SISTEMA DE CLASIFICACIÓN DE PÉPTIDOS ANTIBACTERIANOS UTILIZANDO MÁQUINAS DE SOPORTE VECTORIAL.

Francy Liliana Camacho Urrea

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Bucaramanga, 2013

SISTEMA DE CLASIFICACIÓN DE PÉPTIDOS ANTIBACTERIANOS UTILIZANDO MÁQUINAS DE SOPORTE VECTORIAL.

Francy Liliana Camacho Urrea

Trabajo de grado presentado para optar por el título de Ingeniera de Sistemas e Informática

Directores del Proyecto:
Lola Xiomara Bautista, *Mpe.*
Nydia Paola Rondón, *MeI*
Rodrigo Gonzalo Torres Sáez, *PhD*
Daniel Sierra Bueno, *PhD*

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Bucaramanga, 2013

A mis padres

Agradecimientos

A mis padres y hermanos por el apoyo incondicional que han brindado durante tantos años. Han sido mi mayor motivación para alcanzar mis metas.

A mis directores Lola, Daniel y Paola, por su orientación y por su paciencia.

Al profesor Rodrigo Torres, por sus enseñanzas y por su ayuda incondicional.

A mis amigos del GIBIM y GIIB por su compañía. Fue grato compartir con ustedes.

A Duvan, Emigdio, José y Johare por todo el tiempo que compartirmos. Me siento feliz de contar con grandes amigos como ustedes.

A Darío José por orientar mi camino en momentos donde estaba confundida.

A todos aquellos que no nombré, pero que fueron parte importante de mi vida.

Glosario

- **Anfipático:** Son aquellas moléculas que poseen un extremo polar (hidrofílico o afines al agua) y un extremo no polar (hidrofóbico o repelido por el agua).
- **APD:** La APD (Antimicrobial Peptide Database) es un repositorio de datos, dedicado al glosario, nomenclatura, clasificación, predicción y estadísticas de péptidos antimicrobianos.
- **Catiónico:** Hace referencia a la carga positiva de un átomo o molécula.
- **Péptidos antibacterianos:** Son proteínas de origen natural que tienen propiedades antibióticas y que han sido fabricados por la naturaleza para actuar como medio de defensa en contra de enfermedades producidas por bacterias.
- **Resistencia Antibacteriana:** Es la capacidad de una bacteria para resistir los efectos de un medicamento.
- **Péptido:** Son un tipo de moléculas formadas por la unión de varios aminoácidos mediante enlaces peptídicos.
- **Polipéptido:** Es el nombre utilizado para designar un péptido de tamaño suficientemente grande.

Índice general

Índice de figuras

Índice de cuadros

Índice de anexos

Resumen

TITULO: Sistema de clasificación de péptidos antibacterianos utilizando máquinas de soporte vectorial [1]

AUTOR: Francy Liliana Camacho Urrea. [2]

PALABRAS CLAVE: Diseño de medicamentos, máquinas de soporte vectorial, péptidos antibacterianos

En los últimos años, el reconocimiento de patrones se ha aplicado en diversas áreas para resolver múltiples problemas. Una de estas áreas es el diseño *in silico* de medicamentos, donde ha sido ampliamente utilizados en el análisis de proteínas. Por ejemplo, para predecir la actividad antibacteriana presente en péptidos (proteínas cortas), los cuales se están utilizando como alternativas a los medicamentos tradicionales. En este trabajo, se propone utilizar herramientas como las máquinas de soporte vectorial(SVM, por sus siglas en inglés) junto con el modelo denominado Relación Cuantitativa entre Estructura-Actividad (QSAR, por sus siglas en inglés), para realizar el reconocimiento de patrones y crear algoritmos que permitan identificar la actividad antibacteriana en péptidos. Para llevar a cabo este proceso, se parte de un conjunto de 2288 secuencias representativas de péptidos con y sin actividad antimicrobiana, para los cuales se codifica información numérica y como resultado se creó un clasificador en cascada que muestra una precisión estimada del 80 %, resultados que permiten inferir que los descriptores utilizados para codificar las secuencias contienen la información suficiente para relacionar los péptidos y su actividad antibacteriana mediante el uso de máquinas de aprendizaje.

[1] Trabajo de grado

[2] Facultad de Ingenierías Físico Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Lola Bautista. Codirector: Rodrigo Torres, Daniel Sierra, Paola Rondón.

Abstract

TITLE: System of classification of antibacterial peptides using support vector machines[3]

AUTHOR: Francy Liliana Camacho[4]

KEY WORDS: Design of drugs, support vector machines, antibacterial peptides.

In recent years, the pattern recognition has been applied in many areas to solve diverse problems. One of those areas is the *in silico* drug design, which has been widely used in the protein analysis. For example, to predict the antibacterial activity in peptides (small proteins), which are being used as alternatives to traditional medicines. In this paper, we propose to use tools such as the support vector machine and Quantitative Structure-Activity Relationship (QSAR) model, for recognition patterns and create algorithms to identify the antibacterial activity in peptides. For this process, we begins with a set of 2288 representative sequences of peptides with and without antimicrobial activity, and then we encoded numerical information and such as results, we developed a cascade classifier that have a estimated accuracy of 80 %. This result shows that descriptors used for encoded the sequences, has sufficient information to correlated the peptides with antibacterial activity using learning machines.

[3]Research work

[4]Faculty of Physical-Mechanical Engineerings. Systems engineering and informatics department. Advisor: Lola Bautista. Co-advisor: Rodrigo Torres, Daniel Sierra, Paola Rondón

Introducción

A pesar de los avances en los últimos años en el tratamiento de enfermedades, el creciente aumento de la resistencia de los microorganismos a los medicamentos es catalogado por la Organización Mundial de la Salud (OMS) como una de las tres más grandes amenazas para la salud humana [2], por ello, la búsqueda de nuevos fármacos se ha dirigido al estudio de los péptidos antibacterianos, gracias a su baja toxicidad, rápida acción antibacteriana y por la poca resistencia que generan las bacterias al uso de éstos. Sin embargo, el uso de péptidos antibacterianos no es muy común debido a los altos costos de los recursos que se utilizan el proceso de síntesis [3].

Para soportar este problema, existen modelos *in silicio* como Relación Cuantitativa entre Estructura-Actividad (QSAR) cuya idea se basa en que, la secuencia o estructura de péptidos pueden ser descritos a través de parámetros físico-químicos denominados descriptores, que se correlacionan, a su vez, con la actividad biológica de los mismos [4], [5], [6] y cuya finalidad es utilizar herramientas computacionales que permitan a los científicos simular posibles péptidos con actividad antibacteriana.

Por tal motivo, en este estudio se planteó si es posible clasificar péptidos antibacterianos a partir de diez descriptores (punto isoeléctrico, hidrofobicidad, tamaño del péptido, hélice α , hoja β ,tendencia de la estructura a girar, agregación *in vitro* e *in vivo*, carga y el peso molecular) y mediante el uso de máquinas de soporte vectorial. Como resultado, se obtuvo que el clasificador generado es capaz de identificar péptidos con actividad antibacteriana, con una precisión del 80 %.

Para mostrar el trabajo desarrollado, este documento presenta en la primera parte el marco conceptual, seguido de la metodología llevada a cabo en el

trabajo y los métodos de verificación junto con los resultados. En la tercera parte, se presentan las conclusiones del estudio y finalmente los anexos que contienen los cuadros de resultados, la herramienta software, el manual de usuario e información detallada para usar dicha herramienta.

1. Marco de referencia

1.1. Péptidos Antibacterianos

Los péptidos antibacterianos son proteínas cortas (6 a 100 aminoácidos) que hacen parte del sistema inmune del huésped y se encuentran en una amplia variedad de seres vivos, la mayor parte de ellos son anfipáticos y catiónicos, cuya carga oscila entre +2 a +9, características que les permiten interactuar con la membrana negativa de las bacterias.

Han sido estudiados como alternativas a los medicamentos tradicionales, gracias a que presentan varias ventajas como lo son baja toxicidad, rápida acción y baja probabilidad de generar resistencia bacteriana. No obstante, tienen desventajas como la rápida degradación, toxicología desconocida y los altos costos de producción [2]

Como son proteínas, se caracterizan por tener cuatro formas estructurales, que se pueden visualizar en la figura 1:

- Estructura primaria, se refiere a la secuencia lineal de aminoácidos.
- Estructura secundaria, es la alteración que sufren los aminoácidos cercanos en la secuencia debida a los enlaces de hidrógeno.
- Estructura terciaria, hace referencia a la distribución de los aminoácidos en el espacio.

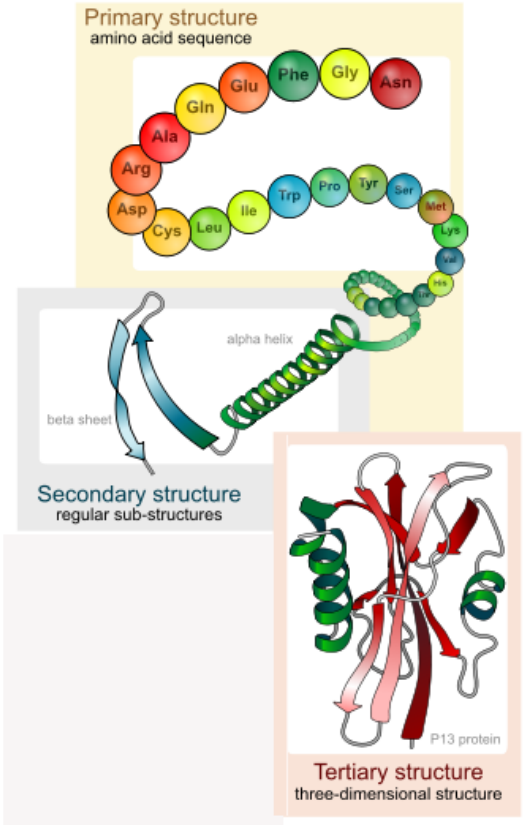


Figura 1: Niveles estructurales de una proteína.[Tomado de Wikipedia.org]

1.2. Relación Cuantitativa Entre Estructura-Actividad(QSAR)

Como se mencionó anteriormente, el modelo denominado relación cuantitativa entre estructura-actividad es uno de los más usados y se basa en que, la secuencia o estructura de un péptido puede ser descrito a través de sus propiedades físico-químicas, denominadas descriptores moleculares y que éstas a su vez se correlacionan con la actividad antibacteriana presente en el péptido [4], [5], [6]. Para llevar a cabo QSAR, se debe tener en cuenta tres aspectos importantes: la extracción de los descriptores, la selección de los descriptores relevantes y por último, el mapeo de los descriptores con la actividad [5]. A continuación se presenta una breve descripción de los aspectos mencionados anteriormente.

1.2.1. Descriptores moleculares

Los descriptores moleculares son el resultado final de un procedimiento lógico y matemático que transforma la información química presente en una molécula, en una medida cuantitativa. No obstante, existen más de 1500 descriptores obtenidos a partir de la estructura primaria, secundaria y terciaria, y el uso de éstos varía en los estudios. Debido a ello, en este trabajo se pretende probar diez descriptores moleculares extraídos a partir de la secuencia de aminoácidos como lo son: carga, peso molecular, punto isoeléctrico, hidrofobicidad, tamaño del péptido, hélice , hoja , tendencia de la estructura a girar y agregación *in vivo* e *in vitro* [7], [8], [9], [10], [11].

1.2.2. Selección de descriptores relevantes

Como se nombraba anteriormente, la cantidad de descriptores es amplia y algunos son apropiados para el proceso de predicción, sin embargo, en ciertos modelos se usan herramientas estadísticas que buscan reducir el número de propiedades físico-químicas usadas, es decir escoger aquellos descriptores que no tengan una relación lineal con otros, y con ello, tratar de mejorar la predicción; entre ellas están la relación de Fisher, prueba de Smirnov-Kolmogorov [5], el coeficiente de correlación de Pearson, muy usado en estudios QSAR [5], [8], [7], entre otros.

También existen otras herramientas que trabajan conjuntamente con el algoritmo que hace la predicción, es decir, se evalúan diferentes subconjuntos de descriptores en el algoritmo y de acuerdo al error obtenido por éste se descartan descriptores, de tal forma que al final, se obtiene el subconjunto con la precisión más alta. Entre éstas herramientas están los algoritmos genéticos, el algoritmo de recocido simulado, selección secuencial hacia adelante, entre otros [5].

1.2.3. Mapeo de descriptores y la actividad

Finalmente, se procede a relacionar los descriptores con la actividad presente en el péptido. Para ello existen métodos como los lineales, donde la actividad es predicha a partir de una relación lineal de los descriptores, por ejemplo, la regresión lineal múltiple (MLR, por sus siglas en inglés), el análisis discriminante lineal (LDA, por sus siglas en inglés), mínimos cuadrados parciales, entre otros. [5]; éste último se usa especialmente cuando existe una cantidad abundante de descriptores [5]; [8], [12].

Por otro lado, están los métodos no lineales, en los que se usa una función, no lineal que relaciona los descriptores y la actividad. Son capaces de aprender a partir de un conjunto de datos de entrenamiento (péptidos con y sin actividad antibacteriana) y responder ante un nuevo dato que no ha sido mostrado previamente (un nuevo péptido al que se le quiere determinar la actividad), y son más precisos para grandes conjuntos de datos. Entre ellos están, las redes neuronales (NN por sus siglas en inglés) y las máquinas de soporte vectorial. Las NN son muy usadas en QSAR y la precisión reportada varía entre 80 % y 94 % [4], [9], [12], [13], [11]. Por su parte, las SVM han sido usadas en la predicción de péptidos antibacterianos y se ha reportado precisiones del 75 % [11] y del 91.5 % [10].

1.3. Máquinas de Soporte Vectorial

Son algoritmos de optimización asociados al aprendizaje supervisado, empleados en tareas de clasificación. A partir de un conjunto de entrenamiento (péptidos con su respectiva actividad, también denominado etiqueta) crean modelos matemáticos, capaces de clasificar un nuevo elemento sin conocer de antemano la clase a la que pertenece. Para generar dicho modelo se parte de un conjunto de entrenamiento $(x_1, y1), (x_2, y2) \cdots (x_n, yn)$, donde x_i es la representación vectorial de los ejemplos de entrenamiento, y_i, la etiqueta que representa la clase a la que corresponde y n la cantidad de individuos. Dicha máquina necesita resolver el problema de optimización mostrado en la ecuación 1.

$$\begin{aligned} \min_{W,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{n} \xi_i \\ Sujeto\,a : y_i(W^T \phi(x_i) + b) \geq 1 - \xi_i, \\ \xi_i \geq 0 \end{aligned} \tag{1}$$

Donde

- $W^T \phi(x_i) + b$ representa el hiperplano de separación
- W es el vector de pesos.
- b es el bias
- ϕ es la función kernel que permite llevar los datos a un espacio N-dimensional mayor en el cual puedan ser separables.
- C es un parámetro positivo que especifica el usuario y que controla el equilibrio entre la complejidad de la máquina y el número de puntos no separables por un hiperplano.
- ξ_i se denomina variable de holgura, que mide la desviación de un punto x_i del punto de separación $W^T \phi(x_i) + b$ (Ver Figura 2).

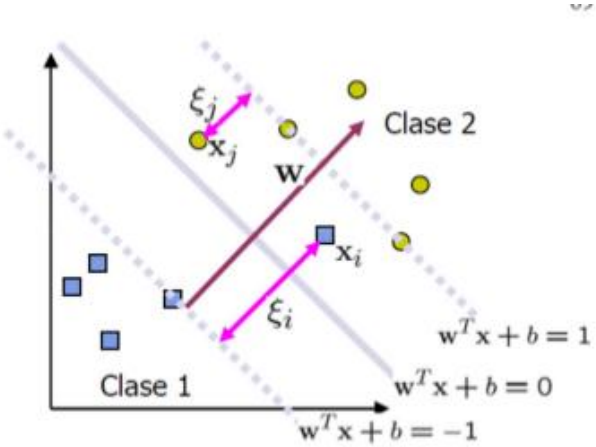


Figura 2: Representación gráfica del modelo de optimización a resolver. [Tomado de [1]]

Los vectores x_i son mapeados a un espacio N-dimensional mayor por la función ϕ que puede ser de tipo lineal, sin embargo, en muchos casos, una función lineal no permite separar adecuadamente los datos, para lo cual se emplean otro tipo de funciones como las funciones kernel, representadas por $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$. Existen varias funciones kernel básicas:

- Función kernel lineal:

$$K(x_i, x_j) = (x_i)^T (x_j) \tag{2}$$

- Función kernel polinomial:

$$K(x_i, x_j) = \gamma(x_i^T x_j + r)^d, \gamma > 0 \tag{3}$$

- Función de Base Radial (FBR):

$$K(x_i, x_j) = (-\gamma \|x_i - x_j\|^2), \gamma > 0. \tag{4}$$

Donde γ, r y d son parámetros libres de los kernels.

De acuerdo a la función kernel que se emplee, existen parámetros libres que se deben ajustar, por ejemplo, C para el modelo SVM y γ para la FBR, cuyo problema consiste en encontrar los valores que deben tomar estos parámetros para que el clasificador f que se genere sea el mejor, es decir, identificar C, γ tales que f sea capaz de predecir adecuadamente los datos de prueba. Las MSV buscan un hiperplano de separación óptimo, donde el margen es el máximo entre los grupos a clasificar. La solución está basada sólo en esos puntos ubicados sobre el margen, a los que se les denomina vectores de soporte [14].

2. Clasificación de péptidos

En este capítulo, se explica todo el proceso llevado a cabo para elegir y extraer el conjunto de datos apropiado para el trabajo, así como el problema de clasificación y la solución.

2.1. El conjunto de datos

El conjunto de datos que será empleado está representado por las secuencias de aminoácidos (cadenas de texto) de péptidos con y sin actividad antimicrobiana (clase positiva y clase negativa, respectivamente). A partir de este conjunto, se hace una división en dos grupos, uno denominado de entrenamiento, utilizado para mostrarle el universo de datos a la máquina y del cual se espera que aprenda, y el otro llamado de prueba, cuya finalidad es evaluar el aprendizaje de dicho algoritmo. Cabe destacar que el conjunto de prueba no se le ha mostrado con anterioridad a la máquina. Para tal fin, se utilizó la base de datos online APD [15], debido a las facilidades que presentaba para obtener los datos apropiados para el trabajo y por ser una de las bases de datos más referenciadas.

Péptidos	Población inicial	Población final
Antibacterianos	1708	958
Anticancer	142	34
AntiVIH	88	27
Insecticidas	21	7
Antiparasitales	45	17
Espermicidas	9	0
Antivirales	126	36
Antifúngicos	748	111
Antiprotisto	3	2
Sin actividad	10014	1144

Cuadro 1: Tamaño del conjunto de datos de los péptidos

A partir de dicha base de datos se extrajeron todas las secuencias pertenecientes a los grupos mostrados en el cuadro 1 [5],teniendo en cuenta que la longitud de las cadenas de texto (representada por x) cumpliera la restricción $6 < x < 100$ aminoácidos. Sin embargo, la APD no provee los péptidos sin actividad antimicrobiana, por tal motivo, dichos datos fueron extraídos de Ping y colaboradores, 2011 [16]. Cabe aclarar que, el conjunto de datos se redujo, como se muestra en la cuadro 1, columna 2, por varias razones, la primera, los péptidos pueden presentar varias actividades al tiempo (es decir un péptido puede ser antibacteriano, antifúngico y antiviral) y para el caso de los antibacterianos, sólo interesan aquellos que ataquen exclusivamente a las bacterias y para los demás péptidos con actividad, sólo se tuvieron en cuenta aquellos que no presentaran actividad antibacteriana.

2.2. Codificación de las secuencias

A partir de las secuencias de aminoácidos, es posible codificar información numérica que representa diversas propiedades físico-químicas de los

[5]En Noviembre de 2012, fecha en la cual se consultó la APD, existía dicha cantidad de secuencias

péptidos (descriptores); sin embargo, existe una amplia variedad de descriptores y para éste trabajo se emplearon diez propiedades, de los cuales, punto isoeléctrico, hidrofobicidad, tamaño del péptido, hélice α , hoja β ,tendencia de la estructura a girar y agregación *in vitro* e *in vivo*, fueron tomadas de Torrent y colaboradores, 2011 [11] y la carga y el peso molecular son empleados en varios estudios [10], [15], [4].

Aa	**id**	**Hydropatía**	**Peso**
L	1	3,8	113,1595
I	2	4,5	113,1595
N	3	-3,5	114,1039
G	4	-0,4	57,052
V	5	4,2	99,1326
E	6	-3,5	129,1155
P	7	-1,6	97,1167
H	8	-3,2	137,1412
K	9	-3,9	128,1742
A	10	1,8	71,0788
Y	11	-1,3	163,1760
W	12	-0,9	186,2133
Q	13	-3,5	128,1308
M	14	1,9	131,1986
S	15	-0,8	87,0782
C	16	2,5	103,1448
T	17	-0,7	101,1051
F	18	2,8	147,1766
R	19	-4,5	156,1876
D	20	-3,5	115,0886

Cuadro 2: Aminoácidos y valores representativos

Para calcular los descriptores hélice α , hoja β ,tendencia de la estructura a girar y agregación *in vitro* se tuvo en cuenta la estructura secundaria utilizando el software Tango [17], para agregación *in vivo* se empleó AGGRESCAN [18], para el tamaño del péptido se usó la función length de MATLAB® y para el punto isoeléctrico, carga neta,hidrofobicidad y peso

molecular, se construyeron las rutinas pertinentes.

Inicialmente, como el conjunto de entrada está formado por cadenas de texto, es necesario convertirlas en información numérica, que pueda ser procesada por la máquina, por ello, se elaboró el algoritmo 1, que le asigna un número a cada aminoácido (ver cuadro 2). Esta información numérica, se convierte en las entradas de los demás algoritmos. En el caso del punto isoeléctrico, éste hace referencia al pH que una sustancia anfótera (que reacciona como una base o un ácido) tiene carga cero y se calcula como la contribución de la carga para cada aminoácido [19], para lo cual se creó el algoritmo 2.

Algoritmo 1 Conversión de texto a señal numérica.

Entrada: Secuencia de aminoácidos representada por $S = \{Aas_1, Aas_2, ..., Aas_m\}$
Salida: Señal numérica dada por $Senal = \{p_1, p_2, ..., p_m\}$
{m es la longitud de la secuencia de aminoácidos}

```
1: para i = 1 hasta m hacer
2:    para j = 1 hasta 20 hacer
3:       //Compara los aminoácidos
4:       si Aas_i=Aa_j entonces
5:          Senal_i = id_j
6:          Interrumpir
7:       fin si
8:    fin para
9: fin para
```

La carga neta se refiere a las interacciones o repulsiones que se llevan a cabo entre algunas partículas subatómicas, que determinan sus interacciones electromagnéticas y se calcula como el aporte de carga de algunos aminoácidos [19] (Ver algoritmo 4).

Algoritmo 2 Punto isoeléctrico. Parte 1.

Entrada: Señal numérica dada por $Senal = \{p_1, p_2, ..., p_m\}$
Salida: Punto isoeléctrico del péptido

{Cálculo del número de aminoácidos existentes para cada uno de los que aportan al pI}

nArg, nLys, nAsp, nGlu, nTyr, nHis, nCys, sump, sumn se inicializan en cero.
para $i = 1$ hasta m **hacer**
  si $p_i = 19$ **entonces**
    nArg=nArg+1
  si no, **si** $p_i = 9$ **entonces**
    nLys=nLys+1
  si no, **si** $p_i = 20$ **entonces**
    nAsp=nAsp+1
  si no, **si** $p_i = 6$ **entonces**
    nGlu=nGlu+1
  si no, **si** $p_i = 11$ **entonces**
    nTyr=nTyr+1
  si no, **si** $p_i = 8$ **entonces**
    nHis=nHis+1
  si no, **si** $p_i = 16$ **entonces**
    nCys=nCys+1
  si no

  fin si
fin para

La hidrofobicidad es la tendencia de las moléculas a repeler el agua o incapacidad de disolverse en ella. Para calcularlo, se cuenta cuántos aminoácidos existen de cada uno en la secuencia de aminoácidos y luego se multiplica por los valores de hidropatía correspondientes (ver cuadro 2, [20],algoritmo 5).

Algoritmo 3 Continuación Punto isoeléctrico. Parte 2.

Entrada: Señal numérica dada por $Senal = \{p_1, p_2, ..., p_m\}$
Salida: Punto isoeléctrico del péptido
{Se calcula la contribución de carga para cada aminoácido}
nA={1, nAsp, nGlu, nCys, nTyr}, nC={3.65, 3.9,4.07, 8.18, 10.46},
pA={nHis, 1, nLys, nArg}, pC={6.04, 8.2, 10.54, 12.48}, $NQ = 0$
mientras $NQ = 0$ **hacer**
  para $i = 1$ hasta 5 **hacer**
    si $i <= 4$ **entonces**
      $qp_i = \frac{pA}{1+10^{pH-pC}}$
      $sump = sump + qp_i$
    fin si
    $qn_i = \frac{nA*-1}{1+10^{nC-pH}}$
    $sumn = sumn + qn_i$
  fin para
  $NQ = sump + sumn$
fin mientras

Algoritmo 4 Carga neta.

Entrada: Señal numérica dada por $Senal = \{p_1, p_2, ..., p_m\}$
Salida: Carga neta del péptido
para $i = 1$ hasta m **hacer**
  si $p_i = 19$ **entonces**
    nArg=nArg+1
  si no, **si** $p_i = 9$ **entonces**
    nLys=nLys+1
  si no, **si** $p_i = 20$ **entonces**
    nAsp=nAsp+1
  si no, **si** $p_i = 6$ **entonces**
    nGlu=nGlu+1
  si no

  fin si
fin para
$C = (nArg + nLys) - (nAsp + nGlu)$

El peso molecular es la suma de los pesos atómicos de cada aminoácido (cuadro 2), que están en la fórmula molecular de un compuesto y para el cual se construyó el algoritmo 6.

Algoritmo 5 Hidrofobicidad media.

Entrada: Secuencia de aminoácidos representada por $S = \{Aas_1, Aas_2, ..., Aas_m\}$
Salida: Hidrofobicidad media del péptido
$\{nAa$ y val se inicializan en $0\}$
para $i = 1$ hasta 20 **hacer**
  para $j = 1$ hasta m **hacer**
    si Aas_j=id_i **entonces**
      $nAa_i = nAa_i + 1$
    fin si
  fin para
fin para
para $i = 1$ hasta 20 **hacer**
  $temp = nAa_i * Hydropatia_i$
  $val = val + temp$
fin para
$meanHyd = \frac{val}{m}$

Algoritmo 6 Peso.

Entrada: Secuencia de aminoácidos representada por $S = \{Aas_1, Aas_2, ..., Aas_m\}$
Salida: Peso del péptido
$\{nAa$ y $weig$ se inicializan en $0\}$
para $i = 1$ hasta 20 **hacer**
  para $j = 1$ hasta m **hacer**
    si Aas_j=id_i **entonces**
      $nAa_i = nAa_i + 1$
    fin si
  fin para
fin para
para $i = 1$ hasta 20 **hacer**
  $temp = nAa_i * Peso_i$
  $weig = weig + temp$
fin para$\{$Se suma el peso del N-Terminal y del C-Terminal$\}$
$Nterm = 1{,}0079$, $Cterm = 17{,}0073$
$weig = weig + Nterm + Cterm$

2.3. Planteamiento del problema

Sea

- $X = \{x_1, x_2, \cdots, x_n\}$ el conjunto de los descriptores de los péptidos, representados en forma vectorial, donde x_i es un péptido representado por un vector de 10 descriptores, donde $i = 1, 2, ..., n$
- n el número total de individuos empleados
- $\beta = \{c_1, c_2\}$ el vector de clases, donde c_1 la clase positiva o de interés y c_2 la clase negativa.

El problema consiste en encontrar un hiperplano de separación óptimo, que clasifique correctamente entre ambas clases, a partir de un conjunto entrenamiento $(x_1, y_1), (x_2, y_2) \cdots (x_m, y_m)$, donde $x_i \in X$ y $y_i \in \beta$.

2.4. Planteamiento de la solución

Para tratar el problema, se planteó un clasificador f conformado por dos clasificadores en cascada f_1 y f_2, tal como se aprecia en la figura 3; donde f_1 clasifica entre péptidos con y sin actividad antimicrobiana, es decir, su función consiste en filtrar los péptidos que poseen actividad, y que luego serán evaluados en f_2 que clasifica entre péptidos con y sin actividad antibacteriana.

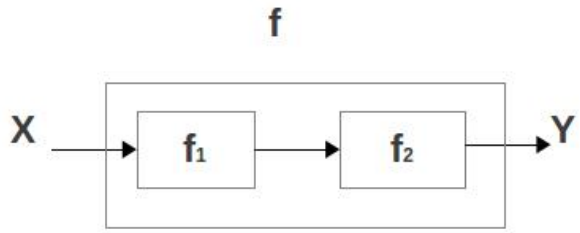


Figura 3: Esquema general de la solución.

Ahora bien, el conjunto de datos para f_1, está conformado por 2288 péptidos, de los cuales 1144 corresponden a c_1 o péptidos con actividad

antimicrobiana, es decir:
$c_1 = \{antibacteriano \cup antifungico \cup antiVIH \cup insecticida \cup anticancer \cup antiparasital \cup antiviral \cup espermicida \cup antiprotisto\}$
Donde el nombre corresponde al conjunto de péptidos con dicha actividad.

Y c_2 es la clase no antimicrobiano o sin actividad, que contiene 1144 péptidos y cuya cantidad fue determinada teniendo en cuenta que, fuese igual a c_1 y se extrajeron con la técnica de muestreo aleatorio. Cabe destacar, que un péptido puede tener varias actividades al tiempo, es decir, está en varios grupos a la vez, por tal motivo, aquellas secuencias repetidas se descartaron y el conjunto total de péptidos antimicrobianos se redujo de 1192 a 1144.

Para la máquina de soporte vectorial, se emplea el conjunto de entrenamiento $(x_1, y_1), (x_2, y_2) \cdots (x_m, y_m)$, donde $x_i \in X$, $y_i \in \beta$ y m el número de péptidos empleados para el entrenamiento. Cabe destacar que, para crear las MSV se usó MATLAB®, que tiene implementadas varias funciones kernel, tres funciones kernel (lineal, polinomial y de base radial, nombradas en el marco teórico) y una cuarta llamada función kernel cuadrática. De dichas funciones, se debe escoger aquella que mejor rendimiento ofrezca ajustando los parámetros libres respectivos (se debe tener en cuenta que para el rendimiento se deben observar la precisión, la sensibilidad, la especificidad y el coeficiente de correlación de Mathews para elegir la mejor función).
Para ajustar dichos parámetros se usó la metodología de Hsu y colaboradores, 2010 [21], en la cual, para la MSV empleando una función de base radial recomiendan el uso de una malla, como se observa en el cuadro 3, variando C (de la MSV) y γ (de la función kernel), de tal forma, que, se establecen rangos $[C_{inicial}, C_{final}]$, $[\gamma_{inicial}, \gamma_{final}]$, el paso δ_C y δ_γ y l la cantidad de muestras empleadas para cada parámetro y con los cuales se desea probar las MSV.

Sin embargo, como la función polinomial tienen más de dos parámetros libres, ésto genera dificultades para conformar las mallas, por lo que se de-

	γ_1	γ_2	.....	γ_l
C_1	$Q_{1,1}$	$Q_{1,2}$	.....	$Q_{1,l}$
C_2	$Q_{2,1}$	$Q_{2,2}$	.....	$Q_{2,l}$
$\vdots$	$\vdots$	$\vdots$	.....	$\vdots$
C_l	$Q_{l,1}$	$Q_{l,2}$	.....	$Q_{l,l}$

Cuadro 3: Malla para escoger los parámetros libres de una función kernel de base radial en una MSV

cidió variar únicamente C y d para la polinomial (Ver cuadro 4).

	d_1	d_2	.....	d_o
C_1	$Q_{1,1}$	$Q_{1,2}$	.....	$Q_{1,l}$
C_2	$Q_{2,1}$	$Q_{2,2}$	.....	$Q_{2,l}$
$\vdots$	$\vdots$	$\vdots$	.....	$\vdots$
C_l	$Q_{l,1}$	$Q_{l,2}$	.....	$Q_{l,o}$

Cuadro 4: Malla para escoger los parámetros libres de una función polinomial en una MSV

Para el caso de la función lineal y cuadrática cuadro 4, como éstas no tienen parámetros libres, sólo se ajusta el parámetro C de la MSV, para ambos casos (Ver cuadro 5).

	C_1	C_2	.....	C_l
Q	Q_1	Q_1	.....	Q_l

Cuadro 5: Malla para escoger el parámetro libre en una MSV

Para f_2, el procedimiento anterior es el mismo, salvo que, el conjunto de datos contiene 372 péptidos, de los cuales, c_1 es la clase péptido antibacteriano y cuyo tamaño es igual a 186 péptidos y c_2 es la clase no antibacteriano, formado por la unión de varios grupos de actividades, $c_2 = \{antifungico \cup antiVIH \cup insecticida \cup anticancer \cup antiparasital \cup antiviral \cup espermicida \cup antiprotisto\}$. Cabe señalar que para c_2, como una secuencia puede tener múltiples actividades, aparece en varios grupos a la vez, por tanto, se descartaron aquellas cadenas de texto repetidas y por

ende, se redujo el conjunto de 234 a 186 secuencias. En la sección de resultados obtenidos, se muestra con más detalle el uso de las mallas y posterior elección de los clasificadores.

3. Resultados

En esta sección, se explican las medidas utilizadas para verificar y validar, y el uso de éstas para elegir los clasificadores con los mejores resultados.

3.1 Medidas de rendimiento

Para verificar el rendimiento de la MSV, existen varias medidas que indican qué tan bueno es el clasificador creado, y para éste trabajo se emplearon: la precisión (Prec, ecuación 5), entendida como la proporción de resultados positivos identificados correctamente en la población, la sensibilidad (Sens, ecuación 6) que mide la capacidad para detectar acertadamente los peptidos antibacterianos. La especificidad (Espec, ecuación 7) que determina la proporción de péptidos no antibacterianos identificados correctamente y el coeficiente de correlación de Mathews (CCM, ecuación 8), que mide la calidad de los clasificadores. Para las medidas anteriores, mientras más cercanas sean a 1 mejor es el rendimiento del clasificador.

$$Prec_i = \frac{VP_i + VN_i}{(VP_i + FP_i + VN_i + FN_i)} \tag{5}$$

$$Sens_i = \frac{VP_i}{(VP_i + FP_i)} \tag{6}$$

$$Espec_i = \frac{VN_i}{(VN_i + FN_i)} \tag{7}$$

$$CCM_i = \frac{VP_iVN_i - FP_iFN_i}{\sqrt{(VP_i + FP_i)(VP_i + FN_i)(VN_i + FP_i)(VN_i + FN_i)}} \tag{8}$$

Para las ecuaciones anteriores, VP (Verdaderos Positivos) y VN (Verdaderos Negativos), representan los péptidos antibacterianos y no antibacterianos clasificados de forma acertada, respectivamente. FN (Falsos Negativos) y FP (Falsos Positivos) son los péptidos no antibacterianos y antibacterianos clasificados en las clases incorrectas.

3.2 Validación

Para validar los clasificadores, se usó validación cruzada (VC, o en inglés es conocido como k-fold cross validation) que consiste en dividir el conjunto de datos X de forma aleatoria, en varios subconjuntos k (para este trabajo se tomó $k = 10$, muy común en la literatura), de los cuales $k-1$ son usados para entrenar y el otro para pruebas y se hace repetitivamente con cada uno de los subconjuntos, como se observa en la figura 4. Esta técnica se usó con el objetivo de garantizar que el conjunto de datos empleado es homogéneo y que los resultados son independientes de los datos usados para el entrenamiento [22].

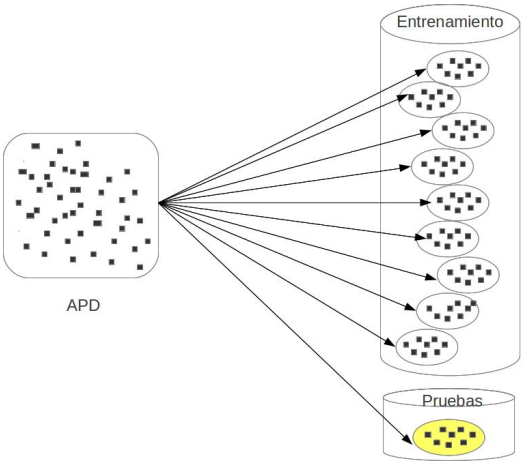


Figura 4: Validación cruzada.

3.3 Resultados obtenidos

Inicialmente, se hace la división del conjunto de datos, donde el 70 % de éstos se emplea para el entrenamiento y el 30 % para pruebas. En el caso de f_1, el conjunto de entrenamiento contiene 1602 péptidos y el de pruebas 686. A partir de dichos conjuntos, se entrena y evalúa la MSV. Por ejemplo, la función kernel lineal, para la cual se establecen los rangos para crear las mallas, en este caso, $C_{inicial} = 0{,}1$ y $C_{final} = 1$, el paso $\delta_C = 0{,}1$ y $l = 10$. Acto seguido, se elige la mejor función teniendo en cuenta que la Prec, Sens, Espec y MCC (Para observar todos los resultados, ver anexos cuadro 13, 14, 15 y 16 respectivamente) sean las mejores dentro del conjunto de máquinas creadas con la función kernel lineal. Para este caso, la mejor de éstas funciones tiene una precisión de 78.7 % para el parámetro libre $C = 0,9$.

	Lineal	*Cuadratica*	*Polinomial*	*FBR*
Prec	0,787	0,843	0,844	0,844
Sens	0,825	0,825	0,851	0,828
Espec	0,749	0,86	0,837	0,860
CCM	0,576	0,686	0,688	0,688
C	0,9	0,5	0,2	1
d	-	-	3	-
γ	-	-	-	1

Cuadro 6: Mejores resultados para cada una de las funciones kernel para f_1 junto con los parámetros libres

Para las funciones kernel cuadrática, polinomial y de base radial (FBR), el procedimiento anterior es el mismo, salvo que varían los parámetros libres y los rangos para evaluarlos. En el caso de la función cuadrática, la Prec, Sens, Espec y el CCM (Mayor información en anexos cuadro 17, 18, 15, 20) con mejores resultados, fue con el parámetro libre $C = 0,5$. Ahora bien, para la función polinomial (mayor información en anexos cuadro 21, 22, 23, 24), la mejor máquina fue aquella con los parámetros libres $C = 0,2$ y $d = 3$. Adicionalmente, con la FBR, la Prec, Sens, Espec y el CCM (mayor información en anexos cuadro 25, 26, 27, 28) y con $C = 1$ y $\gamma = 1$, ofreció los mejores resultados. A partir de lo anterior se compara-

ron los resultados entre las MSVs escogidas y se elegió la función kernel de base radial para f_1, puesto que es la que presenta el mejor rendimiento, como se puede observar en el cuadro 6.
Luego, se realizó validación cruzada con $k = 10$ para f_1, teniendo en cuenta la función FBR junto con los parámetros libres escogidos del paso anterior y se compruebó el rendimiento promedio. Los resultados se observan en anexos cuadro 45.
Es importante aclarar, que mientras mayor sea el rendimiento de f_1 mejor va a ser la predicción de los péptidos antimicrobianos, que luego serán las entradas de f_2.

Ahora bien, para f_2, el conjunto de datos contiene 372 péptidos, donde 260 individuos se usaron para el entrenamiento y 112 para pruebas. Con dicho conjunto, se lleva a cabo todo el procedimiento anterior, donde se mide la Prec, la Sens, la Espec y el CCM para la función lineal (los resultados se observan en anexos cuadro 29, 30, 31, 32), respectivamente), cuadrática (Ver anexos cuadro 33, 34, 35, 36), polinomial (mayor información cuadro 37, 38, 39, 40) y la función de base radial (Ver anexos cuadro 41, 42, 43, 44) y obteniendo como resultado, la elección de la FBR puesto que ofrece los mejores resultados (Ver tabla 7), al igual que f_1. Como se nombraba anteriormente, la importancia de escoger la función kernel para f_2 es establecer su potencial para predecir los péptidos antibacterinos.

	Lineal	*Cuadratica*	*Polinomial*	*FBR*
Prec	0,8	0,8	0,8	0,836
Sens	0,764	0,8	0,8	0,836
Espec	0,836	0,8	0,8	0,836
CCM	0,602	0,60	0,60	0,678
C	0,4	0,2	0,2	0,2
d	-	-	2	-
γ	-	-	-	1

Cuadro 7: Mejores resultados para cada una de las funciones kernel para f_2 junto con los parámetros libres

Adicionalmente, se realizó validación cruzada con $k = 10$ para la MSV elegida anteriormente junto con los parámetros libres y cuyos resultados se observan en anexos, cuadro 46. En el cuadro 8 y 9 se muestra el resumen de los promedios y las desviaciones estándar para las validaciones cruzadas realizadas para f_1 y f_2, respectivamente.

	Prec	*Sens*	*Espec*	*CCM*
f_1	0,855	0,823	0,89	0,711
f_2	0,807	0,844	0,770	0,623

Cuadro 8: Promedio validación cruzada con k=10, para las funciones kernel de base radial en f_1 y f_2

	Prec	*Sens*	*Espec*	*CCM*
f_1	$\pm$ 0,021	$\pm$ 0,025	$\pm$ 0,025	$\pm$ 0,041
f_2	$\pm$ 0,051	$\pm$ 0,089	$\pm$ 0,110	$\pm$ 0,113

Cuadro 9: Desviación estándar para la validación cruzada con k=10, para las funciones kernel de base radial en f_1 y f_2

También se calculó el rendimiento del clasificador f, de tal forma que, se toma el mismo conjunto de pruebas del segundo clasificador y se evalúa en f_1 y si los resultados arrojados son iguales a 1, indica que el péptido evaluado en cuestión es antimicrobiano y puede ser evaluado por f_2, quien decide si la instancia es o no un péptido antibacteriano y cuyos resultados se muestran en el cuadro 10.

	Prec	*Sens*	*Espec*	*CCM*
f	0,8	0,763	0,836	0,601

Cuadro 10: Resultado global para el clasificador f

3.4 Comparación de resultados

La comparación de resultados de este trabajo con otros, se hace de manera global, es decir, de acuerdo a lo reportado en la literatura, se establece

un rango de precisiones para determinar qué tan bueno es el trabajo desarrollado. De acuerdo a lo anterior, se estableció que la precisión oscila entre el 75 % y 94 %, donde se emplean máquinas de soporte (75 % [11] y 92.11 % [23] de precisión),redes neuronales (entre 80 % y 94 % de precisión, [4]; [23]; [24]; [11] ; [13]) y matrices cuantitativas (90.37 % [23]) y en este estudio es del 80 %, lo que indica que es un buen resultado. Sin embargo, la comparación no es del todo justa si se tiene en cuenta que el conjunto de datos, la cantidad y el tipo de descriptores son diferentes en todos los estudios.

Por otro lado, existen varios servidores web que además de almacenar un gran número de información referente a péptidos, ofrecen herramientas de predicción, como el caso de la APD, CAMP, AMPER, BACTIBASE, BAGEL y AntiBP. Sin embargo, la APD, CAMP, AMPER están enfocadas en la predicción de péptidos antimicrobianos, con la BACTIBASE y BAGEL, aunque se basan en los bacteriocins, no cuentan con una buena herramienta de predicción. Con la AntiBP, ésta permite clasificar péptidos antibacterianos pero no se especifica si tienen múltiples actividades o no. Por lo anterior, nuestra herramienta presenta una ventaja adicional, puesto que permite detectar péptidos antibacterianos y a su vez, se asegura de que éstos tengan actividad exclusiva.

4. Discusión

- En este trabajo no se consideró los descriptores de la estructura terciaria, debido a que esperaba hacer un modelo con pocos descriptores y determinar su efectividad. Sin embargo, en un trabajo futuro, que se desarrollará con la beca de jóvenes investigadores, se espera adicionar más descriptores, principalmente aquellos relacionados con la estructura terciaria de los péptidos y con ello, tratar de aumentar el rendimiento.
- Existen otros enfoques que se basan en búsqueda de homología, alineamiento múltiple de secuencias, filogenias, perfiles físicoquímicas, sin embargo, se empleó QSAR porque es uno de los enfoques más ampliamente usados en el diseño de medicamentos, basados en la estructura [25].

5. Conclusiones

-Para el clasificador de péptidos antimicrobianos, se puede afirmar que la MSV creada está realizando una clasificación adecuada, puesto que la precisión mostrada es del 84.4 %, lo cual demuestra que es un valor lo suficientemente alto para asegurar la veracidad de un péptido con y sin actividad antimicrobiana.

-Mediante el ajuste de los parámetros libres de las MSVs, se pudo inferir que la función kernel con mejores resultados en el proceso de clasificación, es la función de base radial, puesto la precisión reportada para ambos clasificadores está por encima del 80 %, con los parámetros $C = 1$, $\gamma = 1$ para f_1 y $C = 0,2$, $\gamma = 1$ para f_2 y cuyos resultados se pueden observan en los cuadros 25 y 41.

-Bajo el supuesto de que diez descriptores eran suficientes para clasificar entre péptidos con y sin actividad antibacteriana, se pudo evidenciar que, éstos permiten realizar una buena clasificación entre éstos péptidos, teniendo en cuenta el esquema de clasificador en cascada. Esta afirmación se da, con base en los resultados obtenidos por el clasificador implementado, donde la precisión alcanzada fue del 80 % .

Bibliografía

[1] Gustavo Betancour. Las máquinas de soporte vectorial (SVMs). *Scientia Et Technica*, XI:67–72, 2005.

[2] George Gilbert, David, Guidos Robert, Boucher Helen, Talbot. The 10 x '20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 50(8):1081–3, April 2010.

[3] Alexandra K Marr, William J Gooderham, and Robert Ew Hancock. Antibacterial peptides for therapeutic use: obstacles and realistic outlook. *Current opinion in pharmacology*, 6(5):468–72, October 2006.

[4] A. Cherkasov. 'Inductive' Descriptors: 10 Successful Years in QSAR. *Current Computer - Aided Drug Design*, 1(1):21–42, January 2005.

[5] Arkadiusz Z Dudek, Tomasz Arodz, and Jorge Gálvez. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial chemistry & high throughput screening*, 9(3):213–28, March 2006.

[6] Olivier Taboureau. Methods for Building Quantitative Structure-Activity Relationship (QSAR) Descriptors and Predictive Models for Computer-Aided Design of Antimicrobial Peptides. In Andrea Giuliani and Andrea C. Rinaldi, editors, *Antimicrobial Peptides, Methods in Molecular Biology*, volume 618 of *Methods in Molecular Biology*, pages 77–86. Humana Press, Totowa, NJ, 2010.

[7] Christopher D Fjell, Hå vard Jenssen, Kai Hilpert, Warren a Cheung, Nelly Panté, Robert E W Hancock, and Artem Cherkasov. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *Journal of medicinal chemistry*, 52(7):2006–15, April 2009.

[8] Hå vard Jenssen, Tore Lejon, Kai Hilpert, Christopher D Fjell, Artem Cherkasov, and Robert E W Hancock. Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward P. aeruginosa. *Chemical biology & drug design*, 70(2):134–42, August 2007.

[9] Kai Hilpert, Christopher D Fjell, and Artem Cherkasov. *Short Linear Cationic Antimicrobial Peptides: Screening, Optiming, and Prediction*, volume 494 of *Methods in Molecular Biology*. Humana Press, Totowa, NJ, 2008.

[10] Shaini Thomas, Shreyas Karnik, Ram Shankar Barai, V K Jayaraman, and Susan Idicula-Thomas. CAMP: a useful resource for research on antimicrobial peptides. *Nucleic acids research*, 38(Database issue):D774–80, January 2010.

[11] Marc Torrent, David Andreu, Victòria M Nogués, and Ester Boix. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PloS one*, 6(2):e16968, January 2011.

[12] Christopher D. Fjell, Robert E.W. Hancock, and Havard Jenssen. Computer-Aided Design of Antimicrobial Peptides. *Current Pharmaceutical Analysis*, 6(2):66–75, May 2010.

[13] Christopher D Fjell, Hå vard Jenssen, Warren a Cheung, Robert E W Hancock, and Artem Cherkasov. Optimization of antibacterial peptides by genetic algorithms and cheminformatics. *Chemical biology & drug design*, 77(1):48–56, January 2011.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[15] Zhe Wang and Guangshun Wang. APD: the Antimicrobial Peptide Database. *Nucleic acids research*, 32(Database issue):D590–2, January 2004.

[16] Ping Wang, Lele Hu, Guiyou Liu, Nan Jiang, Xiaoyun Chen, Jianyong Xu, Wen Zheng, Li Li, Ming Tan, Zugen Chen, Hui Song, Yu-Dong Cai, and Kuo-Chen Chou. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PloS one*, 6(4):e18476, January 2011.

[17] Ana-Maria Fernandez-Escamilla, Frederic Rousseau, Joost Schymkowitz, and Luis Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10):1302–6, October 2004.

[18] Oscar Conchillo-Solé, Natalia S de Groot, Francesc X Avilés, Josep Vendrell, Xavier Daura, and Salvador Ventura. AGGRESCAN: a server for the prediction and evaluation of ”hot spots.of aggregation in polypeptides. *BMC bioinformatics*, 8(i):65, January 2007.

[19] Carlos Polanco González, Marco Aurelio Nuño Maganda, Miguel Arias-Estrada, and Gabriel Del Rio. An FPGA Implementation to Detect Selective Cationic Antibacterial Peptides. *PLoS ONE*, 6(6):e21399, 2011.

[20] J Kyte and R F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–32, January 1982.

[21] Chih-wei Hsu, Chih-chung Chang, and Chih-jen Lin. A Practical Guide to Support Vector Classification. (1):1–16, 2010.

[22] John Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Nueva York, 2004.

[23] Sneh Lata, B K Sharma, and G P S Raghava. Analysis and prediction of antibacterial peptides. *BMC bioinformatics*, 8:263, January 2007.

[24] Artem Cherkasov, Kai Hilpert, Hå vard Jenssen, Christopher D Fjell, Matt Waldbrook, Sarah C Mullaly, Rudolf Volkmer, and Robert E W Hancock. Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS chemical biology*, 4(1):65–74, January 2009.

[25] Riadh Hammami and Ismail Fliss. Current trends in antimicrobial agent research: chemo- and bioinformatics approaches. *Drug discovery today*, 15(13-14):540–6, July 2010.

Anexos

Para este trabajo, se creó una herramienta software (ver figura 5) que clasifica entre péptidos con y sin actividad antibacteriana y que le permite al usuario, a través de una interfaz sencilla, interactuar con las MSVs creadas.

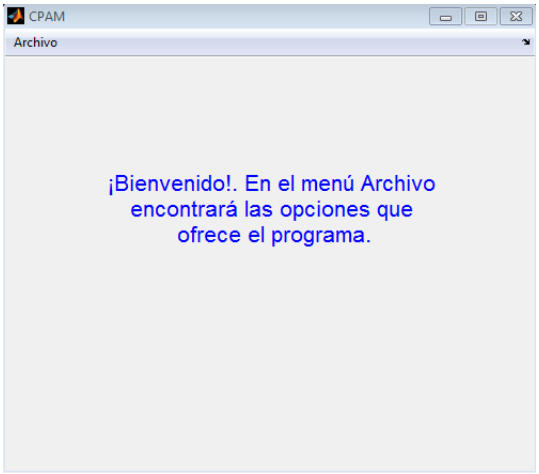


Figura 5: Herramienta software.
abla

A. Manual de usuario

Inicialmente, la herramienta software muestra al usuario un mensaje de bienvenida junto con las opciones que posee el programa, las cuales se encuentran almacenadas en el menú Archivo, en la que se despliegan dos opciones para evaluar las secuencias de aminoácidos de los péptidos.

//Formato para ingresar las secuencias.					
//Secuencia	//In Vivo	//AGG	//TURN	//HELIX	//BETA
LMCTHPLDCSN	-15.6	0	4.803	0	27.211
LMCTHPLDCSN	-15.6	0	4.803	0	27.211

Cuadro 11: Formato para ingresar los datos

La pantalla principal de la herramienta, da la bienvenida al usuario y le muestra el menú Archivo que contiene las opciones que ofrece el programa. La primera opción es **Ingresar manualmente**, donde el usuario debe ingresar una secuencia y los descriptores *in vivo*, AGG, turn, helix, beta correspondientes (ver imagen 6) a ser evaluados. Para calcular el parámetro Agregación In vivo (In Vivo) se utiliza AGGRESCAN [6], donde se ingresa la secuencia en formato fasta y de los resultados arrojados por el software se extrae el valor señalado por la etiqueta **Normalized a4v Sequence Sum for 100 residues (Na4vSS)**. Para los parámetros numéricos restantes, se usa TANGO [7] en donde se ingresa solamente la secuencia y en los parámetros y condiciones que pide el software, se toman los valores por defecto. Ver cuadro 12. En la misma pantalla, aparece el botón **Calcular actividad** que determina si el péptido evaluado es antimicrobiano o no, y si lo es, determina si también es antibacteriano. Finalmente, en el botón **Reiniciar**, permite borrar los cálculos anteriores y realizar otros nuevamente.

Nterm protected	Cterm protected	pH	Temperature	Ionic strength
No protected	No protected	7	298,15	0,02

Cuadro 12: Parámetros y condiciones por defecto de Tango

Como segunda opción, aparece **Cargar archivo** (Ver figura 7), que permite ingresar varias secuencias al tiempo, siguiendo el cuadro 11. El cálculo de los parámetros se hace de la misma forma que la opción anterior y los botones presentan la misma funcionalidad que la opción Ingresar manualmente. Cabe destacar, que el usuario debe asegurarse de que los

[6] http://bioinf.uab.es/aggrescan/
[7] http://tango.crg.es/

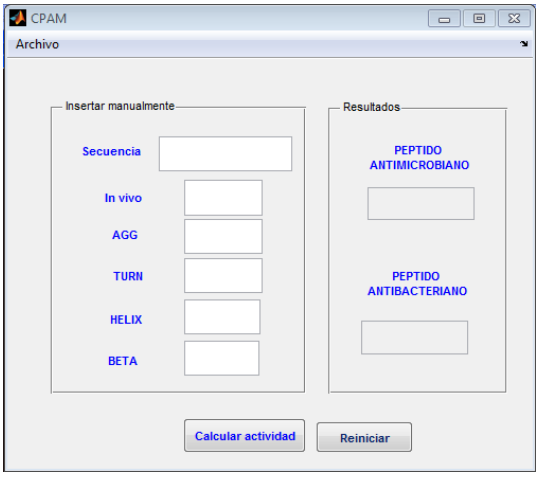


Figura 6: Herramienta software. Opción Ingresar manualmente.

valores estén en las columnas apropiadas para que el software realice el cálculo, de lo contrario mostrará un mensaje de error.

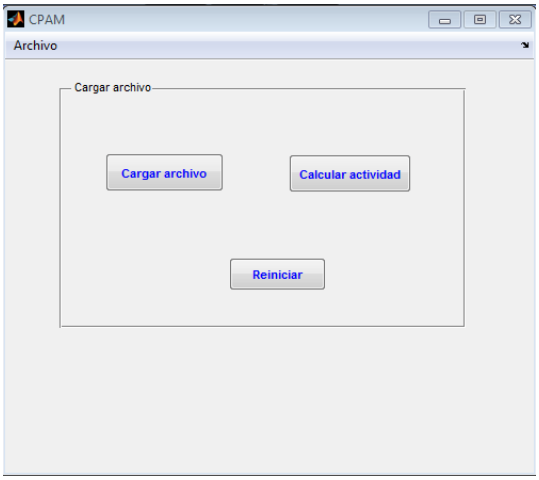


Figura 7: Herramienta software. Opción Cargar archivo

B. Diagrama de casos de uso

En la imagen 8 se muestra el diagrama de caso de uso que describe a la herramienta.

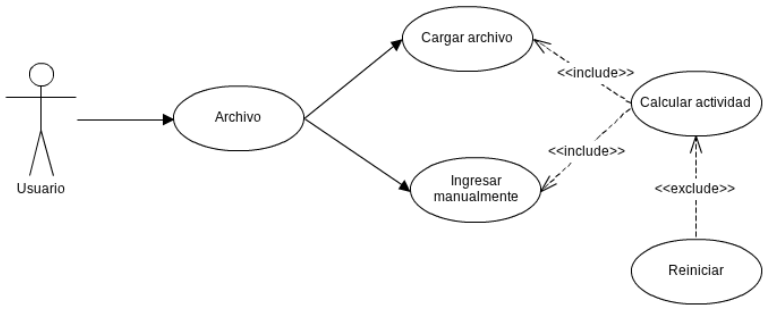


Figura 8: Diagrama de caso de uso de la herramienta software.

C. Especificaciones del caso de uso

ACTOR	
Actor	Usuario
Casos de uso	Archivo, ingresar manualmente, cargar archivo,
-	calcular actividad, reiniciar
Tipo	Iniciador
Descripción	Es el actor que quiere conocer el tipo de actividad
-	de uno o varios péptidos

D. Casos de uso

Caso de uso	Archivo	
Actor	Usuario	
Propósito	Muestra las opciones disponibles en la herramienta	
Descripción	El usuario elige la opción que desea	
Precondición	Ninguna	
Flujo principal	Acciones del actor	Respuesta del sistema
-	1. El usuario elige la	-
-	opción que desea	-
-	-	2. Permite realizar el siguiente
-	-	caso de uso
Subflujos	Ninguno	
Poscondición	Despliega las opción Ingresar manualmente	
-	y cargar archivo	

Caso de uso	Cargar archivo	
Actor	Usuario	
Propósito	Permitir cargar múltiples secuencias de aminoácidos	
Descripción	El usuario ingresa las secuencias a evaluar	
Precondición	Los datos deben tener el formato especificado	
Flujo principal	Acciones del actor	Respuesta del sistema
-	1. El usuario ingresa las	-
-	secuencias en el formato establecido	-
-	-	2. Permite realizar el siguiente caso
-	-	de uso
Sub-flujos	Punto 1: si el archivo no tiene el formato correcto, el sistema	
-	muestra mensaje de error	
Poscondición	El botón Calcular se habilita	

Caso de uso	Ingresar manualmente	
Actor	Usuario	
Propósito	Ingresar una sola secuencia de aminoácidos	
Descripción	El usuario ingresa la secuencia junto con los descriptores especificados	
Precondición	-	
Flujo principal	Acciones del Actor	Respuesta del sistema
-	1. El usuario ingresa los datos	-
-	solicitados	-
-	-	2. Habilita el siguiente caso de uso
Sub-flujos	Punto 1: si los datos se ingresan de forma	
-	incorrecta, el sistema muestra un mensaje de error	
Poscondición	El usuario ya tiene habilitada el siguiente caso de uso	

Caso de uso	Calcular actividad	
Actor	Usuario	
Propósito	Determinar el tipo de actividad de los péptidos evaluados	
Descripción	Carga las MSV y evalúa los datos	
Precondición	Haber ejecutado el caso de uso Cargar archivo o el de Ingresar manualmente	
Flujo principal	Acciones del Actor	Respuesta del sistema
-	1. El usuario oprime el botón Calcular	-
-	actividad	-
-	-	2. Muestra el tipo de actividad
Sub-flujos	-	
Poscondición	El sistema muestra el tipo de actividad al usuario	

Caso de uso	Reiniciar	
Actor	Usuario	
Propósito	Borrar los valores que se muestran en pantalla	
Descripción	Coloca las variables en cero y borra información en pantalla	
Precondición	-	
Flujo principal	Acciones del Actor	Respuesta del sistema
-	1. El usuario oprime el botón Reiniciar	-
-	-	2. Borra información en pantalla
Sub-flujos	-	
Poscondición	El usuario ya puede ingresar datos nuevamente para ser evaluados	

E. Tablas de ajuste de parámetros libres en funciones kernel

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Prec	0,773	0,780	0,784	0,784	0,786	0,786	0,786	0,786	**0,787**	0,787

Cuadro 13: Malla de precisión para la función lineal en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Sens	0,813	0,816	0,822	0,822	0,825	0,825	0,825	0,825	**0,825**	0,825

Cuadro 14: Malla de sensibilidad para la función lineal en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Espec	0,732	0,743	0,746	0,746	0,746	0,746	0,746	0,746	**0,749**	0,749

Cuadro 15: Malla de especificidad para la función lineal en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
CCM	0,547	0,561	0,570	0,570	0,573	0,573	0,573	0,573	**0,576**	0,576

Cuadro 16: Malla de CCM para la función lineal en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Prec	0,831	0,837	0,837	0,841	**0,843**	0,840	0,840	0,840	0,837	0,837

Cuadro 17: Malla de precisión para la función cuadrática en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Sens	0,825	0,822	0,822	0,825	**0,825**	0,822	0,822	0,822	0,822	0,822

Cuadro 18: Malla de sensibilidad para la función cuadrática en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Espec	0,837	0,851	0,851	0,857	**0,860**	0,857	0,857	0,857	0,851	0,851

Cuadro 19: Malla de especificidad para la función cuadrática en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
CCM	0,662	0,674	0,674	0,683	**0,686**	0,680	0,680	0,680	0,674	0,674

Cuadro 20: Malla de CCM para la función cuadrática en f_1

C / d	1	2	3	4	5
0,1	0,773	0,831	0,844	0,843	0,815
0,2	0,780	0,837	**0,844**	0,829	0,818
0,3	0,784	0,837	0,844	0,827	0,815
0,4	0,784	0,841	0,843	0,821	0,810
0,5	0,786	0,843	0,844	0,821	0,808
0,6	0,786	0,840	0,841	0,818	0,803
0,7	0,786	0,840	0,840	0,821	0,808
0,8	0,786	0,840	0,838	0,818	0,805
0,9	0,787	0,837	0,838	0,815	0,803
1	0,787	0,837	0,837	0,812	0,800

Cuadro 21: Malla de precisión para la función polinomial en f_1

C / d	1	2	3	4	5
0,1	0,813	0,825	0,851	0,851	0,845
0,2	0,816	0,822	**0,851**	0,837	0,843
0,3	0,822	0,822	0,851	0,837	0,845
0,4	0,822	0,825	0,848	0,831	0,843
0,5	0,825	0,825	0,854	0,834	0,840
0,6	0,825	0,822	0,848	0,831	0,834
0,7	0,825	0,822	0,845	0,834	0,837
0,8	0,825	0,822	0,845	0,834	0,834
0,9	0,825	0,822	0,845	0,831	0,834
1	0,825	0,822	0,840	0,828	0,831

Cuadro 22: Malla de sensibilidad para la función polinomial en f_1

C / d	1	2	3	4	5
0,1	0,732	0,837	0,837	0,834	0,784
0,2	0,743	0,851	**0,837**	0,822	0,793
0,3	0,746	0,851	0,837	0,816	0,784
0,4	0,746	0,857	0,837	0,810	0,778
0,5	0,746	0,860	0,834	0,808	0,776
0,6	0,746	0,857	0,834	0,805	0,773
0,7	0,746	0,857	0,834	0,808	0,778
0,8	0,746	0,857	0,831	0,802	0,776
0,9	0,749	0,851	0,831	0,799	0,773
1	0,749	0,851	0,834	0,796	0,770

Cuadro 23: Malla de especificidad para la función polinomial en f_1

C / d	1	2	3	4	5
0,1	0,547	0,662	0,688	0,685	0,631
0,2	0,561	0,674	**0,688**	0,659	0,636
0,3	0,570	0,674	0,688	0,653	0,631
0,4	0,570	0,683	0,685	0,642	0,622
0,5	0,573	0,686	0,688	0,642	0,616
0,6	0,573	0,680	0,682	0,636	0,608
0,7	0,573	0,680	0,679	0,642	0,616
0,8	0,573	0,680	0,676	0,636	0,610
0,9	0,576	0,674	0,676	0,630	0,608
1	0,576	0,674	0,673	0,624	0,602

Cuadro 24: Malla de CCM para la función polinomial en f_1

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0,583	0,663	0,729	0,754	0,773	0,803	0,812	0,813	0,810	0,806
0,2	0,586	0,671	0,738	0,758	0,787	0,810	0,821	0,822	0,818	0,813
0,3	0,587	0,675	0,738	0,770	0,792	0,822	0,828	0,827	0,825	0,827
0,4	0,589	0,675	0,736	0,773	0,796	0,822	0,827	0,827	0,829	0,829
0,5	0,590	0,676	0,738	0,777	0,797	0,824	0,824	0,825	0,829	0,832
0,6	0,590	0,678	0,738	0,776	0,802	0,822	0,828	0,827	0,831	0,834
0,7	0,590	0,679	0,739	0,778	0,800	0,821	0,831	0,825	0,829	0,840
0,8	0,590	0,681	0,739	0,777	0,803	0,827	0,829	0,822	0,831	0,841
0,9	0,592	0,684	0,738	0,778	0,806	0,825	0,827	0,824	0,829	0,841
1	0,592	0,685	0,738	0,781	0,808	0,825	0,828	0,822	0,829	**0,844**

Cuadro 25: Malla de precisión para la función de base radial en f_1

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0,166	0,335	0,487	0,577	0,636	0,706	0,735	0,749	0,758	0,773
0,2	0,172	0,350	0,507	0,592	0,662	0,717	0,752	0,767	0,778	0,784
0,3	0,175	0,359	0,510	0,609	0,671	0,738	0,764	0,773	0,787	0,796
0,4	0,178	0,359	0,513	0,615	0,676	0,741	0,764	0,781	0,793	0,799
0,5	0,181	0,362	0,516	0,624	0,682	0,741	0,761	0,781	0,790	0,802
0,6	0,181	0,364	0,516	0,624	0,691	0,743	0,767	0,781	0,793	0,805
0,7	0,181	0,367	0,519	0,630	0,688	0,741	0,773	0,781	0,790	0,813
0,8	0,181	0,370	0,519	0,630	0,694	0,752	0,773	0,778	0,793	0,819
0,9	0,184	0,376	0,519	0,630	0,697	0,749	0,770	0,781	0,796	0,822
1	0,184	0,379	0,519	0,636	0,700	0,746	0,773	0,781	0,796	**0,828**

Cuadro 26: Malla de sensibilidad para la función de base radial en f_1

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	1,000	0,991	0,971	0,930	0,910	0,901	0,889	0,878	0,863	0,840
0,2	1,000	0,991	0,968	0,924	0,913	0,904	0,889	0,878	0,857	0,843
0,3	1,000	0,991	0,965	0,930	0,913	0,907	0,892	0,880	0,863	0,857
0,4	1,000	0,991	0,959	0,930	0,915	0,904	0,889	0,872	0,866	0,860
0,5	1,000	0,991	0,959	0,930	0,913	0,907	0,886	0,869	0,869	0,863
0,6	1,000	0,991	0,959	0,927	0,913	0,901	0,889	0,872	0,869	0,863
0,7	1,000	0,991	0,959	0,927	0,913	0,901	0,889	0,869	0,869	0,866
0,8	1,000	0,991	0,959	0,924	0,913	0,901	0,886	0,866	0,869	0,863
0,9	1,000	0,991	0,956	0,927	0,915	0,901	0,883	0,866	0,863	0,860
1	1,000	0,991	0,956	0,927	0,915	0,904	0,883	0,863	0,863	**0,860**

Cuadro 27: Malla de especificidad para la función de base radial en f_1

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0,301	0,433	0,523	0,542	0,567	0,618	0,631	0,632	0,624	0,614
0,2	0,307	0,445	0,535	0,547	0,593	0,632	0,648	0,648	0,638	0,628
0,3	0,310	0,452	0,534	0,569	0,601	0,654	0,661	0,657	0,652	0,654
0,4	0,312	0,452	0,528	0,574	0,610	0,653	0,658	0,656	0,661	0,660
0,5	0,315	0,454	0,530	0,582	0,611	0,656	0,652	0,653	0,661	0,666
0,6	0,315	0,456	0,530	0,578	0,619	0,652	0,661	0,656	0,664	0,669
0,7	0,315	0,459	0,533	0,583	0,616	0,650	0,666	0,653	0,661	0,680
0,8	0,315	0,461	0,533	0,580	0,621	0,660	0,663	0,647	0,664	0,683
0,9	0,318	0,466	0,528	0,583	0,627	0,658	0,657	0,650	0,660	0,683
1	0,318	0,468	0,528	0,588	0,630	0,658	0,660	0,646	0,660	**0,688**

Cuadro 28: Malla de CCM para la función de base radial en f_1

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Prec	0,773	0,773	0,791	**0,800**	0,800	0,800	0,800	0,800	0,800	0,800

Cuadro 29: Malla de precisión para la función lineal en f_2

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Sens	0,745	0,745	0,764	**0,764**	0,764	0,764	0,764	0,764	0,764	0,764

Cuadro 30: Malla de sensibilidad para la función lineal en f_2

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Espec	0,800	0,800	0,818	**0,836**	0,836	0,836	0,836	0,836	0,836	0,836

Cuadro 31: Malla de especificidad para la función lineal en f_2

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
CCM	0,546	0,546	0,583	**0,602**	0,602	0,602	0,602	0,602	0,602	0,602

Cuadro 32: Malla de CCM para la función lineal en f_2

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Prec	0,791	**0,800**	0,791	0,791	0,800	0,791	0,791	0,791	0,791	0,791

Cuadro 33: Malla de precisión para la función cuadrática en f_2

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Sens	0,800	**0,800**	0,782	0,782	0,782	0,782	0,782	0,782	0,782	0,782

Cuadro 34: Malla de sensibilidad para la función cuadrática en f_2

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Espec	0,782	**0,800**	0,800	0,800	0,818	0,800	0,800	0,800	0,800	0,800

Cuadro 35: Malla de especificidad para la función cuadrática en f_2

C	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
CCM	0,582	**0,600**	0,582	0,582	0,600	0,582	0,582	0,582	0,582	0,582

Cuadro 36: Malla del CCM para la función cuadrática en f_2

C / d	1	2	3	4	5
0,1	0,773	0,791	0,773	0,745	0,709
0,2	0,773	**0,800**	0,782	0,745	0,709
0,3	0,791	0,791	0,791	0,764	0,718
0,4	0,800	0,791	0,791	0,755	0,718
0,5	0,800	0,800	0,782	0,736	0,718
0,6	0,800	0,791	0,782	0,736	0,718
0,7	0,800	0,791	0,782	0,745	0,718
0,8	0,800	0,791	0,773	0,745	0,718
0,9	0,800	0,791	0,773	0,745	0,727
1	0,800	0,791	0,764	0,745	0,727

Cuadro 37: Malla de precisión para la función polinomial en f_2

C / d	1	2	3	4	5
0,1	0,745	0,800	0,764	0,727	0,655
0,2	0,745	**0,800**	0,745	0,691	0,655
0,3	0,764	0,782	0,782	0,709	0,655
0,4	0,764	0,782	0,782	0,709	0,655
0,5	0,764	0,782	0,764	0,673	0,655
0,6	0,764	0,782	0,764	0,673	0,655
0,7	0,764	0,782	0,764	0,691	0,655
0,8	0,764	0,782	0,745	0,691	0,655
0,9	0,764	0,782	0,745	0,691	0,655
1	0,764	0,782	0,727	0,691	0,655

Cuadro 38: Malla de sensibilidad para la función polinomial en f_2V

C / d	1	2	3	4	5
0,1	0,800	0,782	0,782	0,764	0,764
0,2	0,800	**0,800**	0,818	0,800	0,764
0,3	0,818	0,800	0,800	0,818	0,782
0,4	0,836	0,800	0,800	0,800	0,782
0,5	0,836	0,818	0,800	0,800	0,782
0,6	0,836	0,800	0,800	0,800	0,782
0,7	0,836	0,800	0,800	0,800	0,782
0,8	0,836	0,800	0,800	0,800	0,782
0,9	0,836	0,800	0,800	0,800	0,800
1	0,836	0,800	0,800	0,800	0,800

Cuadro 39: Malla de especificidad para la función polinomial en f_2

C / d	1	2	3	4	5
0,1	0,546	0,582	0,546	0,491	0,421
0,2	0,546	**0,600**	0,565	0,494	0,421
0,3	0,583	0,582	0,582	0,530	0,440
0,4	0,602	0,582	0,582	0,511	0,440
0,5	0,602	0,600	0,564	0,477	0,440
0,6	0,602	0,582	0,564	0,477	0,440
0,7	0,602	0,582	0,564	0,494	0,440
0,8	0,602	0,582	0,546	0,494	0,440
0,9	0,602	0,582	0,546	0,494	0,459
1	0,602	0,582	0,529	0,494	0,459

Cuadro 40: Malla del CCM para la función polinomial en f_2

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0,582	0,627	0,691	0,709	0,727	0,736	0,764	0,782	0,809	0,818
0,2	0,582	0,627	0,691	0,718	0,727	0,755	0,773	0,782	0,827	**0,836**
0,3	0,582	0,627	0,691	0,718	0,736	0,764	0,773	0,791	0,800	0,818
0,4	0,582	0,636	0,700	0,718	0,736	0,773	0,773	0,782	0,800	0,818
0,5	0,582	0,636	0,700	0,727	0,745	0,782	0,782	0,782	0,791	0,809
0,6	0,582	0,627	0,700	0,718	0,745	0,791	0,791	0,791	0,791	0,791
0,7	0,582	0,627	0,700	0,727	0,745	0,791	0,800	0,800	0,791	0,791
0,8	0,582	0,627	0,691	0,727	0,745	0,782	0,800	0,809	0,800	0,791
0,9	0,582	0,627	0,691	0,736	0,745	0,791	0,800	0,809	0,800	0,791
1	0,582	0,627	0,691	0,745	0,745	0,791	0,791	0,818	0,800	0,800

Cuadro 41: Malla de precisión para la función de base radial en f_2

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	1,000	1,000	0,982	0,927	0,891	0,891	0,855	0,855	0,818	0,818
0,2	1,000	1,000	0,964	0,927	0,891	0,873	0,855	0,855	0,855	**0,836**
0,3	1,000	1,000	0,964	0,927	0,891	0,873	0,855	0,855	0,836	0,836
0,4	1,000	1,000	0,964	0,927	0,873	0,873	0,855	0,836	0,836	0,836
0,5	1,000	1,000	0,964	0,927	0,873	0,873	0,855	0,836	0,818	0,836
0,6	1,000	0,982	0,964	0,909	0,873	0,873	0,855	0,836	0,818	0,818
0,7	1,000	0,982	0,964	0,909	0,873	0,873	0,873	0,836	0,818	0,818
0,8	1,000	0,982	0,945	0,909	0,873	0,873	0,873	0,836	0,818	0,818
0,9	1,000	0,982	0,945	0,909	0,873	0,873	0,873	0,836	0,818	0,818
1	1,000	0,982	0,945	0,909	0,873	0,873	0,855	0,855	0,818	0,818

Cuadro 42: Malla de sensibilidad para la función de base radial en f_2

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0,164	0,255	0,400	0,491	0,564	0,582	0,673	0,709	0,800	0,818
0,2	0,164	0,255	0,418	0,509	0,564	0,636	0,691	0,709	0,800	**0,836**
0,3	0,164	0,255	0,418	0,509	0,582	0,655	0,691	0,727	0,764	0,800
0,4	0,164	0,273	0,436	0,509	0,600	0,673	0,691	0,727	0,764	0,800
0,5	0,164	0,273	0,436	0,527	0,618	0,691	0,709	0,727	0,764	0,782
0,6	0,164	0,273	0,436	0,527	0,618	0,709	0,727	0,745	0,764	0,764
0,7	0,164	0,273	0,436	0,545	0,618	0,709	0,727	0,764	0,764	0,764
0,8	0,164	0,273	0,436	0,545	0,618	0,691	0,727	0,782	0,782	0,764
0,9	0,164	0,273	0,436	0,564	0,618	0,709	0,727	0,782	0,782	0,764
1	0,164	0,273	0,436	0,582	0,618	0,709	0,727	0,782	0,782	0,782

Cuadro 43: Malla de especificidad para la función base radial en f_2

C / γ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,1	0,299	0,382	0,469	0,465	0,481	0,497	0,536	0,570	0,618	0,636
0,2	0,299	0,382	0,456	0,480	0,481	0,524	0,553	0,570	0,656	**0,673**
0,3	0,299	0,382	0,456	0,480	0,497	0,540	0,553	0,587	0,602	0,637
0,4	0,299	0,397	0,471	0,480	0,491	0,557	0,553	0,567	0,602	0,637
0,5	0,299	0,397	0,471	0,496	0,508	0,573	0,570	0,567	0,583	0,619
0,6	0,299	0,361	0,471	0,472	0,508	0,590	0,587	0,584	0,583	0,583
0,7	0,299	0,361	0,471	0,488	0,508	0,590	0,606	0,602	0,583	0,583
0,8	0,299	0,361	0,444	0,488	0,508	0,573	0,606	0,619	0,600	0,583
0,9	0,299	0,361	0,444	0,504	0,508	0,590	0,606	0,619	0,600	0,583
1	0,299	0,361	0,444	0,520	0,508	0,590	0,587	0,638	0,600	0,600

Cuadro 44: Malla del CCM para la función de base radial en f_2

-	**Prec**	**Sens**	**Espec**	**CCM**
Prueba 1	0,886	0,851	0,922	0,775
Prueba 2	0,860	0,825	0,895	0,721
Prueba 3	0,842	0,825	0,860	0,685
Prueba 4	0,843	0,798	0,887	0,688
Prueba 5	0,873	0,868	0,877	0,746
Prueba 6	0,852	0,800	0,904	0,707
Prueba 7	0,856	0,809	0,904	0,715
Prueba 8	0,817	0,798	0,835	0,634
Prueba 9	0,878	0,852	0,904	0,758
Prueba 10	0,843	0,809	0,877	0,687
Promedio	0,855	0,823	0,886	0,711
Desv. Estándar	±0,021	±0,025	±0,025	±0,041

Cuadro 45: Validación cruzada con $k = 10$ para f_1 con la función de base radial

-	Prec	Sens	Espec	CCM
Prueba 1	0,892	0,895	0,889	0,784
Prueba 2	0,711	0,842	0,579	0,436
Prueba 3	0,838	0,833	0,842	0,675
Prueba 4	0,778	0,722	0,833	0,559
Prueba 5	0,763	0,684	0,842	0,533
Prueba 6	0,816	0,947	0,684	0,655
Prueba 7	0,838	0,947	0,722	0,690
Prueba 8	0,889	0,889	0,889	0,778
Prueba 9	0,757	0,889	0,632	0,536
Prueba 10	0,789	0,789	0,789	0,579
Promedio	0,807	0,844	0,770	0,623
Desv. Estándar	±0,059	±0,089	±0,110	±0,113

Cuadro 46: Validación cruzada con $k = 10$ para la f_2 con la función de base radial

F. Proceso de extracción de secuencias de la APD

La APD está conformada por varios grupos de péptidos con diversas actividades, como se puede observar en la imagen 9.

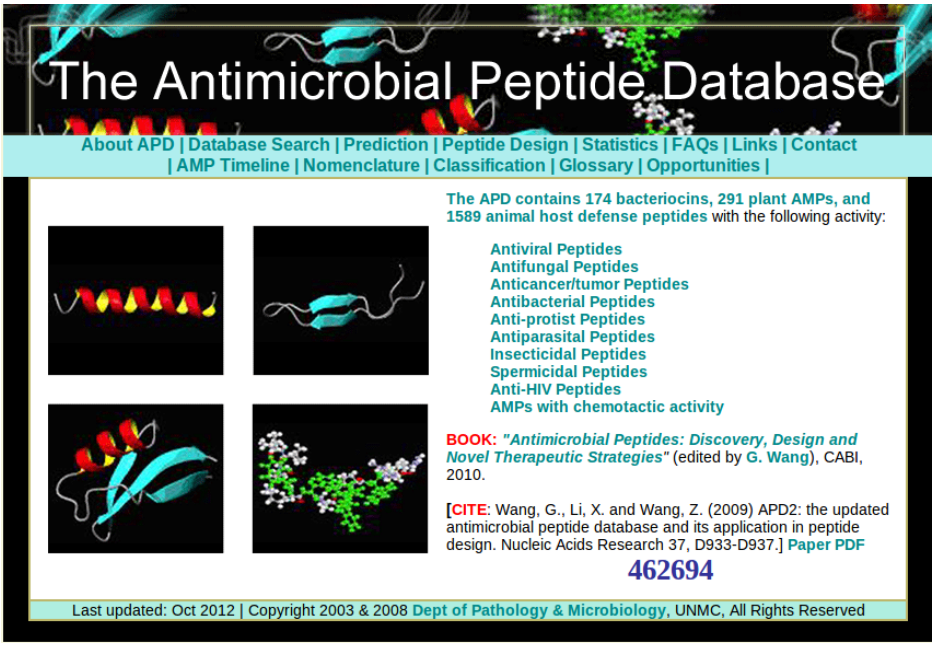


Figura 9: Página principal de la APD.

A continuación, se hace una breve descripión de la navegación por la página web. En este ejemplo, se ingresa al grupo 'Antiviral Peptides', y muestra varias secuencias que presentan dicha actividad agrupadas por páginas y organizadas por un identificador: AP0XXXX y junto a éste se muestra una breve información sobre el péptido. Ver imagen 10. En la barra de direcciones se observa http://aps.unmc.edu/AP/database/antiF.php?page=1, y como son varias páginas, por ejemplo, en este caso son 46 (determinado de forma manual), lo único que varía es el número presente en la dirección web. Esta información es importante, para formar los archivos que contendrán las instrucciones para ser descargadas, pero esto se explicará más adelante.

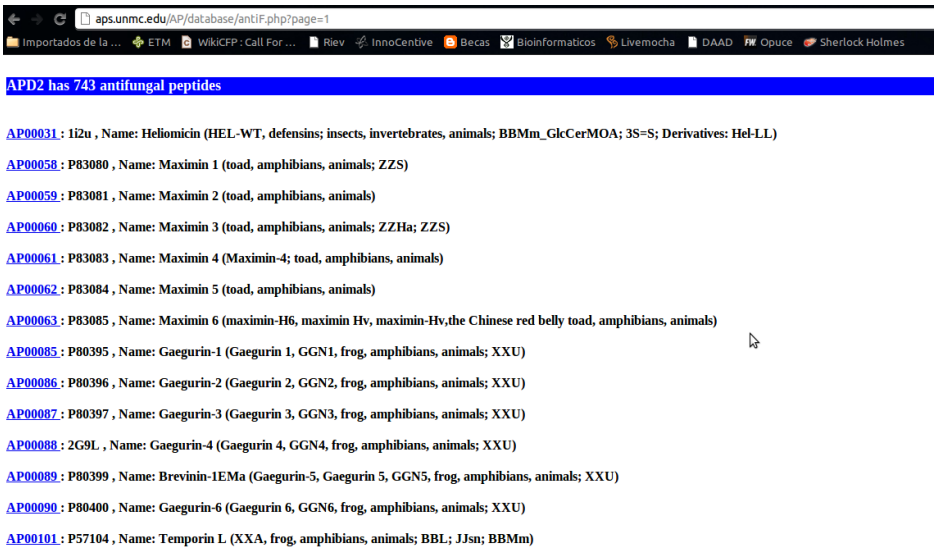


Figura 10: Página para el grupo de péptidos antifúngicos.

Cuando se escoge una secuencia, por ejemplo AP000031, ésta redirecciona a una nueva página web que contiene una descripción detallada del péptido. Ver imagen 11. Al igual que en la páginas por grupos, cada página contiene varios enlaces a los péptidos y tiene el formato http://aps.unmc.edu/AP/database/query_output.php?ID=00031. Como en la dirección web, lo único que cambia es el identificador, ésto será importante para poder realizar la descarga del archivo y posterior selección de la secuencia. Se explicará más adelante.

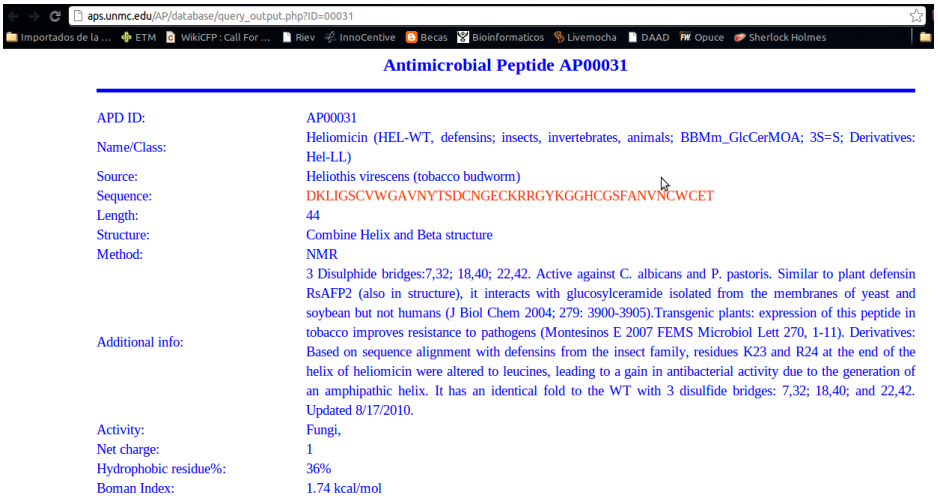


Figura 11: Información del péptido antiviral AP00031.

El procedimiento para la extracción de las secuencias, utilizando los scripts que se encuentran en la carpeta DATABASE APD, es el siguiente:

1. Ejecutar el archivo 'urls.py' en la consola de Ubuntu, escribiendo 'python urls.py'. El resultado son 16 archivos, 8 con nombre de AntiX.sh y 8 denominados NumberPageX, donde la X, es la actividad del péptido (V=antiViral, B=antiBacteriano, F=antiFúngico, H=antiVIH, C=antiCancer, I=Insecticida, S=eSpermicida, P=antiParasital). También se crean 8 carpetas con los nombres de AntiX.

2. Mover los archivos AntiX.sh y NumberPageX a las carpetas AntiX

3. Ubicado en la carpeta AntiX, ejecutar en la consola de Ubuntu los archivos AntiX.sh, que descarga las páginas web que contiene los identificadores de los péptidos.

4. En cada carpeta, existe un archivo con nombre 'DownloadIDsX.py', que se debe ejecutar en la consola de Ubuntu. Este archivo genera 2 archivos, uno se denomina IDsX.sh y el otro, onlyIDsX.

5. Ejecutar en consola el archivo IDsX.sh y para descargar todas las páginas web que contienen las secuencias de aminoácidos de los péptidos.

6. Mover todos los archivos con nombre onlyIDsX afuera de cada una de las carpetas AntiX

7. Ejecutar en MATLAB el archivo 'findRep.m', cuyo fin es detectar los antibacterianos con múltiples actividades. El resultado es 1 archivo llamado Everyb y otro denominado repb.

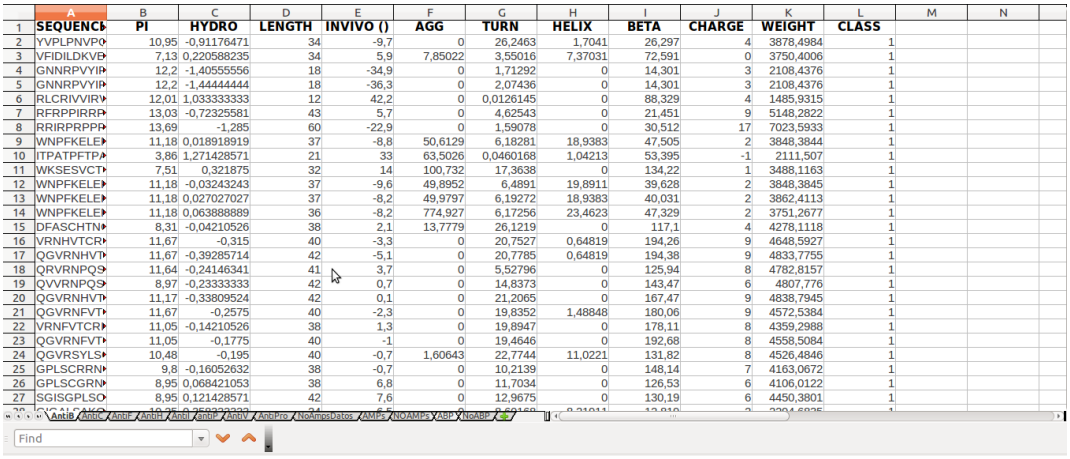
8. Ubicar en MATLAB y sin borrar las variables que se generaron del paso anterior, se ejecuta RepProto2. Éste genera 8 archivos llamados SetX.

9. Copiar en la carpeta antiX los archivos SetX y ejecutar el archivo ExtractSeqNoRep.py, que extrae las secuencias que dentro de sus actividades no son antibacterianas o tienen actividad antibacteriana exclusiva.

10. Finalmente, mover y ejecutar FormatFasta.m al lado de los archivos Sequence y SequencesUX para calcular 5 descriptores y obtener las secuencias en formato fasta para luego usar esta información en AGGRESCAN.

G. Secuencias y descriptores para los péptidos con y sin actividad antimicrobiana

La información de las secuencias y descriptores para todos los grupos se encuentra en el archivo adjunto al libro, con el nombre Information.xls y cuya presentación se muestra en la figura 12. En la hoja con el nombre NAntiX (donde X es B=antibacteriano, C=anticancer, F=antifúngico, H=antiVIH, I=insecticidas, P=antiparasitales, V=antivirales, Pro=Protisto) contiene información de los grupos mostrados anteriormente y en NOAMP está la información de los péptidos sin actividad antimicrobiana. En las hojas con el nombre de AMP y NOAMP son el conjunto de datos para el f_1 y ABP y NOABP son para f_2.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	SEQUENCE	PI	HYDRO	LENGTH	INVIVO ()	AGG	TURN	HELIX	BETA	CHARGE	WEIGHT	CLASS		
2	YVPLPNVPQ	10,95	-0,91176471	34	-9,7	0	26,2463	1,7041	26,297	4	3878,4984	1		
3	VFIDILDKVP	7,13	0,220588235	34	5,9	7,85022	3,55016	7,37031	72,591	0	3750,4006	1		
4	GNNRPVYIP	12,2	-1,40555556	18	-34,9	0	1,71292	0	14,301	3	2108,4376	1		
5	GNNRPVYIP	12,2	-1,44444444	18	-36,3	0	2,07436	0	14,301	3	2108,4376	1		
6	RLCRIVVIRV	12,01	1,033333333	12	42,2	0	0,0126145	0	88,329	4	1485,9315	1		
7	RFRPPIRRP	13,03	-0,72325581	43	5,7	0	4,62543	0	21,451	9	5148,2822	1		
8	RRIRPRPPP	13,69	-1,285	60	-22,9	0	1,59078	0	30,512	17	7023,5933	1		
9	WNPFKELE	11,18	0,018918919	37	-8,8	50,6129	6,18281	18,9383	47,505	2	3848,3844	1		
10	ITPATPFTPA	3,86	1,271428571	21	33	63,5026	0,0460168	1,04213	53,395	-1	2111,507	1		
11	WKSESVCT	7,51	0,321875	32	14	100,732	17,3638	0	134,22	1	3488,1163	1		
12	WNPFKELE	11,18	-0,03243243	37	-9,6	49,8952	6,4891	19,8911	39,628	2	3848,3845	1		
13	WNPFKELE	11,18	0,027027027	37	-8,2	49,9797	6,19272	18,9383	40,031	2	3862,4113	1		
14	WNPFKELE	11,18	0,063888889	36	-8,2	774,927	6,17256	23,4623	47,329	2	3751,2677	1		
15	DFASCHTN	8,31	-0,04210526	38	2,1	13,7779	26,1219	0	117,1	4	4278,1118	1		
16	VRNHVTCR	11,67	-0,315	40	-3,3	0	20,7527	0,64819	194,26	9	4648,5927	1		
17	QGVRNHVT	11,67	-0,39285714	42	-5,1	0	20,7785	0,64819	194,38	9	4833,7755	1		
18	QRVRNPQS	11,64	-0,24146341	41	3,7	0	5,52796	0	125,94	8	4782,8157	1		
19	QVVRNPQS	8,97	-0,23333333	42	0,7	0	14,8373	0	143,47	6	4807,776	1		
20	QGVRNHVT	11,17	-0,33809524	42	0,1	0	21,2065	0	167,47	9	4838,7945	1		
21	QGVRNFVT	11,67	-0,2575	40	-2,3	0	19,8352	1,48848	180,06	9	4572,5384	1		
22	VRNFVTCR	11,05	-0,14210526	38	1,3	0	19,8947	0	178,11	8	4359,2988	1		
23	QGVRNFVT	11,05	-0,1775	40	-1	0	19,4646	0	192,68	8	4558,5084	1		
24	QGVRSYLS	10,48	-0,195	40	-0,7	1,60643	22,7744	11,0221	131,82	8	4526,4846	1		
25	GPLSCRRN	9,8	-0,16052632	38	-0,7	0	10,2139	0	148,14	7	4163,0672	1		
26	GPLSCGRN	8,95	0,068421053	38	6,8	0	11,7034	0	126,53	6	4106,0122	1		
27	SGISGPLSQ	8,95	0,121428571	42	7,6	0	12,9675	0	130,19	6	4450,3801	1		

Figura 12: Archivo de excel con las secuencias y descriptores para los grupos.