

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE CIENCIAS QUÍMICAS Y FARMACIA



Identificación de posibles fármacos para el tratamiento multi-diana de cáncer de próstata a través de los receptores nucleares huérfanos, mediante el uso de máquina de soporte vectorial.

Boris Iván Muñoz Debroy
Químico Farmacéutico

Guatemala, Octubre de 2017

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE CIENCIAS QUÍMICAS Y FARMACIA

Identificación de posibles fármacos para el tratamiento multi-diana de cáncer de próstata a través de los receptores nucleares huérfanos, mediante el uso de máquina de soporte vectorial.

Informe de Tesis

Presentado por:

Boris Iván Muñoz Debroy

Para optar al título de

Químico Farmacéutico

Guatemala, Octubre de 2017

JUNTA DIRECTIVA

Dr. Rubén Dariel Velásquez Miranda	Decano
M.A. Elsa Julieta Salazar Meléndez de Arriaza	Secretaria
MSc. Miriam Carolina Guzmán Quilo	Vocal I
Dr. Juan Francisco Pérez Sabino	Vocal II
Lic. Carlos Manuel Maldonado Aguilera	Vocal III
Br. Andreina Delia Irene López Hernández	Vocal IV
Br. Carol Andrea Betancourt Herrera	Vocal V

AGRADECIMIENTOS

A la Universidad de San Carlos de Guatemala por permitirme la oportunidad de crecer académicamente y desarrollarme como profesional.

A la Facultad de Ciencias Químicas y Farmacia, en especial a la Escuela de Química Farmacéutica por brindarme los medios necesarios para adquirir conocimiento y desarrollar mi perfil profesional.

A mis asesores, Licenciado Omar Ernesto Velásquez González y MSc. Lucrecia Peralta de Madriz por orientarme en la elaboración del trabajo de investigación, su paciencia y excelente disposición todo el tiempo.

A mi revisora, MSc. María Nereida Marroquín Tintí por su colaboración y consejo en la elaboración del trabajo de investigación.

A Rony Letona por orientarme en los intrincados procesos técnicos del presente y brindarme su amistad.

A aquellos catedráticos de la facultad que me brindaron sus conocimientos poniendo su entera dedicación en este proceso.

A aquel personal administrativo que me atendió con amabilidad y ayudo en las gestiones como debe de ser.

A todas aquellas personas que de una manera u otra han contribuido en la culminación de esta etapa de mi carrera.

DEDICATORIA

A Dios por bendecirme con las personas que me rodean y permitirme disfrutar de lo que poseo.

A mis padres, por ser los artífices de este proyecto de vida, por apoyarme siempre de manera incondicional y por brindarme todo lo necesario para llegar hasta este punto.

A mis hermanos por compartir conmigo incontables buenos y malos momentos a lo largo de mi vida.

A mis abuelitos y abuelitas por ser los pilares fundamentales de mi vida y darme su amor y cariño.

A mi compañera de vida por su apoyo incondicional durante la carrera, por ser mi consejera y mi todo.

A mis tíos y primos porque de una u otra forma han contribuido animándome a seguir adelante.

A mis amistades, que siempre me han motivado y disfrutado a mi lado.

ÍNDICE

1. RESUMEN	1
2. INTRODUCCIÓN	3
3. ANTECEDENTES	6
3.1. Mecanismo multi-diana	6
3.2. Receptores huérfanos asociados al cáncer	7
3.3. La máquina de soporte vectorial	9
3.4. Bases de datos	12
4. JUSTIFICACIÓN	13
5. OBJETIVOS	14
6. HIPÓTESIS	15
7. MATERIALES Y MÉTODOS	16
7.1. Universo	16
7.1.1. Población:	16
7.1.2. Muestra:.....	16
7.2. Materiales	16
7.2.1 Materiales y Equipos:.....	16
7.3. Métodos:	17
7.3.1 Obtención y preparación de archivos de trabajo desde bases de datos en línea:	17
7.3.2 Creación de script de programación para la adecuación de los datos y entrenamiento de la Máquina de Soporte Vectorial:.....	19
7.3.3 Preparación y obtención de los descriptores de los receptores del conjunto de entrenamiento y de prueba	19
7.3.4 Preparación y obtención de los descriptores de las moléculas del conjunto de entrenamiento y de prueba	21
7.3.5 Clasificación de las moléculas mediante la máquina de soporte vectorial	23
7.3.6 Análisis de Datos.....	23

7.4 Diseño de la investigación	24
8. RESULTADOS.....	27
8.1. Recuento de la clasificación de las moléculas de estudio para cada receptor nuclear.....	27
8.2. Moléculas con los mejores perfiles multi-diana.....	28
8.3. Características de las moléculas según sus perfiles multi-diana.....	29
8.4. Desempeño y validación del modelo de máquina de soporte vectorial	30
8.5. Características de la molécula con el mejor perfil multi-diana.....	31
9. DISCUSIÓN	32
10. CONCLUSIONES.....	40
11. RECOMENDACIONES	41
12. REFERENCIAS	42
13. ANEXOS.....	46
13.1. Marco Teórico	46
13.2. Kernel de Tanimoto	48
13.3. Screenshots de las páginas y programas utilizados.....	48
13.3.1. Base de datos DUD-E (http://dude.docking.org/)	48
13.3.2. Base de datos RCSB PDB (Protein Data Bank) (https://www.rcsb.org/pdb/home/home.do)....	49
13.3.3. Atomic Charge Calculator (https://webchem.ncbr.muni.cz/Platform/ChargeCalculator)	49
13.3.4. SwissSimilarity (http://www.swiss similarity.ch/)	50
13.3.5. KNIME	51
13.3.6. Script para caracterización de sitio activo	52
13.3.7. Selección de sitio activo y valores de cargas.....	52
13.3.8. Generación de matrices de cargas.....	53
13.3.9. Orientación de moléculas	53
13.3.10. Script para análisis de datos multi-diana	54
13.3.11. Evaluador de desempeño mediante validación cruzada	54

1. RESUMEN

La búsqueda constante de nuevos fármacos representa una fracción esencial de las investigaciones que se llevan a cabo dentro de las ciencias farmacéuticas; este proceso ha evolucionado a lo largo del tiempo, junto con otras disciplinas como la química, medicinal, la bioquímica y las ciencias computacionales principalmente; incorporando así nuevos conocimientos, enfoques y herramientas. La química computacional ha evolucionado permitiendo desarrollar nuevas técnicas y herramientas, siendo estas más poderosas y eficientes en la actualidad.

Mediante estos avances en la química computacional fue posible llevar a cabo el estudio de la identificación de moléculas con potencial farmacológico multi-diana sobre los principales receptores nucleares huérfanos que se encuentran asociados al cáncer de próstata (ERR1, ERR2, Nurr77, ROR1 y ROR3). Para esto se empleó la técnica de Aprendizaje de Máquina (Machine Learning) denominada Máquina de Soporte Vectorial (SVM por sus siglas en inglés), la cual se basa en la construcción de un modelo matemático que permite resolver problemas de clasificación generando un plano separador entre los datos y maximizando su distancia respecto a dicho plano a través de una dimensión superior a la que se encuentran inicialmente los datos. Como descriptores de los sitios activos y de las moléculas se utilizaron matrices de cargas de 1000 puntos representando la magnitud de carga en cada punto equidistante dentro de un espacio de 20 Å cúbicos. Para el entrenamiento del modelo se utilizaron los 11 receptores nucleares (y sus moléculas) pertenecientes a la base de datos DUD-E la cual está constituida por 102 receptores bien estudiados que pertenecen a distintas clases y para los que se presentan dentro de la misma más de 400 moléculas incluyendo activas e inactivas.

El resultado de la validación cruzada aleatoria de la SVM fue de aproximadamente 80% sobre las 4 mediciones calculadas, siendo estas la exhaustividad que refleja la sensibilidad del modelo, la precisión que refleja la dispersión de los resultados, la

especificidad que indica para que elementos de estudio presenta mejor desempeño y el Valor F (F-measure) que refleja la exactitud general. Esto permite observar que el modelo a pesar de tener un margen de error mejorable, este permiten clasificar y correlacionar receptores y ligandos más allá de un resultado aleatorio al superar el 50% en los 4 puntajes. Se evaluaron 1380 compuestos derivados del colesterol obtenidos a través del buscador SwissSimilarity, el cual posee acceso a colecciones de moléculas farmacológicas, comerciales y bioactivas. De esta manera las moléculas en estudio fueron seleccionados debido a la estrecha relación de los compuestos esteroideos con los receptores huérfanos. Más del 94% de las moléculas de prueba fueron clasificadas como activas para los receptores ERR1, ERR3 y Nurr77, mientras que para el receptor ROR1 se obtuvo menos del 84% y para ROR3 menos del 75% distanciándolos del resto. Se determinó que la molécula ZINC70691822 posee el mejor perfil multi-diana al lograr el mejor puntaje sobre los 5 receptores evaluados y presentar las características propias de los compuestos clasificados como activos al poseer una mayor proporción de cargas positivas frente a las negativas y la distribución de cargas requerida, lo que indica que sería un candidato a ligando de dichos receptores huérfanos.

2. INTRODUCCIÓN

El estudio de las interacciones de las sustancias químicas con el organismo es fundamental en la química medicinal y la farmacología permitiendo el descubrimiento y desarrollo de nuevos fármacos. La búsqueda de principios activos es indispensable para lograr avances sobre trastornos o patologías humanas para las que no existe aún un tratamiento efectivo o, que a pesar de tenerlo, este es aún susceptible de mejora. En este sentido algunas terapias actualmente presentan diversos problemas tales como resistencia o dependencia farmacológica, severos efectos adversos, bajos índices terapéuticos, bajos niveles de eficacia o falta de efectividad entre otros.

Un enfoque tradicional que explica el efecto de las sustancias terapéuticas en el organismo es el de fármaco-receptor, el cual representa en forma de analogía al fármaco como una llave con las características específicas que le permiten actuar sobre un tipo específico de receptor, el cual representa la cerradura y que, mediante dicha interacción, se desencadenan los efectos terapéuticos propios del fármaco. Debido a lo anterior, predomina la idea de que un único fármaco suele interactuar sobre un único receptor de forma selectiva, sin embargo nuevos estudios han sugerido que es posible diseñar fármacos que presenten una interacción multi-receptor logrando a través de esto un efecto terapéutico mejorado, superando muchos de los problemas que se plantearon anteriormente.

En el estudio del cáncer se han dilucidado distintos receptores y ligandos relacionados a su desarrollo; gracias a esto ha sido posible el descubrimiento de muchos fármacos que actúan en sus rutas metabólicas y constituyen parte de las terapias de elección. Actualmente se conocen 25 receptores nucleares humanos denominados huérfanos, debido a que no poseen ningún ligando o fármaco asociado. Sobre estos receptores existen estudios que revelan que algunos se encuentran particularmente vinculados al cáncer de próstata (ERR1, ERR3, Nurr77, ROR1 y ROR3), lo que representa

una oportunidad terapéutica si se encuentran compuestos que modulen la actividad de los mismos (Baek & Kim, 2014).

La química computacional posee una amplia gama de metodologías y herramientas esenciales para el diseño racional de fármacos. A través de la misma es posible llevar a cabo procedimientos altamente complejos de forma eficiente que facilitan el descubrimiento de nuevos fármacos. Dentro de la química computacional existen varias metodologías relacionadas bajo el nombre de Aprendizaje de Máquina (Machine Learning) con las cuales se busca llevar a cabo el entrenamiento de un modelo computacional para que sea capaz de realizar asociaciones entre los datos y establecer relaciones lógicas, emulando en parte la acción del aprendizaje humano, siendo finalmente capaces de emitir una respuesta en base a lo “aprendido”. De estas técnicas la denominada Máquina de Soporte Vectorial (Support Vector Machine o SVM por sus siglas en inglés) ha dado muy buenos resultados al ser sumamente eficiente en comparación con otras técnicas similares tales como las redes neuronales, arboles de decisiones o redes bayesianas entre otras. Dicha técnica se basa en la construcción de un modelo matemático a partir de datos de referencia que es capaz de resolver problemas de clasificación generando un plano separador que maximiza la distancia entre los datos a través de una dimensión superior.

En la presente investigación se llevó a cabo la implementación y entrenamiento de una SVM con el fin de identificar sobre una base de datos de 1380 compuestos aquellos con posible actividad multi-diana sobre los receptores nucleares huérfanos ERR1, ERR3, Nurr77, ROR1 y ROR3, utilizando como descriptores las matrices de cargas de los sitios activos de los receptores y de las moléculas. El desempeño de la SVM implementada se evaluó a través de una validación cruzada aleatoria obteniéndose un valor de aproximadamente el 80% sobre los cuatro parámetros medidos: exhaustividad, especificidad, precisión y Valor-F (F-measure), lo cual a pesar de ser mejorable indica que el modelo fue capaz de realizar una correlación entre los datos y clasificar correctamente 8 de cada 10 compuestos. Posterior a esto se evaluaron los perfiles multi-diana de los

mismos dando como resultado más del 94% de las moléculas de prueba clasificadas como activas para los receptores ERR1, ERR3 y Nurr77 y menos del 86% para ROR3 y 75% para ROR1 siendo clara la relación entre los primeros y mostrándose cierta diferencia entre los últimos. Se determinó que la molécula ZINC70691822 posee el mejor perfil multi-diana al lograr el mejor puntaje sobre los 5 receptores evaluados debido a sus características de valores y distribución de carga.

3. ANTECEDENTES

3.1. Mecanismo multi-diana

Lu y colaboradores (2012) realizaron una investigación en la cual observaron que muchos de los fármacos que se usan en la actualidad que se creían solo poseían un único tipo de diana terapéutica, en realidad poseen otras más. Mediante un diseño de nodos relacionaron los receptores involucrados en las patologías más comunes, determinando que estos forman una amplia red en la que se interrelacionan más de lo que se cree, por lo que se recalca la importancia que tienen los fármacos multi-diana. Antes del 2006 solo se conocían en promedio 1.8 receptores para cada fármaco aprobado por la Food & Drugs Administration (FDA), mientras que posterior a esta fecha el número ha alcanzado los 2.5 y va en aumento.

Speck-Planche y colaboradores (2012) resaltan la importancia que existe en la búsqueda sobre diversos receptores y superar el sesgo que existe al limitarse únicamente a un receptor como diana terapéutica objetivo; en su estudio decidieron realizar la búsqueda de agentes con potencial acción contra tumores cerebrales mediante el uso de la bioinformática y la quimioinformática dilucidando los grupos funcionales que proveen de la actividad característica a los agentes antitumorales; finalmente propusieron moléculas conformadas por los principales grupos funcionales encontrados para garantizar un efecto multi-diana.

Koutsoukas y colaboradores (2011) mencionan en su investigación como se han roto los paradigmas en los que los fármacos selectivos para un tipo único de receptor han sido superados por un nuevo mecanismo que se basa en el efecto de una misma molécula sobre varios receptores. El mejor ejemplo de esto son los inhibidores de quinasas que la FDA recientemente aprobó, el Sorafenib y Sunitinib utilizados para el tratamiento de

carcinomas renales y hepatocelulares, actúan sobre los receptores VEGFR, PDGFR, FLT-3 y c-kit modulando distintas vías que resultan en el tratamiento de la misma patología.

Existen estudios que han demostrado de manera *ex vivo* que es posible a través del uso de moléculas multi-diana llevar a cabo el efecto inhibitorio de células cancerígenas. En una investigación Zhou y colaboradores (2012) obtuvieron resultados satisfactorios a través de la aplicación de la molécula AL3810, un potente inhibidor de los receptores tirosina quinasa implicados en la angiogénesis; AL3810 presentó una acción inhibitoria significativa del crecimiento de capilares arteriales sobre modelos de angiogénesis *ex vivo*. Las dosis diarias de AL3810 demostraron un amplio espectro de actividad antitumorígena sobre modelos de cáncer de riñón, páncreas e hígado.

En un estudio realizado por Speck-Planche y colaboradores (2011) se resalta la importancia de la búsqueda de nuevos agentes terapéuticos, particularmente aquellos con un mecanismo multi-diana. En este estudio construyeron una metodología basada en la técnica de Relación Estructura Actividad Cuantitativa (QSAR). Obtuvieron de distintas bases de datos lo referente a moléculas activas contra el cáncer y aplicando la técnica de QSAR llevaron a cabo la identificación de los grupos funcionales responsables de la actividad de estas moléculas. Propusieron que el diseño de nuevos fármacos sobre el cáncer podría basarse en los grupos funcionales encontrados y que esto a su vez garantizará el diseño de fármacos multi-diana, los cuales poseen un espectro terapéutico superior.

3.2. Receptores huérfanos asociados al cáncer

Los receptores de la familia ERR están relacionados al cáncer prostático, principalmente los subtipos β y γ ; el receptor ERR β regula el crecimiento celular en el cáncer prostático de tipo tanto andrógeno dependiente, como andrógeno independiente (Baek & Kim, 2014). Bianco, Sailland & Vanacker (2012) indican que se han llevado a cabo estudios en los cuales se han administrado ligandos sintéticos tanto de los receptores

ERR β y ERR γ a un cultivo celular de cáncer de próstata sobre los cuales se han visto efectos antitumorígenos. Estos receptores actúan como supresores de las células cancerígenas al activar las proteínas p21 las cuales inhiben las quinasas que ayudan a la reproducción celular, lo cual a su vez detiene la transición de la fase G1 a la fase S del ciclo celular, ocasionando la apoptosis.

El receptor ROR α presenta un efecto supresor sobre el cáncer de próstata al reducir la capacidad migratoria e invasiva de las células cancerígenas, lo cual sugiere que posee una acción supresora sobre este tipo de cáncer. El mecanismo por el cual el receptor ROR α ejerce su efecto es mediante el aumento de la producción de la proteína p53, la cual está encargada de responder ante los daños celulares, paralizando el ciclo celular (Baek & Kim, 2014). Según Roshan-Moniri y colaboradores (2014) se han llevado a cabo estudios de cáncer de próstata *in vitro* e *in vivo* utilizando agonistas del receptor ROR α lo cual ha resultado en la reducción de la proliferación, migración y efecto invasivo de las células cancerígenas. También en estos estudios se menciona que el agonista SR1078 activa tanto a ROR α y ROR γ , los que finalmente estabilizan la proteína p53, un importante agente supresor de tumores.

El receptor Nur77 pertenece a la subfamilia de receptores NR4A; este receptor actúa estimulando la producción de la proteína p53 en estados de daño genético y su efecto principal es la promoción de la apoptosis (Baek & Kim, 2014). Otro receptor de esta misma subfamilia es denominado Nurr1; Roshan-Moniri y colaboradores (2014) aseguran que este receptor es una potencial diana terapéutica en cáncer de próstata ya que en estudios de bloqueo genético se ha demostrado que la ausencia de este receptor ocasiona reducción en la proliferación, migración e invasión de las células cancerígenas.

Baek & Kim (2014) indican que los isoflavonoides así como las mercaptopurinas presentan ciertos efectos sobre los receptores nucleares huérfanos y sus derivados podrían proveer moléculas activas sobre estos. Roshan-Moniri y colaboradores mencionan

también algunas moléculas que individualmente presentan algún efecto sobre los receptores huérfanos, entre estas se menciona también derivados mercaptopurínicos así como esteroidales y una clase de moléculas denominadas C-DIMs (1,1-bis(3-indol)-1-(p-fenilsubstituidos) metanos).

Marcov, G. & Laudet (2011) mencionan que los receptores nucleares han evolucionado a lo largo del tiempo desde un mismo receptor antecesor que pudo tener una función sensorial metabólica lo que explica que la familia de estos posee gran diversidad de ligandos. Adicional a lo anterior existe un patrón en los compuestos por los que poseen afinidad los receptores nucleares, reduciendo las posibilidades en cierto margen, dicha afinidad es por los compuestos lipofílicos, principalmente de origen esteroideal y de cadenas poliinsaturadas. Tomando lo anterior en cuenta se decidió reducir la cantidad de compuestos en evaluación a aquellos que tuviesen características en común con el colesterol por su importancia como antecesor de muchas hormonas así como de otros compuestos lipofílicos. Para este motivo se consideró al colesterol desde el punto de vista estructural y de cargas siendo las bases del descriptor utilizado en la presente investigación. Esta consideración se ve reflejada en la selección de búsqueda “Electroshape” que se menciona más adelante. Dado que este criterio de búsqueda se basa en la forma y la carga del compuesto de búsqueda, se acopla perfectamente a las características evaluadas en la presente investigación, así se obtuvo de cada una de las bases de datos aquellas moléculas con los mejores puntajes de similitud con el colesterol.

3.3. La máquina de soporte vectorial

Según Heikamp & Bajorath (2014) la metodología de máquina de soporte vectorial se encuentra actualmente dentro de las mejores técnicas catalogadas bajo el nombre de aprendizaje de máquina y presenta una amplia gama de aplicaciones entre las cuales se encuentra la investigación de fármacos. La técnica de máquina de soporte vectorial presenta la capacidad de predecir las propiedades de los elementos que evalúa en base a

los datos que le son proporcionados, de esta manera se puede anticipar a los resultados de otras metodologías proveyendo de información útil de manera bastante eficiente.

Ma y colaboradores (2010) llevaron a cabo una investigación en la que a través de la máquina de soporte vectorial buscaron moléculas con actividad sobre los receptores de tipo quinasa que intervienen en el metabolismo del cáncer. Demostraron que esta técnica presenta excelentes capacidades en cuanto a realizar predicciones de moléculas activas así como una baja proporción de falsos positivos y una capacidad de cómputo elevada.

Xu y colaboradores (2007) determinaron que era posible mediante la técnica de máquina de soporte vectorial, encontrar proteínas susceptibles de ser usadas como dianas terapéuticas. En este estudio llevaron a cabo la construcción de una máquina de soporte vectorial que alimentaron con datos referentes a 512 proteínas conocidas como ejemplos de dianas terapéuticas. Utilizaron como descriptores las secuencias de cada proteína, las categorías funcionales a las que pertenecen y la especificidad de los tejidos en los que se encuentran. Al alimentar la máquina de soporte vectorial con estos datos lograron determinar que mediante esta técnica sí es posible catalogar nuevas proteínas como posibles receptores farmacológicos.

Existen estudios en los que la metodología de máquina de soporte vectorial ha sido utilizada exitosamente para llevar a cabo tamizajes sobre librerías en búsqueda de nuevos posibles fármacos. Liew, y colaboradores (2009) realizaron la construcción y entrenamiento de una máquina de soporte vectorial capaz de identificar nuevos inhibidores para la proteína Tirosina Quinasa específica de Linfocitos (Lck). El entrenamiento fue realizado con 820 compuestos positivos y 70 negativos para los cuales el parámetro de clasificación fue en base a la Concentración Inhibitoria del 50% de la población estudiada (CI_{50}). La máquina de soporte vectorial construida y entrenada presentó una sensibilidad mayor al 83%, una especificidad mayor al 99% y un valor de

falsos positivos de solo el 0.52%. Mediante esta técnica fue posible encontrar nuevos posibles inhibidores de Lck.

Wassermann, Geppert & Bajorath (2009) indican que al utilizar la SVM para predecir la interacción entre un receptor y un ligando, el resultado de clasificación de dicha interacción se encuentra fuertemente influenciado por un efecto denominado “vecino más cercano” (nearest neighbor effect), en el cual aquellas parejas de receptores y ligandos más parecidos a las parejas en evaluación son los que aportan más información para el resultado de la clasificación. De esto deriva también que se obtendrá un mejor resultado en tanto mayores sean las similitudes. Por esta razón se tomaron en cuenta principalmente los receptores nucleares de la base de datos utilizada como referencia que es mencionada más adelante, maximizando de esta manera la relevancia del conjunto de datos de entrenamiento.

Meslamani & Rognan (2011) realizaron un estudio comparativo utilizando diferentes modelos de SVM, con distintos descriptores tomando en cuenta parejas de receptores y ligandos. En el estudio lograron determinar que aquellos descriptores que toman en consideración la información tridimensional de los receptores dan lugar a mejores resultados. En el estudio también revelan que considerar la información de ambos, ligando y receptor, mejora considerablemente el desempeño del modelo. Tomando en cuenta lo anterior se consideró incluir un descriptor que se basa en la información tridimensional, en el caso de la presente investigación el descriptor utilizado es una matriz de cargas que se genera en base a la posición en el plano tridimensional de cada átomo que conforma el sitio activo, y de cada átomo que conforma el compuesto evaluado.

En su estudio Meslamani & Rognan (2011) mencionan que el modelo de SVM implementado con el kernel de Tanimoto presentó el mejor desempeño frente a los demás kernels que utilizaron. De igual manera Wassermann, Geppert & Bajorath (2009)

observaron mejores resultados al utilizar el kernel de Tanimoto y la información de los pares de receptores y ligandos, operando el kernel sobre el par de descriptores de receptores y de ligandos por aparte y realizando el producto sobre ambos resultados para dar lugar al dato final dentro de la optimización de la SVM. Considerando los datos anteriores decidió utilizar el kernel de Tanimoto con el fin de obtener los mejores resultados. Los descriptores de la proteína y del compuesto considerados en cada evaluación fueron concatenados para dar lugar a un solo descriptor compuesto por 2000 valores de carga, los primeros mil representando la matriz tridimensional del receptor y los siguientes mil representando la del compuesto en evaluación. Posteriormente cada par de ligandos y receptores fueron evaluados por la SVM separando dentro del kernel implementado las características de cada descriptor concatenado en sus dos partes respectivas para llevar a cabo las operaciones con el kernel de Tanimoto sobre cada parte y finalmente obtener el cociente entre ambos resultados.

3.4. Bases de datos

Actualmente existe una enorme cantidad de bases de datos que proporcionan desde la información genética hasta los archivos que contienen las estructuras tridimensionales de compuestos, proteínas, péptidos y demás elementos de importancia en las ciencias de interés. Directory of Useful Decoys Enhanced (DUD-E) es una de estas bases de datos que proporciona todos los archivos necesarios para llevar a cabo distintos estudios aunque en principio sus autores la presentan como una base de datos para benchmarking. Esta base de datos proporciona un total de 102 proteínas de distintas familias incluyendo un conjunto tanto de moléculas activas como inactivas para cada uno de los receptores. Dentro de las proteínas presentadas en esta base de datos existe diversidad como Quinasas, Proteasas y Nucleares entre otras (Mysinger, Carchia, Irwin & Shoichet, 2012).

4. JUSTIFICACIÓN

En la actualidad existen diversas terapias farmacológicas dirigidas al tratamiento del cáncer de próstata, sin embargo muchas de estas presentan varias deficiencias que van desde serios efectos adversos hasta la generación de resistencia farmacológica; por lo que se considera conveniente continuar la búsqueda de fármacos que actúen, no solo a través de nuevas dianas terapéuticas como los son los receptores huérfanos, sino también evaluando la posibilidad de utilizar nuevos enfoques como en este caso el mecanismo multi-diana.

Asimismo se observa la importancia de innovar a través de la aplicación de las herramientas que brinda la química computacional y el desarrollo de metodologías que aprovechen al máximo su potencial, por lo que se buscó dilucidar sustancias que podrían resultar en los constituyentes de nuevas terapias farmacológicas empleando los receptores nucleares huérfanos asociados al cáncer de próstata. Por lo anterior se escogieron aquellos receptores huérfanos que guardan relación específicamente con el cáncer de próstata el cual engloba la mayor cantidad observada (cinco) en estos estudios siendo esencial para aplicar un enfoque multi-diana.

Los compuestos preseleccionados para la evaluación de potencial farmacológico fueron principalmente aquellos que poseen similitudes con el colesterol debido a que los receptores nucleares poseen una marcada afinidad por los compuestos esteroideos.

5. OBJETIVOS

5.1. General

- Identificar potenciales fármacos multi-diana para el tratamiento del cáncer de próstata a través de los receptores nucleares huérfanos mediante la técnica de máquina de soporte vectorial.

5.2. Específicos

- Implementar un modelo computacional de una máquina de soporte vectorial que permita clasificar compuestos como activos o inactivos para los receptores huérfanos ERR1, ERR3, Nurr77, ROR1 y ROR3.
- Generar descriptores basados en matrices de cargas de los sitios activos de los receptores y de las estructuras de las moléculas para determinar si esas propiedades permiten clasificarlas como activas o inactivas a través de la máquina de soporte vectorial.
- Identificar las moléculas con el mejor perfil multi-diana, mediante el análisis de los resultados obtenidos para cada receptor mediante la clasificación realizada por el modelo de máquina de soporte vectorial implementado.
- Evaluar el desempeño de la máquina de soporte vectorial implementada mediante el análisis de su comportamiento respecto a la base de datos de referencia mediante validación cruzada.

6. HIPÓTESIS

Debido a que el diseño de la investigación es de tipo exploratorio, no se llevó a cabo el planteamiento de una hipótesis.

7. MATERIALES Y MÉTODOS

7.1. Universo

7.1.1. Población:

La base de datos de los receptores humanos que se encuentran en la Protein Data Bank en línea.

7.1.2. Muestra:

La base de datos conformada por 5 de los receptores nucleares humanos denominados huérfanos que se encuentran asociados al cáncer de próstata obtenidos a través de la Protein Data Bank en línea.

7.2. Materiales

7.2.1 Materiales y Equipos:

- Computadora (CPU: Pentium Dual-Core 2.50 GHz, RAM: 3 GB, HDD: 500GB, GPU: Nvidia GeForce 9500 GT)
- Computadora (CPU: Core i7-6700HQ 2.60 GHz, RAM: 8 GB, HDD: 1TB, SSD: 128GB, GPU: AMD Radeon R9 M375 4GB)
- Internet
- Hojas de papel bond y lápiz
- Impresora

- **Infraestructura de la Unidad de Química Computacional:**
 - Computadoras (CPU: FX8 4GHz, RAM: 8GB, HDD: 1TB, GPU: Nvidia GeForce GTX770)

7.3. Métodos:

7.3.1 Obtención y preparación de archivos de trabajo desde bases de datos en línea:

7.3.1.1 Obtención de proteínas y moléculas de referencia:

Los datos de las proteínas de referencia y sus respectivos ligandos fueron obtenidos a través del Directory of Useful Decoys Enhanced (DUD-E) encontrado en "<http://dude.docking.org/>" (Ver anexo 13.3.1); asimismo las moléculas de referencia fueron obtenidas de esta misma base de datos para cada proteína; en total la base de datos está constituida por 102 proteínas de distintas familias en formato pdb, de las cuales se utilizaron las 11 nucleares como referencia (Mysinger, Carchia, Irwin & Shoichet, 2012).

7.3.1.2 Obtención de proteínas diana:

Los archivos en formato pdb de las proteínas de estudio fueron: ERR1 (3k6p), ERR3 (2zkc), ROR1 (4s15), ROR3 (4ypq) y Nurr77 (4jgv) y los ligandos utilizados en formato sdf fueron 5fb, bisfenol, 4D8, 4f1 y T94 respectivamente; Todos estos archivos fueron obtenidos a través de RCSB PDB Protein Data Bank (Ver anexo 13.3.2) (<http://www.rcsb.org/pdb/>) (Berman, et al., 2000).

7.3.1.3 Generación de cargas atómicas:

Los datos de cargas atómicas utilizados para los receptores fueron obtenidos mediante el sitio web Atomic Charge Calculator (<https://webchem.ncbr.muni.cz/Platform/ChargeCalculator>), del cual se utilizaron los archivos pqr que genera como resultados la página. Los archivos pdb de los receptores fueron cargados a la página y la configuración de trabajo fué la recomendada por el sitio: EEM parameter: EX-NPA_6-31Gd_gas, Computation Methods: Full EEM, Ignore Waters = False, Precision = Double (Ver anexo 13.3.3 a y b) (Ionescu et al., 2015).

7.3.1.4 Preparación de carpetas y archivos de trabajo:

Los archivos de trabajo se manejaron de manera similar a como se encuentran en la base de datos DUD-E, para esto se preparó cada carpeta nombrada por el receptor que contenía, aplicándose este proceso tanto para los receptores de referencia como para los de prueba; dentro de las carpetas fueron colocados los siguientes archivos:

- Archivo pdb del receptor en el formato: “receptor.pdb”
- Archivo pqr del receptor en el formato: “receptor.pqr”
- Ligando del receptor en el formato “crystal_ligand.mol2”

7.3.1.5 Obtención de moléculas para evaluación de actividad sobre los receptores de estudio:

Las moléculas sobre las que se evaluó una potencial actividad sobre los receptores fueron obtenidas en formato SMILES desde SwissSimilarity. Se utilizó la molécula de colesterol como referencia y se aplicó el criterio de búsqueda “Electroshape” sobre las bases de datos ChEMBL, ChEBI, HMDB y Zinc Drug-Like (Ver anexo 13.3.4 a y b) (Zoete, Daina, Bovigny & Michielin, 2016).

7.3.1.6 Preparación de moléculas:

Las moléculas fueron preparadas con KNIME (Ver anexo 13.3.5) utilizando sus extensiones de RDKit de la siguiente manera:

- Se leyeron en formato SMILES desde un archivo csv
- Se convirtieron a moléculas RDKit
- Se adicionaron hidrógenos
- Se optimizaron las geometrías con MMFF94
- Se filtraron las repetidas y con errores
- Se convirtieron con Open Babel a formato mol2
- Se guardaron en archivos individuales

7.3.2 Creación de script de programación para la adecuación de los datos y entrenamiento de la Máquina de Soporte Vectorial:

Para el procesamiento de los datos utilizados en la construcción, entrenamiento y obtención de resultados de la máquina de soporte vectorial fue necesario escribir un script en el lenguaje de programación Python 2.7 utilizando el entorno de desarrollo Spyder. El script se realizó con los siguientes fines específicos:

- Lectura y escritura de archivos en formato pdb, mol2 y pqr.
- Manejo de coordenadas atómicas para rotación y traslación sobre los ejes X, Y y Z.
- Selección de átomos en una región delimitada.
- Generación de matrices de cargas.
- Comparación de matrices de cargas.
- Escritura de archivos con datos de trabajo.
- Búsqueda de las rotaciones necesarias para orientar moléculas y proteínas en base a su distribución de cargas.
- Implementación y entrenamiento de Máquinas de Soporte Vectorial.
- Obtención de resultados de las Máquinas de Soporte Vectorial generadas.
- Automatización de todos los procesos anteriormente mencionados.

(Ver anexo 13.3.6)

7.3.3 Preparación y obtención de los descriptores de los receptores del conjunto de entrenamiento y de prueba

7.3.3.1 Ubicación y centrado del sitio activo

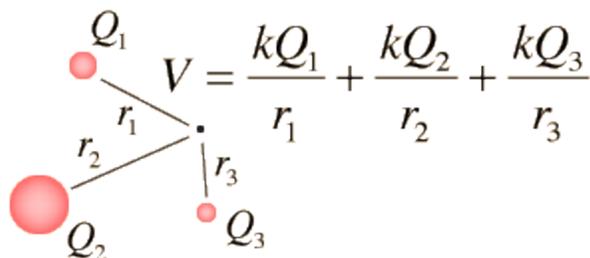
Se estableció el directorio global contenedor de los subdirectorios con los archivos en los formatos anteriormente mencionados; posteriormente con el script en base a las coordenadas del ligando de cada receptor se ubicó el centro de cada sitio activo respectivamente y modificaron las coordenadas de los receptores para centrarlos sobre dichos sitios activos; el centro de cada sitio activo quedó fijado en el mismo centro del plano de referencia en X, Y y Z = 0.0 (Ver anexo 13.3.7).

7.3.3.2 Selección de los principales átomos del sitio activo

Con el script se ubicaron los átomos constituyentes del sitio activo centrado, tomando como regla de decisión seleccionar aquellos átomos que se encontrasen dentro de una esfera establecida con radio de 10 Å y centro en X, Y y Z = 0.0. Se estableció el radio de 10 Å para la esfera ya que dentro del volumen definido por la misma se pudieron contener cada una de las moléculas evaluadas, es decir ninguna sobrepasó los límites de la esfera previos cálculos realizados y por lo mismo se estableció que dentro de este rango de distancia quedaron contenidos todos los átomos que llevan a cabo las interacciones más importantes entre el sitio activo y la molécula (Ver anexo 13.3.7).

7.3.3.3 Generación de matrices de carga

Una vez seleccionados los átomos constituyentes del sitio activo el script procedió a generar matrices de cargas en los procesos en que fueron necesarios; Las matrices utilizadas fueron de 20 Å cúbicos con una resolución de 2 Å (1000 puntos) con centro en 0.0 sobre los tres ejes del plano de referencia; cada punto de la matriz de carga se generó tomando en cuenta la influencia que tuvo cada átomo seleccionado en el sistema. Para cada punto de la matriz de cargas fue calculada la carga relativa generada por el sistema entero, para esto se midió la distancia de cada átomo a cada punto de la matriz y se tomó en consideración el dato de carga obtenido previamente mediante la calculadora de cargas atómicas (Ver anexo 13.3.8). Finalmente se operó como lo indica la ecuación:



$$V = \frac{kQ_1}{r_1} + \frac{kQ_2}{r_2} + \frac{kQ_3}{r_3}$$

7.3.3.4 Orientación de receptores

Todos los receptores fueron orientados hacia una misma posición, en base a las comparaciones entre sus matrices de cargas y la del receptor ANDR de la base de datos de referencia manteniendo este último en la misma posición para todos los casos; el procedimiento se basó en dejar este receptor en una posición arbitraria en el espacio y centrando su sitio activo justamente sobre el centro del plano de referencia, así los sitios activos posteriormente fueron acomodados de manera que las matrices de cargas de su sitio activo se orientaran en la misma dirección. Este proceso se realizó con el fin de normalizar los datos, como parte de los requisitos esenciales para trabajar con la SVM.

7.3.3.5 Almacenamiento de matrices de carga

Las matrices de carga fueron almacenadas dentro de las carpetas de su respectivo receptor bajo el título “m_cargas.dat” como un archivo de texto conteniendo la sucesión numérica de los 1000 puntos de la matriz de carga comenzando por X: -10, Y:-10, Z:10 hacia X:10, Y:10 y Z:-10 recorriendo el eje X de izquierda a derecha, el eje Y de arriba hacia abajo y el eje Z del frente hacia atrás; ésta fue la configuración que se utilizó para el plano de referencia en todo momento y la misma para generar una única lista de entrada para la Máquina de Soporte Vectorial.

7.3.4 Preparación y obtención de los descriptores de las moléculas del conjunto de entrenamiento y de prueba

7.3.4.1 Centrado

Mediante el script se leyeron los archivos de las moléculas y en base a sus coordenadas se centraron en 0.0 sobre los tres ejes de coordenadas del plano de referencia previamente optimizadas en una conformación estable.

7.3.4.2 Generación de matrices de cargas

Una vez centradas las moléculas el script procedió a generar matrices de cargas en un proceso idéntico al utilizado para los receptores obteniendo resultados con las mismas características: matrices de 20 Å cúbicos con una resolución de 2 Å (1000 puntos) con centro en 0.0 sobre los tres ejes del plano de referencia.

7.3.4.3 Orientación de moléculas

Las moléculas fueron alineadas mediante el script en referencia a la molécula de colesterol previamente optimizada, centrada en el plano de referencia y dejada en la misma orientación en todos los procesos en los que participó como referencia; la orientación se basó en traslapar las matrices hasta obtener la menor diferencia posible (Ver anexo 13.3.9).

7.3.4.4 Almacenamiento de matrices de carga

Las matrices de carga fueron almacenadas dentro de las carpetas de su respectiva clasificación (positivas, negativas o prueba); el archivo generado para cada una de estas fue idéntico al de los receptores con la variación del nombre del archivo siendo este el identificador numérico de la molécula seguido de “_mcargas.dat”; este también se almacenó como un archivo de texto con la misma configuración mencionada anteriormente para los receptores.

7.3.4.5 Implementación y entrenamiento de la máquina de soporte vectorial

Los archivos de cargas de las moléculas fueron leídos a través del script y utilizados directamente como un listado de valores de carga. Se generó un par receptor-molécula para cada combinación posible entre los 11 receptores nucleares de referencia, 50 de sus moléculas activas y 50 de sus moléculas inactivas. Cada par se asoció con una etiqueta de clasificación arbitraria la cual fue

1 para los pares de moléculas activas sobre el receptor y 0 para las inactivas. La máquina de soporte vectorial se implementó en el lenguaje de programación Python versión 2.7 utilizando la librería scikit-learn versión 0.18.1. (Pedregosa et al., 2011).

7.3.5 Clasificación de las moléculas mediante la máquina de soporte vectorial

Con la Máquina de Soporte Vectorial entrenada para clasificar moléculas activas e inactivas sobre determinado receptor, se procedió a evaluar de manera automatizada los archivos de cargas generados para las 1380 moléculas en estudio. Para esto el script realizó la combinación de pares de descriptores de cada una de las moléculas en estudio junto con cada receptor nuclear evaluado (1380 x 5), dando como resultado 6900 pares de descriptores. Los resultados de la clasificación para cada par molécula-receptor fueron desplegados y almacenados en un listado en formato de archivo de texto.

7.3.6 Análisis de Datos

7.3.6.1 Puntaje del clasificador

Dentro de las capacidades que ofrece la SVM existe la posibilidad de obtener un dato de distancia entre cada muestra y el plano separador del modelo, esto permite tener un valor que representa que tanto se diferencia un elemento para ser parte de una categoría de clasificación, dichos valores fueron utilizados para generar un ranking para cada par molécula-receptor evaluado.

7.3.6.2 Generación de ranking molécula-receptor

Los valores obtenidos para la asociación de cada ligando-receptor fueron obtenidos para cada par evaluado como un decimal entre -1 y 1. Para cada par clasificado como “inactivo” se obtuvo un resultado entre -1 y 0, siendo más fuerte su afinidad por la categoría tanto más se acercase a -1, de igual manera se obtuvo un valor entre 0 y 1 para cualquier par catalogado como “activo” siendo más

cercano a 1 el resultado para los pares que poseen más afinidad por esta categoría. Bajo la anterior condición se ordenaron a través del script las moléculas en forma descendente según su valor de clasificación para cada receptor individualmente y se enumeraron, de esta manera la molécula con valor de clasificación más cercano a 1 obtuvo el número 1 en el ranking y aquella con el valor más cercano a -1 obtuvo el número 1380 para cada receptor (Ver anexo 13.3.10).

7.3.6.3 Análisis Multi-diana

Los valores del ranking obtenidos anteriormente fueron después analizados para buscar las moléculas con el mejor perfil multi-diana, para esto se sumaron los rankings obtenidos sobre cada receptor y se reordenaron nuevamente en base a estos resultados. Las moléculas con los números más pequeños representan las que obtuvieron los mejores perfiles multi-diana así como las que obtuvieron las sumas mayores son las que presentan los peores perfiles en base al nuevo ranking global.

7.3.6.4 Evaluación del desempeño de la SVM mediante validación cruzada

El desempeño de la SVM implementada fue evaluado a través de una validación cruzada aleatoria. Para cada evaluación se tomó de manera aleatoria el 80% de la base de datos de referencia como datos de entrenamiento y el restante 20% como datos de prueba y se realizaron 10 repeticiones (Ver anexo 13.3.11). Los resultados obtenidos fueron clasificados como verdaderos positivos = VP, falsos positivos = FP, verdaderos negativos = VN y falsos negativos.

7.4 Diseño de la investigación

Se realizó un muestreo por conveniencia en búsqueda de moléculas con actividad multi-diana sobre 5 de los receptores asociados al cáncer de próstata (ERR1, ERR2, Nurr77, ROR1, ROR2) utilizando como criterio de selección aquellas moléculas con similitudes de cargas y estructura en relación al colesterol dada la relevancia del compuesto sobre los receptores nucleares.

Se implementó un modelo computacional de una Máquina de Soporte Vectorial con los datos de 11 receptores nucleares conocidos y sus respectivos ligandos obtenidos de la base de datos DUD-E.

Con el modelo mencionado se logró clasificar un conjunto de 1380 moléculas como activas o inactivas sobre cada uno de los receptores nucleares en estudio; en base a lo anterior se determinaron las moléculas que poseen las características suficientes para ser clasificadas como activas multi-diana sobre los cinco receptores en estudio.

Para evaluar el desempeño del modelo se realizó una validación cruzada aleatoria de 10 repeticiones, variando los datos de referencia de la siguiente manera:

- Para cada evaluación se tomó de manera aleatoria el 80% de la base de datos de referencia como datos de entrenamiento y el restante 20% como datos de prueba. Los resultados obtenidos fueron clasificados como verdaderos positivos = VP, falsos positivos = FP, verdaderos negativos = VN y falsos negativos FN y se operaron según las siguientes fórmulas:
 - Exhaustividad = $VP/(VP + FN)$
 - Precisión = $VP/(VP + FP)$
 - Especificidad = $VN/(VN + FP)$
 - Valor-F o medida-F = $2(\text{recuperación} \times \text{precisión}) / (\text{recuperación} + \text{precisión})$.

Los resultados obtenidos de las diez repeticiones fueron promediados y presentados para cada caso (pares de moléculas-receptores activos e inactivos).

Los valores del ranking obtenidos mediante la clasificación de cada compuesto fueron analizados para buscar las moléculas con el mejor perfil multi-diana, para esto se

sumaron los rankings obtenidos sobre cada receptor y se reordenaron nuevamente en base a estos resultados. Las moléculas con los números más pequeños representan las que obtuvieron los mejores perfiles multi-diana así como las que obtuvieron las sumas mayores son las que presentan los peores perfiles.

8. RESULTADOS

8.1. Recuento de la clasificación de las moléculas de estudio para cada receptor nuclear

En la Tabla 1 se muestran los recuentos de la clasificación de las moléculas de prueba para cada receptor; En esta se puede apreciar que 3 de los 5 receptores obtuvieron más del 94% de moléculas clasificadas como activas. De los receptores restantes el más cercano a este resultado fue ROR3 con aproximadamente el 86% y por último ROR1 que presentó una tendencia un poco más marcada obteniendo menos del 75% de las moléculas clasificadas como activas.

Tabla 1. Recuento de moléculas clasificadas como activas e inactivas para cada receptor

Clasificación	ERR1		ERR3		Nurr77		ROR1		ROR3	
	Conteo	(%)								
Inactivas	80	5.80	77	5.58	62	4.49	362	26.23	197	14.28
Activas	1300	94.20	1303	94.42	1318	95.51	1018	73.77	1183	85.72

Fuente: Datos experimentales obtenidos en la unidad de Química Computacional de la Facultad de Ciencias Químicas y Farmacia, USAC Zona

8.2. Moléculas con los mejores perfiles multi-diana

En la Tabla 2 se presentan los resultados obtenidos para las mejores 15 moléculas en base a su perfil multi-diana incluyendo al colesterol como referencia. Las moléculas se presentan en base a su puntaje global al sumar los rankings obtenidos para cada receptor. Se observa cierto patrón sobre 4 de los 5 receptores al poseer rankings bastante aceptables excluyendo al receptor ERR3 para el cual el ranking en promedio se encuentra más alejado del primer lugar. Se puede evidenciar también que el colesterol se encuentra más cerca de las últimas posiciones ubicándose en la No. 1021 del total de las 1380 moléculas, por lo que a pesar de haber sido clasificado como compuesto activo, este se cataloga con baja actividad multi-diana.

Tabla 2. Resumen de las mejores 15 moléculas con acción multi-diana y colesterol

No.	Molécula	Posición en el ranking para cada receptor					Posición Global
		ERR1	ERR3	NURR77	ROR1	ROR3	
1	ZINC70691822	8	95	11	6	6	126
2	ZINC95948325	7	168	13	11	7	206
3	CHEMBL197469	4	426	7	3	8	448
4	ZINC91526913	90	79	230	110	39	548
5	CHEMBL571057	1	627	1	8	11	648
6	CHEMBL486700	151	287	148	38	50	674
7	ZINC76324228	54	309	133	112	134	742
8	ZINC15770651	58	178	479	22	24	761
9	CHEMBL2058273	128	333	126	77	115	779
10	CHEBI__47805	132	250	222	69	171	844
11	ZINC05500601	63	608	15	97	95	878
12	CHEMBL476132	77	371	352	28	51	879
13	CHEMBL224719	61	631	75	72	53	892
14	CHEMBL229664	87	559	128	53	89	916
15	CHEMBL2069831	59	246	247	98	270	920
1021	CHEBI_16113*	712	983	697	1096	924	4412

*CHEBI_16113 = Colesterol

Fuente: Datos experimentales obtenidos en la unidad de Química Computacional de la Facultad de Ciencias Químicas y Farmacia, USAC Zona 12

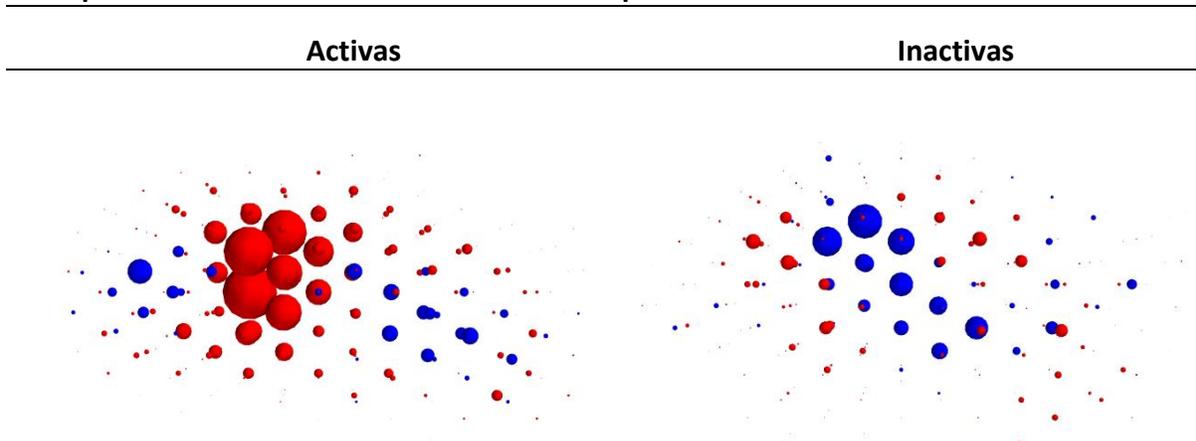
8.3. Características de las moléculas según sus perfiles multi-diana

En la Tabla 3 se muestran los promedios de los valores de cargas obtenidos de las matrices de las moléculas activas e inactivas. Se puede observar que las moléculas activas presentan una mayor cantidad de conteos y magnitudes de cargas positivas, a diferencia de las inactivas. Dentro de las representaciones tridimensionales de las matrices de cargas las esferas rojas indican las cargas positivas y las azules las negativas y el tamaño las magnitudes de los valores de carga. En las moléculas activas se observa una mayor densidad de cargas positivas con ciertas regiones de pequeñas cargas negativas, mientras que las moléculas inactivas presentan una especie de relación inversa.

Tabla 3. Resumen de las características de valores de carga promedio de las moléculas según sus perfiles multi-diana

	Moléculas activas		Moléculas Inactivas	
	Positivos	Negativos	Positivos	Negativos
Conteo de puntos en la matriz de carga	195	47	171	84
Suma de magnitudes totales de carga	3.67	-1.02	1.38	-1.37

Representación tridimensional de matrices promedio de moléculas activas e inactivas



Rojo: Cargas positivas. **Azul:** Cargas Negativas

Fuente: Datos experimentales obtenidos en la unidad de Química Computacional de la Facultad de Ciencias Químicas y Farmacia, USAC
Zona 12

8.4. Desempeño y validación del modelo de máquina de soporte vectorial

En la Tabla 4 se presentan los resultados obtenidos de la validación cruzada aleatoria realizada para determinar el desempeño de la SVM implementada. La exhaustividad refleja la sensibilidad del método; esta junto con la precisión y especificidad presentaron valores muy cercanos lo que indica que el modelo posee la misma capacidad de clasificar moléculas tanto activas como inactivas con cierto margen de error. El valor-F (F-measure) indica la exactitud general del método. En general se observa que el método es capaz de clasificar 8 de cada 10 compuestos correctamente indistintamente de si son activos o inactivos.

Tabla 4. Desempeño del modelo de Máquina de Soporte Vectorial

	Exhaustividad (%)	Precisión (%)	Especificidad (%)	Valor-F
Activas	80	79	79	80
Inactivas	79	80	80	80
Promedio	80	80	80	80

Fuente: Datos experimentales obtenidos en la unidad de Química Computacional de la Facultad de Ciencias Químicas y Farmacia, USAC
Zona 12

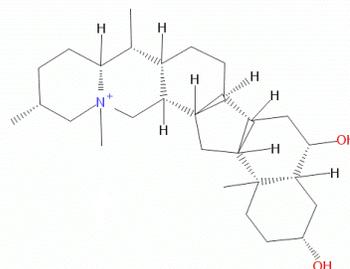
8.5. Características de la molécula con el mejor perfil multi-diana

En el Cuadro 1 se presenta un resumen de las características fisicoquímicas y propiedades de la molécula con el mejor perfil multi-diana encontrado. Dentro de los descriptores se observa una alta densidad de cargas positivas en una distribución característica similar a la obtenida en la Tabla 3 para moléculas activas.

Cuadro 1. Información del compuesto ZINC70691822

Formula molecular	C ₂₈ H ₄₉ N ₁ O ₂
Peso molecular (g/mol)	430.697
Enlaces rotables	0
Donadores de puentes de hidrógenos	2
Aceptores de puentes de hidrógenos	3
Coefficiente de partición	1.37

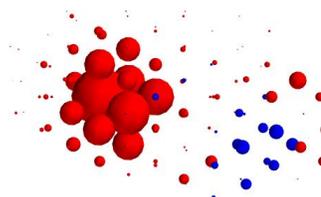
Estructura 2D



Descriptores generados para evaluar dentro del modelo de la SVM

	Positivos	Negativos
Conteo de puntos en la matriz de carga*	113	23
Suma de magnitudes totales de carga*	6.65	-1.13

Representación tridimensional de la matriz de cargas*



*Rojo: Cargas positivas. *Azul: Cargas Negativas

Fuente: ZINC12 Database. *Datos experimentales obtenidos en la unidad de Química Computacional de la Facultad de Ciencias Químicas y Farmacia, USAC Zona 12

9. DISCUSIÓN

Para llevar a cabo la implementación de técnicas de aprendizaje de máquina es esencial la normalización de los datos, con lo cual se asegura el adecuado comportamiento del método al trabajar en las mismas magnitudes y condiciones sobre todos los elementos; asimismo es importante generar un descriptor que cumpla con las características que mejor representen dichos elementos en estudio (Heikamp & Bajorath, 2014). Este proceso fue realizado a través de la generación y orientación de matrices de cargas para cada sitio activo y molécula; para esto se estableció un sitio activo y una molécula de referencia en una posición arbitraria y posteriormente se orientó cada sitio activo y molécula respectivamente para encajar su distribución de cargas a dicha orientación. Este proceso garantizó que los descriptores representados como vectores con mil valores de cargas se encontraran ordenados de la misma manera sin perder el valor intrínseco de posición de cada valor de carga.

El descriptor utilizado consistió en una matriz de cargas de 20 Å cúbicos con una resolución de 2 Å generada a partir de los valores de carga de los átomos presentes en el espacio delimitado por dicha matriz. Cada descriptor fue generado para los receptores y para los compuestos evaluados de manera que cada matriz ocupase la región del sitio activo o la estructura entera de los compuestos; para generar cada matriz se midieron la distancia y el valor de carga de cada átomo hacia cada punto de la misma; se consideró particularmente el uso de estos descriptores ya que abarcan características esenciales del sitio activo y de las moléculas al representar las fuerzas electrostáticas que intervienen en los procesos de acoplamiento e interacción entre los receptores y los compuestos. También en el cálculo de las matrices de carga se encuentra implicada información de la estructura de los compuestos y sitios activos al considerar las distancias de cada átomo constituyente de los mismos hacia los puntos de la matriz en el plano tridimensional.

Como se observa en la Tabla 1, de los 5 receptores 3 obtuvieron más del 90% de moléculas clasificadas como activas, mientras que el receptor ROR3 obtuvo un dato inferior al 86% y ROR1 al 75%. Es bastante probable que esta tendencia de las moléculas de no poseer la misma afinidad por los receptores ROR 1 y ROR3 se deba a que estos están más relacionados entre sí al pertenecer al mismo subtipo, haciéndolos más lejanos al resto del grupo. Por otro lado también los resultados podrían estar influenciados por los aspectos mencionados por Meslamani & Rognan (2011), los cuales indican que la clasificación de la SVM en este tipo de modelos en los que se toma en cuenta la información de los receptores y las moléculas a través del kernel de Tanimoto, los resultados se encuentran principalmente influidos por los efectos del “vecino más cercano” y podría existir una falta de información de receptores con mayores similitudes dentro del conjunto de receptores de referencia dado que está conformado por una base relativamente pequeña constituida por 11 receptores nucleares.

Otro aspecto observable en la Tabla 1 es una elevada clasificación de moléculas activas; este hecho deriva de la composición de la base de datos de moléculas de prueba la cual se construyó sobre la búsqueda de compuestos en SwissSimilarity, utilizando como criterios de selección aquellos compuestos con similitud con el colesterol a nivel de estructura y cargas, lo cual se relaciona directamente con los descriptores utilizados. Esta selección está basada en las propiedades de los receptores nucleares presentadas por Marcov & Laudet (2011) entre las que se mencionan su alta afinidad por compuestos esteroidales al estar estrechamente relacionados con procesos metabólicos y dependientes de respuestas hormonales. Se puede decir entonces que los resultados fueron los esperados al alcanzar altos porcentajes de moléculas clasificadas como activas sobre los receptores nucleares huérfanos en estudio.

La utilización de los receptores nucleares de la base de datos DUD-E parte de la premisa de hacer uso de receptores de referencia que posean la mayor similitud posible con los receptores en estudio como se menciona anteriormente. Dado que estos

pertenecen a la misma familia se pueden considerar como los receptores de elección al proveer de información relevante (Meslamani & Rognan, 2011). La SVM generada fue alimentada con los descriptores de todas las moléculas activas e inactivas de los receptores de referencia previamente preparadas (alineadas en base a su estructura y carga). El proceso de clasificación llevado a cabo por la SVM se basa en encontrar las similitudes entre pares de receptores-ligandos de los datos de referencia y los evaluados. Para esto la SVM utiliza un kernel, el cual es esencialmente una función que se encarga de operar los vectores que contienen los descriptores de los datos en estudio. El kernel utilizado para la implementación de la SVM fue el de Tanimoto (ver anexo 13.2), el cual se basa en comparar un par de elementos y generar un valor decimal que expresa sus similitudes mediante la operación de los productos escalares de los vectores. Esta función kernel toma en cuenta más a profundidad las características del receptor y de la molécula y genera un resultado de clasificación más efectivo en casos comparables según mencionan Wassermann, Geppert & Bajorath (2009) tras contrastar resultados en las mismas condiciones con distintos kernels.

En la Tabla 2 se muestran los resultados obtenidos para las mejores 15 moléculas de la base de datos en estudio constituida por 1380 compuestos incluyendo el colesterol. Para generar el puntaje multi-diana se llevó a cabo la suma de los rankings que las moléculas obtuvieron frente a cada uno de los cinco receptores en estudio; los resultados de clasificación para cada receptor se muestran como una medida de distancia lineal desde la posición que representa el par molécula-ligando en el plano de clasificación de la SVM hacia el plano separador, este dato se puede interpretar como una medida de la pertenencia a la categoría en la que las moléculas se encuentran clasificadas. Este proceso se realizó de manera similar al expuesto por Wassermann, Geppert & Bajorath (2009), lo cual permite generar un ranking sobre un conjunto de datos en una clasificación binaria. En el análisis multi-diana la consideración principal dicta que aquella molécula que muestra buenas interacciones con la mayor cantidad de receptores en estudio a la vez es la que posee el mejor perfil (Lu, Pan, Hu & Wang, 2012). En la tabla 2 también puede

observarse un patrón que muestra que las moléculas con mejores perfiles multi-diana generalmente poseen un buen ranking sobre 4 de los 5 receptores exceptuando al receptor ERR3 para el cual los mismos compuestos presentaron valores de ranking menos favorables en su mayoría. Esto respalda la posibilidad de que los compuestos con potencial farmacológico para el tratamiento de cáncer de próstata podrían actuar con un mecanismo multi-diana sobre los receptores ERR1, Nurr77, ROR1 y ROR3 principalmente y en menor medida sobre el receptor ROR1, ya que se puede observar la existencia de una relación de afinidad de las moléculas hacia los cinco receptores.

Asimismo se presenta el ranking obtenido por la molécula de colesterol en donde se puede observar que se encuentra entre los últimos compuestos (posición 1024) lo cual lo sitúa en una posición menos favorable frente a otros, aunque bajo la clasificación de molécula activa sobre los cinco receptores y con un ranking similar para estos. Se puede apreciar entonces que existen muchos compuestos que a través de sus variaciones poseen mejores perfiles que el colesterol. Esto indica que ciertas modificaciones aportan una diferencia significativa y mejoran la afinidad entre los compuestos y los receptores, pero con una persistente estructura esteroidea; esto resulta consistente con lo mencionado por Marcov & Laudet (2011) quienes indican que los receptores nucleares han evolucionado de un ancestro en común que presentaba afinidad por los compuestos esteroidales y que a lo largo del tiempo ha evolucionado dividiéndose en distintos receptores relacionados los cuales mantienen dicha afinidad, para lo cual ha existido consiguientemente una evolución de los compuestos, adquiriendo nuevos grupos funcionales y variando su estructura sin perder los rasgos esteroidales.

La Tabla 3 muestra que las moléculas activas generalmente presentan dentro de sus matrices de cargas un mayor conteo de cargas positivas frente a negativas (195 a 47), así como mayores magnitudes de cargas positivas en una proporción de 3.6 a 1. A diferencia de estas, las moléculas clasificadas como inactivas presentan un estrechamiento en esta proporción, pese a mantener un conteo de cargas positivas

mayor, esta disminuye a 171 a 84 aproximadamente y las magnitudes llegan a ser prácticamente iguales (aproximadamente 1). Se observa en general que las moléculas activas deberán poseer una mayor cantidad de cargas positivas lo que indica que la actividad de las moléculas a pesar de necesitar grupos funcionales con elementos o puntos electronegativos como pueden ser oxígeno, azufre, halógenos, o algunos enlaces conjugados, estos no deben sobrepasar cierto límite, permitiendo predominar siempre valores de cargas positivas. Lo anterior se puede relacionar a lo mencionado por Kojetin & Burris (2014) quienes indican que en ciertos estudios se ha confirmado la afinidad que los receptores poseen por el colesterol y que derivados de este con grupos electronegativos como el oxígeno se perfilan como compuestos potentes sobre los receptores ROR. También Gallastegui, Mackinnon, Fletterick & Estébanez-Perpiñá (2015) mencionan que para los receptores nucleares ERR se han descubierto moléculas con grupos funcionales oxigenados y aminas terciarias que les confieren actividad. A pesar de que las diferencias numéricas no son tan marcadas entre los compuestos activos e inactivos existe también un factor de estructura esencial, el cual es la ubicación o distribución de dichas cargas. La distribución de cargas de las matrices de los compuestos activos e inactivos poseen diferencias observables a simple vista en sus representaciones tridimensionales, una mayor densidad de cargas positivas se acumula en el lado derecho de la gráfica para las moléculas activas, rodeadas de cargas positivas más pequeñas y cierto patrón de cargas negativas al lado izquierdo, posiblemente siendo el patrón que se requiere para interactuar de manera correcta con los receptores nucleares. Por otro lado en la gráfica que representa la matriz de cargas promedio de una molécula inactiva, se observa un patrón distinto, casi inverso en el que la mayoría de las cargas negativas reemplaza la ubicación de las cargas positivas.

La Tabla 4 presenta la evaluación del desempeño de la SVM. Para esto se llevó a cabo una validación cruzada aleatoria en la cual se dividieron aleatoriamente los datos de referencias en grupos de entrenamiento y prueba, tomando el 80% de estos como datos de entrenamiento para implementar la SVM y el restante 20% para su evaluación cada

vez, realizándose un total de 10 repeticiones. Como se observa la SVM presento puntajes muy similares en todos los aspectos tanto para la clasificación de moléculas activas como inactivas. La exhaustividad se basa en la suma de todas las clasificaciones verdaderas positivas, reflejando una medición de la sensibilidad del método; esta medición junto con la precisión y especificidad presentaron valores muy cercanos lo que indica que el modelo posee la misma capacidad de clasificar moléculas tanto activas como inactivas con cierto margen de error. El valor-F (F-measure) indica la exactitud general del método, el cual a su vez refleja los parámetros anteriores presentando aproximadamente el mismo puntaje. En general se observa que el método es capaz de clasificar correctamente 8 de cada 10 compuestos correctamente indistintamente de si son activos o inactivos. Estos resultados demuestran una verdadera capacidad de clasificación del modelo; al superar el 50% en todas las mediciones el modelo presenta una verdadera capacidad de clasificación y no simplemente un resultado al azar. Cabe destacar que el modelo implementado fue capaz de obtener los puntajes mencionados únicamente utilizando los descriptores basados en valores de carga y posición de los átomos constituyentes de los elementos en estudio representados a través de las matrices de cargas. Contrastando con otros estudios que emplean metodologías similares se observan puntajes cercanos a los presentados por Meslamani & Rognan (2011) quienes obtuvieron datos que van desde el 70% hasta un 86% en Valor-F. Asimismo Wasserman, Geppert & Bajorath (2009) obtuvieron desde valores nulos hasta un máximo de 95% de exhaustividad, comparando distintos modelos a través de la variación de parámetros tales como las bases de datos empleadas y el tipo de kernel. Un factor importante que diferencia los estudios mencionados con el presente es que utilizaron bases de datos más extensas y de receptores de diferentes categorías, los cuales en algunos casos son más comunes y de familias más grandes. Esto pudo favorecer los resultados en dichos estudios ya que como mencionan Wasserman, Geppert & Bajorath (2009) el desempeño de la SVM utilizando particularmente el kernel de Tanimoto depende en gran medida de la relación existente entre los elementos de referencia y los elementos estudiados ya que cuanto mayor sean las similitudes mayor es el valor de entrenamiento que aportaran al modelo de clasificación.

Como podría esperarse los descriptores jugaron un papel importante. Al tomar en cuenta valores de carga y estructura tridimensional se logró establecer un descriptor bastante aceptable, sin embargo el desempeño del modelo podría ser mejorado si se evalúan a fondo otras características como los grupos funcionales, donadores y aceptores de puentes de hidrógenos, lipofilia, enlaces rotables, entre otros. Esto pude observarse en los estudios de Wasserman, Geppert & Bajorath (2009) y Meslamani & Rognan (2011) en los cuales sus descriptores se basan en la generación de una “huella digital” de los elementos de estudio tomando en cuenta no solo información tridimensional sino también las características mencionadas antes.

Otro aspecto que influencia de manera directa el desempeño de la SVM es el kernel utilizado. Dada la complejidad de los descriptores, los datos pueden comportarse de maneras distintas y el kernel se encargará de establecer la dimensión en la que los datos se deben evaluar para poder generar el mejor plano separador entre los mismos. En este caso se decidió utilizar el kernel de Tanimoto el cual ha demostrado ser superior frente a otros al tomar en cuenta la información de dos elementos y devolver un valor que representa su similitud. Según Wassermann, Geppert & Bajorath (2009) este presenta mejores resultados debido a que la clasificación generada por la SVM se encuentra fuertemente influenciada por las similitudes entre los datos de entrenamiento y los de prueba, siendo tal que entre más similares sean los aspectos de los mismos, mejor será el desempeño. Meslamani & Rognan (2011) obtuvieron resultados similares en condiciones comparables demostrando también que el kernel de elección es el de Tanimoto.

En el Cuadro 1 se exponen las características fisicoquímicas y los valores y representación gráfica de la distribución de la matriz de carga del compuesto ZINC70691822 clasificado como el compuesto con el mejor perfil multi-diana. En el conteo de puntos de cargas se puede observar que su matriz manifiesta la proporción necesaria para clasificar un compuesto como activo, siendo bastante mayor la cantidad de cargas

positivas frente a las negativas (113 a 23). Asimismo la suma de las magnitudes de carga posee una proporción aún mayor, predominando los valores de carga positivos (6.65 a -1.13). Finalmente se observa en la representación gráfica de la matriz de cargas una distribución que encaja con la descrita anteriormente para las moléculas activas, presentando una densidad de cargas positivas en el lado izquierdo y algunos pequeños valores de cargas negativas dispersas en el lado derecho. Estas propiedades reflejan las características deseadas en una molécula activa según lo establecido por el clasificador y detallado en la Tabla 3. Como se mencionó antes, es posible que estas sean características a nivel de estructura y cargas atómicas necesarias para interactuar con los sitios activos de los receptores nucleares resultando en una estabilización de ambas estructuras.

En general se observa que es posible llevar a cabo la clasificación de moléculas mediante la técnica de máquina de soporte vectorial en búsqueda de una actividad sobre determinados receptores mediante la descripción de sus principales características, en particular aquellas que están involucradas en las interacciones entre un receptor y un ligando, obteniendo así un desempeño bastante aceptable que eventualmente puede mejorarse considerando otras cuestiones como las bases de datos de referencia, la cantidad y tipos de descriptores e incluso los elementos base a la hora de la implementación como por ejemplo el kernel utilizado. Se puede asegurar entonces que la técnica de máquina de soporte vectorial puede ser utilizada para llevar a cabo la búsqueda de potenciales fármacos, que el comportamiento multi-diana podría estar involucrado en las interacciones de potenciales fármacos y los receptores nucleares evaluados y que la molécula determinada con el mejor perfil multi-diana podría proponerse como un posible ligando o precursor de terapias contra el cáncer de próstata al actuar sobre los 5 receptores huérfanos estudiados.

10. CONCLUSIONES

10.1. Más del 94% de las moléculas fueron clasificadas como activas para 3 de los 5 receptores nucleares huérfanos en estudio, mientras que para los restantes 2 fueron inferiores al 86% y 75%, evidenciando que la preselección de moléculas basadas en la estructura y distribución de cargas del colesterol aportaron un parámetro de delimitación adecuado de la base de datos.

10.2. Al obtener resultados de ranking similares para 4 de los 5 receptores se puede observar una acción multi-diana que podría ser el mecanismo real que describiría un fármaco al actuar en el tratamiento del cáncer de próstata a través de los receptores nucleares huérfanos ERR1, ERR3, Nurr77, ROR1 y ROR3.

10.3. La máquina de soporte vectorial implementada presentó aproximadamente un 80% sobre los cuatro aspectos de desempeño evaluados en la validación cruzada, lo que demuestra que es capaz de clasificar correctamente 8 de cada 10 compuestos sin importar si pertenecen verdaderamente a la categoría de activos o inactivos.

10.4. La molécula con el identificador ZINC70691822 presentó el mejor perfil multi-diana al poseer una sumatoria total de puntajes sobre cada receptor que supera a los demás compuestos; por tal motivo podría considerarse este compuesto como un potencial fármaco para el tratamiento del cáncer de próstata.

10.5. Dado el buen desempeño de la máquina de soporte vectorial implementada, se pueden considerar como efectivos los descriptores utilizados.

11. RECOMENDACIONES

- 11.1** Comparar el desempeño de otras técnicas de Aprendizaje de Máquina tales como: redes neuronales, árboles de decisión, redes bayesianas, reglas de asociación, entre otros y comparar los resultados y desempeño de las mismas.

- 11.2** Comparar los resultados mediante la implementación de distintos descriptores, que tomen en cuenta otros aspectos de los receptores y moléculas como pueden ser grupos funcionales, enlaces rotatorios, donadores y aceptores de hidrógenos, lipofilia o coeficiente de partición por ejemplo.

- 11.3** Evaluar si existe alguna mejora en el desempeño del modelo variando aspectos como el tipo de descriptores, las implementaciones a través de distintos tipos de kernels, o la variación de la selección de las bases de datos de referencia.

12. REFERENCIAS

- Baek, S. & Kim, K. (2014). Emerging Roles of Orphan Nuclear Receptors in Cancer. *The Annual Review of Physiology*, 76(19), 1-19.
- Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. & Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1), 235-242. doi: <https://doi.org/10.1093/nar/28.1.235>
- Bianco, S., Sailland, J., & Vanacker, J. (2012). ERRs and cancers: Effect on metabolism and on proliferation and migration capacities. *Journal of Steroid Biochemistry and Molecular Biology* 130, 180-185.
- Gallastegi, N., Mackinnon, J., Fletterick, R. & Estébanez-Perpiñá, E. (2015). Advances in our Structural understanding of orphan nuclear receptors. *Trends in Biochemical Sciences*, 40(1), 25-35. <https://doi.org/10.1016/j.tibs.2014.11.002>
- Heikamp, K. & Bajorath, J. (2014). Support vector machines for drug discovery. *Expert Opinion in Drug Discovery*, 9(1), 93-104.
- Ionescu, C., Sehnal, D., Falginella, F., Pant, P., Pravda, L., Bouchal, T., Varekova, R., Geidl, S. & Koca, J. (2015). AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules. *Journal of Cheminformatics*, 7(50). doi: 10.1186/s13321-015-0099-x
- Kojetin, D. & Burris, T. (2014). REV-ERB and ROR nuclear receptors as drug targets. *Nature Reviews Drug Discovery*, 13, 197-216. doi:10.1038/nrd4100

- Korcsmáros, T., Szalay, M., Böde, C., Kovács, I. & Csermely, P. (2007). How to design multi-target drugs: Target search options in cellular networks. *Expert Opinion on Drug Discovery*, 2, 1-10.
- Koutsoukas, A., Simms, B., Kirchmair, J., Bond, P., Whitmore, A., Zimmer, S., et al. (2011). From in silico target prediction to multi-target drug design: Current databases, methods and applications. *Journal of Proteomics*, 74, 2554-2574.
- Li, Y., Tan, C., Gao, C., Zhang, C. Luan, X., Chen, X., Liu, H., Chen, Y & Jiang, Y. (2011). Discovery of benzimidazole derivatives as novel multi-target EGFR, VEGFR-2 and PDGFR kinase inhibitors. *Bioorganic & Medicinal Chemistry*, 19, 4529-4535.
- Liew, C., Ma, X., Liu, X., & Yap, C. (2009). SVM Model for Virtual Screening of Lck Inhibitors. *Journal of Chemical Information and Modeling*, 49, 877-885.
- Lu, J., Pan, W., Hu, Y., Wang, Y. (2012). Multi-Target Drugs: The Trend of Drug Research and Development. *PLoS ONE* 7(6). 1-6.
- Ma, X., Shi, Z., Tan, C., Jiang, Y., Go, M., Low, B. & Chen, Y. (2010). In-Silico Approaches to Multi-target Drug Discovery. *Pharmaceutical Research*, 27(5), 739-749.
- Ma, X., Wang, R., Tan, C., Jiang, Y., Lu, T., Rao, H., Li, X., Go, M., Low, B., & Chen, Y. (2010). Virtual Screening of Selective Multitarget Kinase Inhibitors by Combinatorial Support Vector Machines. *Molecular pharmaceutics*, 7(5), 1545-1560.
- Marcov, G. & Laudet, V. (2011). Origin and evolution of the ligand-binding ability of nuclear receptors. *Molecular and Cellular Endocrinology*, 334, 21-30. doi: 10.1016/j.mce.2010.10.017
- Meslamani, J. & Rognan, D. (2011). Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel. *Journal of Chemical Information and Modeling*, 51, 1593-1603. doi: dx.doi.org/10.1021/ci200166t

- Mysinger, M., Carchia, M., Irwin, J. & Shoichet, B. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55, 6582-6594. doi: 10.1021/jm300687e
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(10), 2825-2830.
- Pettersen, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E. & Ferrin, T. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*, 25(13), 1605-12.
- Roshan-Moniri, M., Hsing, M., Butler, M., Cherkasov, A. & Rennie, P. (2014). Orphan nuclear receptors as drug targets for the treatment of prostate and breast cancers. *Cancer Treatment Reviews* 40, 1137-1152.
- Safe, S., Jin, U., Hedrick, E., Reeder, A. & Lee, S. (2014). Minireview: Role of Orphan Nuclear Receptors in Cancer and Potential as Drug Targets. *Molecular Endocrinology* 28, 157-172.
- Speck-Planche, A., Kleandrova, V., Luan, F. & Cordeiro, M. (2011). Multi-target drug discovery in anti-cancer therapy: Fragment-based approach toward the design of potent and versatile anti-prostate cancer agents. *Bioorganic & Medicinal Chemistry*, 19, 6239-6244.
- Speck-Planche, A., Kleandrova, V., Luan, F. & Cordeiro, M. (2012). Chemoinformatics in Multi-target Drug Discovery for Anti-cancer Therapy: In Silico Design of Potent and Versatile Anti-brain Tumor Agents. *Anti-Cancer Agents in Medicinal Chemistry*, 12, 678-685.
- Wassermann, A., Geppert, H. & Bajorath (2009). Ligand Prediction for Orphan Targets Using Support Vector Machines and Various Target-Ligand Kernels Is Dominated by Nearest

Neighbor Effects. *Journal of Chemical Information and Modeling*, 49, 2155-2167.

doi:10.1021/ci9002624

Xu, H., Xu, H., Lin, M., Wan, W., Li, Z., Huang, J., et al. (2007). Learning the drug target-likeness of a protein. *Proteomics*, 7, 4255-4263.

Zhou, Y., Chen, Y., Tong, L., Xie, H., Wen, W., Zhang, J., et al. (2012). AL3810, a multi-tyrosine kinase inhibitor, exhibits potent anti-angiogenic and anti-tumour activity via targeting VEGFR, FGFR and PDGFR. *Journal of cellular and molecular medicine*, 16(10), 2321-2330. doi: 10.1111/j.1582-4934.2012.01541.x

Zoete, V., Daina, A., Bovigny, C. & Michielin, O. (2016). SwissSimilarity: A Web Tool for Low to Ultra High Throughput Ligand-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 56(8), 1399-1404. doi: 10.1021/acs.jcim.6b00174

13. ANEXOS

13.1. Marco Teórico

Históricamente el descubrimiento de medicamentos se ha basado únicamente en la percepción de los cambios que originaban las plantas medicinales sobre el organismo humano. Por lo que se sabe lo único que había interesado era el efecto que tenía la sustancia administrada, logrando el efecto de devolver la salud al organismo, sin embargo cada vez más la ciencia se ha interesado por el mecanismo de acción de estas sustancias y la serie de efectos que esto desencadena sobre el organismo humano (Koutsoukas, et al., 2011).

Las desventajas que presentan los agentes terapéuticos diseñados para actuar sobre un único tipo de receptor frecuentemente son tales como eficacia reducida, perfiles de seguridad indeseados y resistencia farmacológica. Esta última puede ser debida a varios mecanismos compensatorios que presentan los sistemas biológicos así como una elevada robustez. Por otro lado el diseño de fármacos multi-diana permite superar estas deficiencias al estar dirigidos a múltiples receptores bien seleccionados que tienen un impacto mayor en los sistemas biológicos. Los fármacos multi-diana actúan modulando tanto la actividad de los receptores terapéuticos primarios como la de los elementos que puedan contrarrestar sus efectos o desarrollar resistencia. (Ma, et al., 2010).

Los receptores nucleares conforman una superfamilia de receptores caracterizados por ser activados mediante ligandos para llevar a cabo su función y modular su actividad; estos receptores intervienen en una gran cantidad de procesos esenciales que permiten la homeostasis celular y tienen roles importantes en varias patologías incluyendo el cáncer. Los miembros de esta superfamilia han sido catalogados en subgrupos según el tipo de ligando con el que interactúan. Los receptores huérfanos son la tercera clase de

receptores nucleares y son denominados así ya que no se han identificado ligandos que modulen su actividad (Safe, Jin, Hedrick, Reeder & Lee, 2014).

Las estrategias multi-diana no son por completo nuevas, más bien, han sido redescubiertas. Al respecto se puede mencionar que las arañas y serpientes poseen venenos que están constituidos de múltiples componentes que les proporcionan sus potentes efectos tóxicos. Las plantas también se consideran en este punto al generar estrategias combinadas contra amenazas del ambiente; estas son capaces de generar diversas sustancias con un amplio espectro de acción para protegerse de patógenos. Históricamente la medicina se ha constituido de productos terapéuticos fabricados a partir de mezclas de varios extractos naturales. Esto demuestra el beneficio que ha presentado para diversas culturas el uso de productos que ejercen su efecto sobre múltiples objetivos al estar constituido de varias sustancias terapéuticas. Los agentes farmacológicos que atacan un solo receptor pueden no siempre tener el efecto deseado, en especial, si se trata de sistemas complejos que generalmente presentan rutas de respaldo que se activan en el momento preciso, evitando cambios en su funcionamiento normal. Este es el caso de lo que se conoce como redes celulares, las cuales presentan diversidad de rutas e interconexiones que forman sistemas robustos que funcionan a pesar de las modificaciones u obstaculizaciones que se les apliquen a los mismos (Korcsmáros, Szalay, Böde, Kovács, & Csermely, 2007).

El amplio interés que ha impulsado recientemente estudios sobre fármacos con efecto multi-diana o multi-receptor sobre el cáncer se debe a que se ha encontrado que los fármacos que actúan de esta manera poseen un efecto incrementado, moléculas en estudio que tiene efecto sobre los tres principales receptores de Tirosin quinasa involucrados en el cáncer (EGFR, VEGFR-2 y PDGFR) han demostrado efectos significativamente mejores que cualquier otro fármaco utilizado y requieren de dosis notoriamente menores (Li, et al., 2011).

13.2. Kernel de Tanimoto

El kernel de Tanimoto se define por la comparación de dos elementos. En este caso sean \mathbf{u} , \mathbf{v} dos moléculas y d un número entero. Considerando el plano de características ϕ_d y su kernel correspondiente k_d . El kernel de Tanimoto k_d^t se define como:

$$k_d^t(\mathbf{u}, \mathbf{v}) = \frac{k_d(\mathbf{u}, \mathbf{v})}{k_d(\mathbf{u}, \mathbf{u}) + k_d(\mathbf{v}, \mathbf{v}) - k_d(\mathbf{u}, \mathbf{v})}$$

13.3. Screenshots de las páginas y programas utilizados.

13.3.1. Base de datos DUD-E (<http://dude.docking.org/>)

The screenshot shows the homepage of the DUD-E database. At the top, there is a navigation bar with the UCSF logo and links for 'About UCSF', 'Search UCSF', and 'UCSF Medical Center'. The main header features a stylized graphic of a face and the text 'D U D • E A Database of Useful Decoys: Enhanced'. Below this is a navigation menu with links: Home, Targets, Subsets, Generate, Other, FAQ, Revisions, and Thanks. The main content area starts with a 'Welcome to DUD•E' message, followed by a list of statistics: 22,886 active compounds and their affinities against 102 targets, an average of 224 ligands per target, and 50 decoys for each active having similar physico-chemical properties but dissimilar 2-D topology. A section titled 'mol2 and SDF format now available in all packages for actives and decoys. [July 14]' is also present. The footer includes a thank you message to NIGMS for financial support and a note about the research tool's purpose.

13.3.2. Base de datos RCSB PDB (Protein Data Bank)

(<https://www.rcsb.org/pdb/home/home.do>)

The screenshot shows the RCSB Protein Data Bank homepage. At the top, there is a navigation bar with links for Deposit, Search, Visualize, Analyze, Download, Learn, and More. A search bar is prominently displayed in the center, with a 'Go' button. Below the search bar, there is a 'Welcome' section with a 'A Structural View of Biology' article. To the right of this article is a 'Video: How Enzymes Work' player. Further right is the 'August Molecule of the Month' section, which features a 3D model of a protein structure labeled 'Glutathione Transferases'. The page also includes a sidebar with navigation options and a footer with links for 'Latest Entries', 'Features & Highlights', 'News', and 'Publications'.

13.3.3. Atomic Charge Calculator

(<https://webchem.ncbr.muni.cz/Platform/ChargeCalculator>)

a) Página principal y carga de archivos

The screenshot shows the AtomicChargeCalculator website. The page title is 'AtomicChargeCalculator' and the subtitle is 'Conformationally Dependent Atomic Charges of Quantum Mechanics Quality'. The main content area includes a description of the tool, a 'Submit a Computation' button, and a file upload section with a 'Molecule(s):' label and a 'Select file' button. There are also links for 'Quick Help', 'Samples', and 'Command Line Version'.

b) Configuración para cálculo de cargas

The screenshot shows the AtomicChargeCalculator web interface. The browser address bar displays the URL: <https://webchem.ncbr.muni.cz/Platform/ChargeCalculator/Result/cf6fd28f-58a4-474e-8f15-4b95421df1fd>. The page title is "AtomicChargeCalculator Computation Setup".

Below the title, there is a green button labeled "Guide me through the main stages of the setup".

The "Molecules" section contains a table with the following data:

Id	#Atoms	Total Charge	Message
receptor	2270	0	

Below the table, there is a field for "Total Charge for All Molecules" with the value "0" and an "Apply" button.

The "EEM Parameter Sets" section shows 1 selected set: "EX-NPA_6-31Gd_gas (HF/6-31G*/NPA for Biomacromolecules) by Ionescu, C. M., Geiß, S., Svobodová Váreková, R., Koca, J. [more]".

The "Computation Methods" section shows 1 selected method: "Full EEM" with options "Ignore Waters = false, Precision = Double".

The "Computation Summary and Execution" section states: "The computation on 1 molecule, with 1 EEM parameter set, using 1 method, will require 1 job". Below this, there is a blue "Compute" button.

13.3.4. SwissSimilarity (<http://www.swisssimilarity.ch/>)

a) Página inicial de carga de molécula de prueba

The screenshot shows the SwissSimilarity web interface. The browser address bar displays the URL: www.swisssimilarity.ch. The page title is "SwissSimilarity".

The page features a navigation bar with links: Click2Drug, SwissDock, SwissParam, SwissSidechain, SwissBioStere, SwissTargetPrediction, SwissADME, SwissSimilarity, and About us.

The main content area includes a header for "Swiss Institute of Bioinformatics" and a description of the tool's purpose: "This website allows you to perform ligand-based virtual screening of several libraries of small molecules, using different approaches. If you use this web tool, please, cite the following paper: Zoete, V., Daina, A., Bovigny, C., & Michielin, O. SwissSimilarity: A Web Tool for Low to Ultra High Throughput Ligand-Based Virtual Screening. J. Chem. Inf. Model., 2016, 56(8), 1399-1404."

The "Choose a reference small molecule" section contains a text input field for SMILES, a "Clear" button, and a dropdown menu for "Examples".

The "Choose a method and a library to screen" section contains a text input field for the library name and a dropdown menu for the screening method.

The "Perform the screening" section contains a red "Submit" button.

Below the "Submit" button, there is a note: "Will use method FP2 to screen the library of virtual compounds accessible from Sigma-Aldrich compounds through click-chemistry reactions (Takes a couple of seconds to several minutes depending on the library and the screening tool)".

On the right side of the page, there is a 3D ball-and-stick model of a cholesterol molecule, with a legend for the atoms: H (white), C (grey), N (blue), O (red), S (yellow), F (light green), P (orange), Cl (green), Br (brown), I (purple).

b) Página de configuración de búsqueda

#	Combined	FP2 fingerprints	Electroshape	Spectrophores	Shape-IT	Align-IT
Drugs						
Approved	1'516	<input type="checkbox"/>				
Experimental	4'788	<input type="checkbox"/>				
Investigational	504	<input type="checkbox"/>				
Withdrawn	161	<input type="checkbox"/>				
Nutraceuticals	78	<input type="checkbox"/>				
Illicit	169	<input type="checkbox"/>				
Bioactive compounds						
Ligands from the PDB	19'500	<input type="checkbox"/>				
ChEMBL (activity<10µM)	177'000	<input type="checkbox"/>				
ChEBI	27'950	<input type="checkbox"/>				
Kinase inhibitors (ChEMBL)	53'800	<input type="checkbox"/>				
GPCR Ligands (ChEMBL)	140'300	<input type="checkbox"/>				
GPCR Ligands (GLASS)	290'700	<input type="checkbox"/>				
HMDB	39'060	<input type="checkbox"/>				
Commercially available						
Zinc Drug-Like	10'639'400	<input type="checkbox"/>				
Zinc Lead-Like	4'328'000	<input type="checkbox"/>				
Zinc Fragment-Like	705'300	<input type="checkbox"/>				
AldrichCPR	214'000	<input type="checkbox"/>				
Asinex	693'000	<input type="checkbox"/>				
AsisChem	241'000	<input type="checkbox"/>				
ChemBridge	1'022'000	<input type="checkbox"/>				
ChemDiv	1'746'000	<input type="checkbox"/>				
Enamine	2'661'000	<input type="checkbox"/>				
InnovaPharm	367'000	<input type="checkbox"/>				
Maybridge	54'300	<input type="checkbox"/>				
Otava	376'000	<input type="checkbox"/>				
Selleckchem	1'900	<input type="checkbox"/>				
Sigma-Aldrich	65'000	<input type="checkbox"/>				
SPFC-S	10'639'400	<input type="checkbox"/>				

13.3.5. KNIME

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow with the following nodes:

- File Reader (ChEMBL)
- File Reader (ChEBI)
- Concatenate (Node 5)
- Concatenate (Node 15)
- RDKIT From Molecule (Node 6)
- RDKIT

The Node Description panel for the CIR node is open, showing the following text:

CIR

This node uses the network service **CIR (Chemical Identifier Resolver)** by the CADD Group at the NCINIH as a resolver for different chemical structure identifiers and allows one to convert a given structure identifier into another representation or structure identifier.

Dialog Options

Column including the structure identifier

The column in the input data table that contains the molecule representation.

The input structure identifier type can

The KNIME Console at the bottom shows the following output:

```

*** Welcome to the KNIME Analytics Platform v3.3.1.v201612192115 ***
*** Copyright by KNIME GmbH, Konstanz, Germany ***
*****
Log file is located at: C:\Users\boby\.metadata\knime\knime.log
WARN MoI2 Writer 0:17 Output file 'C:\Program Files\KNIME\0' exists
  
```

13.3.6. Script para caracterización de sitio activo

The screenshot shows the Spyder Python IDE with a script named '1. CSA 2.2 (proteinas).py' and an IPython terminal. The script defines various constants and parameters for a protein simulation, including mathematical constants like PI and RAD, and specific parameters for the active site like pHe, pWi, boAr, boAb, boIz, boDe, ArcW, and ArcH. It also defines center coordinates cx, cy, and cz.

```

1# -*- coding: utf-8 -*-
2# python 2.7
3from Tkinter import *
4import tkMessageBox
5import math as mt
6import numpy as np
7import platform
8import os
9import time
10import random
11import copy
12from multiprocessing import Pool, freeze_support
13
14OS = platform.system()
15#Constantes matematicas
16PI = mt.pi
17RAD = PI * 2
18paso = RAD / 200
19
20#constantes que definen el tamaño del plano de graficacion (Canvas)
21pHe = 500
22pWi = 800
23boAr = pHe / 7
24boAb = boAr * 6
25boIz = pWi / 7
26boDe = boIz * 6
27ArcW = (pWi / 7) * 5
28ArcH = (pHe / 7) * 5
29
30#centros utilizados para graficos y calculos
31cx = pWi / 2
32cy = pHe / 2
33cz = 500
34
  
```

The IPython terminal shows the Python version (2.7.13) and the IPython version (5.1.0). It also displays the IPython help text and the current prompt 'In [1]:'.

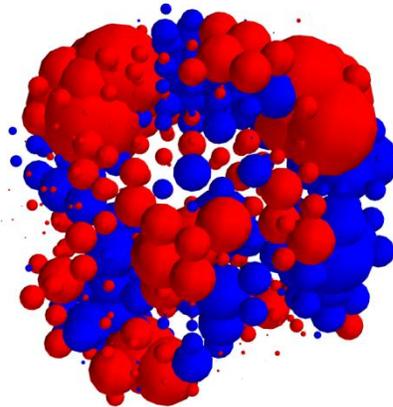
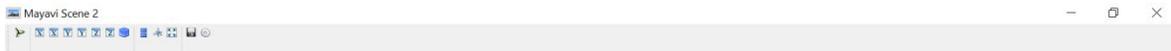
13.3.7. Selección de sitio activo y valores de cargas

The screenshot shows the Bobby-Gl ver. 1.7 software interface. The main window displays a 3D protein structure with a green rectangular box highlighting the active site. The control panel on the right includes a checkbox for 'bloquear proteína', an 'eje X' slider, a 'Comandos' section with buttons for 'zoom in', 'zoom out', 'arriba', 'abajo', 'izquierda', and 'derecha', and a 'Selección de residuos:' section with a list of residues.

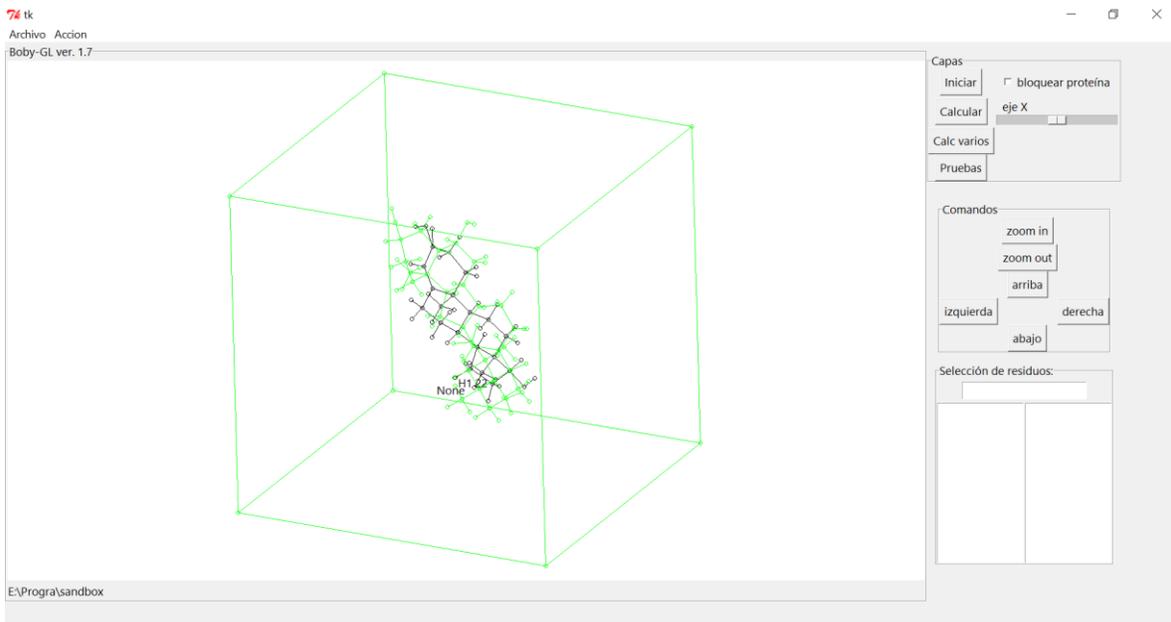
Selección de residuos:

VAL866	NE2 HIS874
LYS808	CE1 HIS874
ARG786	CD2 HIS874
TRP751	CD2 LEU873
GLY750	CD1 LEU873
MET749	CG LEU873
ALA748	CB LEU873
PHE747	HN ALA870
VAL746	CB ALA870
MET745	CA ALA870

13.3.8. Generación de matrices de cargas



13.3.9. Orientación de moléculas



13.3.10. Script para análisis de datos multi-diana

```

1  -*- coding: utf-8 -*-
2  """
3  Created on Wed Aug 09 19:55:08 2017
4
5  @author: baby
6  """
7  from Tesis_svm_func_lib import fragmentar
8
9
10 def ordenar_por(x, att):
11     t = x.len
12
13     for i in range(t):
14         for j in range(t):
15             if not j+i >= t:
16                 if att[j] < att[j + i]:
17                     x.switch(j, j + i)
18     return x
19
20
21 class attbag():
22
23     def __init__(self):
24         self.val = []
25         self.etiq = []
26         self.len = 0
27         self.punt = []
28         self.npg = []
29
30
31     def append(self, valor, etiqueta):
32         self.val.append(valor)
33         self.etiq.append(etiqueta)
34         self.punt.append(0)

```

```

X: -4.965 -4.263 -4.064
Y: 4.748 4.147 2.672
Z: -1.963 -0.880 -1.203
penalty 216.753023189

odist: [0.0, 0.0]

matriz: 0.0826127175419
rmsd: 3
0.232504049104
X: -5.171 -4.327 -4.182
Y: 4.684 4.140 2.646
Z: -1.540 -0.530 -0.786
penalty 216.754361486

odist: [27.83205758, 0.35230453]

10
10
guardado

tiempo total: 0.052300001907 minutos

In [16]: runfile('E:/Progra/prograTesis/5. Ordenamiento
multi-diana.py', wdir='E:/Progra/prograTesis')
cargando archivos..
ordenando datos por valores..
obteniendo puntajes..
ordenando datos por etiquetas
guardando datos... guardado y listo!!

In [17]:

```

13.3.11. Evaluador de desempeño mediante validación cruzada

```

414
415
416 X, Y, e = cargar_mcargas_ref(rutabr)
417
418 X = np.array(X)
419
420 """
421 Xpru, epru, reso = cargar_mcargas_pru(rutabp)
422
423 print 'in cargando SVM...'
424 clf = svm.SVC(kernel = tani_kernel)
425 clf.fit(X,Y)
426
427
428 Xresn, Xress = obtener_resultados(Xpru)
429
430
431 reson, resos, rese = ordenar_resultados(Xresn, Xress, epru, reso)
432
433
434 guardar_res(reson, resos, rese, rutagn, rutags)
435 """
436
437 def evaluacion(rep=10):
438     print generando + str(10) + ' repeticion(es)..
439     res, dat = evaluar_svm(X, Y, tani_kernel, rep)
440
441     imprimir_resultados(res, dat)
442
443
444
445
446

```

```

Kernel: Custom-kernel
TP TN FP FN
[59, 55, 15, 11]
      recall precision especific. F-measure
1      0.84      0.80      0.79      0.82
0      0.79      0.83      0.84      0.81
promedio      0.81      0.82      0.81      0.81

In [96]: runfile('E:/Progra/prograTesis/4. CSA-SVM-K.py',
wdir='E:/Progra/prograTesis')
Reloaded modules: Tesis_svm_lib

Cargando matrices de cargas de REFERENCIA . . . . .

In [97]: evaluacion()
generando 10 repeticion(es)..

Generando puntajes del modelo . . . . .

RESULTADOS:

Repeticiones: 10
Kernel: Custom-kernel
TP TN FP FN
[562, 547, 153, 138]
      recall precision especific. F-measure
1      0.80      0.79      0.79      0.80
0      0.79      0.80      0.80      0.80
promedio      0.80      0.80      0.80      0.80

In [98]:

```



Boris Iván Muñoz Debroy

Estudiante



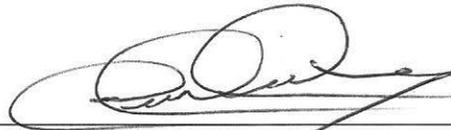
Lic. Omar Ernesto Velásquez González

Asesor



Licda. Lucrecia Margarita Peralta de Madriz

Asesora



Licda. María Nereida Marroquín Tintí

Revisora



Licda. Raquel Azucena Pérez Obregón

Directora de Escuela



Dr. Rubén Dariel Velásquez Miranda

Decano