Facultad de Ingeniería

Escuela de Ingeniería en Bioinformática

Modelo predictivo de epítopos proteicos utilizando inteligencia artificial.

Erwin Rodman Hövelmeyer López

Tutor: Mauricio Arenas.

Co-Tutor: José Reyes.

Co-Tutor: Felipe del Canto.

Profesor Informante: Felipe Besoain

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

Índice de contenidos.

Índice de Figuras.

Introducción.

Materiales y metodología.

Resultados.

Índice de tablas.

Resultados.

Discusión.

Resumen.

El sistema inmunitario, encargado de detectar y eliminar agentes patógenos en el organismo, reconoce regiones específicas en las proteínas para proceder con su eliminación. Los linfocitos B y T se unen a los epítopos, los cuales corresponden a segmentos de proteínas que permiten al sistema inmune reconocer el patógeno. Se han implementado distintas técnicas para predecir regiones o segmentos de una proteína, que son reconocidas por el sistema inmune y así, probar estos epítopos tentativos mediante procesos experimentales. El proyecto plantea el desarrollo de un sistema de detección mediante inteligencia artificial, a través de la implementación de descriptores moleculares. Las medidas físico-químicas y geométricas que se obtuvieron al estudiar las secuencias de antígenos, fueron utilizadas para generar el conjunto de datos y luego, definir los atributos más importantes para la predicción del modelo. Los resultados obtenidos, fueron comparados con el método de predicción actual *DTU,* para evaluar la capacidad del modelo en reconocer epítopos de nuevas proteínas.

Introducción.

1.1 Sistema inmunitario

El sistema inmunitario está conformado por tejidos, células y moléculas que actúan como resistencia ante agentes patógenos. La coordinación para evitar las infecciones por microorganismos es llamada respuesta inmunitaria (Abbas y col.,2014,pp.1). La principal función del sistema inmunitario es prevenir la posible infección de un patógeno y, además, contrarrestar aquellas infecciones en curso, de este modo, cumple la labor de proteger la vida del hospedero. Sin embargo, en ciertos casos, el proceso defensivo es insuficiente, permitiendo al patógeno afectar en gran medida la vida del huésped. Los sistemas celulares encargados de eliminar a microbios requieren una serie de señales o marcadores moleculares para reconocerlos como agentes invasores, sin embargo, los antígenos y sus receptores tienen la capacidad de generar anticuerpos que afectan el funcionamiento del hospedero, donde además, tienen una amplia diversidad de aminoácidos en el bolsillo de anclaje, con pocos sitios conservados, lo que dificulta el diseño y aplicación de técnicas sintéticas que permitan estimular o potenciar una respuesta humoral(Abbas y col.,2014,pp.3).

El mecanismo inmunitario está regulado por la sucesión de receptores intermediarios, tales como, las células presentadoras de antígenos (APC) presentes en los tejidos epiteliales, quienes interactúan con los materiales endocitados buscando sustancias nocivas, en conjunto con los receptores de reconocimiento de patrones (PRR) y receptores de tipo toll (TLR), expresados por las células dendríticas.

1.2 Sistema inmune Innato y adaptativo

La respuesta inmunológica se puede dividir, de acuerdo a la temporalidad, en dos partes, innato y adaptativa. En un comienzo el sistema inmune innato presenta el primer bloqueo contra las infecciones, los agentes patógenos deben cruzar barreras físicas, para contrarrestar la actividad de antibióticos naturales, y resistir el

ataque de células fagocitadas en un lapsus de 1 a 12 horas(Abbas y col.,2014,pp.3-4).

Luego de un periodo, el patógeno al atravesar un epitelio o al ser captado por células especializadas, es reconocido por los receptores de antígenos de los linfocitos, ya sea anticuerpos en la superficie de la célula B, o receptores de células T que reconocen antígenos presentados en el contexto del complejo mayor de histocompatibilidad (MHC). El mecanismo inmunitario está regulado por la sucesión de receptores intermediarios, tales como, las células presentadoras de antígenos (APC) presentes en los tejidos epiteliales, quienes interactúan con los materiales endocitados buscando sustancias nocivas, en conjunto con los receptores de reconocimiento de patrones (PRR) y receptores de tipo toll (TLR), expresados por las células dendríticas. El conjunto de receptores permite identificar sitios conservados en los antígenos y, de esta forma, iniciar el proceso de eliminación de este(Zinsli y col.,2021). Las células encargadas de eliminar agentes infecciosos corresponden a los linfocitos de tipo B o T y sus variantes están determinadas por el tipo de receptor que presentan y el tipo de antígeno que reconocen. Los linfocitos B tienen la capacidad de reconocer proteínas, ácidos nucleicos, polisacáridos entre otros. Estos reconocen antígenos solubles o de la superficie celular guiados por señales de tipo TLR, luego secretan anticuerpos que recubren a microbios, inhibiendo su actividad y estimulando a la vez, el ataque de otros mediadores inmunológicos como el sistema del complemento o células fagociticas.

Los linfocitos T tienen una gran labor dentro de las funciones del sistema inmune adaptativo. El timocito, correspondiente a su fase inmadura, tendrá la función de especializarse en linfocito T CD4+ y CD8+, dependiendo el rol de purificación de la célula o como receptor activo para generar una apoptosis en la célula parasitada(Manijeh y col.,2013).

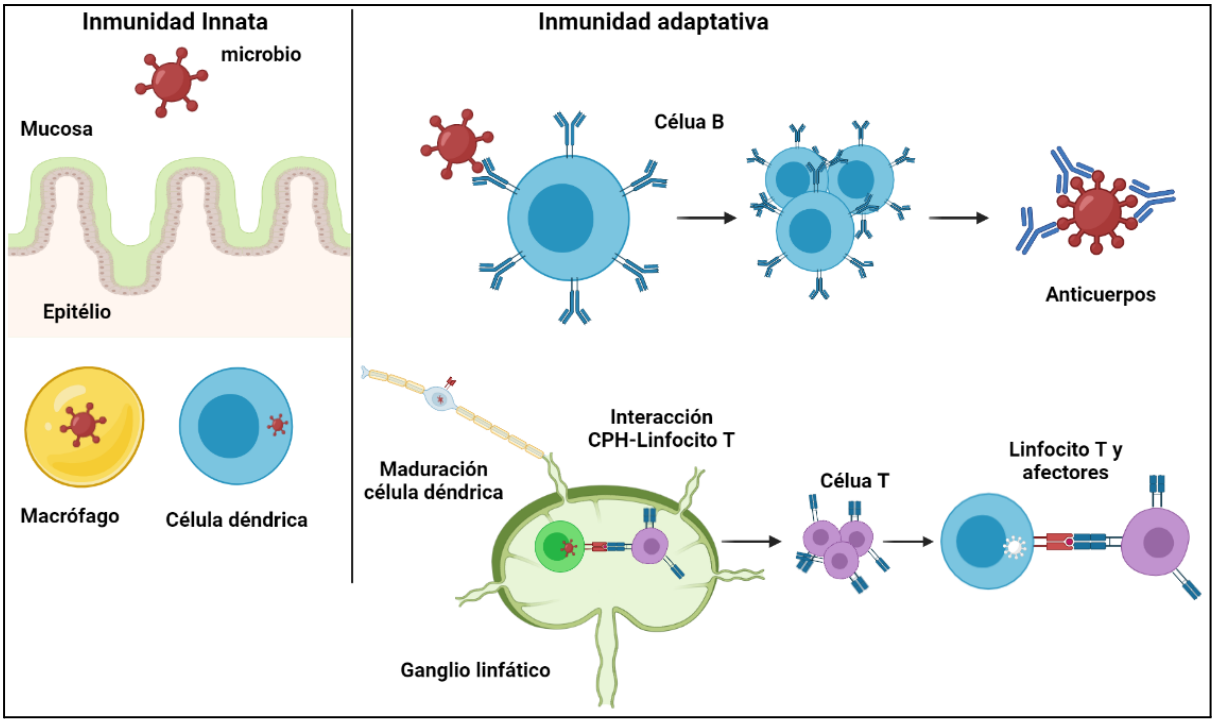


Figura 1. Sistema inmunitario innato y adaptativo. El sistema inmunitario innato brinda la primera barrera de defensa contra las infecciones, los agentes patógenos son fagocitados por células dendríticas ubicadas en el epitelio, mientras que los macrofagos resguardan las infecciones para ser secretados. Por otro lado, el sistema inmune adaptativo presenta dos rutas defensivas contra los patógenos. Los anticuerpos presentes en la pared celular del linfocito B interactúan con los antígenos, luego los anticuerpos son secretados por la célula, inhibiendo la actividad patógena y amplificando la respuesta inmune. Las células dendríticas que fagocitan el patógeno, viajan por el drenaje dendrítico hasta llegar a los ganglios linfáticos, la activación de señales siguientes a la fagocitosis promueven la expresión del complejo CPH. Se promueve la fragmentación del antígeno y es presentado al linfocito T virgen, anclado a los aminoácidos poliformicos del CPH.

Las células APC son las encargadas de fragmentar los antígenos y presentarlos mediante las moléculas MHC a los linfocitos T, CD4+ o CD8+, dependiendo la ruta de eliminación. Las células T sólo reconocen los péptidos que son presentados por CPH, ya que, interactúa con el complejo CPH 1 y 2 como sitio de anclaje, uniéndose al epítopo del agente patógeno(Abbas y col.,2014,pp.54). El complejo CPH forma una hendidura de unión con el péptido antígeno y posee 2

aminoácidos polimórficos, como sitio de anclaje al linfocito T(Abbas y col,2014,pp.50).

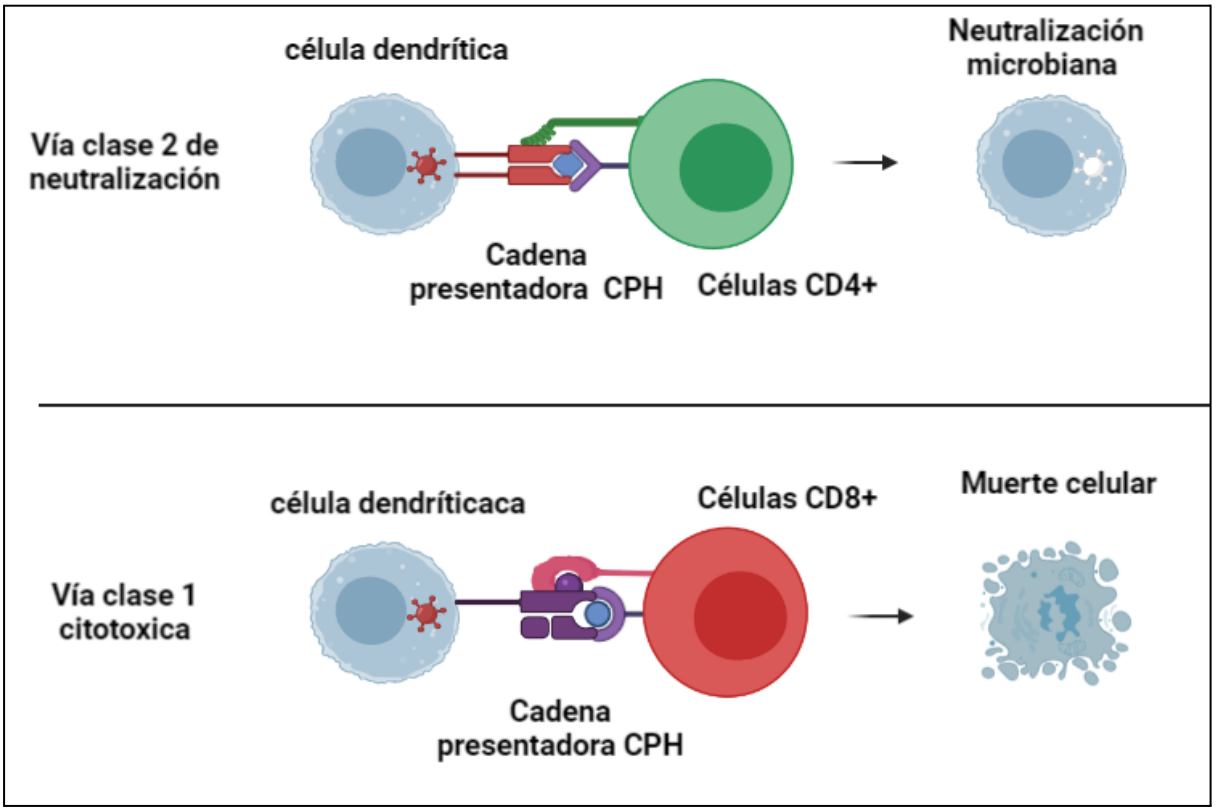


Figura 2. Eliminación de patógenos por linfocito T. La vía de la clase 2 del complejo principal de histocompatibilidad (CPH) convierte los antígenos proteicos que se introducen por endocitosis en las vesículas de las células presentadoras de antígenos en péptidos que se unen a las moléculas de la clase 2 del CPH para su exposición y presentación por los linfocitos T CD4+, quien, genera un proceso colaborativo para la producción de anticuerpos y células citotóxicas. La vía de la clase 1 del CPH convierte las proteínas patógenas, en péptidos que se unen a las moléculas de la clase I del CPH para su exposición y presentación a los linfocitos T CD8+, LTC, linfocito T citotóxico. Este complejo elimina los patógenos en la célula dendrítica y luego se produce una apósitos u otros procesos de muerte celular.

2. Reconocimiento de epítopos para linfocitos T y B.

Una de las respuestas que se ha implementado para reconocer y regular la interacción del antígeno con los linfocitos T, corresponde a la detección del epítopo

mediante la observación de los residuos presentados por el complejo CPH en unión a las células T.

Vaxign es una herramienta para la predicción de epítopos, la cual, utiliza el programa *Vaxitope*, donde, analiza los residuos que unen a MHC de clase I y clase II, y mide los aminoácidos aglutinantes mediante matrices de puntuación PSSM(He y col.,2010). De esta forma, se comparan las secuencias de antígeno, con la base de datos del programa, determinando la probabilidad que corresponda a un epítopo.

Por otro lado, el reconocimiento de epítopos que interactúan con los linfocitos B, se basa en los estudios de anticuerpos presentes en la célula(Liang y col.,2010). Las proteínas pertenecientes a las células B, son analizadas mediante estructuras 3D obtenidas de la cristalografía. En ellas, se identifican las secciones de las proteínas que interactúan con los antígenos.

El principal problema que se presenta al realizar predicciones de epítopos en linfocitos B, se debe en gran parte a que sus secuencias no son continuas, es decir, los aminoácidos del epítopo no se encuentran uno al lado del otro, más bien, la conformación estructural de la proteína permiten que se encuentren dentro de una región específica(Rahman y col.,2019), por lo que, no es posible tratar de predecir los epítopos mediante técnicas de secuencia, en gran parte de ellos.

3. Descriptores moleculares.

La implementación de técnicas computacionales para describir características tanto en estructura primaria y terciaria en las proteínas, otorgan datos medibles en las interacciones con un receptor o ligando. Se han desarrollado métodos que permiten analizar la geometría de péptidos y proteínas, así como, cálculos de energía libre y volumen, entre otros(Rahman y col.,2019), los cuales, permiten generar modelos de predicción y así también, para el desarrollo de nuevos fármacos. Por ejemplo, se han utilizado los descriptores moleculares para comprender las interacciones que ocurren en los bolsillos de unión receptor-ligando. De esta forma, se miden las características fisicoquímicas de los aminoácidos presentes en el bolsillo, simuladas como esferas cargadas y juntas en un espacio

optimizado, donde, la compatibilidad de cargas y espacios del receptor, en conjunto al ligado son medidas en base a su afinidad e interacción(Le Guilloux y col.,2009).

Las distintas bases de datos que contienen tanto secuencias como estructuras 3D de proteínas, han permitido el análisis cuantitativo de los aminoácidos para el desarrollo de bibliotecas de péptidos funcionales(Yang y col.,2010). Estos han sido utilizados para estudiar componentes del sistema inmune, dado que, las interacciones que ocurren entre los péptidos y los linfocitos, pueden ser medidas y estudiadas, en base a los aminoácidos que componen sus secuencias (Lozano y Scior,2012).

4. Machine learning

El machine learning corresponde a una serie de algoritmos que dotan a las computadoras la capacidad de identificar patrones y realizar predicciones, por lo que, tienen la posibilidad de analizar los datos de entrada y buscar características de correlación o independencias entre ellos. Los distintos algoritmos que miden y estudian los datos, permiten el aprendizaje de los modelos implementados, para predecir sobre nuevos datos que se incorporen, y así, generar información útil . Existen dos formas de aprendizajes conocidos(Basogain,2017).

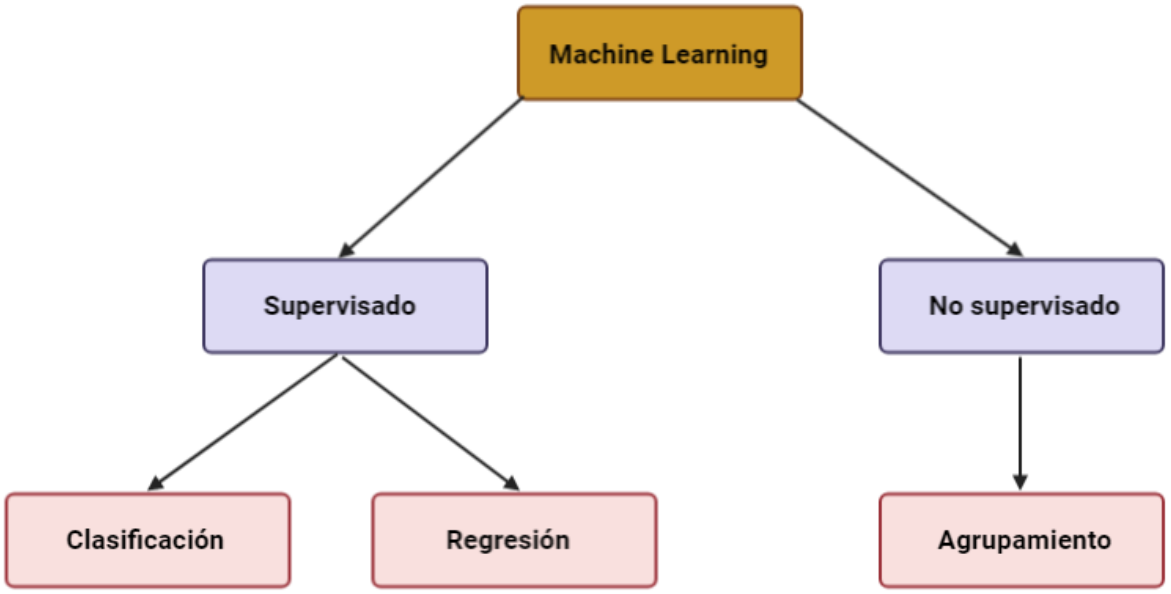


Figura 3. Modelos de aprendizaje. Machine learning presenta dos modelos de aprendizaje definidos por la presencia de una categoría. En el modelo supervisado existe el método de clasificación el cual, estima la clase asociado a un ejemplo, basados en sistemas probabilísticos como naive bayes o por cernía como lo es neighbors, entre otros. El método de regresión estima rangos de valores que puede tomar el vector de clase, tales como, el modelo lineal, polinomial, logística entre otros. Los modelos no supervisados carecen de un vector de clase, por lo que, agrupa ejemplos de atributos definidos por patrones en común, de esta forma se distinguen diferencias en cada grupo, los sistemas que utiliza para definir el agrupamiento consiste en k-medias o árboles de decisión.

Los modelos de aprendizaje no supervisados carecen de un vector de clasificación, por esto, se centra en organizar mediante distintos grupos los datos. Este método analiza los patrones de los atributos para generar una agrupación de los objetos de entrada. Por otro lado, los modelos de aprendizaje supervisados si poseen un vector de clasificación. El sistema de clases asociado a los datos, le permite al modelo tomar decisiones o hacer predicciones en base a las características que presenten sus atributos, y así , lograr clasificar nuevos ejemplos.(Bishop,2006,pp.136-144).

4.1. Método de predicción mediante inteligencia artificial.

Dadas las distintas bases de datos biológicas disponibles, junto al cálculo de descriptores moleculares, aplicadas a secuencias de aminoácidos, permiten la implementación de técnicas como machine learning, para realizar predicciones de epítopos. El modelado de algoritmos mediante un aprendizaje supervisado, plantea una metodología para determinar dos categorías en el conjunto de datos a trabajar (epítopo o no epítopo), el cual, depende de los valores que presentan sus atributos(Singh y col.,2013). El diseño de un algoritmo predictivo basado en un modelo de clasificación, plantea una forma de reconocer características distintivas en el conjunto de atributos, con tal de, comprender las diferencias que existen en las secuencias asociados a epítopos.(Sanchez y col.,2017)

Hipótesis y objetivos.

Hipótesis:

Mediante el análisis de epítopos proteicos conocidos a través de descriptores moleculares y en combinación con métodos de inteligencia artificial, será posible generar un sistema de predicción para identificar epítopos en nuevas proteínas.

Objetivos:

Objetivo General: Desarrollar un sistema de predicción para identificar posibles epítopos en proteínas asociadas a virus de influenza A.

Objetivos Específicos:

1. Identificación y análisis de bases de datos para epítopos proteicos.

2. Implementar descriptores moleculares para caracterizar proteína a nivel de estructura primaria y formación del conjunto de datos.

3. Limpieza, estandarización y selección de atributos óptimos para el conjunto de datos.

4. Implementar y comparar modelos predictivos en inteligencia artificial, evaluando exactitud y sensibilidad.

Materiales y Metodología

1. Identificación y análisis de bases de datos para epítopos proteicos.

1.1. Búsqueda de bases de datos asociadas a epítopos en virus.

La construcción de un conjunto de epítopos con cadenas de aminoácidos conocidos es el primer paso del proceso de minería de datos para este estudio. Esto consiste en reconocer y seleccionar los datos siguiendo criterios guiados por el problema a abordar. La base de datos de *immune epitope database analysis resource* (http://www.iedb.org/home_v3.php) proveniente del centro de investigación inmunológico(Fleri y col.,2017), recolecta una serie de información asociados a epítopos reconocidos en virus. Sin embargo, se utilizaron solo los datos asociadas al virus de influenza A, debido al impacto que presenta en la sociedad.

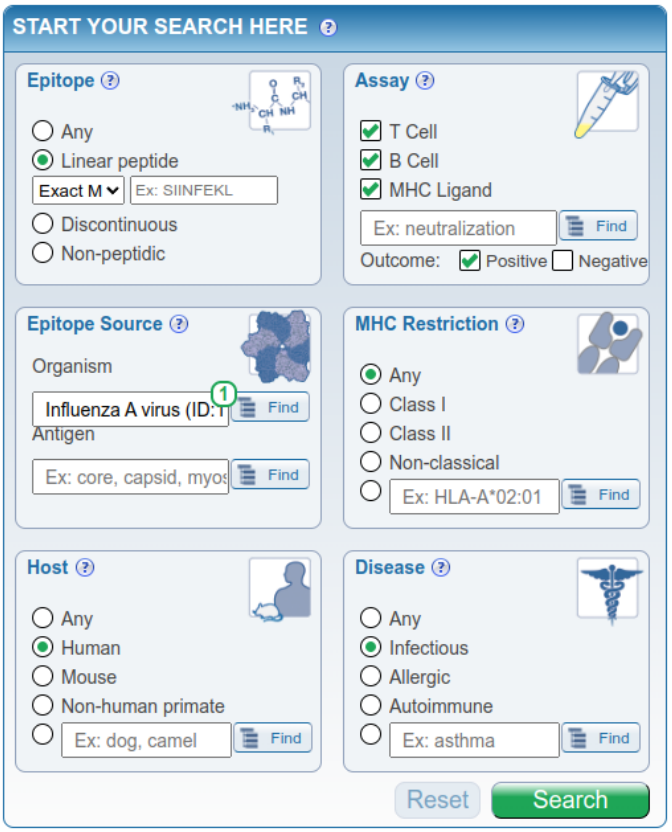


Figura 4. Filtros para la búsqueda de epítopos. La selección de epítopos pertenecientes a influenza de virus A, corresponden a epítopos lineales reportados en células T y B, además de MHC, de tipo infeccioso en humanos.

Los filtros aplicados corresponden a epítopos lineales, en linfocitos T, B y al MHC I y II, así también, la selección del organismo estudiado, en proteínas patógenas asociadas a humanos provenientes del virus de influenza A, con la finalidad que el modelo de clasificación logre reconocer epítopos, en solo un grupo específico de proteínas.

Se obtuvo un total de 256 epítopos, asociados a 10 antígenos provenientes de; Hemagglutinin, Matrix protein 1, Matrix protein 2, Neuraminidase, Non Structural Protein, Nucleoprotein, Polymerase basic protein 1, Polymerase basic protein 2, Nuclear export protein, 1,RNA-direct RNA polymerase catalytic subunit.

1.2 Búsqueda y descarga de secuencias proteicas asociadas a epítopos.

La búsqueda y reconocimiento de proteínas antígenas asociadas a virus, permite analizar las regiones que presentan epítopos. Por esto, se realizó la descarga de secuencias proteicas, mediante la base de datos de proteínas *Uniprot* (https://www.uniprot.org/), donde, el nombre del antígeno se utilizó en la búsqueda de la proteína en la base de datos. Se analizó sólo la estructura primaria de las proteínas por lo que, se descargaron las secuencias en formato *fasta*.

Los códigos *UP* de las proteínas descargadas son las siguientes; P06452, P03485, P03468, P03466, P03428, P03433, P06821, P03508, P3496, P03431

1.3 Mapeo de epítopos en proteínas obtenidas.

Una vez obtenido los archivos fasta con las secuencias de las proteínas, es necesario determinar la cantidad de epítopos que presenta, dado que, una proteína patógena presenta diversas regiones que interactúan con uno o más antígenos (OKT y col.,2010). La búsqueda de los epítopos en los archivos permite reconocer los tanto la cantidad que se presenta como el largo de su secuencia, para luego, generar los descriptores moleculares que le otorgan características medibles al sitio (Voss y col, 2006).

La búsqueda de los epítopos presentes en las distintas regiones de cada proteína, se realizó mediante alineamientos por identidad de un 100%, a través del programa *Jalview.* La herramienta selecciona cada fragmento de epítopo y lo alinea con la secuencia objetivo, sin gaps o similitud, por lo que, busca a lo largo de la proteína el lugar exacto donde se encuentra el fragmento.

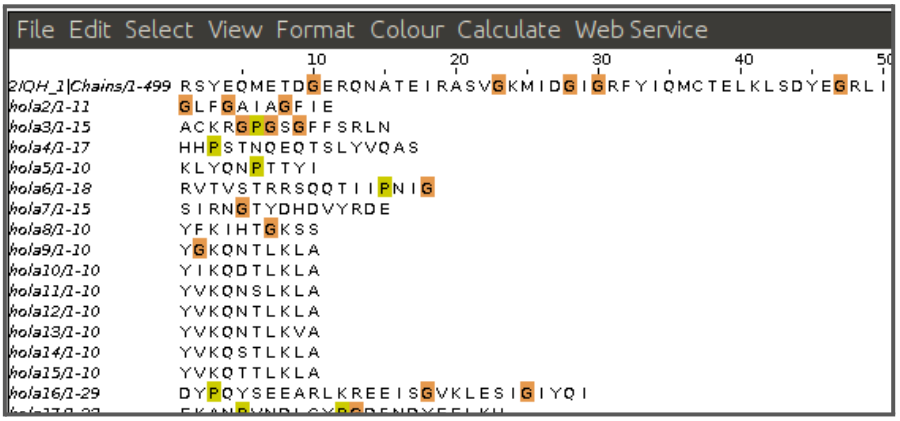


Figura 5. Alineamiento múltiple de epítopos. Se presenta la herramienta jalview, donde, se ingresó el archivo multifasta con; Secuencias de perteneciente a los 256 epítopos descargados, y se incluye la secuencia de la proteína para su alineamiento.

De esta forma, la opción *alineamiento de pares de bases* genera una búsqueda entre todas las secuencias ingresadas. Mediante esta búsqueda de epítopos en la proteína se genera una mapeo de los aminoácidos, obteniendo la cantidad efectiva que se encuentran en la secuencia de proteína, para los epítopos lineales (Liang y col,2010).

2. Implementar descriptores moleculares para caracterizar proteína a nivel de estructura primaria.

2.1 Fragmentación de regiones proteicas.

Para lograr comparar una región de proteína que pertenece a un epítopo con quienes no lo son, se establecen rangos de aminoácidos en forma de secuencias cortas(Manijeh y col.,2013). Al reconocer anteriormente los largos promedio que

presentan los epítopos, esa cantidad se utilizó para definir el número de aminoácidos que componen los fragmentos.

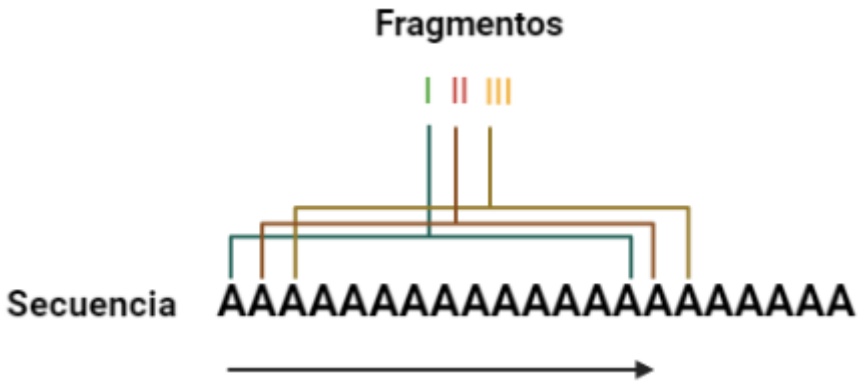


Figura 6. Fragmentación de secuencia. Una vez definido el largo que tendrán los fragmentos, se procede a recorrer la secuencia copiando los aminoácidos seleccionados en un archivo fasta. Los indicadores presentes en cada lectura, muestran el número de secuencias que se generan a medida que se avanza, hasta llegar al último aminoácido. De esta forma, se obtienen fragmentos de secuencia acorde al largo que presentan los epítopos en la proteína

Al final del ensayo se presenta el *script 1,* donde, se produce la fragmentación de la secuencia como se muestra en la figura 6. El archivo *output.fasta* contiene la totalidad de los fragmentos generados con epítopos y no epítopos.

2.2 Implementar descriptores moleculares asociados a estructuras proteicas.

Las regiones analizadas que presentan epítopos tienen una característica diferenciable, la cual, le brinda una antigenicidad distinta a otras regiones para interactuar con el sistema inmune (Basogain,2017). Se emplearon diversos cálculos en las secuencias proteicas, por ejemplo. Para obtener la carga neta en las secuencias, se utilizó la ecuación de henderson, donde, mide las cargas de los aminoácidos en su secuencia, asignando el ph del medio como 7, y determina el punto isoeléctrico para cargas neutras (Kiraga,2008). Además, se midió el índice de inestabilidad de las secuencias basadas en el cálculo de *Guruprasad,* donde, se evalúa la composición de los aminoácidos(Guruprasad y col.,1990).

Así como en ambos casos presentados, se evaluaron 73 descriptores moleculares presentados en anexos, los cuales, miden distintas características de las secuencias analizadas. El cálculo de descriptores utiliza las secuencias de proteínas en formato fasta proveniente del *script1,* donde se aplican los modelos matemáticos, físicos y químicos. Por lo tanto, se generó un dato asociado a cada fragmento en formato csv, para las 10 proteínas de estudio.

2.3 Construcción del set de datos.

Cada dato generado a través de los distintos descriptores moleculares, son incluidos en un conjunto de datos para implementar los modelos predictivos de inteligencia artificial. En los pasos previos se obtuvieron distintas lecturas de archivos y bases de datos, las cuales, almacenan características asociadas a epítopos, proteínas y los descriptores moleculares generados. Por lo tanto, en este punto se incluye cada atributo en una hoja de cálculos en formato *csv* para formar un conjunto de datos completo .

La herramienta *jupyter* es un desarrollador de código compatible con librerías de *Sklearn*(Pedregosa y col.,2012). En esta, se importaron las librerías de *pandas* y *csv* para la lectura y manejo de datos para la hoja de cálculos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3351 entries, 0 to 3350
Data columns (total 74 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   SeqIn                           3351 non-null   object
 1   NumTiny                         3351 non-null   float64
 2   NumSmall                        3351 non-null   float64
 3   NumAliphatic                    3351 non-null   float64
 4   NumAromatic                     3351 non-null   float64
 5   NumNonPolar                     3351 non-null   float64
 6   NumPolar                        3351 non-null   float64
 7   NumCharged                      3351 non-null   float64
 8   NumBasic                        3351 non-null   float64
 9   NumAcidic                       3351 non-null   float64
 10  PorcTiny                        3351 non-null   float64
```

```
59  t2                              4554 non-null   float64
60  t3                              4554 non-null   float64
61  t4                              4554 non-null   float64
62  t5                              4554 non-null   float64
63  z1                              4554 non-null   float64
64  z2                              4554 non-null   float64
65  z3                              4554 non-null   float64
66  z4                              4554 non-null   float64
67  z5                              4554 non-null   float64
68  HydrophobicityIndex             4554 non-null   float64
69  AlphaAndTurnPropensities        4554 non-null   float64
70  BulkyProperties                 4554 non-null   float64
71  CompositionalCharacteristicIndex 4554 non-null   float64
72  LocalFlexibility                4554 non-null   float64
73  ElectronicProperties            4554 non-null   float64
```

Figura 7. Lectura del conjunto de datos. Todos los ejemplos de las 10 proteínas junto a sus atributos se visualizan en una tabla, la cual, presentan los siguientes datos; *SeqIn*, correspondiente a los datos nominales como “0” secuencia no epitope y “1” como secuencia asociada a un epítopo. Además, se presentan los 73 atributos asociados al cálculo de características físico-químicas, como datos de tipo ordinal.

3. Limpieza, estandarización y selección de atributos óptimos para el set de datos.

3.1 Preprocesamiento de datos.

La construcción de un set no redundante, consiste en normalizar, limpiar y transformar los datos(Gupta, y col.,2016), donde, se eliminan o transforman datos que presenten elementos fuera de formato, datos faltantes(Leiva,2015). En este punto, un conjunto de datos representativo para analizar las distintas regiones de epítopos, es aquel que posee al menos un epítopo por proteína y se normalizan las características físico-químicas en rangos de valores óptimos, para lograr predecir sobre ellas.

Mediante las librerías de *sklean preprocessing*(Pedregosa y col.,2012), se realizó la transformación de los datos nominales para el atributo “SeqIn” con valores de 0 y 1, a través de la función *LavelEncode.* Se realizó la búsqueda de datos faltantes mediante la función *isnull* de *pandas, y* finalmente se analizaron los rangos de valores que presenta cada atributo. Este paso resulta de gran importancia dada la relación que existe entre la dispersión de sus datos, y los rangos de valores. El peso que ejerce un atributo con un mayor rango de valores, aun siendo homogéneo, es por mucho superior con aquellos atributos que presentan gran variabilidad, al tener valores tan bajos no influyen en los modelos de predicción(Gupta y col.,2019). Se utilizó la función *preprocessing.normalize,* con tal de, disminuir el rango de valores entre 0-1, manteniendo la dispersión de datos entre cada atributo. De esta forma, los modelos predictivos no se verán sesgados por el peso que ejerce cada atributo.

3.2 Selección de atributos representativos.

La importancia de realizar una selección de atributos, recae en la identificación de los datos que presentan una alta diferenciación entre los atributos, de este modo, es posible obtener mejores predicciones, sin caer en la maldición de la dimensionalidad (Alfredo,2012).

Para lograr generar un conjunto de datos que permitan una mejor exactitud en la clasificación, es necesario recurrir a la metodología de categorización. Esto ayuda en la búsqueda de atributos más relevantes en base a la correlación que tienen entre ellos (Anabel,2016), donde, se realizan evaluaciones de los atributos y los valora desde el mejor al peor.

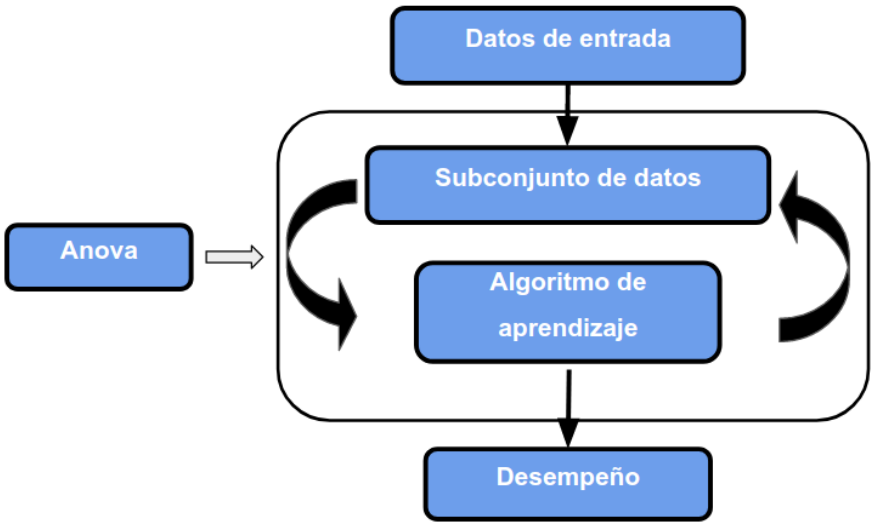


Figura 8. Selección de atributos. La totalidad de atributos se ve reducido a subconjuntos de atributos, donde, se observa la relación de sus varianzas, estos valores definen la disposición de los datos a pertenecer a distintas clases, por lo que, el método de selección refleja el desempeño del modelo.

La selección de atributos se realizó utilizando el método *ANOVA*, a través de la función *feature selection* presentes en las librerías de *SKLEARN*. Este método, realiza una puntuación de los atributos basados en su varianza, y limita las variables booleanas con la siguiente fórmula(Chandrashekar,2014).

$$\mathbf{Var\ [X] = p\ (1 - p)}$$

Ecuación 1. Limitación de variables booleanas. La varianza de los datos booleanos en cada atributo, se limita para evitar la influencia de muchos valores iguales a 0, por lo que, el valor *p* asigna una región de corte no mayor a un 20 % de datos con valor 0.

3.3 Definición del set de entrenamiento.

El entrenamiento de los modelos predictivos tienen la función de enseñar al computador cómo se deben tomar las decisiones. Esta es capaz eventualmente de reconocer patrones y lograr clasificar, dependiendo de la variabilidad de valores en sus atributos. Para esto, se debe separar el conjunto de datos en grupos reducidos y proceder a realizar una validación cruzada en una proporción de 80% de entrenamiento y 20 % de prueba(Rahman y col.,2019). Mediante la implementación de las funciones de *train_test_split,* se utilizó 3.672 ejemplos del conjunto total como entrenamiento, mientras que, 918 fueron parte del conjunto de evaluación del modelo.

Luego de esto, y con la finalidad de evaluar el sobre ajuste de los datos se implementó el método de validación cruzada. El total de ejemplos se dividió en subconjuntos de datos en cinco grupos, donde, cuatro de ellos son utilizados para entrenar, y uno para probar, por lo que, este proceso se repite hasta que cada grupo haya sido utilizado una vez.

Figura 9. Validación cruzada. Cada grupo representa un 20% del conjunto total de datos, por lo que, un 80% se utilizó como datos de entrenamiento, y el 20 % de ellos como prueba. Sin embargo, se realizaron 5 ciclos para que, en cada iteración el grupo de prueba cambie.

4 Implementar y comparar modelos predictivos en inteligencia artificial, evaluando exactitud y sensibilidad.

4.1 Implementación de modelos de clasificación: Random forest.

El modelo de predicción utilizado es de tipo supervisado, esto se debe a la definición de una clase para cada ejemplo(Basogain,2017). La página de sklearn tiene a disposición distintos códigos de implementación para distintos modelos, sin embargo, random forest como modelo de clasificación destaca por sobre otros, basado en sus valores de desempeño, diseño de gráficos, cálculos matemáticos entre otros.

El método se define por ser un conjunto de condiciones organizadas en una estructura jerárquica. Este modelo integra un número de árboles de decisión, quienes, independientemente realizan la operación de clasificación (https://scikit-learn.org/stable/modules/ensemble.html#forest) (Pedregosa y col.,2012). Las decisiones se toman siguiendo las condiciones que se cumplen desde la raíz, hasta alguna de las hojas del árbol, donde, cada árbol en el conjunto se construye a partir de una muestra extraída con reemplazo siendo la raíz el primer atributo seleccionado para tomar una decisión y todos los sucesivos pasan a ser hojas del árbol.

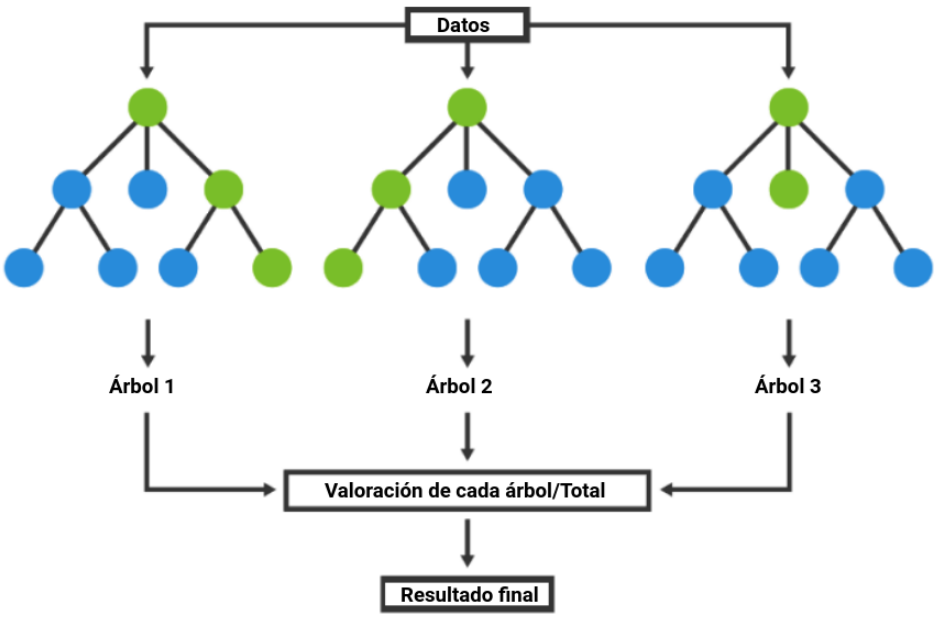


Figura 10. Método Random forest. Se implementa un número designado de árboles de decisión para realizar predicciones independientes, estos reciben como entrada el conjunto de datos, y proceden a clasificar los ejemplos dependiendo del valor de sus atributos(Turing,1992). Luego de esto, se evalúan los resultados de cada árbol basados en los resultados de sus hojas, y se compara con el resto de los árboles. De esta forma se obtiene un resultado consenso

Los argumentos de ajuste que se evalúan en el desempeño de los resultados, corresponden a; *n_estimatro,* respondiendo al número de árboles que genera el modelo para realizar las predicciones; *class_weight,* argumento que define el tipo de prioridad o peso que tienen las variables dependientes, este permite dar importancia en la selección de clases o bien, definir una igualdad entre estos.

4.2 Evaluación de modelos: Exactitud, sensibilidad y especificidad.

Los métodos de evaluación son empleados para estimar y optimizar los resultados entregados por los modelos implementados. Son incorporados en la evaluación cruzada como medidas de análisis de datos, donde, se utilizarán tres medidas de correlación para evaluar el desempeño de los modelos de predicción (Danziger y col.,2017). Dependiendo del resultado al clasificar, los ejemplos son asignados a matrices de confusión asignadas de la siguiente manera.

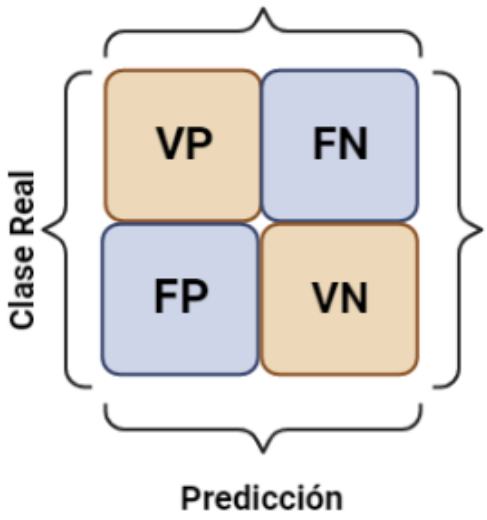


Figura 11. Matriz de confusión. Las clases reales muestran la asignación inicial que tienen los datos, mientras que la predicción presenta el valor asignado por el modelo, de esta forma se obtiene que; Verdaderos positivos (VP), corresponden a valores positivos correctamente asignados tanto en la clase real como predicción; Verdaderos negativos (VN), son los valores negativos correctamente asignados tanto en la clase real como predicción; Falsos positivos (FP) son aquellos que tienen una asignación inicial como negativos, sin embargo, fueron clasificados

como positivos; Falsos negativos (FN), son aquellos que tienen una asignación inicial como positivo, sin embargo, fueron clasificados como negativos.

Accuracy (Exactitud): Mide los resultados globales, basados en los aciertos que han sido bien clasificados por el modelo. Integra los resultados que fueron correctamente medidos como verdaderos positivos y verdaderos negativos y los divide por el número total de casos medidos, ya sean, VP, VN, FP y FN(Pedregosa y col.,2012), midiendo la cantidad de predicciones positivas que fueron correctas *(VP+VN)/(VP+VN+FP+FN).*

Recall (sensibilidad) y Precisión (especificidad): Corresponden a dos valores que evalúan los resultados basados en la fracción de verdaderos positivos y verdaderos negativos(Pedregosa y col.,2012).

- Sensibilidad mide la tasa de casos positivos que fueron clasificados, VP/(VP+FN)

- Especificidad mide la tasa de casos negativos que fueron clasificados VN/(VN+FP)

Resultados.

Obj 1. Identificación y análisis de bases de datos para epítopos proteicos.

Los datos obtenidos por *immune epitope database,* presentan sólo las proteínas analizadas mediante secuencia , avaladas por estudios experimentalmente(Singh y col.,2013). A través del filtro aplicado, se creó un conjunto de datos que integran 10 proteínas de un solo tipo de virus. Si bien, no se encuentran las estructuras 3D de todas las proteínas, el solo hecho de utilizar secuencias lineales en el cálculo de descriptores moleculares, facilitó la recopilación de los datos necesarios. De este modo, incluyendo los resultados obtenidos por base de datos, se realizó la búsqueda de los epítopos en las secuencias proteicas. Sin embargo, no todos los epítopos registrados se encontraron en las secuencias, dado que, existen fragmentos descritos y revisados de forma tentativa, y no forman una identidad exacta con las secuencias existentes. Por lo tanto, solo se utilizaron aquellos fragmentos de epítopos que realmente tuvieran una identidad del 100%.

Obj 2. Implementar descriptores moleculares para caracterizar proteína a nivel de estructura primaria y formación del conjunto de datos.

El proceso de caracterización de las proteínas, conlleva la reducción de la secuencia en varios fragmentos. La implementación del *script* de corte reduce la proteína en varias regiones quienes, difieren en sus secuencias aminoacídicas.

```
>entradaProtein
QLRENAEDMGNGCF
>entradaProtein
KIYHKCDNACIESI
>entradaProtein
RNGTYDHDVYRDEA
>entradaEpitopo
GLFGAIAGFIE
>entradaEpitopo
ACKRGPGSGFFSRLN
>entradaEpitopo
RVTVSTRRSQQTIIPNIG
>entradaEpitopo
SIRNGTYDHDVYRDE
```

Figura 12. Archivo de salida script 1. Luego de asignar el largo que deben tener los fragmentos de proteína, el código crea un archivo de salida *output.fasta* el cual, contiene tanto los fragmentos de proteína como los de epítopos en formato fasta. Estos son posteriormente requisitos para el cálculo de descriptores moleculares.

El cálculo de los descriptores moleculares entregaron 73 atributos independientes,, tales como, la carga total, el número de anillos, volúmen, entre otros. Sin embargo, el descriptor *at_instaindex* y *at_mw* no presentó variabilidad en sus datos, por lo tanto, fueron descartados. Dispersión y la variabilidad tanto en los datos como en sus categorías, son necesarios para definir patrones de búsqueda en los procesos de minería de datos, debido a esto, solo se mantuvieron 71 atributos independientes y 1 atributo dependiente.

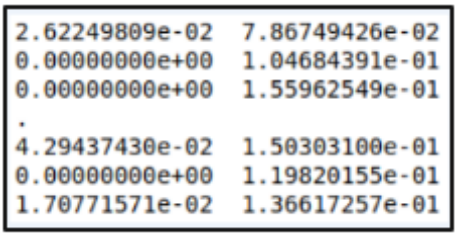
Obj 3. Limpieza, estandarización y selección de atributos óptimos para el set de datos.

El preprocesamiento del conjunto de datos se dividió en 3 puntos principales, nulidad, transformación de datos y normalización, dado que, no existió una variación notable que indicará ruido o mal funcionamiento en los datos obtenidos por parte de los descriptores.

El primer proceso consistió en evaluar la nulidad de los datos. Cada ejemplo del conjunto de datos se revisó con la función de *pandas,* donde, todos los valores fueron reconocidos, por lo tanto, no hubo error por falta de datos. El siguiente paso a evaluar consistió en la transformación de los datos. A través de la función *proprocessing.transform* los datos nominales por parte del atributo clase (epítopo, o no epítopo), se cambió por valores 0 para epítopo y 1 para no epítopo.

Finalmente, para la normalización de los datos se utilizó la librería *preprocessing.normalize,* la cual, reduce la magnitud de los datos en rangos de 0-1 manteniendo su dispersión. Mediante ecuación de normalización, la función utiliza los valores máximos y mínimos en cada atributo, sin embargo, se mantiene la

variabilidad para todos los atributos en decimales. De esta forma, se evitan sesgos por datos excesivamente grandes o muy pequeños.



```
2.62249809e-02  7.86749426e-02
0.00000000e+00  1.04684391e-01
0.00000000e+00  1.55962549e-01
.
4.29437430e-02  1.50303100e-01
0.00000000e+00  1.19820155e-01
1.70771571e-02  1.36617257e-01
```

Figura 12. Normalización de datos. La aplicación de la normalización, reduce los valores iniciales, con valores desde uno hasta cinco decimales, sin embargo, la dispersión de los datos se mantiene, debido al rango 0-1, de esta forma, cada atributo tendrá un valor mínimo y máximo de referencia.

Luego de realizar el preprocesamiento de los datos, se analizaron los atributos óptimos para el modelo. ANOVA corresponde a un selector de atributos, el cual, categoriza cada atributo basados en los puntajes obtenidos por su varianza, y dispersión de datos. El listado de categorías obtenidas por parte de la selección de atributos, fueron utilizados en la evaluación del modelo y definir el número de atributos óptimos en el siguiente objetivo.

Obj 4. Evaluación de modelos: Exactitud, sensibilidad y especificidad.

- **Primer ciclo de evaluación, análisis de entrenamiento.**

El conjunto de datos utilizados está compuesto por 10 proteínas asociadas a epítopos del virus influenza A. Estas fueron utilizadas para construir los modelos de entrenamiento y determinar cómo se ajustan a los datos, además, de predecir sobre el subconjunto de prueba. Se realizó un primer ciclo de evaluación, definiendo los siguientes criterios de entrenamiento.

-Selección de subconjuntos: 80% entrenamiento, 20% prueba.
-n_estimators: 100 árboles.
-class_weight: balanceado.
-Clase 0: 130 datos de fragmentos de epítopos.
-Clase 1: 4955 datos de fragmentos no epítopos.
-Selección de atributos: 71 atributos independientes

Mediante la implementación del modelo *Random Forest* se obtuvieron los siguientes resultados:

Exactitud	Sensibilidad	Especificidad
97.72%	15.38%	99.69%

Tabla 1. Resultados del primer ciclo. Como se planteó en la metodología, se calculó la *exactitud, sensibilidad* y *especificidad,* basados en las tasas de VP, FP, VP y NP.

Los valores de exactitud con un 97.72% indican que, existe una alta eficiencia global de los resultados basados en las matrices de confusión, así también, los valores de especificidad. Sin embargo, la sensibilidad que mide la tasa de verdaderos positivos indica que, hay un desbalance respecto a la especificidad. Por lo tanto, el bajo reconocimiento de la clase minoritaria (clase 1), indica que no se logra un buen entrenamiento y requiere un balance de clases.

- **Segundo ciclo de evaluación, Análisis de entrenamiento con balanceo de clases.**

Debido a la baja sensibilidad que se obtuvo con el modelo inicial de entrenamiento, fué necesario aplicar técnicas de balance de clases. Mediante la función *RandomOverSampler*(Pedregosa y col.,2012), fue posible incrementar el número de ejemplos por parte de la clase minoritaria, a través de nuevos datos variables con índices similares a los preexistentes.

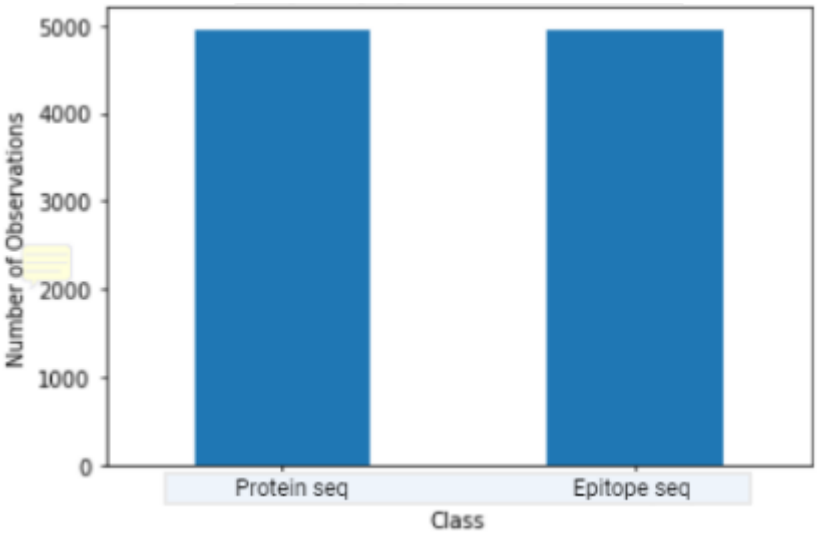


Figura 13. Balanceo de clases. La incorporación de nuevos datos por parte de la clase minoritaria, igualar el número de ejemplos que se presentan por parte de los fragmentos no epítopos, quedando con un total de 4.955 datos por clase.

Luego de implementar un cambio en el total de datos en la clase minoritaria se realizará el segundo ciclo de entrenamiento con los siguientes criterios.

-Selección de subconjuntos: 80% entrenamiento, 20% prueba.

-n_estimators: 100 árboles.

-class_weight: balanceado.

-Clase 0: 4995 datos de fragmentos de epítopos.

-Clase 1: 4955 datos de fragmentos no epítopos.

-Selección de atributos: 71 atributos independientes

Los resultados obtenidos del segundo ciclo son los siguientes.

Exactitud	Sensibilidad	Especificidad
99.37%	100%	98.9%

Tabla 2. Resultados del segundo ciclo. Los valores de exactitud, sensibilidad y especificidad, obtuvieron altos porcentajes de predicción con solo 11 casos de error por parte de *FP.*

- **Tercer ciclo, análisis de entrenamiento, selección de árboles óptimos.**

La asignación de estimadores indica la cantidad de árboles de decisión que se utilizaron para predecir las clases. Anteriormente, se asignaron 100 estimadores para definir el rendimiento de las predicciones, sin embargo, se debe evaluar el número a utilizar, debido al tiempo de cálculo que demora, así como, en los valores de rendimiento. Para definir el óptimo es necesario evaluar la variación en su exactitud, dado que, al no presentar cambios a medida que se aumenta la cantidad de árboles, solo se incrementa el tiempo de cálculo sin afectar su rendimiento.

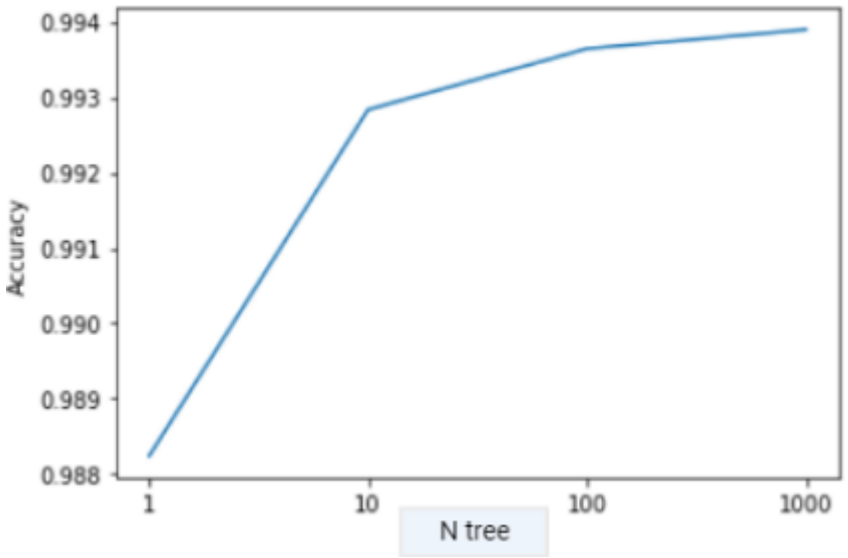


Figura 14. Selección de árboles de decisión. Los cambios en la exactitud, presentan una leve diferencia al aumentar el número de árboles de decisión, en orden de 0.006. Sin embargo, desde los 1000 clasificadores existe un notorio tiempo de cálculo, por lo que, se mantendrán los análisis con 100 árboles.

-**Cuarto ciclo, análisis de entrenamiento, selección de atributos.**

La selección de atributos categoriza los atributos basados en la dispersión de sus datos, para luego, evaluar la correlación que existen entre ellos. De esta forma, se implementó el método de anova comparando los valores de exactitud.

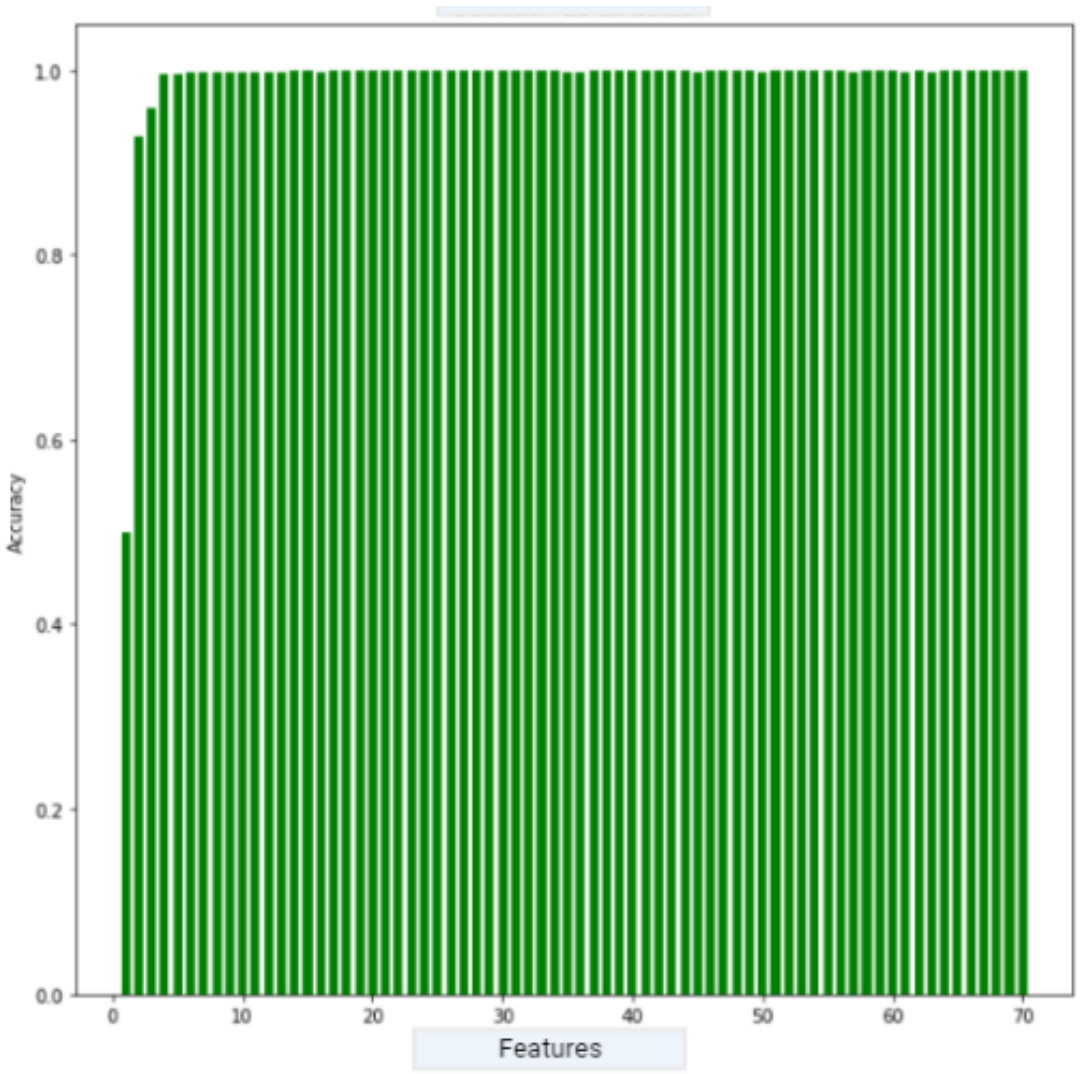


Figura 15. Análisis de elección de atributos. Los resultados de exactitud presentados, fueron cíclicamente calculados desde el mejor atributo hasta el peor, dada la categoría obtenida por anova. De tal forma, se analizaron los cuatro primeros atributos, dado que, los valores de exactitud posteriores se mantienen constantes a lo largo de los ciclos. atributo 1 (A1) *at_lengthpep,* 49.9%; atributo 2 (A1+A2), *NumNonPolar,* 92.6%; atributo 3 (A1+A2+A3),*NumAromatic*,95.7%; atributo 4 (A1+A2+A3+A4), *PartialSpecificVolume,* 99.4%.

Como se observa en el gráfico, solo con 4 de los 71 atributos se logran un alto porcentaje de exactitud, es decir, se obtienen los resultados óptimos. Sin embargo, no explica la correlación que existe entre ellos, o bien, si existe una variabilidad distintiva entre estos cuatro atributos. Por lo tanto, se generó el siguiente gráfico de dispersión.

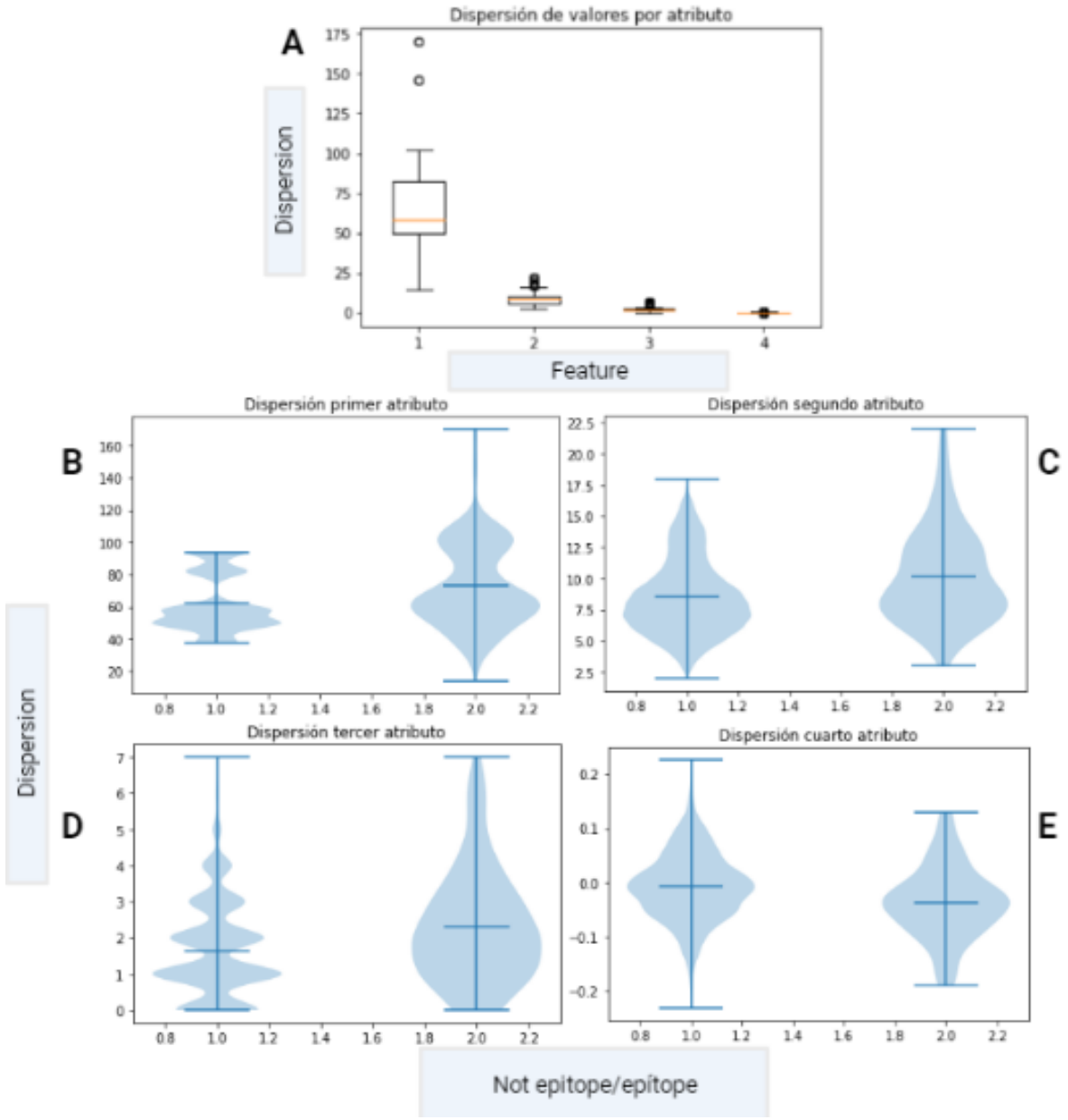


Figura 16. Dispersión de atributos. El gráfico A de dispersión de valores por atributo, se demuestra como los valores del primer atributo son mayores al segundo y sucesivamente hasta al cuarto, por lo que, concuerda el grado de importancia que asigna el método de anova. Por otra parte, al analizar los la dispersión de datos para cada atributo en los gráficos B, C, D y E, se muestran en la posición 1 el rango de valores que tiene la clase *no epítopo,* mientras que en la posición 2, se muestra el rango de valores para la clase *epítopo.* En cada gráfico, ambas clases tienen sus medias muy cercanas, por lo tanto, es correcta la correlación de estos cuatro atributos, como un mecanismo necesario para predecir las categorías, dado que,

independientemente, no logran reconocer los patrones necesarios para separar ambas clases.

- **Análisis en nuevas proteínas.**

Ya definidos los parámetros y pruebas del método de entrenamiento, con tal de comprender si el modelo es capaz de predecir epítopos en nuevas proteínas, se procedió a predecir los epítopos de una proteína aislada del conjunto de datos y comparar sus evaluaciones. De las diez proteínas iniciales, se extrajo una del total y se entrenó con las otras nueve. Este proceso se utilizó para medir por un lado la proteína *Hematogglutinin* (6BKN) y luego se evaluó la proteína *Matrix Protein 1* (7JM3). Ambas fueron seleccionadas para comprender el ajuste de los datos, y determinar cómo el modelo varía al evaluar distintas proteínas, que no están presentes en el entrenamiento.

	Entrenamiento	Predicción
6BKN	Exactitud: 98.83% Sensibilidad: 100% Especificidad: 97.78%	Exactitud: 97.97% Sensibilidad: 18.18% Especificidad: 99.79%
7JM3	Exactitud: 98.88% Sensibilidad: 100% Especificidad: 99.13%	Exactitud: 94.71% Sensibilidad: 85.71% Especificidad:94.6%

Tabla 3. Predicción de casos aislados. Los entrenamientos presentados se realizaron con 9 proteínas del conjunto. Se utilizaron 4.590 datos para el entrenamiento en 6BKN, mientras que, 4.837 para 7JM3.

Como se observó en la tabla, el cambio de sensibilidad se debe a la presencia de fragmentos no epítopos, que fueron clasificados como epítopos, es decir, existen falsos positivos que reducen la sensibilidad. Por lo tanto, se realizó una tabla para evaluar la probabilidad de reconocimiento de epítopos.

	100% - 80%	79%-51%	Epítopos totales
VP (6BKN)	2	0	2
FP (6BKN)	1	0	1
VP (7JM3)	6	0	6
FP (7JM3)	1	10	11

Tabla 4. Probabilidad de epítopos reconocidos. Los porcentajes presentados indican la probabilidad de clasificación del modelo, es decir, la certeza de pertenecer a dicha clase. En la totalidad de ejemplos pertenecientes a la clase 0 (fragmento de epítopo) en 6BKN, 2 de los 11 epítopos originales fueron clasificados como tal, con una probabilidad superior al 80%. En el caso de los falsos positivos, solo 1 de los 484 (fragmentos no epítopos) fue clasificado como un epítopo, con probabilidad superior al 80%. Por otro lado, 7JM3 logró clasificar los 6 epítopos originales con una probabilidad superior al 80%, mientras que, los datos pertenecientes a la clase 1 (fragmentos no epítopos), 10 de ellos fueron mal clasificados, con una probabilidad por debajo del 80%, mientras que 1 de ellos tuvo una probabilidad superior al 80%.

Finalmente, con tal de reconocer los epítopos reconocidos por el modelo, se realizó una tabla comparativa entre los epítopos originales de las proteínas, y los predichos por el modelo.

Epítopos predichos (6BKN)	Epítopos reales (6BKN)
KEFSEVEGRIQDLEKYV	KEFSEVEGRIQDLEKYV
GLFGAIAGFIE	GLFGAIAGFIE
GLFGAIAGFIENGW	
	ACKRGPGSGFFSRLN
	SIRNGTYDHDVYRDE

	RVTVSTRRSQQTIIPNIG
	IEKTNEKFHQIEK
	GFTWTGVTQNG
	DKLYIWGVHHPSTN
	LTKSGSTYPVLNVT
	YVKQNTLKLA
	GTLVKTITDDQIEV

Tabla 5. Predicción de epítopos en 6BKN por RF. Los tres epítopos reconocidos por el modelo, forman parte de los epítopos originales de la proteína. El fragmento en recuadro amarillo, posee tres aminoácidos adicionales en su cadena, no siendo exactamente idéntico al original, sin embargo, forma parte se la secuencia del epítopo. Por lo tanto, el modelo logra reconocer estas regiones, a la hora de predecir en nuevas proteínas.

Epítopos predichos (7JM3)	Epítopos Reales (7JM3)
SLLTEVETYVLSIIPSGPL	
GILGFVFTLT	GILGFVFTLT
KGILGFVFTLT	
LYRKLKREITF	
GLQRRRFVQNA	
LGFVFTLTVPS	
MGAVTTEVAFG	
SLLTEVETYVLSIIPSGPL	SLLTEVETYVLSIIPSGPL
QKRMGVQMQRFK	QKRMGVQMQRFK
FVFTLTVPSER	FVFTLTVPSER

LTVPSERGLQR	
RGLQRRRFVQN	
GAVTTEVAFGL	
RMVLASTTAK	RMVLASTTAK
GILGFVFTL	GILGFVFTL
TLTVPSERGLQ	
AVTTEVAFGLV	
SERGLQRRRFV	

Tabla 6. Predicción de epítopos en 7JM3 por RF. De los 17 epítopos reconocidos por el modelo, 7 de ellos forman parte de los epítopos originales de la proteína. El fragmento en recuadro amarillo, posee un aminoácidos adicional en su cadena, sin embargo, reconoce la regíon original. Los 10 fragmentos adicionales, corresponden a falsos positivos que el modelo reconoció como epítopos. Teniendo en cuenta los resultados obtenidos en la tabla 4**,** estos fragmentos, son aquellos reconocidos con una probabilidad inferior al 80%, a diferencia de los epítopos bien predichos.

Discusión.

El modelo de entrenamiento entregó como resultado un alto porcentaje de exactitud, sensibilidad y especificidad, como se presentó en la figura 18, además, con solo 4 de los 71 atributos iniciales. Pero, al evaluar los resultados en las proteínas de prueba, se obtuvieron valores menores en la sensibilidad. Esto demuestra la baja diferenciación que existe entre los fragmentos de epítopos, de quienes no lo son. Sin embargo, aun teniendo una reducción de verdaderos positivos, la comparación de los vectores de probabilidad presentados en la figura 19, muestra una clasificación de alta confianza dado los altos porcentajes con los que fueron asignados. Entonces, aquellos que son clasificados como epítopos por sobre el 80% se tiene una alta seguridad que lo sea.

Teniendo en cuenta los resultados obtenidos por el modelo, para comprender la eficiencia que tiene, es necesario comparar con los sistemas de predicción ya existentes. DTU Technologies of health, es una herramienta dentro de los servicios que presta la página de este laboratorio (https://services.healthtech.dtu.dk/service.php?BepiPred-2.0)(Odorico y Pellequer,2003,pp.20-23). Este programa brinda servicios de predicción de epítopos, basados en el porcentaje de probabilidad del fragmento. Por lo tanto, al ser similar la respuesta de los resultados pero distinta en su metodología, resulta un buen punto comparativo. Entonces si se observa la siguiente tabla.

		100%-80%	**79%-50%**
DTU	**6BKN**	**0**	**16**
	7JM3	**0**	**7**
Modelo RF	**6BKN**	**3**	**0**
	7JM3	**7**	**10**

Tabla 7. Comparación de modelos. Se analizaron las predicciones realizadas a las proteínas 6BKN y 7JM3, con el modelo *RF* y el programa de predicción *DTU*. Los

porcentajes asociados a las predicciones, son la probabilidad asignada por los programas, de corresponder a secuencias de epítopo, es decir, los 3 secuencias de 6BKN presentados por el modelo RF son clasificados como epítopos con una probabilidad de 80%-100%, al igual que las 7 secuencias de 7JM3, mientras que, 10 epítopos fueron clasificados con 50%-79%. Por otro lado, el método de predicción DTU entregó como resultado 16 epítopos para 6BKN y 7 para 7JM3, con porcentajes del 50% al 79%.

Al realizar la comparación entre el modelo de predicción actual contra el modelo de Random forest, se obtuvo una clara diferenciación en la cantidad de epítopos planteados, además, los porcentajes de cada secuencia también marcan un grado de confianza, dado que, los epítopos reconocidos por el modelo de RF, en la tabla 7 presentan un alto porcentaje de certeza al clasificar.

Para comparar los resultados obtenidos en la búsqueda de epítopos en 6BKN y 7JM3, por parte del programa *DTU,* se reconocieron los siguientes epítopos.

Epítopos Reales (6BKN)	Epítopos predichos por DTU (6BKN)
GLFGAIAGFIE	VQSSSTGKICNNPHRIL
ACKRGPGSGFFSRLN	HCDVFQNE
RVTVSTRRSQQTIIPNIG	CYPYDVPDYAS
GFTWTGVTQNG	TLEFITEGFTWTGVTQNGGSNACKRGPG
IEKTNEKFHQIEK	KSGSTYPVLNVTMP
KEFSEVEGRIQDLEKYV	TNQDQTSLYVQASG
DKLYIWGVHHPSTN	RRSQQTIIPNIGSRPWVRGLSS
YVKQNTLKLA	SSIMRSDAP
GTLVKTITDDQIEV	TPNGSIPNDKPFQNVNKI
LTKSGSTYPVLNVT	ATGMRNVPEKQTR

SIRNGTYDHDVYRDE	IESIRNGTYDHDVYRDEALNNR
	NEKFHQIEKEFSE
	KYVED
	KLFEK
	ENAEDMGN
	GWEGMIDGWYGFRHQNSEGTGQAADLKSTQA

Tabla 8. Predicción de epítopos en 6BKN por DTU. La búsqueda de epítopos se realizó con una probabilidad del 50% o más, dado que, no reconoce epítopos en la proteína con porcentajes superiores al 80%, como se muestra en la tabla 7. Por lo tanto, se realizó la comparativa entre los epítopos originales de la proteína, contra los predichos. Las secuencias encontradas que pertenecen a la proteína original fueron marcadas en color verde, sin embargo, los epítopos predichos presentan un aumento de aminoácidos en comparación con la secuencia de referencia.

Epítopos Reales (7JM3)	Epítopos predichos por DTU (7JM3)
GILGFVFTL	EDVFAGKNTDLEV
FVFTLTVPSER	SERGLQRRRFVQNALNGNGDPNNMDK
GILGFVFTLT	TFHGAKEISLSY
RMVLASTTAK	ADSQHRSHRQMVTTTNPLIRHEN
QKRMGVQMQRFK	MEQMAGSSEQAAEA
SLLTEVETYVLSIIPSGPL	VQAMRTIGTHPSSSAGLKND
SLL	QAYQKRMGVQM

Tabla 9. Predicción de epítopos en 7JM3 por DTU. La búsqueda de epítopos en 7JM3 se realizó de la misma forma que en 6BKN, debido a la falta de epítopos con porcentajes mayores al 80%. A diferencia con el caso anterior, todos los epítopos tentativos reconocidos por *DTU* no se presentan en la realidad.

Al comparar los epítopos predichos en las tablas 7 y 8 en la proteína 6BKN. Ambos modelos, logran reconocer dos epítopos presentes en la secuencia original, sin embargo, las diferencias de cobertura favorecen el modelo de *random forest,* debido al exceso de aminoácidos adicionales por parte de la predicción realizada en *DTU.* Otro punto favorable para el modelo de *RF*, corresponde a los porcentajes de predicción al clasificar los epítopos. *DTU,* reconoció ambos epítopos con bajos porcentajes como se muestra en la tabla 7**,** mientras que, todos los epítopos reconocidos por RF son correctos en secuencia, y tienen un porcentaje superior al 80%, por lo que, entrega una mayor confianza en los resultados.

Al evaluar las predicciones de epítopos en la proteína 7JM3, como se presenta en la tabla 9**,** *DTU* no logró reconocer ninguna secuencia de epítopos que pertenezca a la proteína, en cambio, se obtuvieron 7 fragmentos que no corresponden con los epítopos originales. Además, como se muestra en la tabla 7, todas las secuencias obtenidas por el programa, tienen porcentajes inferiores a un 80%, por lo que, no da certeza en los resultados. Por otro lado, al evaluar la proteína 7JM3 con el modelo de predicción *RF,* la tabla 7 muestra el reconocimiento de 17 epítopos. Las 7 secuencias obtenidas con porcentajes mayores al 80% son aquellas que existen como epítopo en la proteína, mientras que, las 10 secuencias que se presentan tanto en la tabla 9 como en la tabla 7**,** son aquellas con porcentajes entre 50%-79% y no corresponden a los epítopos originales de la proteína.

Una de las grandes diferencias que se pueden reconocer en ambos métodos, corresponde a la forma y las mediciones que se utilizan para la construcción del conjunto de datos. *DTU* para entrenar los modelos de predicción, recorre las estructuras 3D de las proteínas, y ubica los epítopos registrados en su secuencia mediante cadenas de markov ocultas (HHM), luego mide a 4A de distancia desde los *CA.* Si la secuencia reconocida tiene un 70% o menos de identidad en comparación a los epítopos de la base de datos, es descartada(Jespersen y col.,2017). Por lo tanto, los datos utilizados para la construcción del conjunto de datos fue de carácter estructural, la cual, mide su volúmen y polaridad, además de, identidad y similitud de sus aminoácidos.

El análisis aplicado a las secuencia de proteínas y no en su estructura 3D, marca una diferencia en los resultados obtenidos por el modelo de *RF,* dado que, el modelo emplea el cálculo de descriptores moleculares, y no se basa en la identidad de sus aminoácidos, más bien, en las mediciones que estos tienen. Además, *DTU* implementó en sus datos los valores de volúmen y polaridad para la construcción del conjunto datos, sin embargo, dejó fuera otras medidas que implican diferencias en las secuencias de epítopo, tales como, formación de anillos, largo de sus cadenas o sus cargas, entre otros.

Otra diferencia que separa la metodología de ambos modelos, consiste en la cantidad de datos analizados. *DTU* se basó en las estructuras proteicas presentes en *IMDB y PDB* , utilizando 155 proteínas antígenas que no pertenecen a un organismo en específico(Jespersen y col.,2017). Esto difiere al modelo de *RF,* dado que, solo se analizaron 10 proteínas antígenas, pertenecientes al virus de influenza A, de infección a humanos. La especificidad a la hora de seleccionar los epítopos patógenos, produce un mayor resultado en el entrenamiento del modelo, y así también, en la predicción realizada a proteínas que no se encuentran en el conjunto de datos, como se muestra en la tabla 7.

Conclusión.

La hipótesis plantea el uso de descriptores moleculares, como datos para generar un modelo de predicción de epítopos, mediante inteligencia artificial. Por lo tanto, basados en los resultados obtenidos, el modelo de random forest generado, permite reconocer las características que hacen diferente a un epítopo, e identificarlos en nuevas proteínas. Si bien, el modelo no plantea un reconocimiento total de las secuencias de epítopos en una proteína de influenza virus A, posee una gran exactitud en su clasificación, y además con una alta certeza, en comparación a los programas preexistentes.

El proyecto busca como eje central el reconocimiento de epítopos en nuevas proteínas. Al utilizar influenza de virus A como único organismo de estudio, se dejó fuera una gran cantidad de antígenos. La propuesta para estudios posteriores, implica la aplicación del modelo empleado para aumentar la cantidad de casos de estudio, de esta forma, generar un espectro de detección mayor al que existe, teniendo en cuenta la homología entre proteínas y sus organismos. Junto a esto, los epítopos lineales son una parte de las regiones reconocidas por el sistema inmune. Lograr mejorar el modelo, donde, sea capaz de analizar las estructuras proteicas, permitirá predecir sobre los epítopos discontinuos, los cuales, son una gran problemática en la actualidad.

Referencias.

Abul K. Abbas, Andrew H. Lichtman, Shiv Pillai. Inmunologia Basica de Abbas 4ta Edición. ELSEVIER; 2014.

Zinsli LV, Stierlin N, Loessner MJ, Schmelcher M. Deimmunization of protein therapeutics – Recent advances in experimental and computational epitope prediction and deletion. Comput Struct Biotechnol J. 2021;19:315-29.

He Y, Xiang Z, Mobley HLT. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol. 2010;2010:297505.

Liang S, Zheng D, Standley DM, Yao B, Zacharias M, Zhang C. EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. BMC Bioinformatics. 2010;11(1):381.

Rahman MS, Rahman MdK, Saha S, Kaykobad M, Rahman MS. Antigenic: An improved prediction model of protective antigens. Artif Intell Med. marzo de 2019;94:28-41.

Voss NR, Gerstein M, Steitz TA, Moore PB. The Geometry of the Ribosomal Polypeptide Exit Tunnel. J Mol Biol. julio de 2006;360(4):893-906.

Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. diciembre de 2009;10(1):168.

Yang L, Shu M, Ma K, Mei H, Jiang Y, Li Z. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. Amino Acids. marzo de 2010;38(3):805-16

Lozano-Aponte J, Scior T. ¿Qué sabe Ud. acerca de…QSAR? Rev Mex Cienc Farm. 2012;3.

Xabier Basogain Olabe. REDES NEURONALES ARTIFICIALES Y SUS APLICACIONES. Escuela Superior de Ingeniería de Bilbao, EHU; 2017.

Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006. 738 p. (Information science and statistics).

Singh, H., Ansari, H. R., & Raghava, G. P. S. (2013). Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PloS One*, *8*(5), e62216. https://doi.org/10.1371/journal.pone.0062216

Sanchez-Trincado, J. L., Gomez-Perosanz, M., & Reche, P. A. (2017). Fundamentals and methods for T- and B-cell Epitope prediction. *Journal of Immunology Research*, *2017*, 1–14. https://doi.org/10.1155/2017/2680160

Fleri W, Paul S, Dhanda S, Mahajan S, Xu X, Peters B, Sette A. The Immune Epitope Database and Analysis Resource in Epitope Discovery and Synthetic Vaccine Design. Frontiers in Immunology. 2017. 1664-3224

Block OKT, Rodrigo WWSI, Quinn M, Jin X, Rose RC, Schlesinger JJ. A tetravalent recombinant dengue domain III protein vaccine stimulates neutralizing and enhancing antibodies in mice. Vaccine. 2010;28(51):8085–94.

Manijeh M, Mehrnaz K, Violaine M, Hassan M, Abbas J, Mohammad R. In silico design of discontinuous peptides representative of B and T-cell epitopes from HER2-ECD as potential novel cancer peptide vaccines. Asian Pac J Cancer Prev. 2013;14(10):5973–81.

C. Leiva, "Estudio de interacciones entre fármacos y proteínas del sistema monoaminérgico y/o receptores nicotínicos utilizando técnicas de minería de datos," 2015.

Gupta S, Madhu MK, Sharma AK, Sharma VK. ProInflam: a webserver for the prediction of proinflammatory antigenicity of peptides and proteins. J Transl Med Internet. 2016;14(1). Available from: http://dx.doi.org/10.1186/s12967-016-0928-3

Alfredo P. Discriminación entre sitios de unión a metales mediante el uso de Máquinas de Vectores de Soporte y la combinación de diversos tipos de información. 2012

Rolando F, Pellón-C, Luis A. Nuevos descriptores atómicos y moleculares para estudios de estructura-actividad: Aplicaciones, 2003

Anabel G.Métodos de selección de atributos para clasificación supervisada basados en la teoría de la información. 2016

Chandrashekar G. A survey on feature selection methods. 2014. 40: 16-28

Turing, A.M. Collected works of AM Turing — Mechanical Intelligence. . Elsevier Science Publishers.(1992).

Samuel A, Zeng J, Wang Y, Rainer K, Richard H. Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants. 2007. 1367-4803.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. In arXiv [cs.LG]. http://arxiv.org/abs/1201.0490

M. Odorico and J. L. Pellequer, “BEPITOPE: predicting the location of continuous epitopes and patterns in proteins,” Journal of Molecular Recognition, vol. 16, no. 1, pp. 20–22, 2003.

Winter, G., & Fields, S. (1980). Cloning of influenza cDNA into M13: the sequence of the RNA segment encoding the A/PR/8/34 matrix protein. *Nucleic Acids Research*, *8*(9), 1965–1974. https://doi.org/10.1093/nar/8.9.1965

Kiraga, J. (2008). Analysis and computer simulations of variability of isoelectric point of proteins in the proteomes.

Guruprasad, K., Reddy, B. V. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Engineering, Design & Selection: PEDS, 4(2), 155–161. https://doi.org/10.1093/protein/4.2.155

Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Research*, *45*(W1), W24–W29. https://doi.org/10.1093/nar/gkx346

Anexos.

- Código de fragmentación de secuencia

```
txt1=input()
close=True
count1=0
count2=0
index=0
temp=""
array=[]
f = open(txt1, "r")
cadena=f.readlines()
f.close()

while close:
        for i in cadena:
                count1=len(i)

                for j in range(len(i)):
                        if(i[j]!='\n'):
                                if(j>=index):
                                        temp=temp+i[j]
                                        count2=count2+1
                                if(count2==20):
                                        array.append(temp)
                                        count2=0
                                        temp=''
        index=index+1
        count2=0
        temp=''
        if(index==count1):
                close=False
f = open("output.txt", "w")
for i in array:
        f.write(i+'\n')
f.close()

f = open("output.txt", "r")
cadena=f.readlines()
f.close()
new=[]
for i in cadena:
        insert=True
        for j in new:
                if(i==j):
                        insert=False
        if(insert):
                new.append(i)

f = open("output.txt", "w")
for i in new:
        f.write(i)
f.close()
```

- Tabla 1 .Descriptores calculados

NumTiny	pequeño (A + C + G + S + T)	Rice, Peter, Ian Longden, and Alan Bleasby. "EMBOSS: the European molecular biology open software suite." Trends in genetics 16.6 (2000): 276-277.
NumSmall	Vmuy pequeño (A + B + C + D + G + N + P + S + T +	
NumAliphatic	alifático (A + I + L + V)	
NumAromatic	aromático (F + H + W + Y)	
NumNonPolar	no polar (A + C + F + G + I + L	

	+ M + P + V + W + Y)	
NumPolar	polar (D+E+H+K+N+Q+R+S+T+Z)	
NumCharged	cargado (B + D + E + H + K + R + Z)	
NumBasic	básico (H + K + R)	
NumAcidic	ácido (B + D + E + Z)	
PorcTiny	Porcentajes por cada aminoácido	
PorcSmall	Porcentajes por cada aminoácido	
PorcAliphatic	Porcentajes por cada aminoácido	
PorcAromatic	Porcentajes por cada aminoácido	
PorcNonPolar	Porcentajes por cada aminoácido	
PorcPolar	Porcentajes por cada aminoácido	
PorcCharged	Porcentajes por cada aminoácido	
PorcBasic	Porcentajes por cada aminoácido	
PorcAcidic	Porcentajes por cada aminoácido	
at_index	El índice alifático Ikai de una proteína se define como el volumen relativo ocupado por las cadenas laterales alifáticas (alanina, valina, isoleucina y leucina). Puede considerarse como un factor positivo para el aumento de la termoestabilidad de las proteínas globulares.	Ikai (1980). Thermostability and aliphatic index of globular proteins. Journal of Biochemistry,88(6), 1895-1898.
at_boman	Esta función calcula el índice potencial de interacción de proteínas propuesto por Boman (2003) basado en la secuencia de aminoácidos de una proteína. El índice es igual a la suma de los valores de solubilidad para todos los residuos en una secuencia, podría dar una estimación general del potencial de un péptido para unirse a membranas u otras proteínas como receptores, para normalizarlo se divide por el número de residuos . Una proteína tiene un alto potencial de unión si el valor del índice es superior a 2,48.	Boman, H. G. (2003). Antibacterial peptides: basic facts and emerging concepts. Journal of Internal Medicine, 254(3), 197-215.

at_charge	Esta función calcula la carga neta de una secuencia de proteína basada en la ecuación de Henderson -Hasselbalch descrita por Moore, DS (1985). La carga neta se puede calcular a pH definido (pH=7). La carga neta (la suma algebraica de todos los grupos con carga presentes) de cualquier aminoácido, péptido o proteína, dependerá del pH del ambiente acuoso circundante. Cuando el pH8. de una solución de aminoácido o proteínas cambia la carga neta de los aminoácidos también cambiará. Este fenómeno se puede observar durante la titulación de cualquier aminoácido o proteína. Cuando la carga neta de un aminoácido o de una proteína es cero el pH será equivalente al punto isoeléctrico:pi. El pI es el valor de pH en el que una molécula posee una carga neta cero y, por tanto, carece de movilidad en un campo eléctrico. De acuerdo con esto una molécula tiene carga neta positiva en una disolución de pH<pI, y negativa cuando pH>pI.	Kiraga, J. (2008) Analysis and computer simulations of variability of isoelectric point of proteins in the proteomes. PhD thesis, University of Wroclaw, Poland. Bjellqvist, B., Hughes, G.J., Pasquali, Ch., Paquet, N., Ravier, F., Sanchez, J.Ch., Frutige,r S., Hochstrasser D. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis, 14:1023-1031. Dawson, R. M. C.; Elliot, D . C.; Elliot, W. H.; Jones, K. M. Data for biochemical research . Oxford University Press, 1989; p. 592. EMBOSS data are http://emboss.sourceforge.net/apps/release/5.0/emboss/ from apps/iep.html. Nelson, D. L.; Cox, M. M. Lehninger Principles of Biochemistry, Fourth Edition; W. H. Freeman, 2004; p . 1100. Murray, R.K., Granner, D.K., Rodwell, V.W. (2006) Harpers illustrated Biochemistry. 27th edition. Published by The McGraw-Hill Companies. Rodwell, J. Heterogeneity of component bands in isoelectric focusing patterns. Analytical Biochemistry, 1982, 119 (2), 440-449. Sillero, A., Maldonado, A. (2006) Isoelectric point determination of proteins and other macromolecules: oscillating method. Comput Biol Med., 36:157-166. Solomon, T.W.G. (1998) Fundamentals of Organic Chemistry, 5th edition. Published by Wiley. Stryer L. (1999) Biochemia. czwarta edycja. Wydawnictwo Naukowe PWN.
at_pi	El punto isoeléctrico (pI) es el pH al que una molécula o superficie en particular no tiene carga eléctrica neta.	
at_InstaIndex	Esta función calcula el índice de inestabilidad propuesto por Guruprasad (1990). Este índice predice la estabilidad de una proteína en función de su composición de aminoácidos, una proteína cuyo índice de inestabilidad es menor que 40 se predice como estable, un valor superior a 40 predice que la proteína puede ser inestable.	Guruprasad K, Reddy BV, Pandit MW (1990). "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence". Protein Eng. 4 (2): 155 - 61. doi:10.1093/protein/4.2.155
at_lengthpep	Esta función cuenta el número de aminoácidos de la secuencia de una proteína	Kidera, A., Konishi, Y., Oka, M., Ooi, T., & Scheraga, H. A . (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. Journal of Protein Chemistry, 4(1), 23-55.
at_mw	Esta función calcula el peso molecular de la secuencia de una proteína. Este cálculo es la suma de las masas de cada aminoácido usando la escala disponible en la herramienta Compute pI/mw	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.
at_hmoment_alpha	Esta función calcula el momento basado en Eisenberg, D., Weiss, RM y Terwilliger, TC (1984). El momento hidrofóbico es una medida cuantitativa de la anfifilicidad (capacidad para interactuar con	Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. Proceedings of the National Academy of Sciences, 81(1), 140-144.

	lipidos y agua) perpendicular al eje de cualquier estructura peptídica periódica, como la hélice a o la lámina b.	
at_hmoment_sheet	Se puede calcular para una secuencia de aminoácidos de N residuos y sus hidrofobicidades asociadas Hn.	
HelixBend Preference	preferencia de hélice / curva	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. Molecular Informatics, 26(6), 754-763.
SideChainSize	tamaño de cadena lateral	
ExtendedStructure Preference	preferencia de estructura extendida	
Hidrophobicity	hidrofobicidad	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. Molecular Informatics, 26(6), 754-763
DoubleBend Preference	preferencia de doble curvatura	
PartialSpecific Volume	volumen específico parcial	
FlatExtended Preference	preferencia extendida plana	
OccurrenceIn AlphaRegion	Suceso en la región alfa	
pKC	pK-C	
Surrounding Hidrophobicity	hidrofobicidad circundante	
Blosum1,Blosum2, Blosum3, Blosum4 Blosum5, Blosum6 Blosum8, Blosum7, Blosum9,Blosum10	Descriptores numéricos interpretables del espacio de aminoácidos. Sirve para simular cambios evolutivos en secuencias de proteínas. Empleando matrices como BLOSUM62 los algoritmos pueden calcular la similitud entre secuencias, que es la suma de puntuaciones de parejas de residuos	Georgiev, A. G. (2009). Interpretable numerical descriptors of amino acid space. Journal of Computational Biology, 16(5), 703-723.

	alineados tras restar penalizaciones por inserciones y deleciones	
MsWhim1	Las puntuaciones MS-WHIM se derivaron de 36 propiedades de potencial electrostático derivadas de la estructura tridimensional de los 20 aminoácidos naturales (X)	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.
MsWhim2	Y	
MsWhim3	Z	
st1, st2, st3, st4, st5, st6, st7, st8	Las escalas st tienen en cuenta 827 propiedades que son principalmente constitucionales, topológicas, geométricas, hidrofóbicas, electrónicas y estéricas de un conjunto total de 167 AAs.	Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P ., Apostolov, R., ... & Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics, 29 (7), 845-854.
t1, t2, t3, t4, t5	Las escalas T se basan en 67 descriptores topológicos comunes de 135 aminoácidos. Estos descriptores topológicos se basan solo en la tabla de conectividad de aminoácidos, y no consideran explícitamente las propiedades 3D de cada estructura.	Tian F, Zhou P, Li Z: T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. J Mol Struct. 2007, 830: 106-115. 10.1016/j.molstruc.2006.07.004.
z1	Las escalas Z se basan en las propiedades fisicoquímicas de los AA, incluidos los datos de RMN y los datos de cromatografía en capa fina (TLC). lipofilia	Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S: New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 1998, 41:2481-2491.
z2	propiedades estéricas (volumen estérico / polarización)	
z3	propiedades electrónicas (polaridad / carga)	
z4	Relacionan electronegatividad, calor de formación, electrofilicidad y dureza.	
z5	Relacionan electronegatividad, calor de formación, electrofilicidad y dureza.	
kHydrophobicityInde	hidrofobicidad: Las cadenas alifáticas de Ala, Val , Leu y Ile (y Gly) contienen átomos no polares, y por lo tanto interaccionan menos favorablemente con el agua que con otros grupos apolares. Una característica general de las proteínas globulares es que dichos residuos hidrofóbicos se encuentran en el interior de la proteína mientras que los residuos polares se encuentran en la superficie. En este aspecto, el plegamiento de la proteína puede ser	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. Molecular Informatics, 26(6), 754-763.

	comparado con la formación de micelas de lípidos en solución acuosa: la cadena se ordena de forma que los grupos apolares queden internos y los grupos polares expuestos.	
AlphaAndTurn Propensities	Inclinación/propensiones alfa y de giro	
BulkyProperties	Propiedades voluminosas dentro de la secuencia	
Compositional CharacteristicIndex	características de composición: Escala de análisis factorial de información generalizada de aminoácidos como fuente de un nuevo conjunto de descriptores para dilucidar la estructura y las relaciones de actividad de los péptidos antimicrobianos catiónicos.	
LocalFlexibility	flexibilidad local	
ElectronicProperties	propiedades electrónicas	

- **Primer ciclo.**

```
[[  2  26]
 [  2 987]]
[[  6  18]
 [  3 990]]
[[  5  16]
 [  3 993]]
[[  6  18]
 [  3 990]]
[[  5  23]
 [  0 989]]
[[  4  19]
 [  3 991]]
[[  4  25]
 [  2 986]]
[[  3  25]
 [  4 985]]
[[  3  21]
 [  1 992]]
[[  7  18]
 [  1 991]]
```

- **Segundo ciclo**

```
[[995   0]
 [ 11 976]]
[[1010    0]
 [   8  964]]
[[1004    0]
 [  16  962]]
[[975   0]
 [ 14 993]]
[[ 969    0]
 [  12 1001]]
[[979   0]
 [ 10 993]]
[[991   0]
 [ 11 980]]
[[989   0]
 [ 15 978]]
[[981   0]
 [ 16 985]]
[[988   0]
 [ 10 984]]
```

- **Matriz 1**

-Entrenamiento sin 6BKN

```
[[923   0]
 [ 23 940]]
[[946   0]
 [ 26 914]]
[[935   0]
 [ 27 924]]
[[946   0]
 [ 21 919]]
[[965   0]
 [  8 913]]
[[948   0]
 [ 16 922]]
[[957   0]
 [ 23 906]]
[[963   0]
 [ 24 899]]
[[927   0]
 [ 25 934]]
[[948   0]
 [ 17 921]]
```

-Evaluación 6BKN

```
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
[[  2   9]
 [  1 483]]
```

-evaluación 7JM3

```
[[  6   1]
 [ 10 231]]
[[  6   1]
 [ 13 228]]
[[  6   1]
 [  9 232]]
[[  6   1]
 [ 10 231]]
[[  6   1]
 [ 13 228]]
[[  6   1]
 [ 15 226]]
[[  6   1]
 [ 10 231]]
[[  6   1]
 [ 14 227]]
[[  6   1]
 [ 15 226]]
```