Articulo **Revista de Ciencias de la Salud**
Diciembre 2020 Vol.7 No.23 18-28

SVM en paralelo para optimizar el tiempo de respuesta en la predicción de co-receptores de los virus R5, X4 Y R5X4 (mutado) que causan el Sida (VIH-1) en células CD4

Parallel SVM to optimize response time in the prediction of co-receptors of the R5, X4 and R5X4 viruses (mutated) that cause AIDS (HIV-1) on CD4 cells

MEDINA-VELOZ, Gricelda†*, LUNA-ROSAS, Francisco Javier, MARTÍNEZ-VALADEZ, Kelly Carolina y TAVAREZ-AVENDAÑO, Juan Felipe

Universidad Tecnológica del Norte de Aguascalientes
Instituto Tecnológico de Aguascalientes

ID 1er Autor: *Gricelda, Medina-Veloz* / **ORC ID**: 0000-0002-1955-3620, **arXiv Author ID:** GrisArvix18**, CVU CONACYT ID**: 228438

ID 1er Coautor: *Francisco Javier, Luna-Rosas* / **ORC ID**: 0000-0001-6821-4046, **arXiv Author ID:** arXivFco19, **CVU CONACYT ID**: 87098

ID 2do Coautor: *Kelly Carolina, Martínez-Valadez*

ID 3er Coautor: *Juan Felipe, Tavarez-Avendaño* / **ORC ID**: 0000-0001-8336-1546

DOI: 10.35429/JOHS.2020.23.7.18.28 Recibido Octubre 10, 2020; Aceptado Diciembre 28, 2020

Resumen

En particular, los virus R5 HIV-1 usan CCR5 como un co-receptor para la entrada viral, los virus X4 HIV-1 usan el CXCR4, mientras algunos extraños virus conocidos como R5X4 o D-tropic, tienen la habilidad de utilizar ambos co-receptores. Los virus X4 y R5X4 son asociados con un rápido progreso en el VIH-1. En este artículo se realizarán una serie de experimentos para implementar una máquina de aprendizaje supervisado en paralelo que permita optimizar el tiempo de respuesta en la predicción de co-receptores (CCR5, CXCR4) del Virus que causan el sida (VIH-1) en células CD4. Para implementar la máquina de aprendizaje supervisado en paralelo utilizaremos Snow en R. Snow provee el soporte para fácilmente ejecutar funciones en R en paralelo. La mayoría de las funciones en paralelo en Snow son variaciones de la función estándar lapply(). Para implementar las funciones en paralelo, Snow utiliza una arquitectura, maestro/esclavo donde el maestro envía tareas a los trabajadores, y los trabajadores ejecutan las tareas y retornan los resultados al maestro.

SVM en Paralelo, VIH, Tiempo de Respuesta

Abstract

In particular, the R5 HIV-1 viruses use CCR5 as co-receptor for the virus entrance, the X4 virus HIV-1 use the CXCR4, while some strange viruses known as R5X4 or D-tropic, have the ability to use both co-receptors. The X4 and R5X4 viruses are associated with rapid progress in HIV-1. In this article a series of experiments will be carried out to implement a Supervised Learning Machine in Parallel that allows optimizing the response time in the prediction of co-receptors (CCR5, CXCR4) of the virus that cause AIDS (HIV-1) in CD4 cells. To implement the machine in parallel we will use Snow in R. Snow provides the support to easily execute functions in R in parallel. Most functions in parallel in Snow are variations of the standard lapply () function. To implement the functions in parallel, Snow uses a master / slave architecture where the teacher sends tasks to the workers, and the workers execute the tasks and return the results to the teacheR.

Parallel SVM, HIV, Response time

Citación:. MEDINA-VELOZ, Gricelda, LUNA-ROSAS, Francisco Javier, MARTÍNEZ-VALADEZ, Kelly Carolina y TAVAREZ-AVENDAÑO, Juan Felipe. SVM en paralelo para optimizar el tiempo de respuesta en la predicción de co-receptores de los virus R5, X4 Y R5X4 (mutado) que causan el Sida (VIH-1) en células CD4. Revista de Ciencias de la Salud. 2020. 7-23: 18-28.

*Correspondencia al Autor (correo electrónico: gricelda.medina@utna.edu.mx)
† Investigador contribuyendo como primer Autor

1. Introducción

El SIDA es un problema de Salud pública importante en el mundo siendo cada vez más alto el nivel de incidencia en mujeres. Urge implementar medidas preventivas para evitar la transmisión del VIH, aun no hay una vacuna segura, efectiva y lista para usarse contra el SIDA. Mientras los antiretrovirales actuales mejoran la salud del enfermo, en unos años estos desarrollaran resistencia a estos medicamentos.

El tropismo del Virus de Inmunodeficiencia Humana se define como la atracción altamente específica del virus hacia el tejido del huésped, determinado en parte por los marcadores de superficie de las células de este (por ejemplo las células CD4). Los virus desarrollan una habilidad específica para atacar las células en forma selectiva, así como los órganos del huésped y a menudo, ciertas poblaciones de células que se encuentran en los órganos del cuerpo del huésped [1].

En 1983 se aisló por primera vez el retrovirus de la familia de los lentivirus en pacientes con SIDA [2]. A partir de esta fecha, la comunidad científica ha intensificado su búsqueda en conocer más la biología molecular y patogénesis del VIH. Los pacientes con SIDA presentan disminución de linfocitos CD4+ según progresa su enfermedad, así en 1984 se planteó que era precisamente la molécula CD4, el receptor específico para que el virus del VIH entrara a la célula [2].

En 1986 se demostró que la proteína gp120, de la envoltura viral, se acoplaba al CD4 y ambas moléculas co-precipitaban como un complejo inmune, demostrando así la unión gp120-CD4. La expresión del CD4 en la membrana es necesaria pero no suficiente para que el virus se funda con la célula, la búsqueda del posible coreceptor se extendió por unos 10 años.

Actualmente, a partir de 1996 se han identificado el CCR5 y el CXCR4 como principales quimosinas (citocinas quimiotácticas) co-receptoras para el VIH lo cual ha llevado a entender el tropismo viral y la patogénesis en el ámbito molecular [1].

2. Antecedentes

En 1981, el comité internacional sobre la taxonomía de los virus (ICTV) [2] propuso la definición siguiente: “Una especie del virus es un concepto que será representado normalmente por un grupo de cadenas de una variedad de fuentes, o una población de cadenas de una fuente particular, que tienen en común un sistema de propiedades correlacionadas estables que diferencian un grupo de otros grupos de cadenas” [2]. Hoy, el ICTV reconoce más de 3.600 especies del virus [2].

Los virus animales se clasifican en seis clases: I, II III, IV, V, y VI.

2.1 Clase VI. Retroviridae

Es un grupo de virus de RNA que infectan animales y seres humanos. Aunque el nombre de retrovirus fue asignado como tal hasta 1974, los retrovirus fueron descubiertos mucho antes, en 1908 y 1911 [2]. La mayoría de los virus introducen su RNA a la célula huésped y actúa como mRNA o es transcrito en mRNA. El Retroviridae es diferente, lleva con ellos una enzima única llamada transcriptasa reversa. Esta enzima es una polimerasa RNA-dependiente del DNA que convierte el RNA viral en DNA. Esta RNA viral tiene extremos pegajosos únicos que le permite integrar en el anfitrión su propio DNA. Puede integrarse en el genoma celular y permanecer durante muchísimo tiempo inactivo. Está implicado en la producción de tumores y, entre sus virus más conocidos, se encuentra el VIH.

Como se puede observar en la Figura 1, la estructura general de un retrovirus está formada principalmente por una bicapa lípido-proteica formada por dos subunidades proteicas que son codificadas por el gen env (en la sección 2.3 se describen los genes) del virus y componentes lipídicos y proteicos propios de la membrana celular. También son llamadas glicoproteínas específicas e importantes que causan la infección. Contiene además una cápside esférica o cónica formada por 3 subunidades proteicas codificadas por el gen gag. Dentro se encuentran las enzimas virales necesarias para el proceso de replicación viral: proteasa (gen pro) y la transcriptasa inversa e integrasa, codificadas por el gen pol.

ISSN- 2410-3551

Algunos Retrovirus aislados causan leucemia o el sarcoma en células huésped de vertebrados, por eso algunas veces son llamados virus del sarcoma-leucemia [2].

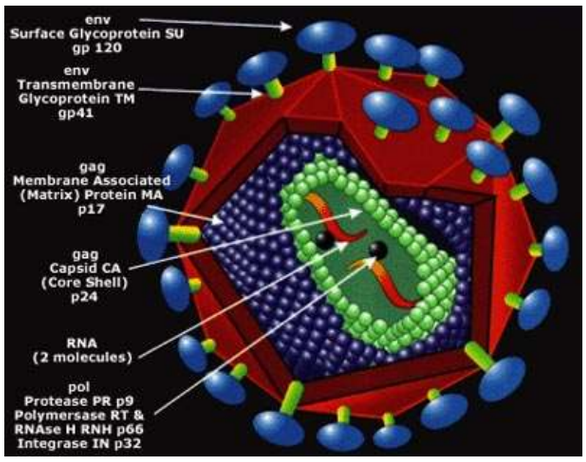

Figura 1 Estructura del Retrovirus
Fuente: Courtesy of Stanford University School of Medicine.
http://www.stanford.edu/group/nolan/tutorials/ret_6_gpe desc.html (USA-2015)

Algunos retrovirus causan cáncer directamente, integrando genes llamados oncógenes en el DNA de la célula huésped, causando la transformación maligna de células normales en las células de cáncer, éstos se llaman virus transformadores agudos. Otros causan cáncer indirectamente activando un proto-oncogen del huésped, éstos se llaman virus transformadores no-agudos Otra característica importante es que algunos retrovirus son citotóxicos para ciertas células, inflándolas. El más notable es el virus del síndrome de la inmunodeficiencia en humanos que destruye los linfocitos CD4 T que infecta.

2.2 Clasificación de Retrovirus

Los retrovirus se clasifican actualmente en 7 géneros [2], lo cual se puede apreciar en la Figura 2, donde se observan los distintos géneros de retrovirus de acuerdo a la familia que pertenecen. Tal es el caso de VIH perteneciente a la familia de Lentivirus.

Las familias conocidas y clasificadas hasta el momento por ICTV son:

a. Alpharetrovirus, Betaretrovirus, Epsilonretrovirus, Gammaretrovirus: Contienen genomas simples.

b. Lentivirus, Spumavirus y Deltaretrovirus. Contienen genomas complejos.

Sólo los retrovirus con genoma simple y spumaviruses llegan a ser retrovirus endógenos en sus huéspedes; Esto es, retrovirus que se encuentran latentes dentro de la célula huésped. Los Lentivirus son retrovirus citopáticos (retrovirus que dañan la célula) que causan fundamentalmente cuadros de inmunodeficiencia, síndromes neurológicos y enfermedades autoinmunes de evolución lenta.

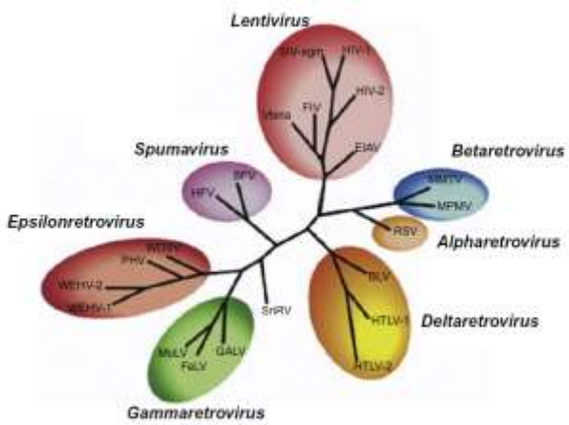

Figura 2 Análisis Filogenético de Retroviruses
Fuente: Courtesy of International Committee on Taxonomy of Viruses (ICTV-2015).
http://ictvonline.org/index.asp

2.3 Ciclo vital de retrovirus

La naturaleza del genoma retroviral fue descubierta en los años 1960´s. [3]. Todos los genomas retrovirales consisten en dos cadenas (filamentos positivos) idénticas de moléculas de DNA o RNA, por lo que son los únicos virus diploides conocidos. El genoma consiste en 2 moléculas idénticas lineales de ssRNA (+) entre 7-11 kilobases. En los extremos del genoma se localizan las regiones extensas terminales repetidas LTR (long terminal repeats), con secuencia redundante (R) que juega un papel primordial durante el proceso de retrotranscripción.

La conservación del genoma diploide de los retrovirus tiene un papel dominante en el ciclo vital del virión. El genoma diploide parece estar ligado físicamente y el sitio en el cual el acoplamiento ocurre se le llama Sitio del Acoplamiento del Dímero (DLS), el cual contribuye como señal para la encapsulación de retrovirus.

ISSN- 2410-3551

En la Figura 3 se aprecia la acción infectiva del retrovirus de VIH (ciclo vital), que se encuentra en una célula infectada denominada célula huésped.

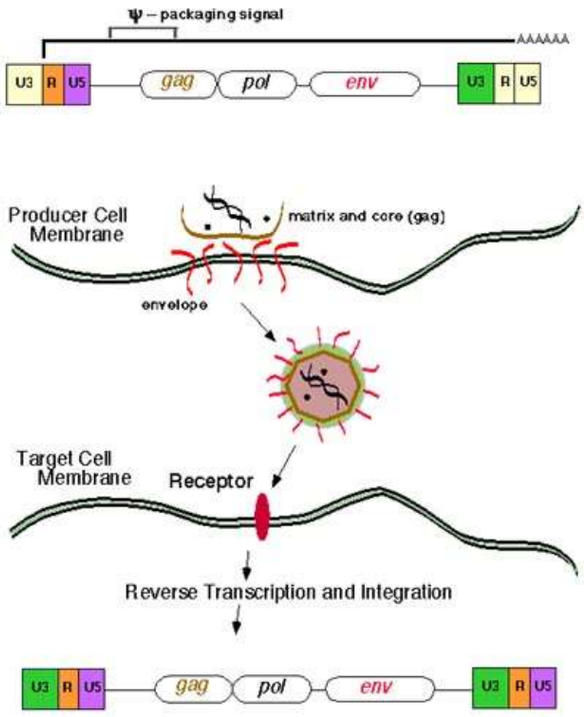

Figura 3 Ciclo vital de retrovirus
Fuente: Courtesy of Stanford University School of Medicine.
http://www.stanford.edu/group/nolan/tutorials/ret_6_gpe desc.html (USA-2015).

El ciclo vital retroviral comienza en el núcleo de una célula infectada, parte superior de la Figura 3, donde se muestra la membrana de la célula huésped (envelope) y el genoma del virus (genes gag, pol, env). En esta etapa del ciclo vital el genoma retroviral es un elemento del DNA integrado en el DNA de la célula huésped. El genoma del virus es de aproximadamente 8-12 kilobases del DNA (depende de la especie retroviral).

El genoma del virus aprovecha los elementos disponibles en la célula huésped para forma la cápside que encierra el corazón del virus encapsulando genes y otros elementos (matriz and core), enseguida el virus sale de la célula huésped como partícula libre (parte central de la Figura 3) y busca otras células sanas para infectarlas, lo cual se observa en la parte inferior de la Figura 3 (Target cell membrane).

La partícula libre puede infectar las nuevas células uniéndose a un receptor de la superficie de la célula (receptor).

La especificidad de la interacción del virus con la célula huésped es determinada en gran parte por las proteínas de la cápside del retrovirus. La infección conduce a la inyección de nucleoproteínas del virus, que consiste sobre todo en proteínas derivadas del genoma gag, RNA genómico integral, y la enzima transcriptasa reversa. Una vez dentro de la célula, el complejo núcleoproteico a través de la enzima transcriptasa reversa, recién producida, inicia la creación de una copia de la doble-cadena del DNA del genoma del retrovirus con objeto de integrarla en el cromosoma de la célula huésped (parte inferior de la Figura 3). Al término de la transcripción reversa, la enzima viral integrasa busca en el DNA de la célula huésped un lugar apropiado, en donde corta un segmento de DNA del anfitrión y pega el DNA de doble cadena recién copiado en el DNA de la célula huésped, el retrovirus ahora está preparado para iniciar una nueva ronda de replicado.

2.4 Retrovirus más comunes en humanos

En el hombre se han descrito tres retrovirus patógenos, el HTLV-I es un retrovirus oncogénico que causa un cuadro de leucemia-linfomas de células T y un cuadro neurológico en el 5% de los pacientes que infecta. El HTLV-II que se ha aislado de algunos casos de leucemias. Y el último, el HTLV-III conocido como VIH, descubrimiento en 1983, agente causal del síndrome de inmunodeficiencia adquirida (SIDA). El virus de la inmunodeficiencia humana (VIH) es un retrovirus no transformante perteneciente a la familia de los Lentivirus [2].

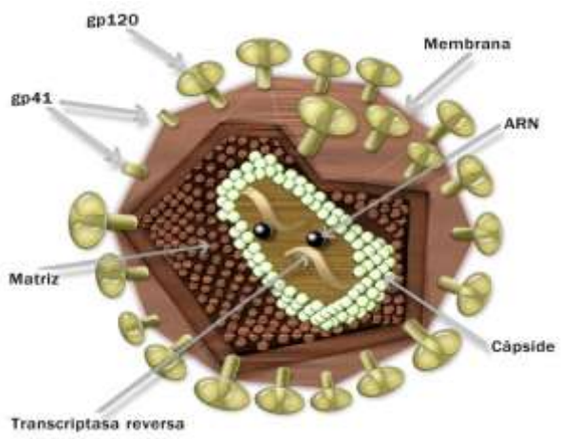

Figura 4 Estructura del VIH
Fuente: Courtesy of Stanford University School of Medicine.
http://www.stanford.edu/group/nolan/tutorials/ret_6_gpe desc.html (USA-2015).

ISSN- 2410-3551

En la Figura 4 se aprecia la estructura del retrovirus de la inmunodeficiencia humana (VIH) [2], con todos sus elementos, semejantes a los descritos en la Figura 1. Como podemos observar en la Figura 4, la proteína principal del VIH asociada con la envoltura es la gp120/41. Esta funciona como antireceptor viral o proteína de adherencia. La gp41 cruza la envoltura, la gp120 está presente sobre la superficie exterior y está unida no covalentemente a la gp41. El precursor de gp120/41 (gp160) es sintetizado en el retículo endoplásmico y es transportado por medio del aparato de Golgi a la superficie de célula.

Actualmente se conocen dos tipos de VIH; El VIH-1, causante de la inmensa mayoría de los casos de la actual pandemia del Síndrome de Inmunodeficiencia Adquirida (SIDA), el cual fue aislado por primera vez en el Instituto Pasteur de París en 1983 y el VIH-2, menos patógeno, con una baja transmisión y por tanto menos extendido. Causa SIDA también en un porcentaje muy pequeño de los sujetos infectados y se encuentra limitado esencialmente a países del área de África occidental. Fue también aislado en el Instituto Pasteur en 1986 [2].

2.5 Daños celulares causantes por VIH

Las células CD4 (llamadas T-4 también) son un tipo de linfocito (glóbulo blanco), y existen una gran variedad de clases. Las células T-4 o CD4 son las células "ayudantes," las que dirigen el ataque contra las infecciones en el sistema inmune en coordinación con otras células llamadas células T-8 o CD8 son las células "supresoras," las que finalizan una respuesta inmunológica. Las células CD8 también pueden ser "asesinas," que matan a células cancerosas y a células infectadas por virus [2], [3].

Las células humanas que infecta el VIH con más frecuencia son las CD4, y cuando ellas se multiplican para combatir infecciones, también hacen más copias del VIH involuntariamente. En infecciones prolongadas por VIH, el número de células CD4 disminuye. Este es un signo de que el sistema inmune se ha debilitado. Cuanto más bajo sea el recuento de células CD4, más posibilidades hay que el individuo se enferme. El síndrome de inmunodeficiencia humana o SIDA, es la expresión final de la infección por el VIH.

Entonces, la infección por este virus ocasiona la destrucción del sistema inmune principalmente. Estas manifestaciones clínicas se deben al tropismo tanto macrofágico como linfocitario del virus [2], [3]. Presenta una preferencia para infectar a linfocitos CD4, en los que la replicación es activa y muy agresiva, lo que provoca como característica de la infección una profunda inmunosupresión [2], [3].

Las células humanas que infecta el VIH con más frecuencia son las CD4, y cuando ellas se multiplican para combatir infecciones, también hacen más copias del VIH involuntariamente. En infecciones prolongadas por VIH, el número de células CD4 disminuye. Este es un signo de que el sistema inmune se ha debilitado. Cuanto más bajo sea el recuento de células CD4, más posibilidades hay que el individuo se enferme.

El síndrome de inmunodeficiencia humana o SIDA, es la expresión final de la infección por el VIH. Entonces, la infección por este virus ocasiona la destrucción del sistema inmune principalmente. Estas manifestaciones clínicas se deben al tropismo tanto macrofágico como linfocitario del virus [2], [3]. Presenta una preferencia para infectar a linfocitos CD4, en los que la replicación es activa y muy agresiva, lo que provoca como característica de la infección una profunda inmunosupresión [2], [3].

2.6 Detección del VIH en humanos

El análisis del VIH se refiere a las pruebas que determinan si un individuo está o no infectado con el virus de inmunodeficiencia humano (VIH), que causa el SIDA. Existen varios tipos de análisis que generalmente se practican en muestras de sangre obtenidas de individuos en estudio, aunque también se utilizan muestras de orina y otros fluidos corporales, inclusive en raspados de mejilla [2], [3].

Los diferentes tipos de análisis en la muestra son:

a. Detección de los anticuerpos y/o antígenos del VIH.
b. Análisis de carga viral (recuento de virus)
c. Recuento de células CD4 y CD8

ISSN- 2410-3551

Enseguida se describen algunos métodos de análisis anteriormente mencionados.

2.6.1 Análisis de detección de anticuerpos

Estos análisis buscan "anticuerpos" contra el VIH en la sangre, saliva u orina. Los anticuerpos son proteínas producidas por el sistema inmune para combatir a un germen específico en este caso contra VIH los cuales demoran entre dos y tres meses en aparecer después que se ha infectado el organismo. Los análisis de anticuerpos contra el VIH tienen un 99.5% de precisión [4]. Para obtener un resultado, el análisis debe ser hecho dos o más veces. El primer análisis usa enzimas para detectar anticuerpos, este análisis se llama "MEIA" o "ELISA" que es un enzimoinmunoanálisis de micro partículas [4]. Antes de que se reporte un análisis MEIA positivo, los resultados se confirman con otro análisis llamado "Western Blot" [4], que es un análisis de inmunotransferencia. Esta técnica permite caracterizar los anticuerpos dirigidos contra cada proteína vírica, confirmando así la seropositividad o bien identificando posibles reacciones inespecíficas.

2.6.2 Análisis de carga viral

El análisis de carga viral mide la cantidad de VIH en la sangre. Existen diferentes técnicas. El método PCR (polymerase chain reaction) utiliza una enzima para multiplicar al VIH de la muestra de sangre. Luego una reacción química marca al virus. Los marcadores son medidos y se calcula la cantidad de virus [5], [6].

2.6.3 Análisis de células CD4

Consiste en realizar un conteo de células CD4 y CD8 en la muestra. Se especifica el número de células por milímetro cúbico de sangre (mm3). No existe un acuerdo acerca de cuál es el nivel promedio normal de células CD4. El recuento normal de CD4 es entre 500 y 1600 células y el de CD8 es entre 375 y 1100 células. Las células CD4 pueden disminuir drásticamente en personas VIH+ y en algunos casos pueden llegar a cero. Debido a que el recuento de células CD4 varía mucho, también se analiza el porcentaje de células CD4. Este porcentaje se refiere al total de linfocitos. Si el análisis indica que existe un 34% de CD4, significa que el 34% de sus linfocitos son células CD4.

El porcentaje es más estable que el número de células CD4. El rango normal es entre 20% y 40%. Un porcentaje debajo de 14% indica daño serio al sistema inmune. Es una señal del SIDA en personas infectadas con VIH [7].

3. Materiales y Metodos

3.1 Máquinas de aprendizaje supervisado

El aprendizaje supervisado (Machine Learning) toma un conjunto de datos y respuestas conocidas, y busca la manera de construir un modelo predictivo que genera predicciones razonables o adecuadas a los nuevos datos que se ingresen. Es decir, dada una base de datos $D = \{t1, t2, \ldots, tn\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C1, C2, \ldots, Cm\}$, el problema de la clasificación/predicción es encontrar una función $f: D \rightarrow C$ tal que cada ti sea asignada a una clase Cj. $f: D \rightarrow C$ podría ser un Método KNN (Vecino más Cercano), un Método de Árbol de Decisión, una Máquina de Soporte Vectorial, un Modelo Bayesiano, un Método de Bosques Aleatorios (Random Forest), y un Método de Potenciación (Boosting) [8].

3.2 SVM en paralelo

No es de sorprenderse que R ha llegado a ser el favorito en la era de Big Data Analytics (McCallum y Weston 2011). Snow provee el soporte para fácilmente ejecutar funciones en R en paralelo. La mayoría de las funciones en paralelo en Snow son variaciones de la función estándar lapply(). Para implementar las funciones en paralelo, Snow utiliza una arquitectura, maestro/esclavo donde el maestro envía tareas a los trabajadores y los trabajadores ejecutan las tareas y retornan los resultados al maestro.

Una importante característica de Snow es que este puede ser usado con diferentes mecanismos de transporte para comunicarse entre el maestro y los trabajadores. Snow puede ser usado con conexiones de Socket, MPI, PVM o NetWorkSpaces. Los Sockets no requieren paquetes adicionales y son los más portables.

Ahora estamos listos para usar Snow y La Máquina de Soporte Vectorial (por sus siglas en inglés, SVM) para optimizar el tiempo de respuesta en el diagnóstico de cáncer de mama.

ISSN- 2410-3551

La SVM en Paralelo, es un clasificador y su objetivo es encontrar un modelo para predecir la clase a la que pertenecería cada espectro de cáncer de mama (sano y dañado), esta predicción se debe hacer con la mayor precisión posible [8].

4. Resultados y Discusión

4.1 Colección de datos. Secuencias aisladas de HIV representando los tres tropismos virales

M-trópico R5 (Tabla 2), T-trópico X4 (Tabla 3) y Dual-trópico R5X4 (Tabla 1), identificados por Lamers en [9], National Center for Biotechnology Information (NCBI. http://www.ncbi.nlm.nih.gov/) [10], UniProtKB-2019). https://www.uniprot.org/uniprot/ [11], The HIVDatabases. https://www.hiv.lanl.gov/content/index [12].

R5X4 Virus			
AB014795	U08445	AF259019	AF112925
AF062029	AF355674	AF259025	M17451
AF062031	AF355647	AF259021	K02007
AF062033	AF355630	AF259041	U39362
AF107771	AF355690	AF258970	AF069140
U08680	M91819	AF258978	AF458235
U08682	AF035532	AF021607	AF005494
U08444	AF035533	AF204137	

Tabla 1 Número de Acceso para las Diferentes Secuencias de los Virus R5X4

R5 Virus			
AF062012	AY010852	M38429	U08453
L03698	U08670	U27443	AF307755
AF231045	U08798	U79719	AF307750
AY669778	AY669715	U04909	AY043176
U08810	U08710	U04918	AY158534
U51296	U16217	U40908	AX455917
AF407161	M26727	U08450	AY043173
AB253421	AJ418532	AF112542	AF307757
U08645	AJ418479	M63929	U08803
U08647	AJ418495	U66221	U88824
AB253429	AJ418514	AF491737	U69657
AY288084	AJ418521	U08779	AF355326
AF307753	U23487	L22084	U88826
AF411964	U04900	U27413	U08368
U08823	AF022258	AF005495	U27426
AF411965	AF258957	U52953	AJ006022
U92051	AF021477	AF321523	U08795
AF355318	U08716	L22940	
AY010759	U39259	U45485	
AY010804	AF204137	AB023804	

Tabla 2 Número de Acceso para las Diferentes Secuencias de los Virus R5

X4 Virus			
AB014785	X01762	AF258981	U27408
AB014791	L31963	AF259003	AF411966
AB014796	U08447	AF021618	U27399
AB014810	AF355660	AF128989	U08822
U48267	AF355748	M17449	U08738
U08666	AF355742	AF075720	U08740
AF069692	AF355706	U48207	U08193
AF355319	AF180915	U72495	AF355330
AF355336	AF180903	AY189526	
M14100	AF035534	AF034375	
A04321	AF259050	AF034376	

Tabla 3 Número de Acceso para las Diferentes Secuencias de los Virus X4

Tipo	Virus
R5	AF062012
R5	AF231045
R5	U08810
R5	AF407161
R5	AB253421
R5	U08645
X4	AB014785
X4	AB014791
X4	AB014796
X4	AB014810
X4	U08666
X4	AF069672
R5X4	AB014795
R5X4	AF062029
R5X4	AF062031
R5X4	AF062033
R5X4	U08680
R5X4	U08682

Tabla 4 Proteína gp120 de diferentes virus (R5, X4, R5X4)

El ADN y los aminoácidos que conforman la proteína gp120 de cada virus fueron obtenidos de la bases de datos: NCBI-2020 http://www.ncbi.nlm.nih.gov/ [9], UniProtKB-2020 https://www.uniprot.org/uniprot/ [10], The HIV Databases-2020 https://www.hiv.lanl.gov/content/index [11]. La Tabla 4 muestra la categoría del virus, el número de acceso y los aminoácidos de va rios virus seleccionados de cada uno de los grupos que representa los tres tropismos virales (R5, X4 y R5X4).

ISSN- 2410-3551

4.2 Generación de características

Diseño de software propio fue usado para calcular estadísticas generales por posición, algunas propiedades fueron calculadas de las propiedades de los aminoácidos acorde a las Tablas 5 y 6 con software propio (por ejemplo, tipo de aminoácido, carga, volumen (A3), masa(daltons), HP Escala, Área de Superficie, Alpha Hélix, B-strand y Turn) y las propiedades restantes fueron extraídas de Expasy (https://web.expasy.org/protparam/ [13] y https://www.protpi.ch/Calculator [14] (por ejemplo, Molecular Weight, Theoretical PI, Amino Acid Composition, Atomic Composition (Carbon, Hidrogeno, Nitrogeno, Oxigeno, Sulfuro), Instability Index, Aliphatic Index, etc.). Las propiedades de los aminoácidos fueron posteriormente agrupadas en tablas (Data Sets) y cargados al software de R (ver Tablas 7 y 8).

Amino acid residues	Charge	Volume(A3)	Masa(Daltons)	HP Scale
Alanine(A) Arginine(I)				
Aspargine(N)	0	67	71.09	1.8
Aspartic	+1	148	156.19	-4.5
Acid(D)	0	96	114.11	-3.5
Cystine(C)	-1	91	115.09	-3.5
Glutamine(Q)	0	86	103.15	2.5
Glutamic	0	114	128.14	-3.5
Acid(E)	-1	109	129.12	-3.5
Glycine(G)	0	48	57.05	-0.4
Histidine(H)	0	118	137.14	-3.2
Isoleucine(I)	0	124	113.16	4.5
Leucine(L)	0	124	113.16	3.8
Lysine(K)	+1	135	128.17	-3.9
Methionine(M)	0	124	131.19	1.9
Phenylalanine(F)	0	135	147.18	2.8
Proline(P)	0	90	97.12	-1.6
Serine(S)	0	73	87.08	-0.8
Threonine(T)	0	93	101.11	-0.7
Tryptophane(W)	0	163	186.21	-0.9
Tyrosine(Y)	0	141	163.18	-1.3
Valine(V)	0	105	99.14	4.2

Tabla 5.1 Propiedades de los Aminoácidos [8]

Amino acid residues	Surface Area	Alpha Helix	B-strand	Turn
Alanine(A) Arginine(I)				
Aspargine(N)	0.74	1.41	0.72	0.82
Aspartic	0.64	1.21	0.84	0.90
Acid(D)	0.63	0.76	0.48	1.34
Cystine(C)	0.62	0.99	0.39	1.24
Glutamine(Q)	0.91	0.66	1.40	0.54
Glutamic	0.62	1.27	0.98	0.84
Acid(E)	0.62	1.59	0.52	1.01
Glycine(G)	0.72	0.43	0.58	1.77
Histidine(H)	0.78	1.05	0.8	0.81
Isoleucine(I)	0.88	1.09	1.67	0.47
Leucine(L)	0.85	1.34	1.22	0.57
Lysine(K)	0.52	1.23	0.69	1.07
Methionine(M)	0.85	1.30	1.14	0.52
Phenylalanine(F)	0.88	1.16	1.33	0.59
Proline(P)	0.64	0.34	0.31	1.32
Serine(S)	0.66	0.57	0.96	1.22
Threonine(T)	0.7	0.76	1.17	0.90
Tryptophane(W)	0.85	1.02	1.35	0.65
Tyrosine(Y)	0.76	0.74	1.45	0.76
Valine(V)	0.86	0.90	1.87	0.41

Tabla 6 Propiedades de los Aminoácidos, 2D structure propensity [8]

Tipo	Virus	Aminoácidos	PI	negatively	positively	Carbon	Hydrogen	Nitrogen	Oxygen	Sulfur	number of atoms	Aliphatic
R5	AF062012	[illegible]	8.87	10	13	[illegible]	[illegible]	194	179	8	[illegible]	88.81
R5	AF231045	MSPGQVFMT	8.8	2	4	165	264	52	53	1	538	68.80
R5	U08810	[illegible]	6.77	15	15	668	1061	189	198	4	2120	101.34
R5	AF407161	SCSGKLICTTS	5.37	14	12	659	1044	182	201	3	2089	101.34
R5	AB253421	SCSGKLICTTV	5.25	13	11	659	1039	179	198	3	2078	105.2
R5	U08645	[illegible]	10.4	7	18	460	787	145	131	2	1503	92.17
R5	U08647	GQAFYATGD	10.6	5	17	456	759	145	128	2	1488	93.37
R5	AB253429	NCSGKLICPT	5.88	16	15	672	1072	184	197	3	2128	109.76
R5	AY288084	SCSGKICTTTV	5.83	14	15	667	1053	185	200	3	2112	101.34
R5	AF307753	RESVVNGPGQ	7.82	6	8	326	521	90	104	2	1054	78.13
X4	AB014785	[illegible]	10.09	8	17	541	855	161	158	4	1719	92
X4	AB014791	[illegible]	9.8	9	10	540	888	150	160	30	1749	92
X4	AB014796	[illegible]	9.86	9	10	540	880	160	161	4	1754	94.76
X4	AB014810	[illegible]	10.02	8	10	539	886	162	157	3	1747	95.51
X4	U48267	[illegible]	9.17	18	25	1051	1673	287	312	11	3345	78.20
X4	U08666	[illegible]	11.45	3	15	456	745	149	136	3	1489	85.85
X4	AF069672	CSGKLICTTNV	6.76	12	13	660	1052	180	195	4	2091	105.2
X4	AF355319	[illegible]	9.51	15	26	964	1547	265	309	8	3113	78.46
X4	AF355336	[illegible]	9.08	18	22	1008	1575	285	319	9	3212	83.44
X4	M14100	[illegible]	8.75	7	8	566	894	180	158	3	1781	113.8
R5X4	AB014795	[illegible]	9.86	9	16	538	886	160	159	2	1745	99.39
R5X4	AF062029	[illegible]	8.92	9	16	751	1275	221	219	7	2533	88.88
R5X4	AF062031	[illegible]	9.00	12	18	758	1213	209	225	8	2415	86.69
R5X4	AF062033	[illegible]	9.41	10	29	760	1228	230	229	7	2447	78.92
R5X4	[illegible]IDGP120:AA	QCEIKQ[illegible]	9.17	77	101	4352	6678	1224	1260	30	13753	88.69
R5X4	U08680	[illegible]	9.81	7	13	454	781	138	137	3	1471	92.57
R5X4	U08682	[illegible]	10.22	8	14	462	752	142	135	2	1493	93.36
R5X4	[illegible]IDGP120:AAB	[illegible]	8.62	83	91	4376	6917	1214	1266	42	13674	91.11
R5X4	U08445	VQEVGKAMYA	8.84	78	92	4261	6747	1188	1235	37	13471	94.35
R5X4	AF355674	[illegible]	9.83	9	15	541	677	187	167	3	1753	85.82

Tabla 7 Estadísticas de las Propiedades Moleculares para los Virus R5, X4 y R5X4

Tipo	Virus	hydropathicity	Volume (A3)	Mass (daltons)	HP scale	Surface area	alpha helix	B-strand	Turn	Isoelectric Point	Charge at PH 7.0	Molecular Weight
R5	AF062012	-0.339	12185	13009.2	-40.4	85.94	112.2	116.8	113.8	8.358	3.098	13026.82
R5	AF231045	-0.514	3561	3870.46	-18	25.43	29.94	35.05	34.31	8.613	1.848	3888.35
R5	U08810	-0.518	14096	14998.5	-65.8	91.61	137.1	128.2	111.3	7.045	0.033	15016.12
R5	AF407161	-0.53	13904	14811.2	-67.3	91.11	136.7	127.4	112.3	5.328	-1.97	14828.82
R5	AB253421	-0.38	13830	14716.2	-48.3	91.83	135.6	128.8	112.1	5.315	-1.971	14733.76
R5	U08645	-0.483	9944	10443.4	-44.4	65.8	93.42	92.48	83.43	10.409	11.182	10461.13
R5	U08647	-0.438	9853	10339.3	-40.3	65.86	94.2	90.71	83.7	10.611	12.177	10357.02
R5	AB253429	-0.452	14137	14959.6	-57.4	91.52	137.8	127.8	111.4	5.96	-0.966	14977.2
R5	AY288084	-0.567	14055	14998.4	-72	91.47	136.7	128.1	111.6	5.912	-0.969	15016.04
R5	AF307753	-0.441	7021	7561.72	-30.4	49.88	63.72	67.42	65.87	8.02	1.186	7579.47
X4	AB014785	-0.506	11632	12259.5	-54.1	76.76	105.9	108	98	10.094	9.19	12277.2
X4	AB014791	-0.496	11593	12232.4	-53.1	76.74	106.3	107.7	98.11	9.789	7.193	12250.13
X4	AB014796	-0.491	11628	12278.5	-52.5	76.94	106.9	108.5	97.58	9.853	7.193	12296.16
X4	AB014810	-0.448	11587	12226.4	-47.9	77.27	105.8	108.7	97.71	10.021	8.355	12244.13
X4	U48267	-0.391	22314	23812.9	-83.3	154.23	204.3	206.9	202.9	8.972	7.576	23830.35
X4	U08666	-0.539	9888	10569.4	-50.1	67.26	88.31	96.29	85.24	11.54	10.18	10587.04
X4	AF069672	-0.397	13916	14739.3	-50.4	91.71	136.1	128.2	112.4	7.038	0.026	14756.95
X4	AF355319	-0.651	21180	22853.2	-133	147.25	189.9	197.2	201.8	9.434	9.235	22870.65
X4	AF355336	-0.66	21477	23311.6	-138	150.04	194.3	200.9	205	8.852	6.235	23329
X4	M14100	-0.175	11851	12546.8	-18.9	79.48	115.1	112.2	92.53	8.544	2.355	12564.49
R5X4	AB014795	-0.426	11560	12186.4	-45.6	77.15	105.6	108.4	98.36	9.854	7.192	12204.04
R5X4	AF062029	-0.283	16869	17912.1	-46.2	117.81	154	159.5	154.8	8.649	5.4	17929.59
R5X4	AF062031	-0.289	16090	17095.2	-44.5	111.41	145.1	151.6	145.1	8.825	6.396	17112.75
R5X4	AF062033	-0.406	16292	17376.4	-64.2	113.82	144.5	153.8	152.5	9.326	9.392	17393.92
R5X4	AF107771,IDGP120:AAD49789	-0.252	91729	97742.9	-217	627.38	857	864.9	781.6	8.981	25.263	97758.52
R5X4	U08680	-0.446	9760	10370.1	-41	66.14	89.45	90.85	85.94	9.819	6.019	10387.83
R5X4	U08682	-0.435	9904	10502.3	-40	66.39	90.87	91.64	85.05	10.232	8.018	10520.04
R5X4	U08444,IDGP120:AAB04069.1	-0.242	92450	98534.9	-211	636.23	878.2	868.6	791.9	8.266	11.471	98550.33
R5X4	U08445	-0.194	89836	95587.5	-165	619.02	845.9	850.3	770.7	8.529	15.313	95603.16
R5X4	AF355674	-0.483	11657	12431.5	-53.1	79.31	106.7	110.1	102.4	9.806	6.356	12449.14

Tabla 8 Estadísticas de las Propiedades Moleculares para los Virus R5, X4 y R5X4

4.3 Aprendizaje

4.3.1 Aprendizaje No Supervisado

En este apartado usaremos aprendizaje no supervisado para predecir el tropismo Dual-tropico R5X4, ya que este tropismo puede usar como correceptor a CCR5 o CCRX4 por tal motivo comparte propiedades moleculares con ambos correceptores. Para esto se realizaron diferentes combinaciones de las propiedades moleculares (Tablas 5.1 y 5.2 y, Tablas 6.1 y 6.2) con la finalidad de obtener una predicción del tropismo Dual. La mejor combinación de las propiedades moleculares fue: volumen, oxigeno carbono e hidrogeno. Utilizamos 220 secuencias de virus, 78 corresponden al tropismo R5, 73 al tropismo X4 y 69 al Dual-trópico (R5X4) como se pudieron observar en las Tablas 1, 2 y 3. En la Tabla 9 se muestra un ejemplo de la base de datos utilizada, donde se puede observar cada secuencia de virus. Implementamos PCA para identificar dos grupos de virus. La idea es que PCA identifique las propiedades moleculares del R5 y las propiedades moleculares del X4 y Dual-trópico (R5X4), como Dual-trópico comparte características de R5 y X4 al clasificar dos grupos PCA identifica el coreceptor del Dual-trópico (R5X4) en su co-receptor CCR5 o CXCR4 (Figura 5).

TIPO	VOLUMEN	OXIGENO	CARBON	HIDROGENO
R5	9944	131	460	767
R5	9853	128	456	759
R5	7021	104	328	521
R5	11998	171	568	898
R5	9624	134	456	726
R5	9534	136	445	718
R5X4	73941	1242	4330	6851
R5X4	91065	1212	4424	6668
R5X4	88812	1222	4190	6788
R5X4	90232	1252	4327	6838
R5X4	85881	1238	4346	6853
R5X4	91330	1243	4395	6934
R5X4	92375	1251	4353	6859
X4	22314	313	1051	1673
X4	13916	195	660	1052
X4	21180	309	984	1567
X4	21477	319	1008	1573
X4	13785	202	684	1027
X4	12085	166	325	920
X4	17287	247	813	1292

Tabla 9 Propiedades Moleculares para los Virus R5, X4 y R5X4 que Mejor Identifican los Co-Receptores (CCR5 y CXCR4)

Una vez que el aprendizaje supervisado identifico las mejores propiedades biomoleculares: volumen, oxigeno, carbón e hidrogeno (ver Tabla 9).

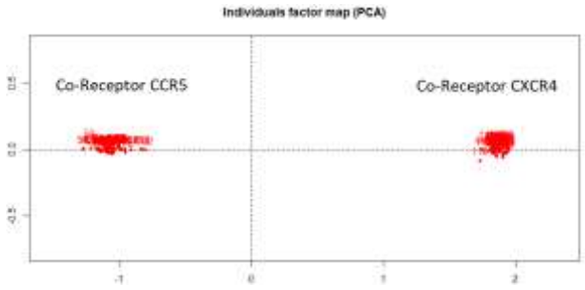

Figura 5 Identificando el Co-Receptor CCR5 o CXCR4 del Virus Dual-trópico (R5X4)

Con este dataset entrenamos cuatro algoritmos de aprendizaje supervisado: KNN, Arboles de Decisión, Naive Bayes y SVM. Para realizar el entrenamiento de cada uno de los algoritmos se utilizaron las mismas 220 secuencias de virus utilizadas en el aprendizaje no supervisado, 78 corresponden al tropismo R5, 73 al tropismo X4 y 69 al Dual-trópico (R5X4) como se pudieron observar en las Tablas 1, 2 y 3.

Utilizamos LOOCV para evaluar el porcentaje de asertividad del algoritmo utilizando la siguiente matriz de confusión:

	Predicción		
	R5	R5X4	X4
R5	16	0	0
R5X4	0	17	2
X4	0	1	14

Tabla 10 Matriz de Confusión

Como podemos observar en la Tabla 10, tenemos 16 secuencias de virus para el R5 de las cuales hubo 16 predicciones correctas y 0 errores, de las 19 secuencias de virus para el Dual-trópico, hubo 17 predicciones correctas y 2 errores, por último, las 15 secuencias de virus para el trópico X4, hubo 14 predicciones correctas y 1 error. Como podemos observar en la Tabla 11, cada clasificador se entrenó con el mismo dataset usando LOOCV y los mejores resultados los obtuvimos con la máquina de soporte vectorial (SVM).

Clasificador	Asertividad	Error
KNN	0.89	0.11
Arboles de Decisión	0.87	0.13
SVM	0.94	0.06
Bayesiano	0.67	0.33

Tabla 11 Predicción de los Clasificadores Usando LOOCV para el tropismo del VIH

ISSN- 2410-3551

3.6 SVM en el paralelo para optimizar la predicción de los co-receptores de los virus R5, X4 Y R5X4. En este apartado se mostrarán los resultados para optimizar el tiempo de respuesta en la predicción de co-receptores de los virus R5, X4 y R5X4 al evaluar la SVM en Paralelo, nuestro modelo clasifica 10000 virus usando la técnica de evaluación LOOCV. Tomando en consideración la matriz de confusión descrita previamente (ver Tabla 10) pudimos evaluar la predicción de R5, X4 y Dual-trópico (R5X4) de este clasificador en paralelo. En la Tabla 10 podemos observar el tiempo de procesamiento en segundos, que tarda la SVM en realizar la predicción en forma secuencial y paralelo, iniciamos con 1000 virus, continuamos con incrementos de 1000 hasta lograr el análisis de 10,000 virus R5, X4 y R5X4, que predicen los co-receptores CCR5 y CXCR4.

Virus	Proceso Secuencial	Proceso Paralelo	Predicción de los Co-Receptores (CCR5 y CXCR4)	
X4, R5 y R5X4	Tiempo en Segundos	Tiempo en Segundos	Asertividad	Error
1000	141.34	80.06	94%	0.06%
2000	464.72	273.36	94%	0.06%
3000	1452.17	806.76	94%	0.06%
4000	3181.52	1871.48	94%	0.06%
5000	5971.47	3317.48	94%	0.06%
6000	9554.35	5307.97	94%	0.06%
7000	14331.57	7961.98	94%	0.06%
8000	20064.19	11560.1	94%	0.06%
9000	26076.8	13376.28	94%	0.06%
10000	33899.84	18833.24	94%	0.06%

Tabla 12 Predicción de Co-Receptores CCR5 y CXCR4 en Procesos Secuenciales y Paralelos

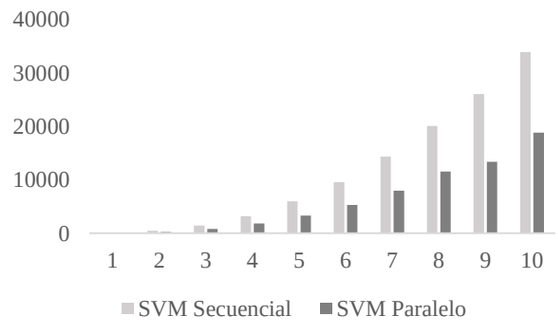

Figura 6 Tiempos de Ejecución de la SVM en Paralelo para Detectar el Tropismo Viral de los Co-Receptores CCR5 y CXCR4

En la gráfica de la Figura 6, claramente observamos que ganamos un porcentaje considerable de tiempo, por lo que podemos decir que la SVM en Paralelo optimiza el tiempo de respuesta obteniendo información verídica y útil para proveer a los especialistas herramientas clínicas importantes para una rápida y eficiente detección automatizada del tropismo viral.

5. Conclusiones

En particular, los virus R5 HIV-1 usan CCR5 como un co-receptor para la entrada viral, los virus X4 HIV-1 usan el CXCR4, mientras algunos extraños virus conocidos como R5X4 o D-tropic, tienen la habilidad de utilizar ambos co-receptores. Los virus X4 y R5X4 son asociados con un rápido progreso en el VIH-1. En este artículo se realizaron una serie de experimentos para implementar una máquina de aprendizaje supervisado en paralelo que permita optimizar el tiempo de respuesta en la predicción de co-receptores (CCR5, CXCR4) del Virus que causan el sida (VIH-1) en células CD4. Para implementar la máquina de aprendizaje supervisado en paralelo utilizamos Snow en R. Snow provee el soporte para fácilmente ejecutar funciones en R en paralelo. La mayoría de las funciones en paralelo en Snow son variaciones de la función estándar lapply(). Para implementar las funciones en paralelo, Snow utiliza una arquitectura, maestro/esclavo donde el maestro envía tareas a los trabajadores, y los trabajadores ejecutan las tareas y retornan los resultados al maestro.

Podemos concluir en esta investigación que logramos obtener una prueba de tropismo que es rápida, segura y de bajo coste accesible para laboratorios de diagnóstico, dando a conocer que el uso de la bioinformática es una herramienta de confianza para el desarrollo de investigaciones biomédicas ya que los ligados naturales del CXCR4 y CCR5 pueden inhibir la entrada viral del VIH.

6. Trabajo futuro

Existen varias rutas a explorar como trabajo futuro a partir de los resultados obtenidos; principalmente los relacionados con las propiedades moleculares de los virus del VIH. Mas virus X4, R5 y Dual-trópico (R5X4) y más herramientas de clasificación en paralelo que permitan optimizar el tiempo de respuesta serán estudiadas para entender mejor el tropismo viral.

ISSN- 2410-3551

7. Referencias

[1] Lara Villegas Humberto H, Ixtepan Liliana del C., Rodríguez Padilla Cristina. El Tropismo y su Identificación. Reporte Técnico, Laboratorio de Bioseguridad Nivel III. Departamento de Inmunología y Virología. Facultad de Ciencias Biológicas. Universidad Autónoma de Nuevo León, México.

[2] International Comittee on Taxonomy of Viruses (ICTV-2019): http://ictvonline.org/index.asp.

[3] Carroll Karen C, Morse Stephen A, Mietzner Timothy A, Miller Steve. Medical microbiology. 27th ed. McGraw-Hill Education; 2016. ISBN 9780-0-71-82498-9.

[4] Denis, F., Leonard, G., Sangare, A., et al. Comparison of 10 Enzyme Immunoassays for Detection of Antibody to Human Immunodeficiency Virus Type 2 in West Africa Sera. J Clin Microbiol 1988; 26(5):1000-4.

[5] Viral Count Analysis (Roche Labs 2015). http://www.roche.com/index.htm.

[6] Viral Count Analysis (Biomerieux Labs 2015). http://www.biomerieux.com/.

[7] Analysis of CD4 cells (Infonet AIDS 2015). http://aidsinfonet.org/.

[8] Theodoridis, S. Y Koutroumbas, K., 2006. Pattern Recognition, Third Edition [en línea]. 2006. S.l.: s.n. ISBN 0123695317. Disponible en: http://www.amazon.com/Pattern-Recognition-Edition-Sergios-Theodoridis/dp/1597492728.

[9] Lamers Susanna L, Salemi Marco, McGrath Michael S, Fogel Gary B. Prediction of R5, X4, and R5X4 HIV-1
Coreceptor Usage with Evolved Neural Networks. IEEE/ACM Transactions on Computational Biology and
Bioinformatics April-June 2008;5(2):253.

[10] The National Center for Biotechnology Information Advances Science and Health by Providing Access to
Biomedical and Genomic Information. Data Base On-Line (NCBI-2020). http://www.ncbi.nlm.nih.gov/.

[11] The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. Data Base On-Line (UniProtKB-2020). https://www.uniprot.org/uniprot/.

[12] The HIV database. The HIV databases contain comprehensive data on HIV genetic sequences and immunological epitopes (The HIV database 2020). https://www.hiv.lanl.gov/content/index

[13] ExPASy. Is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence (ExPASy database 2020). https://web.expasy.org/protparam/

[14] Prot pi. Protein Tool is a web application for calculating physico-chemical parameters of proteins and peptides (Prot pi 2020). https://www.protpi.ch/Calculator.

ISSN- 2410-3551