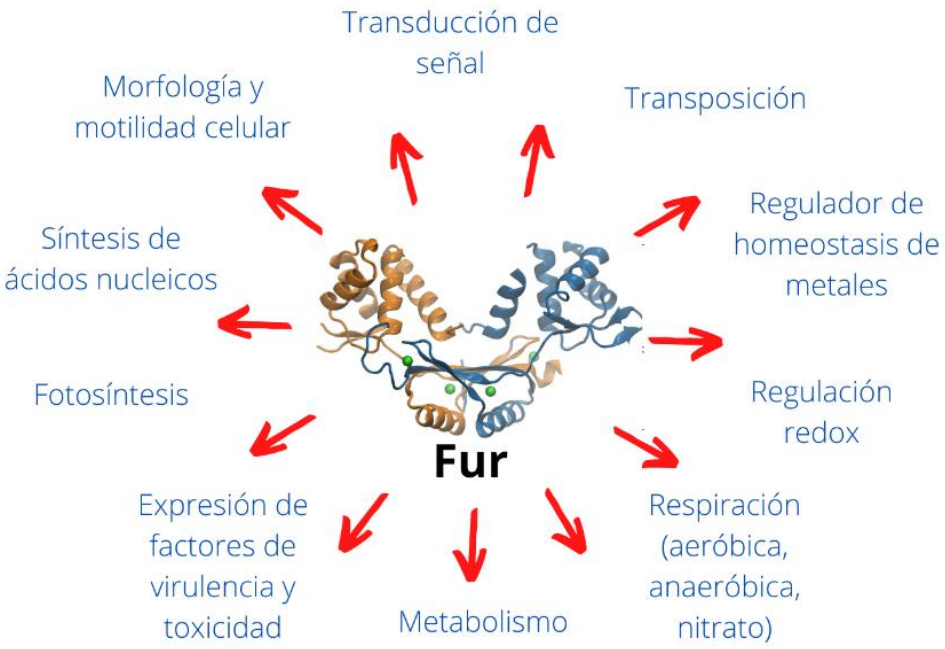

Figura 2. Procesos celulares modulados por reguladores de absorción férricos (Fillat, 2014).

1.3 Sitio de unión al ADN

Un sitio de unión es una región o zona donde ocurren ciertas interacciones importantes para las funciones de las proteínas, como la unión de cofactores, de un sustrato, o algún inhibidor. En el caso de este estudio las proteínas de la superfamilia Fur tienen la capacidad de unirse al ADN. Las proteínas de la familia Fur constan generalmente de un dominio de unión al

ADN o DNA binding domain (DBD) y un dominio de unión a metales, en cada monómero. Además, cuentan con una región bisagra que se encarga de unir los dominios entre monómeros. DBD se compone generalmente de una estructura hélice-giro-hélice, que provee afinidad y puede reconocer una secuencia especifica de ADN.

Estructuralmente las alfas hélices encajan con el surco mayor del ADN y solo son necesarios los cambios mediados por los cofactores para permitir la unión (Pohl et al., 2003). Se ha demostrado que de dos sitios de unión a metales presentes, el cofactor del sitio 1 es suficiente para la unión de Fur al ADN, sin embargo si bien el sitio 2 no sería esencial, modificaciones en él disminuirían la capacidad de unión al ADN en Fur (Deng et al., 2015). También se han identificado aminoácidos conservados que median la unión al ADN mediante contactos específicos de base. Por ejemplo, Arg57 forma enlaces de hidrógeno insertándose en el surco mayor, donde este residuo reconocería bases conservadas de guanina (G) y citosina (C) dentro de un motivo de unión (Fur-Box) (Pinochet-Barros & Helmann, 2018). Por otro lado, mediante el mecanismo de complementariedad superficial (realiza reconocimiento de la forma del ADN) es capaz de distinguir que Lys15 posee una unión favorable en el surco menor (por secuencias ricas en AT) que tiene potencial electrostático negativo (Deng et al., 2015).

También se han documentado proteínas Fur con la capacidad de unirse al ADN en ausencia de hierro. Por ejemplo, en *H. pylori* se produce la represión en ausencia del cofactor y ocurre una desrepresión cuando se une Fe^{2+} a Fur (Lee & Helmann, 2007). Otro ejemplo estudiado es PerR, que presenta el motivo hélice-giro-hélice (wHTH) y su forma de unión al ADN no requiere de un cofactor en el sitio de unión a metal, ya que su interacción se produce directamente con Fur-Box (Makthal et al., 2013).

De acuerdo con las investigaciones y estudios realizados, inicialmente se sabía que las proteínas Fur reconocen genes dentro del regulón, al unirse a una secuencia palindrómica denominada Fur-box que consta de 19 pares de

bases y se ubica dentro de la región promotora de los genes diana (Lee & Helmann, 2007):

5’----GATAATGATwATCATTATC----3’ , donde w = A o T

Luego en búsqueda de un motivo de unión mínima se demostró que este puede ser un motivo de repetición invertido heptamérico (7-1-7) (Pinochet-Barros & Helmann, 2018):

5’----TGATAAwATTATCA----3’ , donde w = A o T

La primera Fur-box descrita ha sido consistente en sitios de unión de estructuras cristalizadas en Mur (Berg et al., 2020), mientras que parálogos Zur y PerR en *B. subtilis* se ajustan con algunas variaciones al patrón 7-1-7 (Lee & Helmann, 2007). En la búsqueda de una Fur-box consenso el enfoque ha cambiado a que existe un patrón funcional dentro de la secuencia. Por lo tanto, más que el lugar o largo de una secuencia, se ha planteado que existe un hexámero independiente de la orientación y el número de repeticiones, y que este otorgaría la capacidad de interacción a Fur (Berg et al., 2020):

NATA/TAT , donde N es cualquier nucleótido.

Además, las Fur-box suelen tener un alto porcentaje de A/T y los complejos de esta familia tienen la capacidad de unirse al ADN en distintas proporciones.

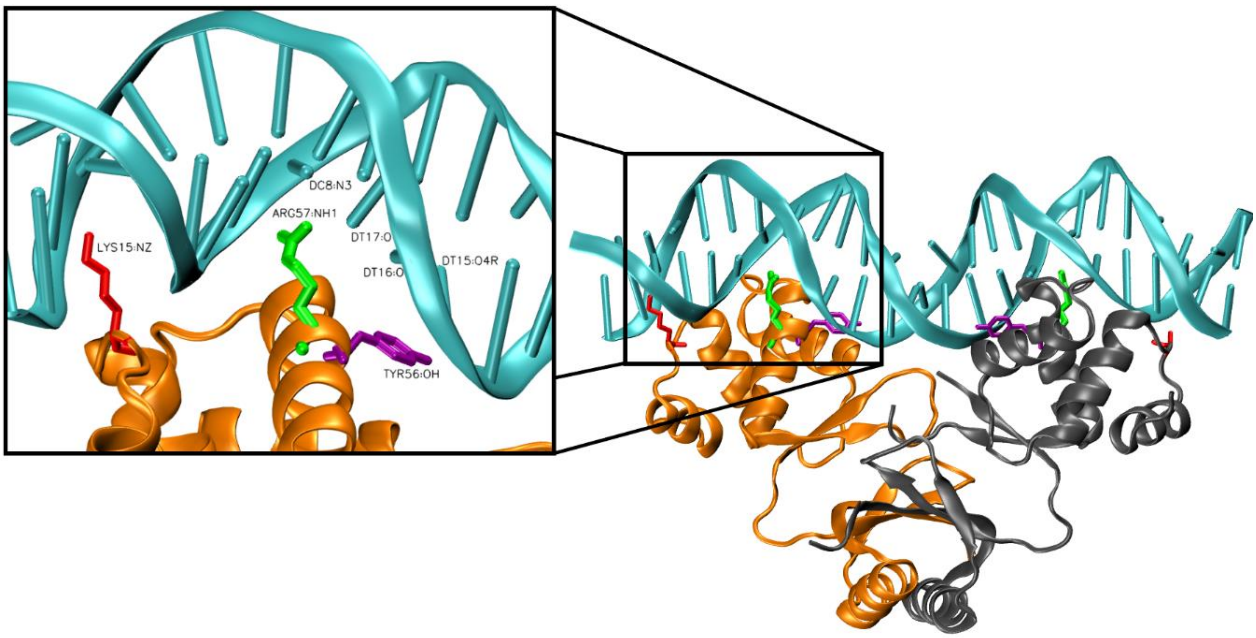

Figura 3. Representación de la estructura MRS2-Fur (código PDB: 4RB2). Se observa un acercamiento en los aminoácidos LYS15 (rojo), ARG57 (verde) y TYR56 (violeta), donde estos tres aminoácidos presentan diferentes modos de reconocimiento del ADN. Como se puede apreciar lisina se inserta en el surco menor, mientras que arginina y tirosina se insertan en el surco mayor, siendo la tirosina capaz de formar interacciones de Van der Waals con una o dos timinas del surco mayor (Deng et al., 2015).

Actualmente se tienen nociones y evidencia científica de interacciones que se producen y conservan en el sitio de unión al ADN entre proteínas de la superfamilia Fur, como la lectura de base mediante arginina y lectura de forma mediante el residuo lisina (ver Figura 3). Esto debido a la favorable interacción que se produce entre al ADN y aminoácidos cargados positivamente, o las posibilidades de dominio Fur-Box dentro de estas proteínas. Esta evidencia y conocimientos han permitido comprender un poco más sobre la unión, conocimientos que aumentan cada vez más debido a la creciente disponibilidad de complejos proteína-ADN disponibles en Protein Data Bank (PDB), (ver Figura 4), pero aún quedan muchas dudas de cómo son capaces de reconocer específicamente la zona de unión al ADN. Se ha visto que esta zona y su modo de unión, es variable en las distintas proteínas que conforman la superfamilia Fur, y en términos de secuencia no son idénticas, a pesar de las similitudes estructurales que poseen. Por lo tanto, el poder predecir el sitio de unión al ADN de proteínas con esta capacidad, como la superfamilia Fur, mediante métodos rápidos y menos

costosos que los realizados en laboratorio, sería un gran aporte para la investigación en esta área.

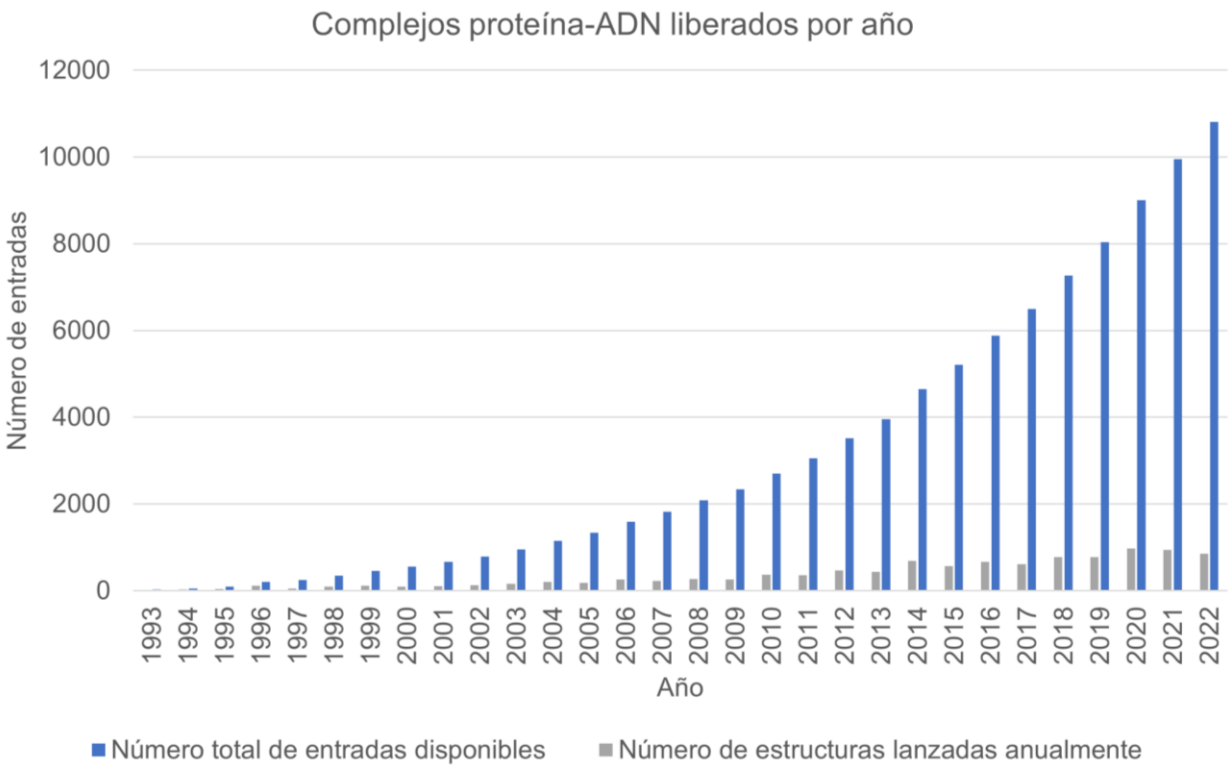

Figura 4. Número de complejos proteína-ADN liberadas en PDB cada año. Datos extraídos desde PDB statistics (https://www.rcsb.org/stats) 29 de septiembre, 2022.

2. Predicción de sitios de unión al ADN

Las predicciones de sitios de unión son importantes ya que permiten entender interacciones entre moléculas que juegan un rol esencial en muchos procesos biológicos. Los enfoques convencionales suelen realizarse en laboratorios húmedos, siendo costosos. Por lo tanto, los métodos computacionales han adquirido cada vez más relevancia en la predicción de sitios de unión de iones metálicos, cofactores, ligandos, interacción proteína-ADN etc. Estos métodos analizados se usan para verificar, y explicar molecularmente, las interacciones predichas en el laboratorio y para predecir nuevas interacciones que pueden ser importantes. El descubrimiento y verificación computacional de estas nuevas interacciones puede tomar menor tiempo y costo que las técnicas convencionales. Entre estos métodos destaca el *machine learning,* una subdisciplina dentro de la inteligencia artificial (IA).

La inteligencia artificial puede definirse como el estudio de algoritmos que dan a las máquinas la capacidad de aprender y resolver problemas, y dentro de esta disciplina el machine learning es el que permite realizar predicciones a partir del reconocimiento de patrones (Hashimoto et al., 2018).

2.1 Machine Learning

Machine Learning (ML) o aprendizaje automático se utiliza en diversas disciplinas, y tiene como objetivo principal entrenar una máquina para que aprenda de una experiencia pasada. Se proporcionan datos recopilados, y estos pueden ser etiquetados (aprendizaje supervisado) y no etiquetados (aprendizaje no supervisado) según la problemática que se quiera resolver (Mehmood & Selwal, 2020).

El aprendizaje supervisado es un aprendizaje donde existen datos etiquetados de entrada y salida, y mediante un algoritmo se analizan los datos de entrada para predecir los de salida, proporcionando así datos más precisos (Mehmood & Selwal, 2020).

En el aprendizaje no supervisado solo se tienen datos y variables de entrada, los datos no están etiquetados y la predicción se realiza heredando las características de los datos de entrada. Los resultados son promedios y la complejidad es menor que en el caso del aprendizaje supervisado (Mehmood & Selwal, 2020).

El ML se utiliza cada vez más por la rapidez y el bajo costo con que un programa puede resolver problemas. Se aplica en problemas que suelen tener unos cuantos resultados y observaciones, pero teóricamente no se conoce por completo la respuesta.

2.2 Métodos de predicción basados en machine learning

La predicción de sitios de unión al ADN basados en ML puede realizarse utilizando solo la información de la secuencia aminoacídica de una proteína. Por ejemplo, Ma et al., 2013 presenta un enfoque de predicción de sitios de

unión al ADN en proteínas, basado en Máquina de Vectores de Soporte y una característica híbrida, utilizando solo la secuencia primaria de la proteína.

También se puede predecir mediante propiedades fisicoquímicas, como lo realizan en la investigación del predictor DNABP, enfocado en la conservación de las propiedades fisicoquímicas de los aminoácidos, y que utiliza el método Random Forest (Ma et al., 2016). O se puede predecir basándose en la estructura tridimensional (3D), como en el método DBPSite, que mediante la estructura de la proteína predice residuos de unión al ADN utilizando Máquina de Vectores de Soporte (Xiong et al., 2011). Algunos de los clasificadores más utilizados son:

Support Vector Machine (SVM, Máquinas de Vectores de Soporte): método de clasificación de aprendizaje supervisado, utilizado en proteínas para predicciones de estructuras, funciones, interacciones ADN-proteína debido a su capacidad de resolver problemas no lineales, y características dimensionales (Si et al., 2015). Se han realizado predicciones con este tipo de clasificador para identificar residuos de unión al ADN mediante la estructura 3D, distinguiendo entre los aminoácidos que no se unen, de los que si se unen al ADN (Xiong et al., 2011).

Artificial neural network (ANN, Redes neuronales artificiales): método de clasificación mediante aprendizaje no supervisado, que está inspirado en las redes neuronales del cerebro y, se componen de una capa de entrada, una o más capas ocultas y la capa de salida (Mehmood & Selwal, 2020). Este algoritmo tiene la ventaja de realizar múltiples pasos de entrenamiento. En la predicción de sitios de unión en proteínas se suelen usar descriptores estructurales como datos de entrada y la complejidad depende de la dimensión del vector de entrada (Si et al., 2015).

Decision Tree (Árbol de decisión): método de clasificación de aprendizaje supervisado, que se basa en nodos y ramas, donde los nodos son los

atributos del grupo de datos que se desea clasificar y las ramas el valor que puede tomar ese nodo (Mehmood & Selwal, 2020). La sensibilidad y complejidad de este clasificador depende de los datos de entrada, y en datos biológicos no siempre satisface las necesidades de los problemas que se quieren predecir. Por lo tanto, se han diseñado variaciones basadas en este clasificador como Random Forest usado en predicciones de sitio de unión al ADN (Wu et al., 2009), árboles de decisión potenciados, y árboles de decisión alternos (Si et al., 2015).

Bayesian learning (Aprendizaje bayesiano): método de clasificación supervisado, basado en el teorema bayesiano, que se basa en la probabilidad de una hipótesis a medida que se obtiene más información (Si et al., 2015). Se ha utilizado para la predicción a partir de la secuencia aminoacídica, donde se entrena el algoritmo para predecir si un aminoácido particular es capaz de unirse al ADN dada su identidad y las identidades de los vecinos en la secuencia (Yan et al., 2006).

Cada uno de estos clasificadores requiere de la construcción de un conjunto de datos para poder realizar una predicción, la eficacia y capacidad de cada predictor depende de qué tan caracterizado se encuentre el set de datos, en proteínas es importante tener características basadas en la secuencia, en su estructura o características fisicoquímicas según el tipo de predicción que se desee realizar (Si et al., 2015). Herramientas como los descriptores de topología, secuencia, estructura, potencial electrostático, o descriptores fisicoquímicos son importantes para un conocimiento general de la proteína, y obtener un set de datos con características que permitan realizar predicciones.

3. Descriptores Moleculares

Los descriptores moleculares han posibilitado el entendimiento y explicación de las reacciones, interacciones, cambios conformacionales etc., de las moléculas. Existe una gran variedad de descriptores donde cada uno de ellos apunta a caracterizar una o más propiedades de la o las

moléculas. Estos los podemos clasificar según: secuencia aminoacídica o estructura tridimensional.

3.1 Basados en secuencia aminoacídica

Los primeros descriptores que se generaron para ayudar a comprender mejor las moléculas y estructuras proteicas complejas se basaron en la secuencia aminoacídica. Se inició con el estudio de los aminoácidos, analizando las propiedades físicas y las características que confieren a la secuencia primaria, para así poder determinar, por ejemplo, la relación que estos tienen con la estabilidad de la estructura en una proteína. Como por ejemplo, en un estudio se analizaron propiedades ortogonales para los 20 aminoácidos naturales, mediante análisis factorial y se obtuvieron 188 propiedades físicas de los aminoácidos (Weng et al., 2013).

Uno de los métodos utilizados en varios descriptores debido a su baja dimensionalidad es el análisis de componentes principales (PCA), que describe aminoácidos individuales a partir de una matriz de propiedades (Van Westen et al., 2013). Desde el método PCA, se puede derivar en otro descriptor denominado escalas de topología estructural (o escala ST) que consiste en el análisis de propiedades principalmente topológicas. Éste consta de 827 variables estructurales y 167 propiedades topológicas de aminoácidos (Yang et al., 2010). Así como las escalas ST también, se derivan descriptores como VHSE (vectores de propiedades hidrofóbicas, estéricas y electrónicas) y escalas Z, entre otros (Van Westen et al., 2013).

También se encuentran descriptores de propiedades fisicoquímicas como FASGAI que se basa en un análisis factorial de propiedades fisicoquímicas (hidrofobicidad, composición, propiedades electrónicas, voluminosidad etc.). Este permite analizar problemas multidimensionales como en el modelado de relación cuantitativa estructura-actividad (QSAR) para la búsqueda de péptidos funcionales (Liang & Li, 2007; Van Westen et al., 2013).

En los estudios realizados, y para tener una mejor caracterización de los cambios y diferencias que producen una determinada respuesta en una proteína se han desarrollado descriptores combinados, con diversos algoritmos y métodos de aprendizaje automático para realizar predicciones. Por ejemplo, Disulfind (Ceroni et al., 2006), es un predictor de puentes disulfuro entre cadenas proteicas, que fusiona características obtenidas a partir de la secuencia aminoacídica como conservación y composición de aminoácidos, descripción de información importante de las cisteínas (cantidad, conservación, tipo de enlace etc.) con el clasificador SVM para automatizar el proceso de predicción de puentes.

3.2 Basados en estructura tridimensional

También se han desarrollado descriptores moleculares basados en las características obtenidas a partir de la distribución espacial de los átomos. Debido a la cantidad creciente y disponibilidad de estructuras proteicas en la base de datos en PDB (ver Figura 5), es posible realizar más y mejores caracterizaciones estructurales en proteínas, ya que la estructura entrega información de propiedades como superficie, información de estructura secundaria, función y evolución biológicamente relevantes.

La identificación y caracterización de sitios de unión en una proteína es constante tema de estudio y los descriptores han contribuido en este ámbito. Un ejemplo interesante es el descriptor de bolsillos y cavidades Fpocket (Le Guilloux et al., 2009), que es un software que a través del concepto de las esferas alfa, logra describir y caracterizar cavidades para discriminar entre las más relevantes y posiblemente sitios de unión. Otro evaluador que entrega diferentes descripciones de estructuras proteicas a través de su estructura tridimensional es Mordred, programa que cuenta con más de 1800 descriptores bidimensionales y tridimensionales, que permiten obtener información sobre relaciones cuantitativas entre estructura y propiedad (Moriwaki et al., 2018).

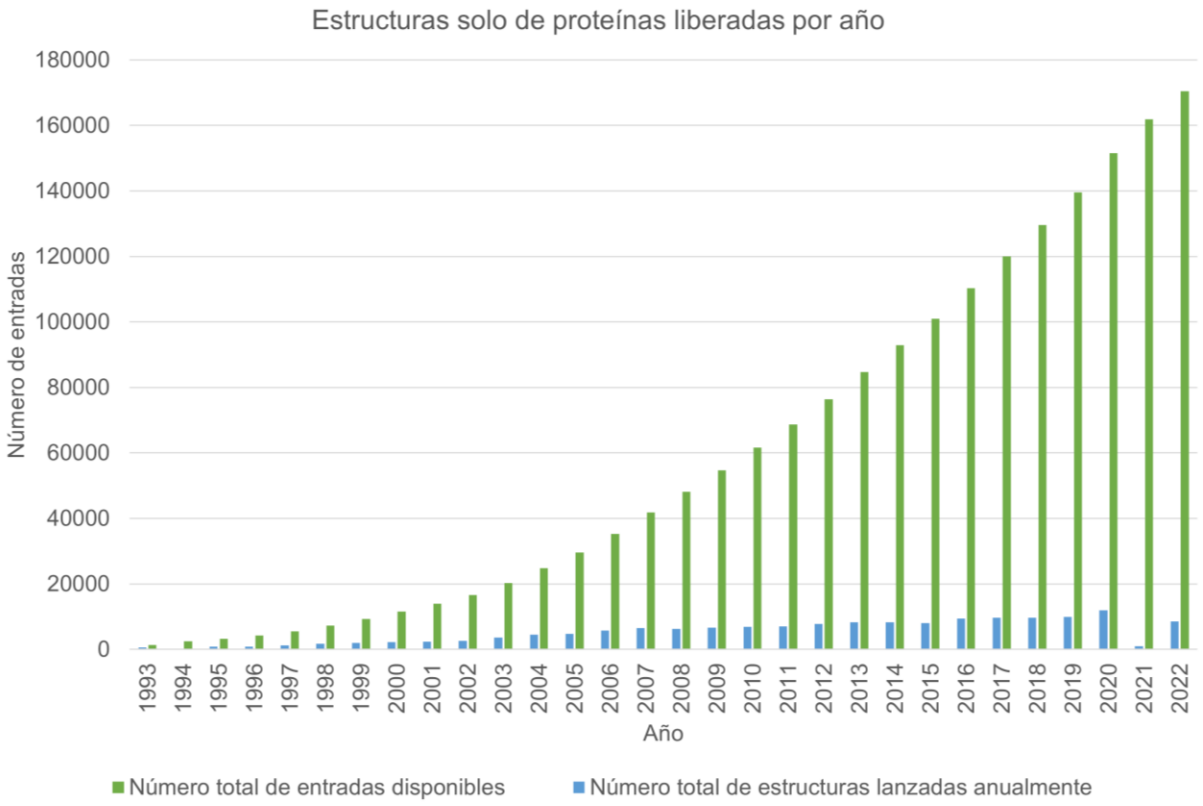

Figura 5. Número de estructuras proteicas liberadas en PDB cada año. Datos extraídos desde PDB statistics (https://www.rcsb.org/stats) 29 de septiembre, 2022.

La diferencia entre los descriptores de secuencia y estructura es que ambos realizan diferentes caracterizaciones, por lo tanto, para el desarrollo de predictores más completos se requiere que se consideren múltiples características como de secuencia, químicas, configuraciones espaciales de los aminoácidos presentes, entre otras de forma que se pueda caracterizar información nueva, no redundante y más completa (Terán et al., 2019). Un software que complementa tanto análisis de secuencia como descriptores 3D es ProtDCal (Ruiz-Blanco et al., 2015), que utiliza diversos descriptores como el análisis de componentes principales y pruebas de entropía de Shannon, índices termodinámicos tanto en secuencia como en estructura 3D, entre otros descriptores. Esta es una herramienta que permite el análisis integrado de secuencia y estructura en una sola plataforma.

La investigación presentada ésta dirigida al estudio del sitio de unión de complejos proteína-ADN, sitio que será caracterizado mediante descriptores moleculares y se evaluarán modelos de predicción.

HIPÓTESIS Y OBJETIVOS

1. Hipótesis

A través de la utilización de técnicas de inteligencia artificial y de descriptores moleculares será posible predecir la zona(s) de unión al ADN en proteínas de la Superfamilia Fur.

2. Objetivo general

Desarrollar un modelo predictivo de zonas de unión al ADN utilizando inteligencia artificial y descriptores moleculares, aplicado en la superfamilia Fur y otras proteínas con capacidad de unión al ADN.

3. Objetivos específicos

1. Construir un set de datos que contenga información de secuencia aminoacídica y/o de estructural de los fragmentos de unión al ADN en proteínas de la superfamilia Fur y otros factores de transcripción.

2. Caracterizar el set de datos con descriptores moleculares a nivel de secuencia aminoacídica y estructural.

3. Evaluar modelos Máquinas de Vectores de Soporte y Random Forest para su aplicación en la predicción de sitios de unión al ADN.

MATERIALES Y MÉTODOS

A continuación, se presentan los materiales y métodos utilizados para alcanzar los objetivos y desarrollo de esta investigación, la Figura 6 muestra el flujo de trabajo de esta metodología.

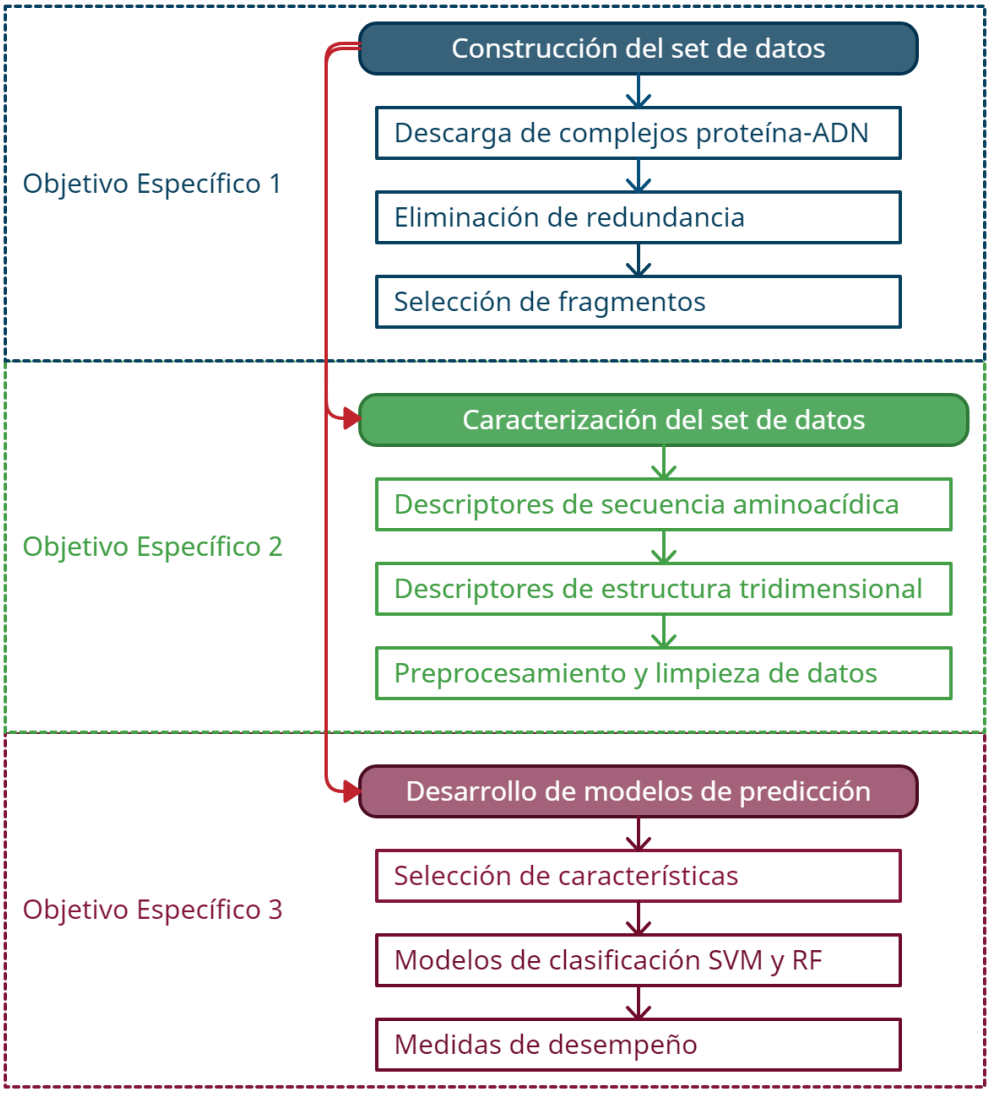

Figura 6. Diagrama de la metodología general que se aplicará en esta investigación.

1. Construcción del set de datos

Para la construcción del set de datos se utilizaron las siguientes herramientas:

Protein Data Bank (PDB): es una base de datos que almacena estructuras tridimensionales de proteínas, ácidos nucleicos y ensambles complejos mediante distintas técnicas como cristalografía, rayos X, resonancia magnética nuclear y múltiples métodos. Cada una de las estructuras es almacenada con un código identificador de 4 letras y un archivo pdb que contiene diversa información como datos experimentales, función, coordenadas atómicas, entre otros. Se puede acceder en https://www.rcsb.org/ .

Visual Molecular Dynamics (VMD): es un programa de visualización molecular que permite observar y analizar diversos sistemas moleculares, mediante diferentes comandos posibilita la visualización de moléculas específicas, distancias, alineamientos, extracción de estructuras etc. Se puede acceder al software en https://www.ks.uiuc.edu/Research/vmd/ .

Molecular Evolutionary Genetics Analysis (MEGA): es un software que permite el análisis de secuencias de proteínas y ADN de diferentes especies y poblaciones, mediante alineamientos, distancias, modelos, filogenia etc. Se puede acceder al software en https://www.megasoftware.net/home .

PDBeFold: aplicación web que permite realizar un análisis comparativo estructural, entrega valores de desviación cuadrática media (RMSD), porcentaje de identidad de secuencia, alineamiento de estructura secundaria etc. Se puede acceder al software en https://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver .

1.1 Descargar estructuras de proteínas unidas al ADN.

La selección de cada uno de los complejos proteína-ADN a utilizar, se realizó mediante la búsqueda avanzada considerando los siguientes filtros:

- Polymer Entity Type: se seleccionaron solo proteínas unidas al ADN.

- Macromolecule Name: se filtró por aquellos pertenecientes a la superfamilia Fur (ferric uptake regulator, zinc uptake regulator, manganese uptake regulator, iron uptake regulatory, Fur family, peroxide stress sensing regulator).

- Polymer Entity Description: se filtró solo factores de transcripción.

- Parent Scientific Name: pertenecientes a Bacteria.

El conjunto de filtros permitió obtener un total de 91 complejos proteína-ADN (formato FASTA y PDB), incluyendo 5 complejos pertenecientes a la superfamilia Fur (códigos 4MTE, 4MTD, 4RB1, 4RB2, 4RB3).

1.2 Eliminación de redundancia del set de datos

Para estudiar las zonas de unión y utilizar modelos de predicción es importante utilizar un set de datos no redundante, por lo que se realizó un análisis filogenético de las proteínas de cada uno de los complejos proteína-ADN obtenidos desde PDB.

Se aplico una separación de los complejos proteína-ADN desde el archivo FASTA, dejando solo la secuencia correspondiente a la proteína, ya que utilizar el ADN unido a las proteínas causa ruido y no permite evaluar la redundancia entre proteínas. Con la obtención de cada una de las proteínas se procedió a utilizar el software MEGA el cual permitió realizar una eliminación de aquellas proteínas con una alta similitud, mediante un alineamiento de secuencias y árbol filogenético, lo que resulto en una reducción de las 91 proteínas iniciales a un total de 63, eliminando así la redundancia del set de datos inicial.

1.3 Selección de fragmentos con y sin capacidad de unión al ADN.

Con la obtención de 63 proteínas no redundantes se procedió a procesar los 63 complejos de las proteínas seleccionadas, para obtener el set de datos final para la caracterización y aplicación en modelos de predicción.

El objetivo de este estudio es la predicción de sitios de unión al ADN, por lo que evaluar la estructura completa de la proteína no es el mejor enfoque, ya que la zona relevante es el fragmento que logra la unión al ADN. Para el entrenamiento de los modelos en machine learning es importante tener un set de datos balanceado (misma cantidad de muestras positivas y negativas), para evitar errores en la clasificación por desbalance, es por esto por lo que el set de datos analizar cuenta con 63 fragmentos positivos (clase 1) y 63 fragmentos negativos (clase 2), un fragmento positivo es aquel que corresponde a una zona de unión al ADN en la proteína, mientras que un fragmento negativo es una zona que no une al ADN.

Todos los fragmentos fueron extraídos utilizando el software VMD, desde el archivo PDB de cada uno de los complejos proteína-ADN. Las zonas de unión (fragmentos positivos) fueron seleccionados mediante observación directa del complejo proteína-ADN y la ratificación de la zona mediante la investigación asociada a cada complejo, mientras que los fragmentos negativos se obtienen de zonas que no unen al ADN. Los fragmentos van desde los 10-50 aminoácidos según la zona de proteína que se une al ADN, la cual es variable en cada complejo. Es importante mencionar que todas las proteínas de los complejos seleccionados tienen diversas cadenas, sin embargo, estas cadenas tienen una alta similitud, por lo que se seleccionó un fragmento positivo y un fragmento negativo por cada complejo proteína-ADN. Por lo tanto, el set de datos final contiene 126 fragmentos (63 positivos y 63 negativos).

2. Caracterización del set de datos

Las herramientas utilizadas para la caracterización del set de datos corresponden a una serie de descriptores de secuencia y al software Protein, Ligand, Geometrical and Physicochemical Properties 3D (PLGP3D) desarrollados he implementados por el grupo de trabajo ProtechLab.

2.1 Descriptores de secuencia aminoacídica

Se caracterizo el set de datos de 126 fragmentos según su secuencia FASTA, entre los descriptores moleculares implementados se encuentran:

- Porcentajes por cada tipo de aminoácido alifático, aromático, polar, no polar, cargado, básico, acido etc. (Rice et al., 2000).
- Propiedades electrónicas, hidrofobicidad, propensiones alfa y de giro, flexibilidad local, características de composición etc. (Liang et al., 2008).
- Propiedades fisicoquímicas, como escalas Z que incluyen propiedades estéricas, electrónicas (polaridad/ carga), electronegatividad, calor de formación etc. (Sandberg et al., 1998).
- Descriptores topológicos, escalas T basados solo en la tabla de conectividad de aminoácidos, no consideran las propiedades 3D de cada estructura (Tian et al., 2007).
- Escalas st, consideran propiedades principalmente constitucionales, topológicas, geométricas, hidrofóbicas y estéricas (Pronk et al., 2013).
- Puntuaciones MS-WHIM, corresponden a derivaciones de 36 propiedades de potencial electroestático (Zaliani & Gancia, 1999).
- Descriptores numéricos interpretables del espacio de aminoácidos, corresponden a matrices Blosum (Georgiev, 2009).
- FASGAI, análisis factorial que incluye hidrofobicidad circundante, suceso en la región alfa, volumen especifico parcial, preferencia de doble curvatura etc. (Liang & Li, 2007).

La ejecución del script que contiene cada uno de estos descriptores moleculares, se realiza en el servidor de ProtechLab, cada uno de los

descriptores entrega un valor numérico que va desde un número entero a un valor decimal según el descriptor al que corresponde, entregando 74 atributos.

2.2 Descriptores de estructura tridimensional

Para obtener una caracterización de la estructura tridimensional de cada uno de los fragmentos se utilizó el programa PLGP3D, este programa permite calcular diversas propiedades geométricas y fisicoquímicas a partir de la estructura tridimensional de una proteína, ésta pensado en la caracterización de sitios de unión a ligando en proteínas, sin embargo su funcionamiento se basa en realizar una caracterización mediante un área o radio de un sitio que se requiera estudiar, por ejemplo, realiza la caracterización de todo lo que rodea al ligando en un radio de 14 Å, permitiendo ingresar las coordenadas especificas del punto central o zona de ligando.

Los datos básicos requeridos para el funcionamiento del programa son: código del archivo PDB, radio a estudiar, número de capas y coordenadas del ligando (opcional). Como nuestro set de datos está basado en fragmentos, para poder realizar una caracterización de estos se utilizó el programa PLGP3D de la siguiente forma:

- Coordenadas: se entregó una coordenada en el espacio en cada uno de los fragmentos, como el punto central.
- Radio: se asignó un radio desde el punto central hasta cubrir todo el fragmento.
- Número de capas: el número de capas permite realizar una división del área de estudio, en este caso solo deseamos caracterizar el fragmento, por lo que el número de capas se asignó en 0.

Los atributos obtenidos en esta caracterización corresponden a distancias, ángulos, torsiones, hidrofobicidad, composición aminoacídica,

composición atómica por capa, interacciones, energías, radios de sitio, aminoácidos cercanos, volumen, Alpha Shapes etc. La ejecución del script se realiza en los servidores de ProtechLab, y entrega 153 atributos.

2.3 Preprocesamiento y limpieza de datos

Un punto importante en esta etapa es la verificación de la integridad de los datos que serán entregados a los modelos predictivos, es por esto por lo que se procedió a realizar un análisis y limpieza de los datos.

Las caracterizaciones realizadas a nivel de secuencia y estructura entregan una serie de atributos, los cuales deben ser analizados para eliminar aquellos que entreguen datos nulos o inconsistentes, ya que la entrega de datos erróneos a los modelos predictivos puede provocar errores de cálculos, por lo tanto, se debe asegurar la coherencia de los datos.

3. Desarrollo de los modelos de predicción

Con el set de datos caracterizado se procedió a desarrollar cada uno de los modelos predictivos a estudiar, la herramienta utilizada para esta etapa es **Scikit Learn** (Barupal & Fiehn, 2019), módulo que integra algoritmos de machine learning basado en el lenguaje de programación Python, permite el análisis predictivo de datos, se puede acceder en https://scikit-learn.org/stable/index.html.

3.1 Selección de características

Luego de tener el set de datos caracterizado y limpio, es importante realizar una selección de aquellos atributos que permiten al modelo clasificar correctamente. Por lo tanto, se aplicó una selección de características con librerías presentes en Scikit Learn en los atributos basados en la secuencia aminoacídica como en los estructurales.

Se utilizó la librería principal SelectKBest que permite hacer una selección de los mejores atributos según las k puntuaciones más altas, donde k es el número de características principales a seleccionar. Esta selección se

complementa con una función predeterminada, en esta investigación se aplicaron dos funciones distintas para la selección de atributos:

Mutual Info Classif*:* permite estimar la información mutua entre dos variables, lo que permite medir la dependencia entre las variables, calcula un puntaje donde valores altos significan mayor dependencia entre las variables y un valor igual a cero si las variables son independientes.

F Classif (ANOVA): mediante un análisis de varianza, utiliza puntajes F para probar la igualdad de medias.

Se desarrollo el código de cada una de estas funciones y se aplicó en los atributos basados en secuencia, como en los atributos basados en estructura tridimensional, para realizar una selección por separado, y así evaluar el set de datos en los modelos de predicción.

3.2 Modelos de clasificación SVM Y RF

Los modelos de clasificación evaluados son máquinas de vectores de soporte y Random Forest, cada uno fue desarrollado utilizando Scikit Learn. Luego de la selección de atributos se obtiene un nuevo set de datos con cada uno de los 126 fragmentos y los mejores atributos seleccionados. Este nuevo set de datos fue entrenado en cada uno de los modelos:

Máquinas de vectores de soporte (SVM): método de aprendizaje supervisado, mediante la búsqueda de un hiperplano o recta separa las clases, maximizando el margen que separa los ejemplos de cada clase, es efectivo en casos donde el número de dimensiones es mayor que el número de muestras puede ser utilizado para clasificación o regresión, en este estudio se utiliza como clasificador, y se desarrolla mediante la librería libsvm de Scikit Learn.

Random Forest: método de aprendizaje supervisado, es un meta estimador que ajusta una serie de árboles de decisión en varias submuestras del conjunto de datos y utiliza el promedio para mejorar su

precisión predictiva. Se utiliza como clasificador y se desarrolla mediante la librería RandomForestClassifier de Scikit Learn.

Ambos modelos son de aprendizaje supervisado y utilizados como clasificadores, donde las clases son: clase 1 (fragmentos positivos, si unen al ADN), clase 2 (fragmentos negativos, no unen al ADN). Cada uno de los modelos fue entrenado con los atributos obtenidos desde la caracterización del set de datos, con el objetivo de poder conocer si estas características permiten predecir si un fragmento es una zona de unión al ADN o no, de esta manera si la performance de los modelos evaluados es positiva en su clasificación, se puede dilucidar que existen diferencias entre las zonas de unión y no unión tanto a nivel de secuencia, como estructural.

Los modelos fueron probados con distintas combinaciones de parámetros hasta lograr la mejor exactitud en cada uno, mediante una la división del set de datos en un 80% de entrenamiento y un 20% testeo, además de realizar una validación cruzada de 10 iteraciones, se calculó la media aritmética de las métricas de desempeño: Accuracy, Recall (sensibilidad) y Precision. Las métricas son calculadas mediante la matriz de confusión que es calculada en cada iteración, ésta permite visualizar el desempeño del clasificador (ver Tabla 1).

		Predicción	
		Clase 1	Clase 2
Real	Clase 1	VP	FN
	Clase 2	FP	VN

Tabla 1. Matriz de confusión. Permite calcular indicadores de eficiencia global en modelos de predicción, las columnas corresponden a la clase de clasificación, donde una predicción correcta se denomina Verdaderos Positivos (VP) para aquellos fragmentos que, si unen al ADN, Verdaderos Negativos (VN) fragmentos que no unen al ADN y aquellos ejemplos que son clasificados erróneamente se denominan Falsos Positivos (FP) o Falsos Negativos (FN).

Las métricas de desempeño utilizadas en cada evaluación son: Accuracy, es la capacidad del modelo de ML de predecir correctamente.

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

Recall, es la capacidad del clasificador de encontrar todas las muestras positivas, mientras más cercano el valor al 100% mejor es el desempeño.

$$Recall = \frac{VP}{VP + FN}$$

Precision, es la capacidad del clasificador de no etiquetar como positiva una muestra que es negativa, un porcentaje más alto indica un mejor desempeño.

$$\text{Precision} = \frac{VP}{VP + FP}$$

RESULTADOS

1. **Resultados objetivo específico 1:** Construir un set de datos que contenga información de secuencia aminoacídica y/o de estructural de los fragmentos de unión al ADN en proteínas de la superfamilia Fur y otros factores de transcripción**.**

1.1 Estructuras utilizadas en el estudio.

Al eliminar la redundancia de los 91 complejos proteína-ADN se obtuvo un total de 63 complejos (ver Figura 7), todos ellos pertenecientes a distintas familias de proteínas bacterianas que se clasifican como factores de transcripción, dentro de este grupo 2 de los complejos pertenecen a la superfamilia Fur (códigos 4RB1 y 4MTD).

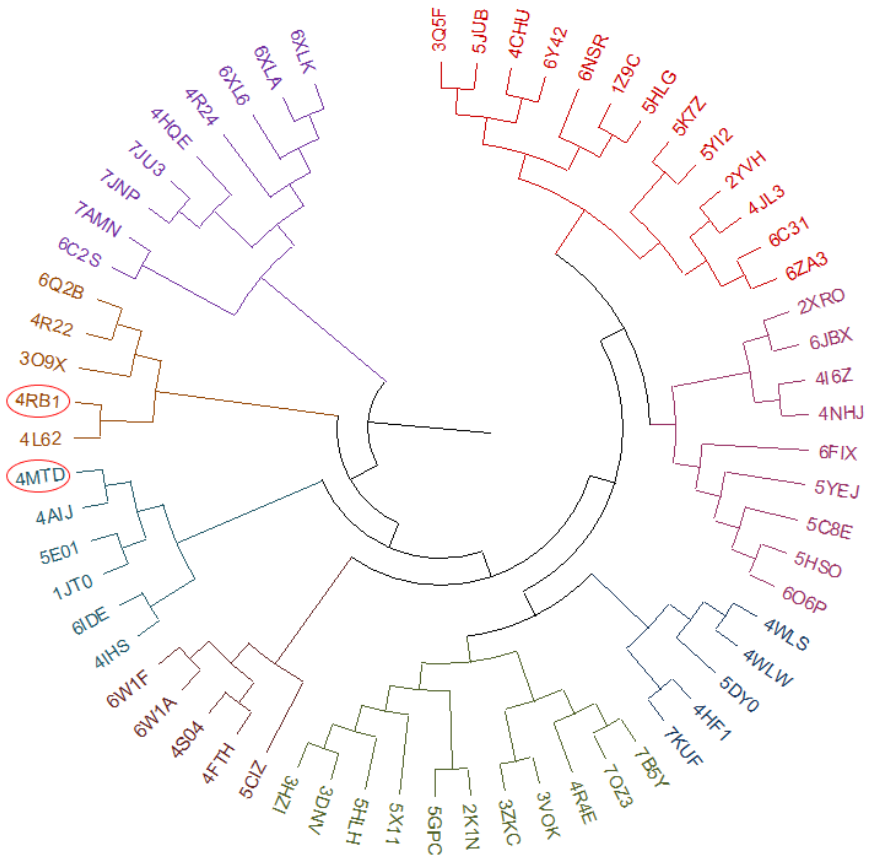

Figura 7. Estructuras utilizadas en la investigación. En la figura se observan los distintos códigos de los complejos proteína-ADN utilizados, se diferencian 8 grupos, y se destacan las proteínas pertenecientes a la superfamilia Fur, con los códigos 4RB1 y 4MTD. Análisis realizado con el software MEGA.

Todas las proteínas son diferentes a nivel de secuencia, cada una tiene distinto estado oligómero, los cuales presentan una alta similitud, y pueden conformar especies diméricas, triméricas, tetrámericas etc. Cada proteína cuenta con dos tipos de dominios, un dominio de unión a ligando y un dominio de unión al ADN. Como se observa en la Figura 7 las dos proteínas pertenecientes a la superfamilia Fur 4RB1 y 4MTD están en distinto grupo, esto se debe a que presentan una baja similitud en su secuencia.

4RB1 corresponde a una proteína dímerica reguladora de magnesio presente en el organismo *Magnetospirillum gryphiswaldense* (Deng et al., 2015), mientras que 4MTD es una proteína de *E. coli* reguladora de zinc, con dos dímeros presentes (Gilston et al., 2014), la proteína 4L62 de *Pseudomonas aeruginosa* es una reguladora que responde ante cambios ambientales y presenta 2 dímeros (Kim et al., 2013), finalmente la proteína dímerica 4AIJ del organismo *Yersinia pseudotuberculosis* responde ante cambios térmicos que modifican su estructura (Quade et al., 2012). Todas estas proteínas presentan en su cadena el dominio de unión al ADN, 4RB1 presenta una alta similitud con 4RB2 de la representación de unión al ADN en la Figura 3 de este documento, como se observa en la matriz de distancia p de aminoácidos (ver Tabla 2), las proteínas 4MTD-4AIJ tienen una mayor cercanía que 4MTD-4RB1.

	4MTD	4RB1	4L62	4AIJ
4MTD				
4RB1	0,834			
4L62	0,881	0,917		
4AIJ	0,784	0,859	0,874	

Tabla 2. Matriz de distancia p de aminoácidos. La siguiente tabla muestra la proporción (p) de sitios de aminoácidos en que las secuencias a comparar son diferentes. Se observa la distancia entre las proteínas de la superfamilia Fur 4RB1 y 4MTD, y las parejas más cercanas 4L62 y 4AIJ respectivamente, dada su distinta agrupación en el árbol filogenético (ver Figura 7).

Sin embargo, si se realiza un análisis de la zona de unión en estas proteínas se observa que tienen una alta similitud estructural a pesar de las diferencias a nivel de secuencia (ver Figura 8), destacando la estructura

secundaria hélice alfa, motivo estructural que es sabido por estudios previos que tiene una importancia en proteínas con capacidad de unión al ADN. Por ejemplo, uno de los métodos que se ha utilizado para la identificación de proteínas con capacidad de unión al ADN es reconocer motivos estructurales como hélice-giro-hélice, hélice-horquilla-hélice o hélice-bucle-hélice, así como además identificar un potencial electroestático positivo en la zona de unión (Shanahan et al., 2004).

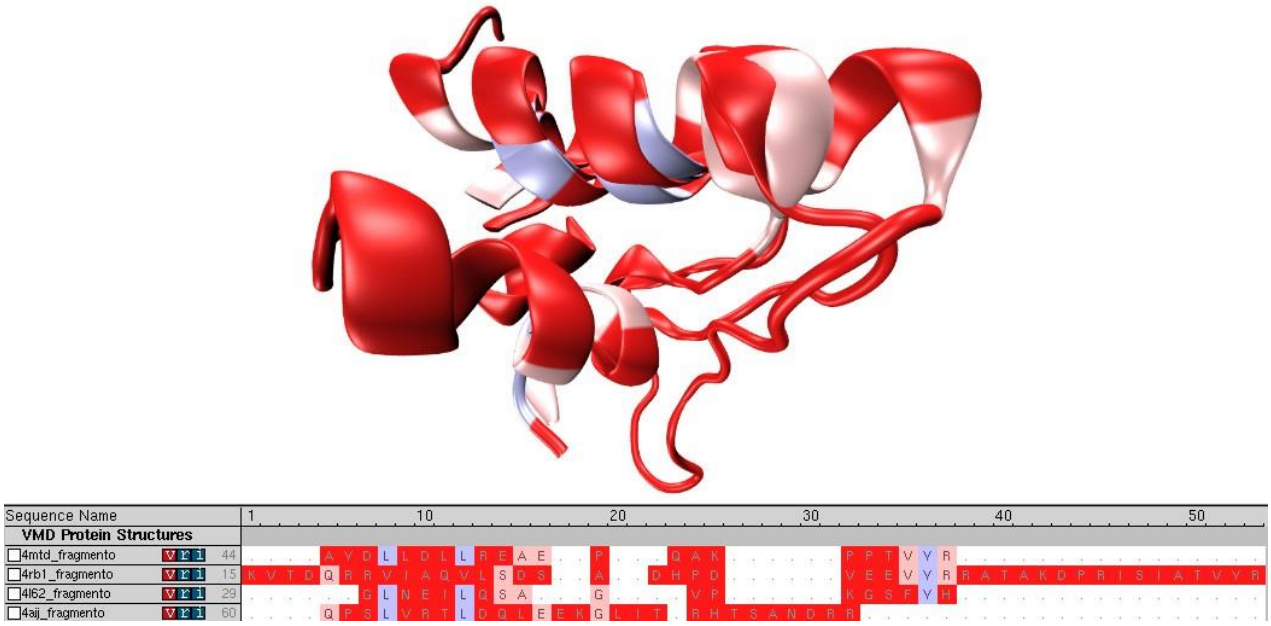

Figura 8. Alineamiento estructural de la zona de unión al ADN. Se observa el alineamiento de la zona de unión al ADN de las proteínas 4MTD, 4RB1, 4L62 y 4AIJ, destacando en color rojo la diferencia aminoacidica, y en color rosado la zona de identidad entre aminoacidos, alineamiento realizado en el software VMD**.**

Además, para un mejor análisis de la similitud estructural entre estos fragmentos de unión al ADN se realiza el cálculo de RMSD de estos 4 fragmentos utilizando el programa PDBeFold, programa que entregó un RMSD general de 0.6868, al tomar como referencia el fragmento 4MTD se puede analizar los RMSD individuales y el porcentaje de identidad de las secuencias, como se observa en la Figura 9.

Para el análisis de estos resultados se establece que un RMSD inferior a 3,5 y un porcentaje de identidad de aminoácidos inferior al 25% índica que estas estructuras son estructuralmente similares, pero a nivel de secuencia diferentes, y como se observa en la gráfica estas zonas de unión comparadas con 4MTD perteneciente a la superfamilia Fur, presentan una alta similitud estructural donde 4RB1 presenta un RMSD de 0.614, 4L62 de 0.983 y 4AIJ de 0.933, todas con un porcentaje de identidad de secuencia inferior o igual al 10%.

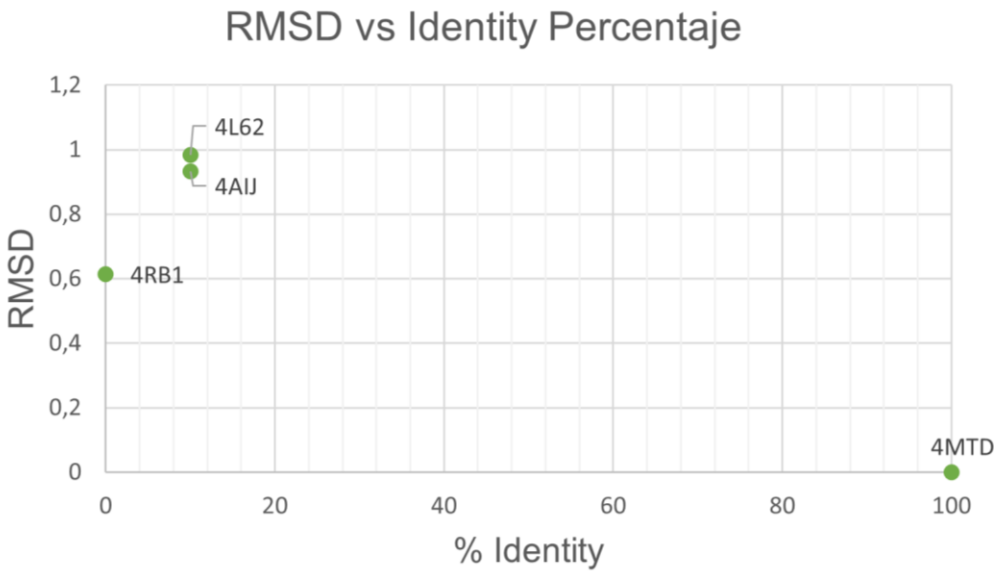

Figura 9. Comparación estructural de zona de unión. La siguiente gráfica muestra el valor de RMSD vs el porcentaje de identidad de los fragmentos 4RB1, 4L62 y 4AIJ, con 4MTD como referencia.

2. **Resultados objetivo específico 2:** Caracterizar el set de datos con descriptores moleculares a nivel de secuencia aminoacídica y estructural**.**

La caracterización a nivel de secuencia de los 126 fragmentos (63 fragmentos positivos y 63 negativos) entregó un set de datos compuesto por 74 atributos, mientras que la caracterización estructural entrego un set de datos de 153 atributos de estructura. El preprocesamiento y limpieza de los datos se realizó en los atributos de secuencia y estructura.

Atributos de secuencia: los descriptores moleculares utilizados para la caracterización a nivel de secuencia corresponden a 74 atributos, dentro de estos se encontró 1 atributo que entregaba datos nulos o inconsistentes, por lo que se realizó una limpieza eliminando el atributo inconsistente, lo que resulto en un set de datos limpio de 73 atributos de secuencia.

Atributos de estructura: los atributos asociados a características estructurales entregados por el programa PGLP3D son 153. Al analizar se encontró que un gran número de ellos entregaba datos nulos, esto se puede justificar ya que la utilización del programa PGLP3D se basó en fragmentos y no proteínas completas, por lo que se esperaba que varios de los atributos que entrega el programa no obtuvieran un resultado correcto. Por lo tanto, se eliminaron los datos nulos y el set de datos resultante es de 37 atributos estructurales encontrándose entre los atributos correctamente caracterizados distancias, ángulos, hidrofobicidad, aminoácidos (polares, no polares, negativos, positivos, neutros), energía, tamaño alfa etc.

Por lo tanto, los sets de datos finales a analizar son:

- 73 atributos de secuencia
- 37 atributos estructurales

3. **Resultados objetivo específico 3:** Evaluar modelos Máquinas de Vectores de Soporte y Random Forest para su aplicación en la predicción de sitios de unión al ADN**.**

3.1 Selección de características y aplicación de modelos.

Debido a la baja cantidad de ejemplos disponibles (126 fragmentos), y la alta cantidad de atributos para entrenar los modelos, es necesario realizar una selección de atributos para evitar el problema denominado *“maldición de la dimensionalidad”*, ya que una alta dimensionalidad afecta en los cálculos de los modelos de predicción (Berisha et al., 2021).

Para la selección se utilizaron las funciones Mutual Information y ANOVA, donde inicialmente se realizó un ranking de los atributos con cada una de estas funciones, para luego eliminar aquellos atributos independientes que no aportarían mayor aprendizaje al modelo. Luego se procedió a una evaluación acumulativa de los atributos utilizando el modelo SVM, para ver el desempeño global de los atributos y finalmente se hizo una selección de aquellos atributos destacados por ambas funciones, para evaluar el desempeño de estos atributos relevantes en los modelos SVM y RF.

3.1.1 Atributos de secuencia

Se analizaron los 73 atributos de secuencia utilizando las funciones de selección de atributos Mutual Information y ANOVA, donde estas funciones realizan un ranking de los atributos dando un puntaje de mejor a peor atributo para entrenar un modelo (ver Anexo 1).

Los resultados de ambas funciones de clasificación muestran aquellos atributos que tienen un mejor puntaje para entrenar los modelos predictivos, se eliminan aquellos atributos que tienen puntaje igual a 0 para Mutual Information, y puntajes cercanos o igual a 0 para ANOVA. Para Mutual Information valores iguales a cero indican que los atributos no tienen mayor relevancia y no realizan un buen aporte al aprendizaje del modelo predictivo, mientras que con ANOVA lo anterior se refleja en puntajes cercanos o iguales a cero, por lo tanto, para Mutual Information se seleccionan 56/73 atributos, y para ANOVA se seleccionan 53/73 atributos.

Dados los atributos seleccionados por cada función se evalúa el rendimiento estos atributos de forma acumulada (ver Figura 10), utilizando el modelo SVM y la métrica de rendimiento Accuracy.

Al observar las gráficas acumuladas se observa en ambos casos que el rendimiento inicial de los atributos rankeados por las funciones no permiten tener un rendimiento superior al 70%, sin embargo, a medida que aumentan los atributos comienza a mejorar el rendimiento y a mantenerse, siendo ANOVA la función que logra un mejor desempeño con el ranking de atributos.

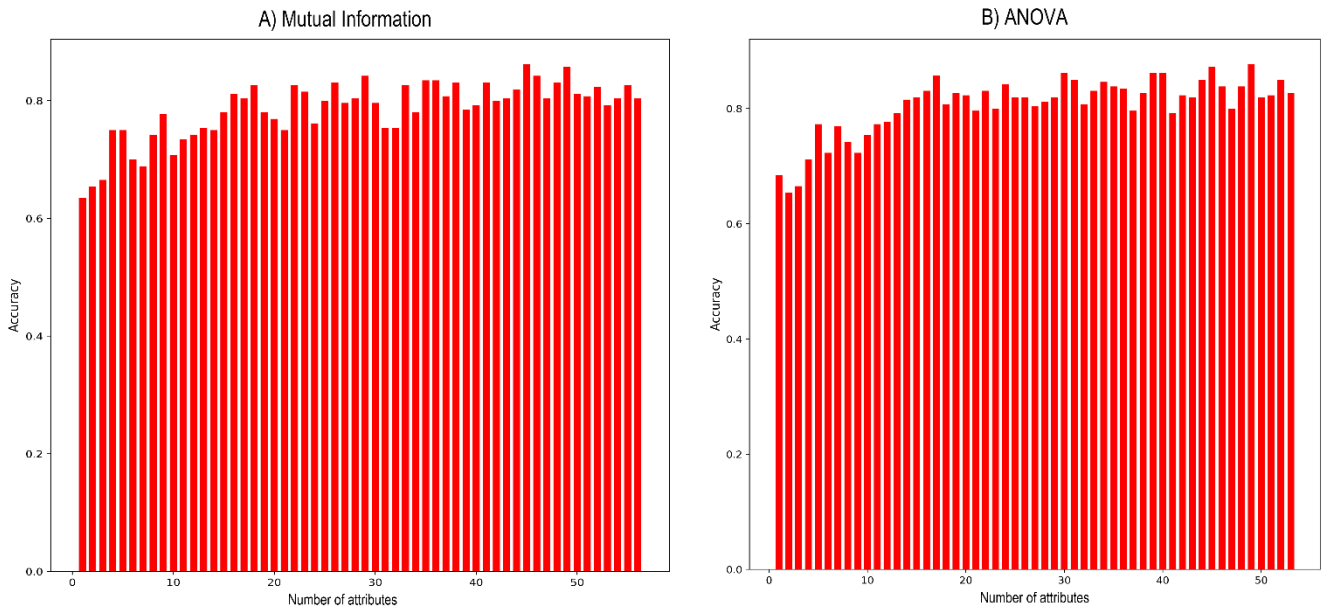

Figura 10. Atributos de secuencia acumulados. La siguiente figura muestra dos gráficas, la gráfica A) corresponde al rendimiento de los 56 atributos acumulados utilizando Mutual Information, mientras que la gráfica B) muestra el rendimiento de los 53 atributos acumulados rankeados por ANOVA. Ambas graficas muestran el rendimiento de Accuracy en el modelo SVM, los valores de accuracy se representan desde 0 a 1, lo que es equivalente en porcentaje a 0%-100%.

Por lo tanto, dado el análisis anterior existe un mínimo de atributos para que el modelo tenga un buen aprendizaje, tanto en *Mutual Information*, como en ANOVA, en donde sobre 10 atributos el rendimiento comienza a mantenerse superior al 70%. Sin embargo, los atributos son diferentes según el ranking que dio cada función, y ante esto se procedió a analizar el orden del ranking de los atributos para seleccionar los más relevantes para entrenar modelos.

Se procede a realizar una selección de atributos que comparten Mutual Information y ANOVA, dentro de las primeras 30 posiciones (ver Anexo 2),

donde se encuentran 16 atributos que comparten ambas funciones dentro del ranking. Debido al mejor rendimiento de ANOVA se considera ese orden y aquellos atributos que comparte con Mutual Information, de estos 16 atributos destacados se buscó el número mínimo de atributos que logra un mejor desempeño en los modelos SVM y RF, los cuales corresponden a:

at_charge, **at_pi**, **NumBasic**, **Blosum7**, **at_lengthpep**, **NumPolar**, **PorcBasic**, **at_mw**, **st8**, **AlphaAndTurnPropensities**, **Blosum1**, **z4**.

Los 12 atributos de secuencia seleccionados como relevantes indican características que efectivamente se han utilizado, o mencionado en otros estudios como características capaces de determinar si existe capacidad de unión al ADN o no. Por ejemplo, el tipo de residuo existente en la zona y la carga que presentan, particularmente los residuos cargados positivamente tienen más probabilidades de interactuar con el ADN (Si et al., 2015). Dentro de estos atributos relevantes encontramos **at_charge** corresponde a la carga neta del fragmento, el atributo **NumBasic** indica el número de aminoácidos básicos, **NumPolar** indica el número de aminoácidos polares, **PorcBasic** indica el porcentaje de aminoácidos básicos y **at_lengthpep** que indica la cantidad de aminoácidos (Rice et al., 2000).

La información evolutiva también se ha utilizado en la predicción de sitios funcionales (Ahmad et al., 2008), característica evolutiva presente en los atributos **Blosum7** y **Blosum1**, donde estos simulan cambios evolutivos y pueden calcular la similitud entre secuencias (Georgiev, 2009). También se encuentran dentro de estas 12 características el atributo **at_mw** que entrega el peso molecular de la secuencia analizada, propiedades topológicas con el atributo **st8** (Pronk et al., 2013), un atributo de **escalas z (z4)** que se relaciona con la electronegatividad (Sandberg et al., 1998), el atributo **at_pi** caracteriza el punto isoeléctrico (Liang et al., 2008) y **AlphaAndTurnPropensities** que indica la inclinación que tendría la secuencia por una estructura del tipo alfa o giro (Liang & Li, 2007).

El rendimiento de estos 12 atributos relevantes se observa en la Figura 11, lo cual indica que el modelo SVM presenta un mejor desempeño, superando el 80% de rendimiento en todas las métricas, mientras que el desempeño del modelo RF es inferior superando solo un 70%. En ambos casos se observa que las medidas Accuracy, Recall y Precision obtuvieron valores similares.

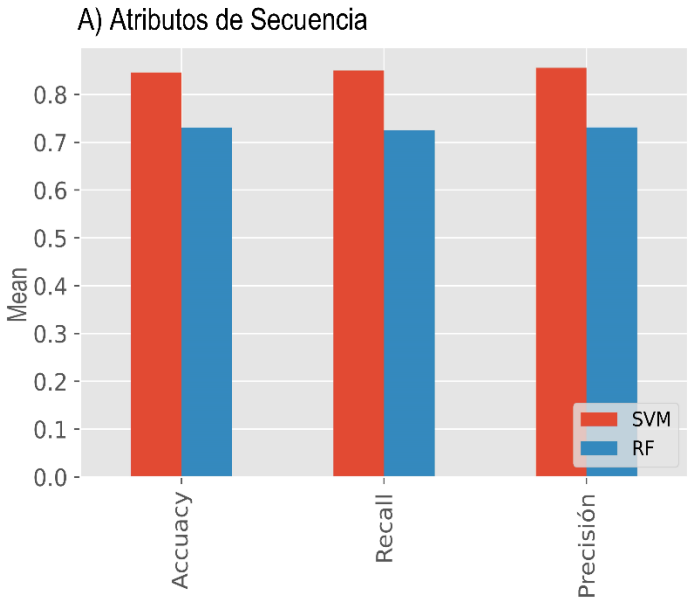

Figura 11. Rendimiento de modelos SVM y RF en atributos de secuencia. La siguiente figura presenta los rendimientos de los 12 atributos de secuencia seleccionados, los valores de las metricas van de 0 a 1, lo que es equivalente en porcentaje a 0%-100%.

Estos 12 atributos seleccionados muestran un buen rendimiento, siendo el modelo destacado SVM, donde su desempeño de SVM indica que estas características son relevantes y permiten la predicción de sitios de unión, siendo capaz de diferenciar entre una zona de unión y una sin capacidad de unión al ADN.

3.1.2 Atributos de estructura

Para el análisis de los atributos estructurales se realizó el mismo procedimiento anterior, donde se analizan los 37 atributos estructurales con las funciones Mutual Information y ANOVA, para obtener el ranking de los atributos (ver Anexo 3). Los resultados dan como atributos seleccionados 32/37 por la función Mutual Information, y 28/37 por la función ANOVA, y se procede a evaluar el rendimiento acumulado de los atributos rankeados (ver Figura 12).

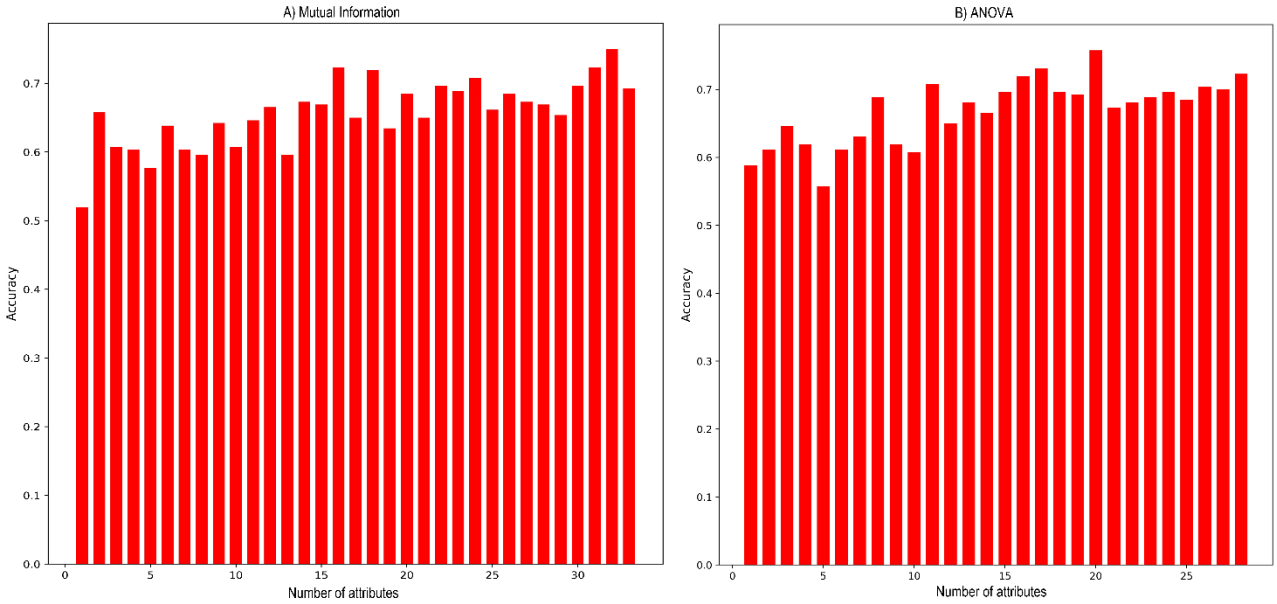

Figura 12. Atributos de estructura acumulados. La siguiente figura muestra dos gráficas, la gráfica A) corresponde al rendimiento de los 32 atributos acumulados utilizando Mutual Information, mientras que la gráfica B) muestra el rendimiento de los 28 atributos acumulados rankeados por ANOVA. Ambas graficas muestran el rendimiento de Accuracy en el modelo SVM, los valores de accuracy se representan desde 0 a 1, lo que es equivalente en porcentaje a 0%-100%.

Las gráficas muestran que los atributos estructurales tienen un menor rendimiento en comparación con el rendimiento obtenido por los atributos de secuencia, independiente de la función de selección de atributos. Los atributos estructurales en promedio general logran utilizando Mutual Information un 65.1% de Accuracy y con ANOVA un 67.2% (promedios estimados desde los atributos acumulados, Figura 12), mientras que los atributos de secuencia alcanzan un promedio general de 78.1% de Accuracy con Mutual Information y un 79.4% con ANOVA (promedios estimados desde

los atributos acumulados, Figura 10). Por lo tanto, al observar las gráficas se prevé que el rendimiento de los modelos con atributos estructurales será inferior al de los atributos de secuencia.

Ante el rendimiento observado con los atributos acumulados, se observa que la función ANOVA logra un mejor rendimiento según el orden en que selecciona los atributos. Debido a la baja cantidad de atributos estructurales, ambas funciones comparten dentro de las primeras 22 posiciones 15 atributos (ver Anexo 4), por lo tanto, se busca el mínimo de atributos que logra un mejor desempeño en los modelos SVM y RF, los cuales corresponden a:

Hidrofobicidad_K, **N-N**, **Hidrofobicidad_H**, **C-O**, **Total_aa_sitio**, **C-C**, **O-O**, **energia_captacion_4aa**, **C-N**.

Dentro de estos 9 atributos relevantes se encuentran características de hidrofobicidad (**Hidrofobicidad_K** y **Hidrofobicidad_H)**, enlaces atómicos como **N-N**, **C-O**, **C-C**, **O-O** y **C-N**, cantidad de aminoácidos en la secuencia (**Total_aa_sitio**), y finalmente un atributo de energía de captación (**energia_captacion_4aa**).

El rendimiento de estos 9 atributos relevantes se observa en la Figura 13, lo que indica que SVM y RF solo con estos 9 atributos relevantes mejora considerablemente a diferencia de lo observado en la Figura 11 con los atributos acumulados. Ambos modelos superan el 70% de rendimiento en todas las métricas evaluadas, las cuales mantienen un equilibrio y demuestran que no existe mayor problema de desbalance o de maldición de la dimensionalidad.

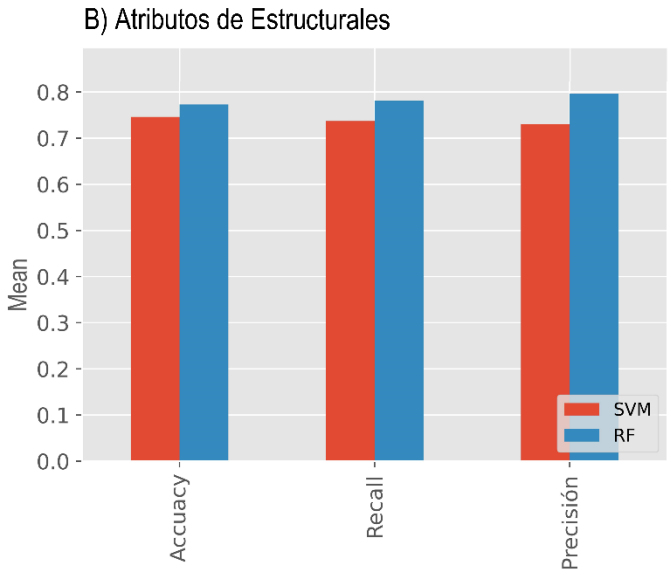

Figura 13. Rendimiento de modelos SVM y RF en atributos estructurales. La siguiente figura muestra los rendimientos de los 9 atributos de estructura seleccionados, los valores van de 0 a 1, lo que es equivalente en porcentaje a 0%-100%.

Ambos modelos tienen un buen rendimiento, sin embargo, a diferencia de los atributos de secuencia, el modelo RF es el modelo que logra el mejor desempeño superando el 75% en todas las métricas evaluadas.

A continuación, se observa la Tabla 3, con los resultados exactos (observados anteriormente de forma gráfica) de ambos modelos (SVM y RF) tanto en los atributos de secuencia como de estructura:

Data set type	Number of attributes	SVM			RF		
		Accuracy	Recall	Precision	Accuracy	Recall	Precision
Sequence	12	84.6%	85.0%	85.5%	73.0%	72.5%	73.1%
Structure	9	74.6%	73.7%	73.0%	77.3%	78.1%	79.6%

Tabla 3. Resultado de los modelos SMV y RF. La siguiente tabla contiene los rendimientos de las métricas Accuracy, Recall y Precision en los modelos SVM y RF, utilizando los atributos óptimos en secuencia y estructura.

Con los resultados finales se aprecia que el conjunto de 12 atributos de secuencia seleccionados, son características óptimas para la predicción de sitios de unión utilizando el modelo SVM, el cual logra un rendimiento del 84.6% en Accuracy, lo que indica la proporción entre los fragmentos

positivos y negativos predichos correctamente, un 85% en Recall, o la capacidad de predecir correctamente fragmentos positivos, y finalmente logra un 85.5% de Precision que indica la capacidad del predictor para clasificar correctamente un fragmento negativo y no clasificarlo como un falso positivo.

Finalmente, el conjunto de 9 atributos de estructura seleccionados, permiten que el modelo RF destaque su rendimiento con un 77.3% en Accuracy, 78.9% en Recall y 79.6% en Precision, resultados bastante cercanos al 80%, por lo que son características que permiten al modelo predecir sitios de unión al ADN.

DISCUSIÓN

El desarrollo y construcción del set de datos analizado, permitió generar un modelo predictivo de sitios de unión al ADN utilizando características de diversos complejos proteína-ADN de origen bacteriano.

Durante este trabajo, el set de datos utilizado se conformó de zonas de unión al ADN de los 63 complejos proteína-ADN obtenidos desde PDB, dentro de los 63 complejos estudiados, se encuentran 2 complejos pertenecientes a la superfamilia Fur, las proteínas 4MTD y 4RB1, mientras que los 61 complejos restantes son diversos factores de transcripción. Como se analizó en el árbol filogenético presentado (Figura 7), los complejos se pueden subdividir en 8 grupos, observando que las proteínas de la superfamilia Fur presentes no se ubican en un mismo grupo. El estudio de esta diferencia indicó que a pesar de estar en distinto grupo estas proteínas de interés presentan una zona de unión estructuralmente similar, así como con aquellas proteínas más cercanas a nivel grupal 4MTD con 4AIJ y 4RB1 con 4L62, donde los 4 fragmentos tienen un bajo nivel de identidad de secuencia, pero una alta similitud estructural con un RMSD general de 0.6868 (valor calculado con PDBeFold).

La caracterización de las 63 zonas de unión se basó en la utilización de descriptores moleculares de secuencia y estructura, de las que se obtuvo 73 atributos de secuencia y 37 atributos estructurales.

Es importante mencionar que una de las dificultades al momento de realizar caracterización estructural es que los softwares actualmente creados como Fpocket (Le Guilloux et al., 2009), VADAR (Willard et al., 2003), ProtDCal (Ruiz-Blanco et al., 2015), y PLGP3D están pensados en analizar proteínas completas y no fragmentos de éstas, por lo que simplemente no permitieron caracterizar fragmentos, o entregaron una cantidad mínima de características, como es el caso de PLGP3D que permitió obtener 37 atributos, encontrando distancias, ángulos,

hidrofobicidades, enlaces etc. Sin embargo, no se encuentran características como potencial electroestático, carga neta, momentos dipolares y cuadripolares etc., características que se han demostrado en estudios previos que son importantes, y que, por ejemplo, en cuanto al potencial electroestático se ha estudiado que la magnitud de momentos de distribución de la carga eléctrica es significativamente diferente en sitios de unión al ADN comparados con los sitios de no unión (Si et al., 2015).

Por lo tanto, en un futuro una mejor caracterización estructural de los fragmentos en la cual se puedan incluir características importantes como el potencial eléctrico, se deben evaluar para saber si son un factor determinante y aumentan mucho más el rendimiento de los modelos predictivos.

Para evitar la *"maldición de la dimensionalidad"* (Berisha et al., 2021), debido a la cantidad de atributos obtenidos en la caracterización, se realizó una selección de atributos utilizando Mutual Information y ANOVA, donde el resultado de esta selección redujo la cantidad de atributos necesarios para lograr un buen rendimiento en los modelos.

Se logró un mínimo de 12 atributos utilizando el set de datos con información de secuencia, obteniendo un rendimiento de Accuracy de 84.6%, Recall de 85.0% y 85.5% de Precision con el modelo SVM, mientras que el modelo RF obtuvo un 73.0% de Accuracy, 72.5 en Recall y 73.1% en Precision. Por otro lado, el set de datos con caracterización estructural logró un mínimo de 9 atributos, con valores de 74.6% en Accuracy, 73.7% en Recall y 73.0% en Precision utilizando el modelo SVM, mientras que el rendimiento obtenido con el modelo RF es de 77.3% en Accuracy, 78.1% en Recall y 79.6% en Precision.

Los modelos de clasificación generados muestran métricas de desempeño equilibradas y con buenos rendimientos, lo que corrobora que no existe desbalance de clases o problemas de dimensionalidad. El buen rendimiento

obtenido demuestra que existen atributos relevantes para entrenar distintos modelos, y que el modelo SVM es el modelo más apto para la predicción dada información de secuencia, mientras que el modelo RF es más idóneo en la predicción utilizando características estructurales.

La predicción de zonas de unión al ADN utilizando información de secuencia se ha investigado desarrollando métodos como HMMPred, que propone una predicción de sitios de unión al ADN basada en características extraídas desde perfiles HMM, con una selección de características mediante la técnica XGBoost y el clasificador SVM, su método alcanza un Accuracy del 83.90% (Sang et al., 2020). También se encuentran estudios como el método RF-SVM que logra un 84.25% de Accuracy, donde este método utiliza RF como un seleccionador de características y SVM como predictor, y su predicción se basa en 174 atributos seleccionados por RF (Yanping Zhang et al., 2022). Comparando el método presentado en esta investigación, la selección de 12 atributos de secuencia mediante las funciones Mutual Information y ANOVA, permiten un rendimiento del 84.6% de Accuracy en SVM, lo que supera levemente los métodos HMMPred y RF-SVM. Sin embargo, es importante tener en conocimiento que los métodos HMMPred y RF-SVM utilizan un conjunto de datos mayor debido a la utilización de diferentes proteínas y no complejos proteína-ADN como fue el caso de este estudio. Por lo tanto, para seguir mejorando el método expuesto sería importante realizar una prueba utilizando un nuevo set de datos de secuencias de proteínas, manteniendo el set de complejos proteína-ADN como set de entrenamiento del predictor SVM y las 12 características seleccionadas.

Por otra parte, la predicción de proteínas con capacidad de unión al ADN también se ha estudiado desde las características estructurales, por ejemplo, el método iDNAProt-ES utiliza características estructurales y evolutivas para identificar la funcionalidad de unión al ADN en proteínas mediante el predictor SVM, logrando un rendimiento de 80.64% de Accuracy

(Chowdhury et al., 2017). También se han evaluado las características estructurales enfocadas en la predicción de zonas de unión como es el caso de DeepDISE, método que mediante aprendizaje profundo y características estructurales logra un 88.0% de rendimiento en Accuracy, un alto rendimiento que obtiene con la utilización de la estructura 3D de las proteínas más la información del tipo de átomo en la superficie (Hendrix et al., 2021). En esa investigación, por el contrario, no se logró un rendimiento sobresaliente utilizando solo características estructurales, y el mejor rendimiento fue de un Accuracy del 77.3% utilizando el predictor RF. En consecuencia, como se mencionó anteriormente es importante en un futuro mejorar los descriptores estructurales enfocados en zonas de unión.

Por último, esta investigación muestra que se puede desarrollar un método rápido y sencillo para predecir zonas de unión al ADN en diferentes grupos de proteínas, utilizando solo características obtenidas desde la zona de unión de complejos proteína-ADN.

CONCLUSIONES

En esta investigación se consiguió evaluar un modelo de predicción de zonas de unión al ADN en proteínas de origen bacteriano, pertenecientes a la Superfamilia Fur y otros factores de transcripción.

Se desarrollo un set de datos de zonas de unión basadas en complejos proteína-ADN, cada zona de unión al ADN fue caracterizada a nivel de secuencia y de estructura, estimando atributos de propiedades de composición de aminoácidos, fisicoquímicas, y estructurales.

A pesar de la baja conservación aminoacídica a nivel de secuencia de cada una de las zonas de unión de las 63 proteínas estudiadas, se logró mediante funciones de selección de características, encontrar 12 atributos de secuencia que permiten discriminar eficientemente una zona de unión al ADN. Mientras que, aun cuando existe similitud estructural entre estas zonas de unión, se pudo encontrar 9 atributos estructurales relevantes que permiten realizar una predicción correcta.

Los métodos de predicción implementados logran rendimientos robustos y equilibrados, sin problema de desbalance o dimensionalidad, siendo Máquinas de Vectores de Soporte el modelo predictivo óptimo para una predicción a nivel de secuencia, y Random Forest a nivel estructural. De lo anterior se concluye que, sí es posible desarrollar un predictor de sitios de unión al ADN utilizando descriptores moleculares y modelos predictivos, ya que se demostró que existen características a nivel de secuencia que permiten diferenciar una zona de unión al ADN con un 85% de rendimiento utilizando el modelo Máquinas de Vectores de Soporte.

En un trabajo futuro sería importante desarrollar una mejor caracterización estructural de la zona de unión, que permita obtener características más informativas para el rendimiento de los predictores. En cuanto al método aplicado con información de secuencia, sería conveniente

realizar una nueva prueba, con el set de datos de complejos proteína-ADN como entrenamiento de los predictores y otro set de datos de solo proteínas como testeo, para evaluar así la solidez y capacidad de generalización del predictor.

REFERENCIAS

Agriesti, F., Roncarati, D., Musiani, F., Del Campo, C., Iurlaro, M., Sparla, F., Ciurli, S., Danielli, A., & Scarlato, V. (2014). FeON-FeOFF: the Helicobacter pylori Fur regulator commutates iron-responsive transcription by discriminative readout of opposed DNA grooves. *Nucleic Acids Research*, *42*(5), 3138–3151. https://doi.org/10.1093/nar/gkt1258

Ahmad, S., Keskin, O., Sarai, A., & Nussinov, R. (2008). Protein-DNA interactions: Structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Research*, *36*(18), 5922–5932. https://doi.org/10.1093/nar/gkn573

Ahn, B. E., Cha, J., Lee, E. J., Han, A. R., Thompson, C. J., & Roe, J. H. (2006). Nur, a nickel-responsive regulator of the Fur family, regulates superoxide dismutases and nickel transport in Streptomyces coelicolor. *Molecular Microbiology*, *59*(6), 1848–1858. https://doi.org/10.1111/j.1365-2958.2006.05065.x

Akcapinar, G. B., & Sezerman, O. U. (2017). Computational approaches for de novo design and redesign of metal-binding sites on proteins. *Bioscience Reports*, *37*(2), 1–12. https://doi.org/10.1042/BSR20160179

Barupal, D. K., & Fiehn, O. (2019). Generating the blood exposome database using a comprehensive text mining and database fusion approach. *Environmental Health Perspectives*, *127*(9), 2825–2830. https://doi.org/10.1289/EHP4713

Bellini, P., & Hemmings, A. M. (2006). In vitro characterization of a bacterial manganese uptake regulator of the fur superfamily. *Biochemistry*, *45*(8), 2686–2698. https://doi.org/10.1021/bi052081n

Berg, K., Pedersen, H. L., & Leiros, I. (2020). Biochemical characterization of ferric uptake regulator (Fur) from Aliivibrio salmonicida. Mapping the DNA sequence specificity through binding studies and structural modelling. *BioMetals*, *33*(4–5), 169–185. https://doi.org/10.1007/s10534-020-00240-6

Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss,

J. (2021). Digital medicine and the curse of dimensionality. *Npj Digital Medicine*, *4*(1), 1–8. https://doi.org/10.1038/s41746-021-00521-5

Bsat, N., Herbig, A., Casillas-Martinez, L., Setlow, P., & Helmann, J. D. (1998). Bacillus subtilis contains multiple Fur homologues: Identification of the iron uptake (Fur) and peroxide regulon (PerR) repressors. *Molecular Microbiology*, *29*(1), 189–198. https://doi.org/10.1046/j.1365-2958.1998.00921.x

Butcher, J., Sarvan, S., Brunzelle, J. S., Couture, J. F., & Stintzi, A. (2012). Structure and regulon of Campylobacter jejuni ferric uptake regulator fur define apo-Fur regulation. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(25), 10047–10052. https://doi.org/10.1073/pnas.1118321109

Ceroni, A., Passerini, A., Vullo, A., & Frasconi, P. (2006). Disulfind: A disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.). https://doi.org/10.1093/nar/gkl266

Chowdhury, S. Y., Shatabda, S., & Dehzangi, A. (2017). IDNAProt-ES: Identification of DNA-binding Proteins Using Evolutionary and Structural Features. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/s41598-017-14945-1

Deng, Z., Wang, Q., Liu, Z., Zhang, M., Machado, A. C. D., Chiu, T. P., Feng, C., Zhang, Q., Yu, L., Qi, L., Zheng, J., Wang, X., Huo, X. M., Qi, X., Li, X., Wu, W., Rohs, R., Li, Y., & Chen, Z. (2015). Mechanistic insights into metal ion activation and operator recognition by the ferric uptake regulator. *Nature Communications*, *6*(May). https://doi.org/10.1038/ncomms8642

Dian, C., Vitale, S., Leonard, G. A., Bahlawane, C., Fauquant, C., Leduc, D., Muller, C., De Reuse, H., Michaud-Soret, I., & Terradot, L. (2011). The structure of the Helicobacter pylori ferric uptake regulator Fur reveals three functional metal binding sites. *Molecular Microbiology*, *79*(5), 1260–1275. https://doi.org/10.1111/j.1365-2958.2010.07517.x

Fillat, M. F. (2014). The fur (ferric uptake regulator) superfamily: Diversity and versatility of key transcriptional regulators. *Archives of Biochemistry and Biophysics*, *546*, 41–52. https://doi.org/10.1016/j.abb.2014.01.029

Georgiev, A. G. (2009). Interpretable numerical descriptors of amino acid space. *Journal of Computational Biology*, *16*(5), 703–723. https://doi.org/10.1089/cmb.2008.0173

Gilston, B. A., Wang, S., Marcus, M. D., Canalizo-Hernández, M. A., Swindell, E. P., Xue, Y., Mondragón, A., & O'Halloran, T. V. (2014). Structural and Mechanistic Basis of Zinc Regulation Across the E. coli Zur Regulon. *PLoS Biology*, *12*(11). https://doi.org/10.1371/journal.pbio.1001987

Hashimoto, D. A., Rosman, G., Rus, D., & Meireles, O. R. (2018). Artificial Intelligence in Surgery: Promises and Perils. In *Annals of Surgery* (Vol. 268, Issue 1, pp. 70–76). NIH Public Access. https://doi.org/10.1097/SLA.0000000000002693

Hendrix, S. G., Chang, K. Y., Ryu, Z., & Xie, Z. R. (2021). Deepdise: Dna binding site prediction using a deep learning method. *International Journal of Molecular Sciences*, *22*(11). https://doi.org/10.3390/ijms22115510

Kaushik, M. S., Singh, P., Tiwari, B., & Mishra, A. K. (2016). Ferric Uptake Regulator (FUR) protein: properties and implications in cyanobacteria. *Annals of Microbiology*, *66*(1), 61–75. https://doi.org/10.1007/s13213-015-1134-x

Kim, Y., Kang, Y., & Choe, J. (2013). Crystal structure of Pseudomonas aeruginosa transcriptional regulator PA2196 bound to its operator DNA. *Biochemical and Biophysical Research Communications*, *440*(2), 317–321. https://doi.org/10.1016/j.bbrc.2013.09.074

Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, *10*. https://doi.org/10.1186/1471-2105-10-168

Lee, J. W., & Helmann, J. D. (2007). Functional specialization within the fur family of metalloregulators. *BioMetals*, *20*(3–4), 485–499. https://doi.org/10.1007/s10534-006-9070-7

Liang, G., Chen, G., Niu, W., & Li, Z. (2008). Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human

amphiphysin-1 SH3 domains and their peptide ligands. *Chemical Biology and Drug Design*, *71*(4), 345–351. https://doi.org/10.1111/j.1747-0285.2008.00641.x

Liang, G., & Li, Z. (2007). Factor analysis scale of generalized amino acid information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides. *QSAR and Combinatorial Science*, *26*(6), 754–763. https://doi.org/10.1002/qsar.200630145

Ma, X., Guo, J., & Sun, X. (2016). DNABP: Identification of DNA-binding proteins based on feature selection using a random forest and predicting binding residues. *PLoS ONE*, *11*(12), 1–20. https://doi.org/10.1371/journal.pone.0167345

Ma, X., Wu, J., & Xue, X. (2013). Identification of DNA-binding proteins using support vector machine with sequence information. *Computational and Mathematical Methods in Medicine*, *2013*. https://doi.org/10.1155/2013/524502

Makthal, N., Rastegari, S., Sanson, M., Ma, Z., Olsen, R. J., Helmann, J. D., Musser, J. M., & Kumaraswami, M. (2013). Crystal structure of peroxide stress regulator from streptococcus pyogenes provides functional insights into the mechanism of oxidative stress sensing. *Journal of Biological Chemistry*, *288*(25), 18311–18324. https://doi.org/10.1074/jbc.M113.456590

Mehmood, R., & Selwal, A. (2020). Fingerprint biometric template security schemes: Attacks and countermeasures. In *Lecture Notes in Electrical Engineering* (Vol. 597). https://doi.org/10.1007/978-3-030-29407-6_33

Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1), 1–14. https://doi.org/10.1186/s13321-018-0258-y

Panina, E. M., Mironov, A. A., & Gelfand, M. S. (2003). Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(17), 9912–9917. https://doi.org/10.1073/pnas.1733691100

Pérard, J., Nader, S., Levert, M., Arnaud, L., Carpentier, P., Siebert, C., Blanquet, F., Cavazza, C., Renesto, P., Schneider, D., Maurin, M., Coves, J., Crouzy, S., & Michaud-Soret, I. (2018). Structural and functional studies of the metalloregulator Fur identify a promoter-binding mechanism and its role in Francisella tularensis virulence. *Communications Biology*, *1*(1). https://doi.org/10.1038/s42003-018-0095-6

Pinochet-Barros, A., & Helmann, J. D. (2018). Redox Sensing by Fe2+ in Bacterial Fur Family Metalloregulators. *Antioxidants and Redox Signaling*, *29*(18), 1858–1871. https://doi.org/10.1089/ars.2017.7359

Pohl, E., Haller, J. C., Mijovilovich, A., Meyer-Klaucke, W., Garman, E., & Vasil, M. L. (2003). Architecture of a protein central to iron homeostasis: Crystal structure and spectroscopic analysis of the ferric uptake regulator. *Molecular Microbiology*, *47*(4), 903–915. https://doi.org/10.1046/j.1365-2958.2003.03337.x

Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., Van Der Spoel, D., Hess, B., & Lindahl, E. (2013). GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, *29*(7), 845–854. https://doi.org/10.1093/bioinformatics/btt055

Quade, N., Mendonca, C., Herbst, K., Heroven, A. K., Ritter, C., Heinz, D. W., & Dersch, P. (2012). Structural basis for intrinsic thermosensing by the master virulence regulator RovA of Yersinia. *Journal of Biological Chemistry*, *287*(43), 35796–35803. https://doi.org/10.1074/jbc.M112.379156

Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Ruiz-Blanco, Y. B., Paz, W., Green, J., & Marrero-Ponce, Y. (2015). ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics*, *16*(1), 1–15. https://doi.org/10.1186/s12859-015-0586-0

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., & Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry*, *41*(14), 2481–2491. https://doi.org/10.1021/jm9700575

Sang, X., Xiao, W., Zheng, H., Yang, Y., & Liu, T. (2020). HMMPred: Accurate Prediction of DNA-Binding Proteins Based on HMM Profiles and XGBoost Feature Selection. *Computational and Mathematical Methods in Medicine*, *2020*. https://doi.org/10.1155/2020/1384749

Shanahan, H. P., Garcia, M. A., Jones, S., & Thornton, J. M. (2004). Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Research*, *32*(16), 4732–4741. https://doi.org/10.1093/nar/gkh803

Si, J., Zhao, R., & Wu, R. (2015). An overview of the prediction of protein DNA-binding sites. *International Journal of Molecular Sciences*, *16*(3), 5194–5215. https://doi.org/10.3390/ijms16035194

Terán, J. E., Marrero-Ponce, Y., Contreras-Torres, E., García-Jacas, C. R., Vivas-Reyes, R., Terán, E., & Torres, F. J. (2019). Tensor Algebra-based Geometrical (3D) Biomacro-Molecular Descriptors for Protein Research: Theory, Applications and Comparison with other Methods. *Scientific Reports*, *9*(1), 1–15. https://doi.org/10.1038/s41598-019-47858-2

Tian, F., Zhou, P., & Li, Z. (2007). T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *Journal of Molecular Structure*, *830*(1–3), 106–115. https://doi.org/10.1016/j.molstruc.2006.07.004

Van Westen, G. J. P., Swier, R. F., Wegner, J. K., Jzerman, A. P. I., Van Vlijmen, H. W. T., & Bender, A. (2013). Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets. *Journal of Cheminformatics*, *5*(9). https://doi.org/10.1186/1758-2946-5-41

Weng, C. C., Chen, S. Y., & Chang, J. C. (2013). Predicting remaining discharge time of a Lithium-ion battery by using residual capacity and workload. *Proceedings of*

the International Symposium on Consumer Electronics, ISCE, *4*(1), 179–180. https://doi.org/10.1109/ISCE.2013.6570172

Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D., & Wishart, D. S. (2003). VADAR: A web server for quantitative evaluation of protein structure quality. *Nucleic Acids Research*, *31*(13), 3316–3319. https://doi.org/10.1093/nar/gkg565

Wu, J., Liu, H., Duan, X., Ding, Y., Wu, H., Bai, Y., & Sun, X. (2009). Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, *25*(1), 30–35. https://doi.org/10.1093/bioinformatics/btn583

Xiong, Y., Xia, J., Zhang, W., & Liu, J. (2011). Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures. *PLoS ONE*, *6*(12). https://doi.org/10.1371/journal.pone.0028440

Yan, C., Terribilini, M., Wu, F., Jernigan, R. L., Dobbs, D., & Honavar, V. (2006). Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, *7*, 1–10. https://doi.org/10.1186/1471-2105-7-262

Yang, L., Shu, M., Ma, K., Mei, H., Jiang, Y., & Li, Z. (2010). ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. *Amino Acids*, *38*(3), 805–816. https://doi.org/10.1007/s00726-009-0287-y

Zaliani, A., & Gancia, E. (1999). MS-WHIM scores for amino acids: A new 3D-description for peptide QSAR and QSPR studies. *Journal of Chemical Information and Computer Sciences*, *39*(3), 525–533. https://doi.org/10.1021/ci980211b

Zhang, Yan, & Zheng, J. (2020). Bioinformatics of metalloproteins and metalloproteomes. *Molecules*, *25*(15), 1–23. https://doi.org/10.3390/molecules25153366

Zhang, Yanping, Ni, J., & Gao, Y. (2022). RF-SVM: Identification of DNA-binding proteins based on comprehensive feature representation methods and support vector machine. *Proteins: Structure, Function and Bioinformatics*, *90*(2), 395–404.

https://doi.org/10.1002/prot.26229

ANEXOS

1. Anexo 1: Ranking de atributos de secuencia.

El siguiente anexo muestra el ranking de los atributos de secuencia realizados por las funciones Mutual Information y ANOVA, en cada lista se observa el orden de los atributos por puntaje y en letra roja aquellos atributos eliminados por corresponder atributos con un bajo puntaje, lo que indica que son variables independientes y no aportan aprendizaje al modelo.

1. Atributos rankeados por Mutual Information

#	**Atributos**	**Puntaje**
1	at_hmoment_sheet	0,148557
2	NumPolar	0,134732
3	NumAliphatic	0,133264
4	at_charge	0,127759
5	st4	0,120256
6	Blosum7	0,119705
7	DoubleBendPreference	0,117646
8	Blosum10	0,117167
9	NumBasic	0,114918
10	PorcBasic	0,110322
11	FlatExtendedPreference	0,108349
12	at_pi	0,104045
13	z4	0,101813
14	SideChainSize	0,099285
15	MsWhim2	0,097111
16	t3	0,094279
17	Blosum1	0,091895
18	PorcAcidic	0,090899
19	Blosum8	0,089298
20	Blosum2	0,08804
21	at_mw	0,079797
22	st8	0,07953
23	Hidrophobicity	0,078956
24	at_lengthpep	0,076696
25	st5	0,07482
26	Blosum5	0,072924
27	Blosum4	0,068555
28	AlphaAndTurnPropensities	0,063865
29	at_hmoment_alpha	0,05829
30	HydrophobicityIndex	0,055725
31	st1	0,055533
32	z5	0,051675
33	NumSmall	0,050144
34	Blosum6	0,049965
35	Blosum3	0,04907
36	z2	0,048314
37	NumTiny	0,048162
38	HelixBendPreference	0,046067
39	at_boman	0,044545
40	NumAcidic	0,044463
41	ElectronicProperties	0,037653
42	NumNonPolar	0,035095
43	MsWhim1	0,027284
44	at_index	0,026385
45	t5	0,024429
46	st6	0,02233
47	t1	0,022031
48	PorcAromatic	0,018827
49	PorcAliphatic	0,016684
50	BulkyProperties	0,0165
51	t4	0,014979
52	PorcSmall	0,009564
53	CompositionalCharacteristicIndex	0,006982
54	ExtendedStructurePreference	0,006788

55	pKC	0,003818
56	OccurrenceInAlphaRegion	0,003679
57	z1	0,003494
58	NumAromatic	0,000186
59	LocalFlexibility	0
60	NumCharged	0
61	PorcTiny	0
62	z3	0
63	SurroundingHidrophobicity	0
64	t2	0
65	PorcNonPolar	0
66	st3	0
67	st2	0
68	PorcPolar	0
69	MsWhim3	0
70	PorcCharged	0
71	Blosum9	0
72	PartialSpecificVolume	0
73	st7	0

2. Atributos rankeados por ANOVA

#	Atributos	Puntaje
1	at_charge	42,16782
2	at_pi	38,34236
3	NumBasic	29,1845
4	Blosum7	28,18563
5	NumTiny	22,38421
6	at_lengthpep	14,84791
7	NumPolar	14,50483
8	PorcBasic	13,73135
9	at_mw	13,5173
10	st8	12,57037
11	NumSmall	12,38642
12	AlphaAndTurnPropensities	11,88152
13	Blosum1	11,32872
14	at_index	11,13972
15	PorcAcidic	10,91515
16	pKC	10,50836
17	z1	10,09197
18	z4	10,06817
19	st6	9,563418
20	HelixBendPreference	9,350411
21	ElectronicProperties	9,206551
22	NumCharged	8,574254
23	Hidrophobicity	8,268428
24	NumNonPolar	7,954024
25	PorcTiny	7,846448
26	at_hmoment_sheet	7,204809
27	Blosum5	7,128812
28	at_hmoment_alpha	5,181145
29	st2	4,918841
30	at_boman	4,684871
31	st7	4,387042
32	t3	4,039177
33	Blosum10	3,997557
34	NumAliphatic	3,566637
35	PorcAliphatic	3,535762
36	MsWhim3	3,521157
37	Blosum9	3,312629
38	t5	3,205579
39	st5	2,600767
40	z3	2,443989
41	NumAromatic	2,429467
42	HydrophobicityIndex	2,241792
43	PartialSpecificVolume	2,135344
44	Blosum3	1,957142
45	ExtendedStructurePreference	1,749584
46	PorcNonPolar	1,586095
47	BulkyProperties	1,572262
48	t4	1,560462
49	Blosum2	1,503109
50	Blosum8	1,218881
51	PorcPolar	1,139152
52	z5	1,033863
53	NumAcidic	1,029831
54	Blosum4	0,732558
55	Blosum6	0,632319
56	CompositionalCharacteristicIndex	0,599098
57	MsWhim1	0,369814
58	t1	0,185693
59	PorcCharged	0,157627
60	SideChainSize	0,143831
61	PorcSmall	0,132287
62	z2	0,111292
63	st4	0,110587
64	t2	0,10507
65	st1	0,06862
66	DoubleBendPreference	0,064511
67	SurroundingHidrophobicity	0,044781
68	st3	0,024715
69	PorcAromatic	0,012447
70	OccurrenceInAlphaRegion	0,007946
71	LocalFlexibility	0,000727
72	FlatExtendedPreference	0,000406
73	MsWhim2	3,9E-06

2. Anexo 2: Selección de atributos de secuencia.

La siguiente lista muestra los atributos relevantes (resaltados) que ambas funciones seleccionan dentro de los primeros 30 atributos.

#	Atributos Mutual Info.	Puntaje
1	at_hmoment_sheet	0,148557
2	NumPolar	0,134732
3	NumAliphatic	0,133264
4	at_charge	0,127759
5	st4	0,120256
6	Blosum7	0,119705
7	DoubleBendPreference	0,117646
8	Blosum10	0,117167
9	NumBasic	0,114918
10	PorcBasic	0,110322
11	FlatExtendedPreference	0,108349
12	at_pi	0,104045
13	z4	0,101813
14	SideChainSize	0,099285
15	MsWhim2	0,097111
16	t3	0,094279
17	Blosum1	0,091895
18	PorcAcidic	0,090899
19	Blosum8	0,089298
20	Blosum2	0,08804
21	at_mw	0,079797
22	st8	0,07953
23	Hidrophobicity	0,078956
24	at_lengthpep	0,076696
25	st5	0,07482
26	Blosum5	0,072924
27	Blosum4	0,068555
28	AlphaAndTurnPropensities	0,063865
29	at_hmoment_alpha	0,05829
30	HydrophobicityIndex	0,055725

#	Atributos ANOVA	Puntaje
1	at_charge	42,16782
2	at_pi	38,34236
3	NumBasic	29,1845
4	Blosum7	28,18563
5	NumTiny	22,38421
6	at_lengthpep	14,84791
7	NumPolar	14,50483
8	PorcBasic	13,73135
9	at_mw	13,5173
10	st8	12,57037
11	NumSmall	12,38642
12	AlphaAndTurnPropensities	11,88152
13	Blosum1	11,32872
14	at_index	11,13972
15	PorcAcidic	10,91515
16	pKC	10,50836
17	z1	10,09197
18	z4	10,06817
19	st6	9,563418
20	HelixBendPreference	9,350411
21	ElectronicProperties	9,206551
22	NumCharged	8,574254
23	Hidrophobicity	8,268428
24	NumNonPolar	7,954024
25	PorcTiny	7,846448
26	at_hmoment_sheet	7,204809
27	Blosum5	7,128812
28	at_hmoment_alpha	5,181145
29	st2	4,918841
30	at_boman	4,684871

3. Anexo 3: Ranking de atributos de estructura.

El siguiente anexo muestra el ranking de los atributos de estructura realizados por las funciones Mutual Information y ANOVA, en cada lista se observa el orden de los atributos por puntaje y en letra roja aquellos atributos eliminados por corresponder atributos con un bajo puntaje, lo que indica que son variables independientes y no aportan aprendizaje al modelo.

1. Atributos rankeados por Mutual Information

#	Atributos	Puntaje
1	Hidrofobicidad_W	0,22046
2	Hidrofobicidad_H	0,19595
3	C-C	0,174725
4	Distancia_AC1-AC3	0,145755
5	Total_aa_sitio	0,135277
6	O-N	0,114187
7	C-O	0,113372
8	Hidrofobicidad_K	0,112061
9	C-N	0,110606
10	Polares_Negativos	0,104435
11	Distancia_CA1-CA3	0,092917
12	Distancia_A+C1	0,091511
13	No_Polares	0,090847
14	Distancia_CA3-CA4	0,082954
15	Angulo_CA3-C_3	0,080995
16	Angulo_CA1-C_1	0,080318
17	Distancia_AC1-AC2	0,079683
18	Distancia_AC2-AC3	0,078887
19	O-O	0,077474
20	Distancia_A+C2	0,075827
21	energia_captacion_4aa	0,075335
22	N-N	0,064477
23	Polares_Neutros	0,061502
24	Optimal alpha	0,059479
25	Polares_Positivos	0,056886
26	Distancia_CA1-CA2	0,048084
27	Distancia_CA1-CA4	0,04778
28	Angulo_CA4-C_4	0,041852
29	Distancia_CA2-CA3	0,040615
30	Distancia_CA2-CA4	0,040325
31	Angulo_CA2-C_2	0,03567
32	Distancia_A+C4	0,018203
33	Distancia_AC1-AC4	0,006583
34	Distancia_AC2-AC4	0
35	Distancia_AC3-AC4	0
36	Smallest alpha	0
37	Distancia_A+C3	0

2. Atributos rankeados por ANOVA

#	Atributos	Puntaje
1	Hidrofobicidad_K	18,53231
2	N-N	18,23628
3	Hidrofobicidad_H	16,796
4	C-O	16,39142
5	Total_aa_sitio	14,74038
6	C-C	14,10943
7	O-O	11,68753
8	energia_captacion_4aa	10,33893
9	C-N	10,2415
10	Polares_Positivos	9,207183
11	Distancia_CA3-CA4	7,195171
12	O-N	5,84777
13	Polares_Negativos	5,705176
14	Angulo_CA4-C_4	5,391129
15	Distancia_AC2-AC4	4,93044
16	Polares_Neutros	4,878569
17	Distancia_CA2-CA4	4,376008
18	Optimal alpha	4,332673
19	Distancia_CA1-CA3	3,923682
20	No_Polares	3,003231
21	Distancia_AC1-AC3	2,930261
22	Smallest alpha	2,066085
23	Angulo_CA1-C_1	1,465504
24	Distancia_AC3-AC4	1,449729
25	Distancia_CA1-CA4	1,407559
26	Distancia_A+C3	1,340702
27	Distancia_CA1-CA2	1,206093
28	Distancia_A+C4	1,177874
29	Angulo_CA2-C_2	0,956276
30	Hidrofobicidad_W	0,708801
31	Distancia_A+C2	0,549166
32	Distancia_A+C1	0,531817
33	Distancia_AC2-AC3	0,381509
34	Distancia_AC1-AC2	0,290521
35	Distancia_CA2-CA3	0,215735
36	Distancia_AC1-AC4	0,010142
37	Angulo_CA3-C_3	0,008872

4. Anexo 4: Selección de atributos de secuencia.

La siguiente lista muestra los atributos relevantes (resaltados) que ambas funciones seleccionan dentro de los primeros 28 atributos.

#	Atributos Mutual Info	Puntaje
1	Hidrofobicidad_W	0,22046
2	Hidrofobicidad_H	0,19595
3	C-C	0,174725
4	Distancia_AC1-AC3	0,145755
5	Total_aa_sitio	0,135277
6	O-N	0,114187
7	C-O	0,113372
8	Hidrofobicidad_K	0,112061
9	C-N	0,110606
10	Polares_Negativos	0,104435
11	Distancia_CA1-CA3	0,092917
12	Distancia_A+C1	0,091511
13	No_Polares	0,090847
14	Distancia_CA3-CA4	0,082954
15	Angulo_CA3-C_3	0,080995
16	Angulo_CA1-C_1	0,080318
17	Distancia_AC1-AC2	0,079683
18	Distancia_AC2-AC3	0,078887
19	O-O	0,077474
20	Distancia_A+C2	0,075827
21	energia_captacion_4aa	0,075335
22	N-N	0,064477
23	Polares_Neutros	0,061502
24	Optimal alpha	0,059479
25	Polares_Positivos	0,056886
26	Distancia_CA1-CA2	0,048084
27	Distancia_CA1-CA4	0,04778
28	Angulo_CA4-C_4	0,041852

#	Atributos ANOVA	Puntaje
1	Hidrofobicidad_K	18,53231
2	N-N	18,23628
3	Hidrofobicidad_H	16,796
4	C-O	16,39142
5	Total_aa_sitio	14,74038
6	C-C	14,10943
7	O-O	11,68753
8	energia_captacion_4aa	10,33893
9	C-N	10,2415
10	Polares_Positivos	9,207183
11	Distancia_CA3-CA4	7,195171
12	O-N	5,84777
13	Polares_Negativos	5,705176
14	Angulo_CA4-C_4	5,391129
15	Distancia_AC2-AC4	4,93044
16	Polares_Neutros	4,878569
17	Distancia_CA2-CA4	4,376008
18	Optimal alpha	4,332673
19	Distancia_CA1-CA3	3,923682
20	No_Polares	3,003231
21	Distancia_AC1-AC3	2,930261
22	Smallest alpha	2,066085
23	Angulo_CA1-C_1	1,465504
24	Distancia_AC3-AC4	1,449729
25	Distancia_CA1-CA4	1,407559
26	Distancia_A+C3	1,340702
27	Distancia_CA1-CA2	1,206093
28	Distancia_A+C4	1,177874