EXÁMEN FINAL

Predicción del uso del correceptor **R5**, **X4** y **R5X4** **VIH-1**

Benemérita Universidad Autónoma de Aguascalientes

Centro de Ciencias Básicas

Asignatura:

"Inteligencia Artificial"

Maestro:

Francisco Javier Luna Rosas

Alumnos:

- José Alfredo Díaz Robledo
- Luis Pablo Esparza Terrones
- Luis Manuel Flores Jiménez
- Juan Francisco Gallo Ramírez

Ingeniería en Computación Inteligente

3er Semestre

Introducción

En el ámbito de la investigación del VIH-1, la identificación y comprensión del tropismo dual R5X4, caracterizado por la capacidad del virus para utilizar los co-receptores CCR5 y CXCR4 durante la entrada en las células CD4 del sistema inmune, representa un área crítica. Este proyecto se propone abordar esta complejidad mediante un enfoque integral que combina el diseño de un dataset estructurado, la extracción de información molecular de secuencias de aminoácidos, y la implementación de técnicas de aprendizaje automático.

La primera fase del proyecto se centra en la creación del dataset, incorporando estadísticas generales y propiedades de aminoácidos calculadas a través de códigos específicos y recursos en línea. Este conjunto de datos será fundamental para la construcción y entrenamiento de modelos de aprendizaje automático en fases posteriores.

La segunda etapa del proyecto implica el uso de aprendizaje supervisado y no supervisado para predecir el tropismo dual R5X4 del VIH-1. Se exploran diversas combinaciones de propiedades moleculares de aminoácidos, y se emplea un enfoque de selección de características aleatorio para encontrar un conjunto óptimo que permita alcanzar una precisión global mínima del 90%.

La evaluación de la precisión del modelo es esencial en la tercera fase. Se utilizan matrices de confusión para analizar los resultados y se buscan características que garanticen la consecución de los objetivos de precisión global establecidos. Este enfoque sistemático y riguroso busca contribuir al entendimiento de las interacciones entre el VIH-1 y las células del sistema inmune, proporcionando una herramienta predictiva robusta para el comportamiento dual-trópico de las variantes R5X4.

En resumen, este proyecto representa un paso significativo hacia la comprensión y predicción de fenómenos clave en la infección por VIH-1, utilizando enfoques multidisciplinarios que integran biología molecular, estadísticas y aprendizaje automático.

Análisis del Proyecto

Para realizar este algoritmo de aprendizaje automático, fue necesario consultar un artículo otorgado por el profesor. El artículo, publicado en la revista IEEE/ACM Transactions on Computational Biology and Bioinformatics, se centra en la predicción del uso de coreceptores R5, X4 y R5X4 en el VIH-1 mediante el empleo de Redes Neuronales Evolutivas. Los autores, Susanna L. Lamers, Marco Salemi, Michael S. McGrath y Gary B. Fogel, presentan una investigación que se relaciona con el análisis y la predicción de tropismos del VIH-1 utilizando enfoques computacionales avanzados, con un énfasis particular en las Redes Neuronales Evolutivas como herramienta predictiva.

En dicho artículo se describe la recopilación de secuencias del bucle V3 del VIH, abarcando un total de 149 aislamientos virales. Estas secuencias representan tres tropismos virales determinados experimentalmente, siendo 77 secuencias R5, 41 secuencias X4 y 31 secuencias R5X4. Estas secuencias fueron obtenidas de diversos subtipos de VIH y fueron identificadas a partir de la base de datos del Laboratorio Nacional de Los Álamos para Secuencias de VIH. Se muestra la tabla contenida en el artículo:

R5X4 (D-tropic)	R5		X4
AB014795 AF062029 AF062031 AF062033 AF107771 U08680 U08682 U08444 U08445 AF355674 AF355647 AF355630 AF355690 M91819 AF035532 AF035533 AF259019 AF259025 AF259021 AF259041 AF258970 AF258978 AF021607 AF204137 AF112925 M17451 K02007 U39362 AF069140 AF458235 AF005494	AF062012 L03698 AF231045 AY669778 U08810 U51296 AF407161 AB253421 U08645 U08647 U08795 AB253429 AY288084 AF307753 AF411964 U08823 AF411965 U92051 AF355318 AY010759 AY010804 AY010852 U08670 U08798 AY669715 U08710 U16217 M26727 AJ418532 AJ418479 AJ418495 AJ418514 AJ418521 U23487 U04900 AF022258 AF258957 AF021477	U08716 U39259 AF204137 M38429 U27443 U79719 U04909 U04918 U04908 U08450 AF112542 M63929 U66221 AF491737 U08779 L22084 U27413 AF005495 U52953 AF321523 L22940 U45485 AB023804 U08453 AF307755 AF307750 AY043176 AY158534, AX455917 AY043173 AF307757 U08803 U88824 U69657 AF355326 U88826 U08368 U27426 AJ006022	AB014785 AB014791 AB014796 AB014810 U48267 U08666 AF069672 AF355319 AF355336 M14100 A04321 X01762 L31963 U08447 AF355660 AF355748 AF355742 AF355706 AF180915 AF180903 AF035534 AF259050 AF258981 AF259003 AF021618 AF128989 M17449 AF075720 U48207 U72495 AY189526 AF034375 AF034376 U27408 AF411966 U27399 U08822 U08738 U08740 U08193 AF355330

Obtención del DataSet

En base a estos identificadores se obtuvo la cadena de aminoácidos en la página https://www.ncbi.nlm.nih.gov usando web scraping. Con la cadena obtenida, se procedió a calcular más características de esta usando propiedades que contienen los aminoácidos, donde dichas propiedades fueron obtenidas en el mismo artículo mencionado. Se muestran de igual forma en tabla:

Amino acid residues	Chemical property	Charge	Volume (A3)	Mass (daltons)	HP Scale	Surface Area	2D structure propensity		
							alpha helix	B-strand	Turn
Alanine (A)	aliphatic	0	67	71.09	1.8	0.74	1.41	0.72	0.82
Arginine I	basic	+1	148	156.19	-4.5	0.64	1.21	0.84	0.90
Asparagine (N)	amide	0	96	114.11	-3.5	0.63	0.76	0.48	1.34
Aspartic Acid (D)	acidic	-1	91	115.09	-3.5	0.62	0.99	0.39	1.24
Cysteine(C)	reactive	0	86	103.15	2.5	0.91	0.66	1.40	0.54
Glutamine (Q)	amide	0	114	128.14	-3.5	0.62	1.27	0.98	0.84
Glutamic Acid (E)	acidic	-1	109	129.12	-3.5	0.62	1.59	0.52	1.01
Glycine (G)	small	0	48	57.05	-0.4	0.72	0.43	0.58	1.77
Histidine (H)	aromatic	0	118	137.14	-3.2	0.78	1.05	0.80	0.81
Isoleucine (I)	aliphatic	0	124	113.16	4.5	0.88	1.09	1.67	0.47
Leucine (L)	aliphatic	0	124	113.16	3.8	0.85	1.34	1.22	0.57
Lysine (K)	basic	+1	135	128.17	-3.9	0.52	1.23	0.69	1.07
Methionine (M)	aliphatic	0	124	131.19	1.9	0.85	1.30	1.14	0.52
Phenylalanine (F)	aromatic	0	135	147.18	2.8	0.88	1.16	1.33	0.59
Proline (P)	cyclic imino	0	90	97.12	-1.6	0.64	0.34	0.31	1.32
Serine (S)	hydroxyl	0	73	87.08	-0.8	0.66	0.57	0.96	1.22
Threonine (T)	hydroxyl	0	93	101.11	-0.7	0.70	0.76	1.17	0.90
Tryptophane (W)	aromatic	0	163	186.21	-0.9	0.85	1.02	1.35	0.65
Tyrosine (Y)	aromatic	0	141	163.18	-1.3	0.76	0.74	1.45	0.76
Valine (V)	aliphatic	0	105	99.14	4.2	0.86	0.90	1.87	0.41

Además, con la cadena obtenida, se extrajeron otras propiedades complementarias de la página https://www.protpi.ch/Calculator/ProteinTool.

En resumen, las propiedades que se obtuvieron para un dataset general fueron:

- **Cantidad de aminoácidos** contenidos en la cadena de: *Alanina, Arginina, Asparagina, Ácido Aspártico, Cisteína, Glutamina, Acido Glutámico, Glicina, Histidina, Isoleucina, Leucina, Lisina, Metionina, Fenilalanina, Prolina, Serina, Treonina, Triptófano, Tirosina y Valina.*
- **Porcentaje de aminoácidos** contenidos en la cadena de: *Alanina, Arginina, Asparagina, Ácido Aspártico, Cisteína, Glutamina, Acido Glutámico, Glicina, Histidina, Isoleucina, Leucina, Lisina, Metionina, Fenilalanina, Prolina, Serina, Treonina, Triptófano, Tirosina y Valina.*
- **Número de aminoácidos totales.**
- **Volumen total.**
- **Masa total.**
- **Área de superficie total.**

- **Punto isoeléctrico.**
- **Carga neta en pH 7.4.**

Estas 46 características fueron las usadas para para el dataset general.

Una vez con el dataset general de los tropismos del tipo R5 y X4, se procedió a realizar un algoritmo de prueba para la combinación de columnas adecuada, es decir, se buscó una combinación de características que una vez entrenada la red neuronal resultará de más del 90% de precisión. Las columnas que nos otorgaron este resultado fueron guardadas en un dataset nuevo, al que denominamos “dataset óptimo”.

Entrenamiento de Red Neuronal

Ya obtenida la combinación de columnas adecuada, se procedió a volver a entrenar la red neuronal con dicho dataset óptimo para finalmente testear los tropismos R5X4.

Predicción de tropismos R5X4

Finalmente se testearon estos tropismos, ya que la finalidad del algoritmo era obtener una predicción del correceptor al que irían dichos tropismos.

Algoritmo completo

Inicialización.

1. **Importación de Librerías:**

Se importan diversas bibliotecas como NumPy, OpenPyXL, Pandas, Requests, BeautifulSoup, Colorama, Selenium, y Scikit-Learn. Estas librerías se utilizan para manipular datos, realizar scraping web, implementar la red neuronal, y evaluar el rendimiento del modelo.

2. **Diccionarios de Características de Aminoácidos:**

Se definen diccionarios que contienen características de aminoácidos como masa, área de superficie y volumen.

Web Scraping

3. **Creación de un DataSet General:**

Se crea un archivo Excel (DataSet.xlsx) que servirá como el conjunto de datos principal para el proyecto.

4. **Carga de Identificadores:**

Se cargan identificadores de tropismos R5, X4 y R5X4 desde un archivo Excel (NumeroAccesoSecuencias.xlsx).

5. **Función de Búsqueda de Datos (Web Scraping):**

La función getIdenData realiza web scraping para obtener la cadena de aminoácidos y otras propiedades relevantes, utilizando identificadores proporcionados.

6. **Web Scraping para Identificadores R5 y X4:**

Se realizan operaciones de web scraping para obtener datos asociados a identificadores R5 y X4, y se almacenan en el DataSet.xlsx.

Red Neuronal

7. **Obtención de Columnas Óptimas:**

Se realiza una combinación aleatoria de columnas del conjunto de datos para obtener un modelo de red neuronal con una precisión superior al 90%.

8. **Entrenamiento de la Red Neuronal:**

Se entrena una red neuronal utilizando las características óptimas seleccionadas.

9. **Índices de Calidad del Modelo:**

Se calculan e imprimen índices de calidad del modelo, como la matriz de confusión, la precisión y el error globales.

10. **Predicción de Tropismos R5X4:**

Se utiliza la red neuronal entrenada para predecir el tropismo de VIH-1 para identificadores de tropismos R5X4.

Implementación

La implementación se realizó mediante el uso de Jupyter Notebooks, usando lenguaje Python. En el link del archivo que se presenta a continuación se expone el algoritmo asi como el análisis a detalle de la red neuronal.

Algoritmo Autorreproducible

- ExámenFinal_IA.html

Archivo con los identificadores

- NumeroAccesoSecuencias.xlsx

DataSet General

- DataSet.xlsx

DataSet Óptimo

- DataSetOptimo.xlsx

Prueba y Resultados

- Con base al algoritmo y al web scraping se obtuvo un dataset general que se muestra a continuación (Se muestran todas sus columnas, pero solo una muestra de sus filas):

#Ala (A)	%Ala (A)	#Arg (R)	%Arg (R)	#Asn (N)	%Asn (N)	#Asp (D)	%Asp (D)	#Cys (C)	%Cys (C)
147	41.17647	0	0	0	0	0	0	55	15.40616
890	34.82003	12	0.469484	0	0	0	0	434	16.97966
46	43.80952	0	0	0	0	1	0.952381	18	17.14286
898	35.00975	1	0.038986	0	0	0	0	432	16.84211
820	34.95311	0	0	0	0	0	0	412	17.56181

#Glu (E)	%Glu (E)	#Gln (Q)	%Gln (Q)	#Gly (G)	%Gly (G)	#His (H)	%His (H)	#Ile (I)	%Ile (I)
0	0	0	0	71	19.88795518	0	0	0	0
0	0	0	0	578	22.61345853	0	0	0	0
0	0	0	0	19	18.0952381	0	0	0	0
0	0	0	0	595	23.19688109	0	0	0	0
0	0	0	0	540	23.01790281	0	0	0	0

#Leu (L)	%Leu (L)	#Lys (K)	%Lys (K)	#Met (M)	%Met (M)	#Phe (F)	%Phe (F)	#Pro (P)	%Pro (P)
0	0	0	0	0	0	0	0	0	0
0	0	1	0.039123631	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

#Ser (S)	%Ser (S)	#Thr (T)	%Thr (T)	#Trp (W)	%Trp (W)	#Tyr (Y)	%Tyr (Y)	#Val (V)	%Val (V)
0	0	84	23.52941176	0	0	0	0	0	0
1	0.039123631	637	24.92175274	1	0.039123631	2	0.078247261	0	0
0	0	21	20	0	0	0	0	0	0
0	0	638	24.87329435	0	0	1	0.038986355	0	0
0	0	574	24.46717818	0	0	0	0	0	0

#Aminoacid	Volume	Mass	Surface Area	Isoelectric point	Net charge at pH 7.4	Correceptor
357	25799	28667.27	268.75	5.113	-6.848	R5
2556	186368	208021.27	1926.83	6.766	-37.866	R5
105	7586	8449.19	79.42	3.819	-3.802	R5
2565	185501	207171.92	1934.04	5.646	-49.717	R5
2346	169674	189635.74	1772.32	4.368	-48.337	R5

- Después de realizar una búsqueda de columnas óptimas se obtuvieron los resultados, donde se muestran las columnas adecuadas para una precisión superior a 90%.:

```
>>> CARACTERÍSTICAS ÓPTIMAS <<<

['#Ala (A)' '#Trp (W)' '%Ala (A)' '#Phe (F)' '%Asn (N)']

- Índices obtenidos:

> Matriz de Confusión:
[[25  0]
 [ 3  8]]

> Precisión Global:
0.9166666666666666

> Error Global:
0.08333333333333337
```

- Así pues, y con las columnas óptimas, se realizó el dataset óptimo. Se muestra a continuación (Se muestran todas sus columnas, pero solo una muestra de sus filas):

#Ala (A)	#Trp (W)	%Ala (A)	#Phe (F)	%Asn (N)	Correceptor
147	0	41.17647059	0	0	R5
890	1	34.8200313	0	0	R5
46	0	43.80952381	0	0	R5
898	0	35.00974659	0	0	R5
820	0	34.95311168	0	0	R5

- Finalmente, con el dataset adecuado para entrenar la red neuronal, se procedió a realizar las predicciones, de los tropismos R5X4 fueron las siguientes:

```
>>> PREDICCION PARA TROPISMOS R5X4: <<<

AB014795        Correceptor: ['X4']
AF062029        Correceptor: ['X4']
AF062031        Correceptor: ['X4']
AF062033        Correceptor: ['X4']
AF107771        Correceptor: ['R5']
U08680  Correceptor: ['R5']
U08682  Correceptor: ['R5']
U08444  Correceptor: ['R5']
U08445  Correceptor: ['R5']
AF355674        Correceptor: ['X4']
AF355647        Correceptor: ['X4']
AF355630        Correceptor: ['X4']
AF355690        Correceptor: ['X4']
M91819  Correceptor: ['R5']
AF035532        Correceptor: ['R5']
AF035533        Correceptor: ['R5']
AF259019        Correceptor: ['X4']
AF259025        Correceptor: ['X4']
AF259021        Correceptor: ['X4']
AF259041        Correceptor: ['X4']
AF258970        Correceptor: ['X4']
AF258978        Correceptor: ['X4']
AF021607        Correceptor: ['R5']
AF204137        Correceptor: ['X4']
AF112925        Correceptor: ['R5']
M17451  Correceptor: ['R5']
K02007  Correceptor: ['R5']
U39362  Correceptor: ['R5']
AF069140        Correceptor: ['R5']
AF458235        Correceptor: ['X4']
AF005494        Correceptor: ['R5']
```

Conclusiones

En este proyecto multidisciplinario de investigación del VIH-1, se ha abordado la complejidad del tropismo dual R5X4 mediante un enfoque integral que combina la biología molecular, la estadística y el aprendizaje automático. La creación de un dataset estructurado, obtenido a través de scraping web, ha permitido la recopilación de información molecular crucial para el diseño y entrenamiento de modelos de aprendizaje automático.

La implementación de una red neuronal, basada en el análisis de un artículo de referencia, ha demostrado ser una herramienta poderosa para la predicción del tropismo dual R5X4 del VIH-1. La selección óptima de características ha llevado a un modelo con una precisión global superior al 90%, lo cual es un logro significativo en la predicción de fenómenos clave en la infección por VIH-1.

La evaluación rigurosa del modelo, utilizando matrices de confusión y otros índices de calidad, ha proporcionado una comprensión profunda de su desempeño y validez. Este enfoque sistemático no solo contribuye al entendimiento de las interacciones VIH-1/células inmunes, sino que también ofrece una herramienta predictiva robusta para el comportamiento dual-trópico de las variantes R5X4.

La implementación del algoritmo en Jupyter Notebooks y la presentación clara de resultados, junto con la disponibilidad de un dataset general y óptimo, hacen que este proyecto sea autorreproducible y accesible para futuras investigaciones. En resumen, este trabajo representa un paso significativo hacia la comprensión y predicción de fenómenos clave en la infección por VIH-1, destacando la eficacia de la integración de enfoques multidisciplinarios en la investigación científica.

Referencias

Lamers, S. L., Salemi, M., McGrath, M. S., & Fogel, G. B. (2008). Prediction of R5, X4, and R5X4 HIV-1 Coreceptor Usage with Evolved Neural Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 291-300.

National Library of Medicine. (s.f.). *National Center for Biotechnology Information*. Obtenido de https://www.ncbi.nlm.nih.gov/

Prot pi. (2023). *Prot pi[pra□t pa□] | Protein Too*. Obtenido de https://www.protpi.ch/Calculator

scikit-learn. (s.f.). *scikit-learn*. Obtenido de Machine Learnign in Python: https://scikit-learn.org/stable/index.html