MODELACIÓN BASADA EN MULTICLASIFICADORES DE COMPUESTOS QUÍMICOS CON ACTIVIDAD ANTIMALÁRICA

Universidad Central “Marta Abreu” de Las Villas
Facultad Química-Farmacia

Tesis en opción al Título de Licenciado en Ciencias Farmacéuticas

Santa Clara 2016

Autora: Ana Yisel Caballero Alfonso
Tutor: Reinaldo Molina Ruiz

FRASE O PENSAMIENTO

Porque Dios da la sabiduría, y de su boca viene el conocimiento y la inteligencia (...) cuando la sabiduría entra al corazón y la ciencia es grata al alma, la inteligencia te preserva. Bienaventurado el hombre que halla la sabiduría y obtiene la ciencia.

Proverbios

DEDICATORIA

A mis padres por todo su esfuerzo, cariño y dedicación.

AGRADECIMIENTOS

A mi Señor Jesús por darme una segunda oportunidad, al Todo y Poderoso Dios, Señor de la ciencia y la sabiduría.

A mi mamá por su entrega durante todos mis años de estudio, por sus oraciones y su espíritu incansable.

A mi papá por su apoyo incondicional y todas las horas que me dedicó.

A mi hermanita por entenderme siempre.

A mi esposo por su amor.

A mis familiares y amigos que me han animado siempre, especialmente a mis tíos y tías.

A mima, pipo y Yanne que siempre me animaron y ayudaron.

A mis profesores por sus enseñanzas y especialmente a los que me dieron su comprensión.

A Daimel por creer en mí.

A mis compañeros del Departamento de Diseño de Fármacos del CBQ por sus consejos y ayuda.

A mis tutores Maikel y Chiqui, por darme su amistad y cariño, por creer en mí, por su comprensión, por enseñarme el camino a seguir, por compartir conmigo estos últimos años y por todo lo aprendido. Los quiero.

Y a todos aquellos que de una manera u otra hicieron que esto fuera posible…

RESUMEN

La malaria o paludismo es una enfermedad causada por parásitos del género *Plasmodium* y se transmite al ser humano por la picadura de mosquitos infectados del género *Anopheles*. Esta patología es considerada como la primera causa de muerte en países tropicales, y aunque existen diversos tratamientos, estos poseen una elevada toxicidad. Los métodos computacionales basados en Relaciones Cuantitativas Estructura-Actividad han sido ampliamente empleados en el diseño de fármacos y son una alternativa racional al descubrimiento de fármacos. El objetivo del presente trabajo fue desarrollar modelos computacionales que permitan la identificación de *hits* antimaláricos. Con este propósito se recopiló de la literatura un conjunto de datos adecuado para la modelación y se sometieron a un estricto proceso de curación. Además se estudió el rendimiento de diversas metodologías de clasificación basadas en multiclasificadores y se seleccionó una de estas como la más apropiada para la identificación de hits antimaláricos en base a su desempeño. Durante el proceso de recopilación se obtuvieron datos de una buena calidad, ensayados por un único laboratorio y siguiendo un mismo protocolo. La curación de estos datos garantizó que estuvieran libres de cualquier error o ambigüedad que afectara la modelación. Por otra parte se encontró que los multiclasificadores que empleaban como método de selección Algoritmo Genético y como forma de agregación de salidas el Voto Mayoritario exhibían un desempeño superior.

ABSTRACT

Malaria or Paludism is a tropical disease caused by parasites of the Plasmodium genre and transmitted to humans through the bite of infected mosquitos of the Anopheles genre. This pathology is considered as the first cause of death in tropical countries and, despite several therapies exist, they have a high toxicity. Computational methods based on Quantitative Structure-Activity Relationship studies have been widely used for drug design and they stand as an alternative for drug discovery. The main goal of this research is to develop computational models for the identification of anti-malaric hit compounds. For this, a dataset suitable for the modeling of the anti-malaric activity of chemical compounds was compiled from the literature and subjected to a thorough curation process. In addition, the performance of a diverse set of ensemble-based classification methodologies was evaluated and one of these ensembles was selected based on its performance as the most suitable for the identification of anti-malaric hits. During the compilation of the dataset it was possible to obtain a high quality dataset that was curated to ensure that no noisy data would affect the modeling process. Among the explored ensemble-based methods, the one combining Genetic Algorithms for the selection of the base classifiers and Majority Vote for their aggregation showed the best performance.

INDICE

INTRODUCCIÓN

La malaria, causa aproximadamente medio millón de muertes anualmente y es uno de los problemas mundiales de salud pública más devastadores. Todos los años se registran aproximadamente entre 124 y 283 millones de casos de malaria y alrededor del 40 % de la población mundial, es decir 3.200 millones de personas de 107 países, corre el riesgo de contraer malaria(1).

El paludismo o malaria es una enfermedad causada por parásitos del género *Plasmodium* que se transmiten al ser humano por la picadura de mosquitos infectados del género *Anopheles y aunque existen 127 especies y subespecies de este género solo* cuatro parasitan al hombre*: Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale.*(2):

Siendo el más mortal de estos el paludismo provocado por *P. falciparum*, especie por la cual la malaria reina como la suprema causa de muerte en las zonas tropicales (África, Asia y Latinoamérica)(2).

Esta patología es una enfermedad febril aguda en la que los síntomas aparecen generalmente entre los 10 y los 15 días luego de la picadura del mosquito infectivo y si no se trata en las primeras 24 horas, el paludismo por *P. falciparum* puede agravarse, llevando a menudo a la muerte(3).

En la lucha del hombre contra esta enfermedad han aparecido varios métodos con el fin de erradicarla. El empleo de terapias combinadas para el tratamiento de la malaria ha sido ampliamente difundido sin embargo estos tratamientos poseen una elevada toxicidad. Por esta razón se necesitan nuevos compuestos que permitan tratar el paludismo de una manera efectiva y que con una toxicidad mínima(4).

El proceso de obtención de un medicamento eficaz y seguro es extremadamente largo y costoso, estimándose que por los métodos convencionales de ¨prueba y error ¨ se obtiene una probabilidad de éxito inferior a uno por cada 10 000 candidatos que inician el proceso. Debido a todo lo anteriormente planteado, el diseño computarizado de fármacos se ha convertido en una forma urgente de solución en la actualidad. La industria farmacéutica se encuentra bajo una creciente presión para encontrar nuevos fármacos en una forma más rápida y eficiente que en el pasado.

Los estudios Quantitative Relations Structure Avtivity (QSAR) engloban las relaciones cuantitativas entre la estructura molecular y la actividad biológica, lo cual constituye uno de los paradigmas de la Química Médica. Esta aproximación ha sido ampliamente empleada en el

diseño de fármacos con diversas actividades y el paludismo no es una excepción. Dichos estudios han surgido como una alternativa más racional y económica, ya que tratan de modelar o predecir determinada actividad biológica, basados en parámetros que caracterizan la estructura molecular.

Problema científico

Los fármacos empleados actualmente en el tratamiento de la malaria poseen una elevada toxicidad y la obtención de compuestos con actividad antimalárica y baja toxicidad por los métodos clásicos de ¨prueba y error¨ resulta compleja y costosa.

Hipótesis

Si se implementan métodos computacionales como los estudios QSAR entonces se podrá racionalizar el proceso de obtención de candidatos a fármacos antimaláricos potentes y seguros.

Objetivo general

Desarrollar modelos computacionales que permitan la identificación de *hits* antimaláricos

Objetivos específicos

Recopilar de la literatura un conjunto de datos adecuado para la modelación basada en ligandos de la actividad antimalárica y la toxicidad de compuestos químicos.

Obtener un conjunto de datos curado y apropiado para la modelación computacional de hits antimaláricos.

Estudiar el desempeño de métodos computacionales basados en multiclasificadores para la identificación de hits antimaláricos.

Establecer una metodología basada en clasificación para la identificación de hits antimaláricos.

CAPÍTULO 1. MARCO TEÓRICO

1.1 Malaria

1.1.1 Historia de la Malaria

La malaria o paludismo es una de las enfermedades más antiguas, afectando a la humanidad desde las épocas más remotas. Se piensa que el hombre prehistórico debió haber sufrido de malaria. Probablemente esta enfermedad se originó en África y acompañó a las migraciones humanas al Mediterráneo, a la India y a Asia Sur Oriental. Numerosos escritos médicos hacen mención a esta enfermedad a lo largo de la historia. Hacia los años 1880 se descubre las formas asexuadas y gametos del parasito, en 1897 William McCallum comprueba y describe el ciclo sexual del *Plasmodium*, años más tarde se comprueba la transmisión de la enfermedad por el mosquito *Anopheles*. Durante el siglo XX la aplicación de la medicina preventiva para el control de la malaria tuvo grandes logros(5)

La malaria fue declarada por la Organización Mundial de la Salud (OMS) como problema primordial en las reuniones de Santiago de Chile en 1954 y en México en 1955, con lo cual puso en marcha un Programa de erradicación; países como Estados Unidos, Venezuela, Cuba y otros, lograron su total erradicación, pero en la mayoría de países, después de logros prometedores, la situación se ha deteriorado por diversas causas (6).

1.1.2 Etiología

La importancia de la malaria radica en que es una de las enfermedades parasitarias que genera mayor morbilidad y mortalidad en el ser humano(7).

Este género puede dividirse en nueve subgéneros, de los cuales tres existen en mamíferos, cuatro en aves y dos en reptiles. La mayoría infectan a los animales y sólo 4 son peligrosos para el hombre (2):

- *Plasmodium vivax*. Es el agente responsable del paludismo vivax o malaria terciana, así denominado porque su ciclo febril dura 48 horas. Es el parásito que predomina en la

mayor parte de las zonas donde el paludismo es endémico y es el responsable del 43% de los casos de malaria(2, 7).

- *Plasmodium malariae*. Es el agente etiológico del paludismo cuartano, con ciclo febril de 72 horas. Es menos frecuente que el anterior atribuyéndosele el 7% de los casos de malaria y se localiza en las zonas templadas y subtropicales(2, 7).

- *Plasmodium ovale.* Se conoce desde 1922 y se encuentra mayoritariamente en África tropical aunque también aparece en Sudamérica y Asia. Este parasito tiene una muy baja incidencia(2, 7).

- *Plasmodium falciparum*. Es el agente causal del paludismo terciario maligno. Se distribuye mayoritariamente en los trópicos y zonas subtropicales y es la especie más agresiva de este género(2, 7).

El Plasmodium falciparum es el más virulento de las especies de *Plasmodium* que afectan a humanos. La malaria por esta especie reina como la suprema causa de muerte en las zonas tropicales (África, Asia y Latinoamérica), da cuenta de alrededor del 50% de los casos de malaria y de la práctica totalidad de los casos de muerte por malaria. Los merozoitos del parasito pueden invadir todo tipo de eritrocitos dado que presenta redundancia de moléculas implicadas en la invasión, por lo que la malaria ocasionada presenta mayores niveles de parasitemia que los producidos por otras especies. Se llegan a observar parasitemias de hasta el 65%, es decir, el 65% de los eritrocitos contienen parásitos. Una densidad del 25% es normalmente fatal.(2).

1.1.3 Signos clínicos

Las manifestaciones clínicas se deben sólo a las formas asexuadas de la esquizogonia hemática, ya que ni las exoeritrocíticas (intrahepáticas), ni las sanguíneas sexuadas (gametocitos) producen síntoma alguno. Estas están influidas por factores dependientes del parásito y del hospedador, por ejemplo: P. ovale y P. vivax parasitan sólo los hematíes más jóvenes, P. malariae los más viejos y P. falciparum tiene capacidad de parasitar los de todas las edades siendo sus parasitemias mucho mayores. En cuanto al hospedador, el grado de

inmunidad, innata o adquirida, que este pueda presentar (semi-inmunidad), influye en las manifestaciones del paludismo(3).

1.1.3.1 Malaria no complicada

Los primeros síntomas de este tipo de malaria son inespecíficos y similares a los de una enfermedad vírica sistémica leve. Comprenden cefalea, decaimiento, fatiga, molestias abdominales y dolores musculares y articulares, generalmente seguidos de fiebre, escalofríos, sudoración, anorexia, vómitos y creciente malestar general. Sin tratamiento la enfermedad evoluciona a una malaria grave en pocas horas (3).

1.1.3.2 Malaria grave

Se manifiesta generalmente con uno o más de los siguientes signos: coma (malaria cerebral), acidosis metabólica, anemia grave, hipoglucemia, insuficiencia renal aguda o edema pulmonar agudo. En esta fase de la enfermedad, la tasa de letalidad entre quienes reciben tratamiento suele ser del 10 al 20%. Sin embargo, sin tratamiento, la malaria grave es mortal en la mayoría de los casos. (3).

1.1.4 Epidemiología

El paludismo es una enfermedad prevenible y curable, gracias al aumento de las medidas de prevención y control la carga de la enfermedad se está reduciendo notablemente en muchos lugares. No obstante, los esfuerzos son insuficientes, ya que según las últimas estimaciones, en 2013 se produjeron 198 millones de casos de paludismo (con un margen de incertidumbre que oscila entre 124 millones y 283 millones) que ocasionaron la muerte de unas 584 000 personas (con un margen de incertidumbre que oscila entre 367 000 y 755 000) y aunque la tasa de mortalidad por malaria se ha reducido en más de un 47% desde el año 2000 a nivel mundial, y en un 54% en la Región de África es importante destacar que la mayoría de las muertes se producen entre niños que viven en África, donde cada minuto muere un niño a causa del paludismo(1).

1.1.4 Ciclo biológico

Plasmodium es un protozoo con un ciclo de vida bastante complejo que comprende una fase asexual o esquizogonia en el hospedador humano, y una breve fase sexual obligatoria en el mosquito vector. En el humano, la infección comienza con la picadura de la hembra del género *Anopheles* infectada, la cual introduce las formas del parásito denominadas esporozoítos(8), los cuales muestran una selectividad particular por las células hepáticas, en donde el parásito se desarrolla y genera el esquizonte hepático (ciclo pre-eritrocítico), que en unas dos semanas se rompe, dejando en libertad a miles de merozoítos hepáticos (10.000 a 30.000 merozoítos) que entran en el torrente circulatorio e invaden los glóbulos rojos(9).

Una vez que el parásito invade los eritrocitos comienza una nueva fase, denominada ciclo eritrocítico. La invasión es extremadamente rápida (~30 segundos) y continúa con la formación de una vacuola parasitófora donde se localiza el parásito(10), que en esta fase se denomina anillo. A partir de 15 h post-invasión existe un progresivo incremento en la actividad metabólica y biosintética dentro de la célula infectada y el parásito pasa al estadio de trofozoíto caracterizado al microscopio por presentar un pigmento característico denominado hemozoína(11, 12).

Aproximadamente a las 40 h post-invasión se puede observar en el parásito el comienzo de la división esquizogónica. El esquizonte dará lugar a 20-30 merozoítos que son liberados cuando la célula hospedadora finalmente se escinde a 48 h post-invasión. Cada uno de los merozoítos liberados es capaz de invadir otro eritrocito y, por tanto, repetir el ciclo. Tras varias esquizogonias, las divisiones en los eritrocitos se sincronizan, por lo que el hospedador sufre de ataques febriles periódicos(13, 14).

La duración del ciclo asexual en el hombre varía dependiendo de la especie, de 48-50 h en *P. vivax*, *P. ovale* y *P. falciparum* y 72 h en *P. malariae*, resultando en el patrón de fiebres tercianas o cuartanas observado en los distintos tipos de malaria humana(15). A pesar de que la mayoría de los merozoítos que invaden los eritrocitos desarrollan el ciclo asexual, una pequeña fracción puede dar lugar a células sexuales masculinas y femeninas que se denominan gametocitos. Cuando un mosquito vuelve a picar al individuo, los gametocitos que ingieren con la sangre, debido a la menor temperatura y a la presencia de ácido xanturénico en

el estómago del insecto, se desarrollan en gametos que luego dan lugar a los macro y microgametocitos(12, 16).

La fecundación, que se produce en el estómago del mosquito, genera un cigoto diploide, que se transforma en un ooquineto que, tras atravesar las paredes del estómago, se localiza en la lámina basal y segrega una cubierta ooquística. Este ooquiste crece y experimenta una esporogonia que rinde un gran número de esporozoítos, por ello a esta fase del ciclo en el interior del mosquito se la conoce como ciclo esporogónico. La mayoría de estos esporozoítos se dirigen al interior de las glándulas salivales, desde donde se inyectarán a un nuevo hospedador vertebrado a través de una nueva picadura e iniciando nuevamente el ciclo(12, 17). El ciclo biológico del *Plasmodium* puede observarse en la Figura 1.1 del Anexo 1.

1.1.5 Fármacos antimaláricos. Clasificación

De acuerdo con su eficacia frente a las diversas etapas por las que transcurre el ciclo vital del plasmodio, los antimaláricos se pueden clasificar del siguiente modo:

- Cura clínica:

Los fármacos curan el ataque clínico de malaria porque eliminan las formas asexuadas del parásito, ya que se comportan como esquizonticidas sanguíneos(18):

- Cloroquina
- Hidroxicloroquina
- Amodiaquina
- Quinina
- Pirimetamina
- Pirimetamina/sulfadoxina
- Pirimetamina/Dapsona
- Cloroguanida
- Mefloquina
- Halofontrina
- Artemisinina
- Artesunato
- Artemeter

El *P. falciparum* puede desarrollar resistencia a la Cloroquina, en cuyo caso se recurre a la quinina o a derivados de la artemisina(18).

- Cura radical

Pretende suprimir tanto las formas asexuadas sanguíneas como tisulares. En el caso de *P. falciparum* y *P. malariae* basta con un esquizonticida sanguíneo, ya que las formas exoeritrocíticas terminan por desaparecer, pero en el caso del *P. vivax* y *P. ovalae* la eliminación de hipnozoítos requiere la administración de un esquizonticida tisular: la Primaquina o la Pirimetamina pueden ser útiles en este último caso(18).

- Profilaxis clínica:

Se lleva a cabo con los mismos fármacos que se realiza la cura clínica siempre que se administren antes, durante y después de un posible contacto(18).

- Profilaxis causal:

Se emplean los esquizonticidas tisulares que actúan sobre las formas primarias hepáticas; de este modo se evita la posterior invasión en los hematíes y la transmisión ulterior a los mosquitos. Se emplean la Cloroguanida y la Pirimetamina(18).

- Gametocidas:

Destruyen las formas sexuadas eritrocíticas. Tienen esta actividad la Primaquina, sobre todo frente a P. falciparum y la Cloroquina y Quinina frente a *P. vivax* y *P. malariae*(18).

1.1.6 Tratamientos antimaláricos clásicos

Existen distintos tratamientos contra la malaria, pero su eficacia se ve comprometida por el desarrollo de resistencias. Aunque la mayoría de los antimaláricos actúa contra la fase asexual eritrocítica, estos fármacos deberían actuar en la etapa hepática del parásito para reducir el progreso de la infección y evitar recidivas, en la sanguínea para aliviar los síntomas clínicos, y en la gametocítica para inhibir el ciclo de transmisión y proteger a otros seres humanos. Los antimaláricos más comunes afectan a algunas de las dianas clásicas del parásito(19) y sus estructuras se muestran en la Figura 1.2 del Anexo 1.

En cuanto a las combinaciones de antimaláricos, actualmente las terapias de combinación basadas en las Artemisinas siguen siendo la piedra angular de los tratamientos contra *Plasmodium falciparum*, por lo que es comprensible la alarma que generó la detección de parásitos resistentes a las mismas que motivó el lanzamiento del “Plan Mundial de Contención

de la Resistencia a la Artemisinina" en enero de 2011. No obstante, las combinaciones de derivados de artemisinina con otros antimaláricos siguen siendo esencial, como demuestra el hecho de que a finales de 2011 la Agencia Europea del Medicamento aprobara la combinación de Dihidroartemisina y Piperaquina Eurartesim® (Figura 1.3 del Anexo 1). Otra combinación ampliamente utilizada es la de Artemeter, derivado de la Artemisina, y la Lumefantrina (un inhibidor de la polimerización del grupo hemo) desarrollada por la Fundación Novartis (Figura 1.4 del Anexo 1) (19).

1.1.7 Relación entre la estructura y la actividad (REA) de los principales grupos de fármacos antimaláricos.

Una gran cantidad de sustancias farmacéuticas, han sido aisladas de fuentes vegetales, tales como: Quinina; Quinidina, Cinconina, Cinconidina, Artemisina; mientras que otras se han obtenido sintéticamente, tal es el caso de la Cloroquina, Pamaquina y otros. El uso de estos compuestos y sus derivados ha sido ampliamente difundido en la actualidad para luchar contra la amenaza de la malaria en muchos países del mundo, todos estos con una elevada diversidad estructural y aunque existen patrones comunes dentro de cada uno de los grupos antimaláricos: derivados de la quinina, análogos de las 4-aminoquinolinas, 8-aminoquinolinas, derivados de las 9-aminoacridinas, sulfonas, artemisinas y otros, no se ha establecido claramente una relación estructura-actividad entre estos grupos(20). Las estructuras de los compuestos que componen cada uno de estos grupos se muestran en el Anexo 2.

1.1.8 Resistencia a los medicamentos antimaláricos

La resistencia a los medicamentos está documentada para todas las clases de antimaláricos (Amodiaquina, Cloroquina, Mefloquina, Quinina y Sulfadoxina-Pirimetamina) y, más recientemente, a derivados de la Artemisina., y es una seria amenaza para el control de la malaria. El uso generalizado e indiscriminado de antimaláricos ejerce una fuerte presión selectiva para que los parásitos desarrollen altos niveles de resistencia. Es posible prevenir esta resistencia, o desacelerar considerablemente su inicio, combinando antimaláricos con diferentes mecanismos de acción y procurando tasas de cura muy elevadas mediante una total adherencia a la posología correcta de los regímenes terapéuticos(3).

1.2 Quimioinformática y desarrollo racional de fármacos.

1.2.1 Generalidades del proceso de descubrimiento de fármacos

La investigación terapéutica es un proceso en extremo complejo y costoso en términos de tiempo y dinero. Estudios recientes muestran que el tiempo y costo promedio para el desarrollo de un fármaco, desde la identificación del blanco hasta la obtención de la aprobación por la FDA, estaba alrededor de los 15 años y los 900 millones de dólares cifra que puede aumentar o disminuir dependiendo del fármaco en desarrollo en cuestión. En las pasadas décadas sin embargo, los avances tecnológicos, junto con la progresiva reconsideración del proceso en si, ha conducido a una gran revolución del proceso de encontrar un nuevo fármaco(21).

Aunque por muchos años se pensó que el cuello de la botella en el descubrimiento de fármacos era la generación de *hits* el verdadero problema radicaba en la obtención del compuesto líder, etapa que consumía gran cantidad de recursos, no solo monetarios sino también de tiempo, la Quimioinformática y el desarrollo de herramientas computacionales cada vez más eficaces ha permitido la racionalización del proceso de obtención de fármacos actuando precisamente sobre la etapa de obtención del compuesto líder (21).

1.2.2 Criterios para la selección de hits

Cada organismo causante de una enfermedad tiene sus propios desafíos, lo que resulta en diferencias a la hora de establecer criterios para la selección de candidatos a *hits* y *leads*. Luego de años de investigación muchos criterios han sido establecidos para la selección de candidatos a *hits* estos pueden ser clasificados en dos grupos fundamentales: criterios específicos de la enfermedad y criterios específicos para los candidatos. En el proceso de selección de hits los criterios relacionados con la potencia y citotoxicidad admisible, para que el hit sea consistente con el potencial de ofrecer un compuesto *lead*, pueden ser muy útiles en los estudios QSAR(22).

Los *hits* validados como posibles candidatos en el tratamientos de la malaria provocada por *P. falciparum* se rigen por criterios de potencia y citotoxicidad, en cuanto al primero un hit debe tener una concentración inhibitoria media(CI50) menor o igual que 1µM para sepas sensibles y

resistentes, y relacionado con la citotoxicidad se plantea que debe existir una mayor selectividad por el parásito en líneas celulares de mamíferos tal que la selectividad sea 10 veces mayor o sea que los valores de la concentración efectiva media(CE50) deben ser 10 veces mayor que los de la concentración inhibitoria media(CI50)(22).

1.2.3 Proceso de curación del conjunto de datos

En todo proceso de modelación, el pre-procesamiento de los datos desempeña un rol fundamental. Un modelo, por definición, es una representación simplificada de la realidad, con un dominio de aplicación limitado y sujeto a ciertos márgenes de error. En base a estas posibles limitaciones, la idea de construir modelos basados en datos imprecisos y ruidosos no constituye un camino razonable para brindar soluciones potenciales a problemas de la vida real, por lo que se hace necesario implementar protocolos estándar de curación de los datos químicos, y aunque no existe una metodología fija algunas pautas fueron establecidas por Fourches y col (23).

1.2.3.1 Remoción de compuestos inorgánicos

La mayoría de los *softwares* empleados para los estudios QSAR no son capaces de extraer información de estructuras de moléculas inorgánicas, de hecho la mayoría de los *softwares* solo son capaces de trabajar con moléculas orgánicas por lo que estructuras que contengan átomos de: B, Al, Si, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Mo, Ag, Cd, In, Sn, Sb, Te, Gd, Pt, Au, Hg, Tl, Pb, Bi, pueden inducir errores en la codificación de la información estructural. Atendiendo a que muy pocos compuestos con elementos inorgánicos presentan relevancia en el proceso de descubrimiento de fármacos se hace aconsejable su remoción antes del calculo de los descriptores(23).

1.2.3.2 Conversión estructural

Consiste en la conversión estructural de los SMAILS a estructuras o grafos moleculares bidimensionales para esto pueden ser empleados diversos sofwares,por ejemplo: MOE (*Molecular Operating Environment*), Sybyl, OpenBabel entre otros, también se puede emplear el *plugin JChem for Excel* de *ChemAxon*(23).

1.2.3.3 Normalización de quimiotipos específicos

Muchos de los grupos funcionales existentes puede ser representados estructuralmente de diversas maneras, el grupo nitro y similares son una muestra de esto.

Para la Quimioinformática esta situación puede traer serios problemas porque descriptores moleculares calculados para las diferentes formas de representación de un mismo grupo pueden tener diferentes significados y por lo tanto muchas estructuras podrían no ser reconocidas como idénticas. Es importante representar cada grupo funcional de la misma forma, para evitar este tipo de problemas. Un ejemplo de esto se puede observar en la Figura 3.1 del Anexo 3, donde Fourches y colaboradores mostraron cinco representaciones diferentes de la misma molécula (23).

1.2.3.4 Tratamiento de tautómeros

En muchos casos los compuestos pueden existir en distintas formas tautoméricas, un ejemplo clásico de esto es el equilibrio ceto-enólico, las presencia de formas tautómeras puede tener un impacto significativo no deseado en la capacidad de predicción de los modelos(23).

1.2.3.5 Remoción de duplicados

Un análisis estadístico riguroso de un conjunto de datos asume que cada compuesto es único y estructuralmente diferente a los otros compuestos. Comúnmente en bases de datos químicos, sobre todo en aquellas de mayor tamaño se presentan estructuras duplicadas, cuya presencia puede afectar la distribución del conjunto de datos. Como consecuencia, la remoción de los duplicados es de alta prioridad en los estudios de modelación QSAR(23).

1.2.3.6 Inspección manual

Esta es la etapa final del proceso de curación y la que requiere de mayor cuidado. Implica, de ser posible, la inspección de todas las estructuras en el conjunto de datos, o al menos una

muestra representativa de estas. Se verifica que no se hayan producido cambios accidentalmente en la información original y se presta una atención especial a la detección de errores en la estructura de las moléculas (23).

1.2.4 Criterios para la selección de los subconjuntos de entrenamiento, prueba y validación externa.

En la validación de los modelos QSAR se suele dividir los conjuntos de datos de partida en subconjuntos de entrenamiento, de prueba y de validación externa. En varias publicaciones (24-26) se ha alertado sobre la necesidad de validar adecuadamente los modelos obtenidos, y entre los principales aspectos mencionados se encuentra la adecuada partición de los conjuntos de datos. En cuanto a este aspecto, en (24) Tropsha recomienda como se debe proceder. Primero, la selección del subconjunto de validación externa se puede llevar a cabo mediante la selección aleatoria de hasta el 20% de las instancias del conjunto de datos de partida. El resto de las instancias quedan reservadas para la modelación. En el caso del subconjunto de modelación obtenido, varios autores (25-29) coinciden en que este debe ser dividido en subconjuntos de entrenamiento y prueba. Aquí, el objetivo fundamental es que ambos sean representativos de todo el espacio de descriptores del conjunto de datos original. De esta manera, cada punto del subconjunto de prueba estará cercano al menos a un punto del subconjunto de entrenamiento. Para la selección de los subconjuntos de modelación se emplean diversos métodos siempre tratando de garantizar la representatividad de los subconjuntos de entrenamiento y prueba, por lo que la selección aleatoria no se emplea, entre los métodos más utilizados se destacan los métodos de agrupamiento y aun más prometedores los métodos de exclusión de esferas.

1.2.5 Los activity cliffs en la modelación QSAR.

En esencia, un modelo QSAR consiste en correlacionar los aspectos estructurales de las moléculas con su actividad biológica. La modelación QSAR ha constituido una poderosa herramienta en el proceso de diseño de fármacos(30, 31). En un mundo en el cual el desarrollo de nuevas tecnologías en las diferentes ramas de la ciencia ha acrecentado los volúmenes de datos disponibles. A pesar de sus ventajas, todavía existen grandes limitaciones en el empleo de los modelos QSAR. Aun siguiendo cuidadosas metodologías o “buenas prácticas”

reportadas en la literatura para la construcción y evaluación de modelos más fiables(23, 32), en ocasiones no se logra obtener la capacidad predictiva deseada debido a las influencias negativas de diferentes factores en la obtención de modelos QSAR. Entre estos, resalta la naturaleza de los paisajes de estructura-actividad basados en una representación estructural dada, y a consecuencia de estos, la presencia de los *activity cliffs(33)*. Los *activity cliffs* son pares de compuestos que presentan elevada similitud estructural pero grandes diferencias en los valores de actividad biológica o propiedad modelada(34).

1.2.6 Descriptores moleculares

La adecuada representación de la estructura molecular de un compuesto es un factor decisivo para el desarrollo de cualquier estudio de Correlación Estructura-Actividad(35). El descriptor molecular es el resultado final de un procedimiento matemático y lógico, que transforma la información química codificada, en una representación simbólica de una molécula en un número útil o el resultado de un experimento estandarizado.
De acuerdo con la definición dada, los descriptores se pueden clasificar en dos grandes grupos:

1) Mediciones experimentales: Refractividad molar, Momento dipolo, Polarizabilidad, etcétera(36).

2) Descriptores moleculares teóricos: Se derivan de las representaciones simbólicas de las moléculas y pueden ser clasificados a su ves en:

    - Descriptores 0D: Se derivan de la fórmula química de la molécula. Se puede decir que son independientes de la estructura molecular(37, 38).

    - Descriptores 1D: Basados en la información del tipo lista sub-estructural(39).

    - Descriptores 2D: Basados en la representación en dos dimensiones de la molécula, la cual se basa en el conocimiento de la conectividad entre los átomos que la forman en términos de la presencia o ausencia de enlaces químicos entre ellos(40).

- Descriptores 3D: Basados en la representación tridimensional de la molécula como un objeto rígido.

- Descriptores 4D: Derivados de la representación estereo electrónica de la molécula que se relaciona con las propiedades moleculares y que dependen de la interacción de la distribución electrónica con una sonda que caracteriza el ambiente. (campos de interacción molecular)(35).

Las principales características de un descriptor se representan en una taxonomía de tres niveles. La primera de ellas ya se expresó basada en la representación molecular. El segundo nivel tiene que ver con la representación matemática de los descriptores. Aquí se destacan los que se representan en forma de valor escalar, vector, matrices entre otros.
El tercer nivel caracteriza las propiedades de invarianza de los descriptores. Esto se define como la habilidad que tiene el algoritmo de generación del descriptor para ser independiente de las características particulares de la representación molecular(numeración de los átomos, conformación molecular, etcétera.), siendo la primera de estas características la asumida como mínima para cualquier descriptor(41). Los descriptores moleculares además, deben poseer determinados requerimientos básicos como son tener interpretación estructural, una buena correlación con al menos una propiedad, discriminar entre isómeros, ser simples entre otras (42).

1.2.7. Algoritmos genéticos como método de selección de atributos

Los AG son métodos eficientes en la optimización de funciones Estos simulan la evolución natural mediante la modelación de poblaciones dinámicas de soluciones. Los conjuntos de atributos seleccionados constituyen los cromosomas y estos a su vez las poblaciones de soluciones. Frecuentemente estas se codifican utilizando valores binarios para identificar cuales descriptores son seleccionados, y cuales se eliminan para cada miembro de la población generada. Cada cromosoma conduce a un modelo construido utilizando los atributos codificados. Para cada modelo, el error en la predicción sobre el conjunto de datos de entrenamiento es cuantificado y sirve como una función de evaluación o ajuste. Durante el curso de la evolución, los cromosomas se someten al cruce y mutación. Promoviendo la

supervivencia y reproducción de los cromosomas de mayor ajuste, el algoritmo minimiza de manera efectiva la función de error en las siguientes generaciones(43).

El desempeño de los AG depende de varios factores. Los parámetros que rigen el cruce, mutación y supervivencia de los cromosomas deberían ser cuidadosamente escogidos para permitir explorar el espacio de soluciones y prevenir una convergencia temprana hacia poblaciones homogéneas ocupando un mínimo local. La selección de una población inicial también es importante en la selección de atributos mediante este tipo de algoritmos(43).

1.2.8 Método de clasificación

1.2.8.1 Máquinas de soporte vectorial de mínimos cuadrados.

Una Máquina de Soporte Vectorial (SVM) aprende la superficie decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento. *Least-Squares Support Vector Machines* (LS-SVM) son una reformulación de las máquinas estándar de soporte vectorial (SVM), esta nueva formulación asignan los datos de entrada en un espacio de característica de alta dimensión utilizando un mapeo no lineal φ (x) para esto se establece un espacio de características determinado por un hiperplano de separación lineal(44).

1.2.9 Métodos basados en multiclasificadores

Dado que una de las tareas fundamentales en el área de la clasificación supervisada es mejorar la eficacia de los métodos existentes, surge la idea de combinar múltiples clasificadores de la misma forma en que intentamos buscar varias opiniones en los problemas que resuelven las personas a diario. De esta forma una combinación de clasificadores es un clasificador que basa

su decisión en la decisión de otros clasificadores a los cuales nos referiremos como los clasificadores individuales o modelos base.

El objetivo con el desarrollo de combinaciones de clasificadores es que estos tengan un mejor desempeño que los clasificadores individuales, o sea, presentar mejores soluciones a los problemas que no han sido resueltos de manera aceptable por los métodos tradicionales(45).

Sin embargo se hace necesario construir una combinación donde los clasificadores no cometan los mismos errores y en la que estos sean capaces de mejorar el desempeño de un clasificador individual. Existen tres razones por las que el uso de una combinación de clasificadores debe ser considerado, planteadas por Dietterich y Kuncheva y se exponen a continuación. La primera razón es estadística. En la práctica, una vez dado el conjunto de entrenamiento, la tarea es encontrar un clasificador lo más eficaz posible. Sin embargo, puede darse el caso de que en lugar de encontrar un solo clasificador que es el más eficaz, se obtengan varios con una eficacia igual de aceptable, pues todos tienen un error estimado muy semejante o igual. Sin embargo, estos clasificadores pueden generalizar de manera diferente, es decir, cuando se clasifican objetos no presentes en la muestra disponible los clasificadores no tienen por qué trabajar de la misma manera, estos pueden cometer errores sobre elementos distintos. Una opción disponible es tomar aleatoriamente uno de estos clasificadores, entonces el clasificador que se tome puede ser la mejor de las opciones, pero también la peor. Una alternativa puede ser tomarlos todos y de alguna manera promediar los resultados de estos para intentar reducir el riesgo de una clasificación incorrecta(45).

La segunda razón es computacional. En muchos casos los algoritmos de clasificación realizan una búsqueda en un conjunto de clasificadores posibles a partir del conjunto de entrenamiento en cuestión con el objetivo de encontrar un buen clasificador. En la literatura al conjunto de clasificadores posibles se le llama conjunto de las hipótesis. Realizar esta búsqueda de manera extensiva, o sea, tenerlos en cuenta a todos, puede ser muy costoso. Por esta razón, algunos algoritmos de clasificación realizan la búsqueda a través de alguna heurística, y esto trae como consecuencia la posibilidad de quedar atrapado en un óptimo local. Esto implica que en muchos casos aunque la información disponible acerca del problema – conjunto de entrenamiento – permita encontrar un clasificador con una alta eficacia, por problemas de complejidad computacional no será posible acceder a este. Entonces, si se toman varios clasificadores, – cada uno pudo haber sido un óptimo local distinto, por lo que no tienen por qué cometer los

mismos errores – al combinarlos en búsqueda de consenso se puede lograr una mejor eficacia(45).

La tercera razón es representacional. Como se ha visto anteriormente un algoritmo de clasificación define una familia de clasificadores. Existe la posibilidad de que el clasificador óptimo para un problema dado no se encuentre dentro de la familia de clasificadores que define el algoritmo de clasificación que se está utilizando(45).

Es importante siempre tener en cuenta que una combinación de clasificadores es también un clasificador.

1.2.10Selección de clasificadores y combinación de la salida de estos.

Se sabe que para la formación de un buen multiclasificador es necesario tanto definir cómo se van a crear modelos diferentes y a seleccionar un subgrupo de estos, como de que manera se van a combinar los resultados de cada uno de los modelos para producir la predicción final.

La selección de clasificadores es un enfoque que se encarga de seleccionar cuáles son los mejores clasificadores, mientras que la manera en que se combinan los resultados de los clasificadores individuales es una cuestión fundamental en el estudio de las combinaciones de clasificadores. La idea aquí es optimizar la decisión que se va a tomar a partir de las decisiones individuales. Existen diferentes métodos de combinación de las salidas, entre estos se encuentran, Votación Mayoritaria, Votación ponderada por exactitud, sensibilidad y especificidad, Voto de Borda, *AdaBoost* entre otros(45).

1.2.11 Diversidad: la piedra angular de los sistemas basados en multiclasificadores

Si tuviese acceso a un clasificador con capacidad de generalización perfecta, no habría ninguna necesidad de recurrir a técnicas por multiclasificadores, sin embargo la realidad es otra: ruidos, valores atípicos y la superposición de las distribuciones de datos, hacen de este clasificador ideal una propuesta imposible. A lo sumo, podemos esperar clasificadores que asignen correctamente los datos del conjunto de entrenamiento. Por tanto, la estrategia en sistemas de multiclasificadores es crear muchos clasificadores, y combinar sus salidas de tal

manera que esta combinación mejore el rendimiento de un solo clasificador como se planteó anteriormente, esto requiere, sin embargo, que los clasificadores individuales cometan errores en instancias diferentes. La intuición es que si cada clasificador hace diferentes errores, a continuación, una combinación estratégica de estos puede reducir el error total. En concreto, es necesario formar clasificadores cuyos límites de decisión sean adecuadamente diferente a los de los demás. Cuando se obtiene un multiclasificador con estas condiciones , estamos hablando de un multiclasificador diverso(45).

CAPÍTULO 2. MATERIALES Y MÉTODOS.

2.1 Preparación de los datos

2.1.1 Descripción del conjunto de datos

Los datos químicos se obtuvieron de la base de datos de Novartis, estos contenían 5697 estructuras codificadas como SMAILS y valores de potencia expresados como Concentración Efectiva Media (CE50) para una cepa de P. falciparum sensible(3D7), valores de potencia(CE50) para una cepa de P. falciparum resistente a cloroquina, quinina, pirimetamina, cicloguanil y sulfadoxina (W2) y valores de citotoxicidad espresados como Concentración Inhibitoria Media (IC50) sobre líneas celulares de carcinoma hepatocelular(Huh7).

2.1.2 Selección de *hits* y remoción de compuestos que no pueden ser asignados como activos o inactivos

La asignación de los compuestos en los grupos de activos e inactivos, como clase 1 o clase 0 respectivamente se realizó empleando *Jchem for Excel* que forma parte del paquete *ChemAxon*. Los criterios para considerar un compuesto como hit son los discutidos en la sección XXX del Capítulo 1. Es importante destacar que para seleccionar un compuesto como hit este debe cumplir con los criterios de potencia medidos como IC50 para la cepa sensible (3D7) y la cepa resistente (W2) y con el criterio de toxicidad (Huh7) simultáneamente. Las moléculas que no satisfagan paralelamente estos tres criterios son consideradas como inactivas. De esta manera se genera un campo con la etiqueta activos/inactivos. Los compuestos cuyas mediciones no permiten asignarlos a alguna de estas clases se excluyeron del proceso de modelación.

2.1.3 Procedimiento de curación de los conjuntos de datos.

Para la eliminación de compuestos inorgánicos se calcula la fórmula global de los compuestos, y luego estas se filtraron a partir de la búsqueda de los elementos mencionados en el Capítulo 1, sección XXX. Esto permitió la detección y eliminación de compuestos inorgánicos, organometálicos y que contienen elementos traza o poco representados en el conjunto de datos. Este procedimiento se realizó empleando *JChem for Excel.*

El conjunto de datos de partida se encuentra en formato SMILES y su conversión a estructuras bidimensionales se llevó a cabo con *JChem for Excel*. En la normalización de las estructuras se tomaron en cuenta los elementos representados en el conjunto de datos, y en base a estos se escogieron los quimiotipos a normalizar. El paso siguiente consistió en la generación del fichero de configuración SDF con las posibles transformaciones, para lo cual se empleó el módulo *Standardizer* de *ChemAxon* (46). Se normalizaron las estructuras, teniendo en cuenta las siguientes transformaciones: eliminación de los átomos de hidrógeno, aromatización de los anillos, la normalización de los quimiotipos específicos y la curación de las formas tautoméricas. A continuación se realizó una inspección visual para detectar deficiencias en este proceso, y de ser necesario, repetirlo iterativamente. La detección y eliminación de los duplicados fue el último paso en esta etapa, el *software* EdiSDF del proyecto ISIDA (47) se empleó con este fin. Finalmente se realizó un proceso de inspección manual lo más minucioso posible.

2.1.4 Detección y eliminación de activity cliffs del conjunto de modelación.

Para la detección de los *activity cliffs* es necesario establecer dos criterios. El primero se relaciona con la similitud estructural y el segundo con la diferencia de potencia. En cuanto a la métrica de similitud estructural, el coeficiente de Tanimoto es el más ampliamente utilizado en Quimioinformática y es el que se emplea. Por otra parte, los *fingerprints* son las representaciones más idóneas para codificar las similitudes globales de las estructuras, por lo tanto son la forma de representación elegida para esta tarea. Finalmente como criterio de actividad se emplea la clase asignada(48).

La estructura de los compuestos se codificó empleando *fingerprints* del tipo ECFP (por sus siglas en inglés, *Extended Connectivity Fingerprints*). El valor de corte de similitud del coeficiente de Tanimoto para esta representación se tomó igual a 0.55 de acuerdo a recomendaciones encontradas en la literatura (49).

El algoritmo empleado para la detección de compuestos generadores de *activity cliffs* ha sido previamente validado y se encuentra en proceso de publicación. Este algoritmo toma como entrada un conjunto de estructuras químicas que se han asignado a dos grupos y devuelve una lista de los compuestos generadores de *activity cliffs* que se van eliminando en cada ciclo. A continuación se describen los pasos para la detección y eliminación compuestos generadores de *activity cliffs(48).*

Pasos del algoritmo para la detección y eliminación de los *activity cliffs*

Paso 1: Calcular los *fingerprints* para las estructuras.

Paso 2: Calcular la matriz de similitud $S = \sum\sum s_{i,j}$ para el *fingerprint* de tipo ECFP calculado en el paso 1.

Paso 3: Calcular la matriz de distancia entre moléculas $D = \sum\sum d_{i,j}$, a partir de la matriz de similitud obtenida en el paso 2, donde cada valor de distancia está dado por: $d_{i,j} = 1 - s_{i,j}$.

Paso 4: Buscar los pares de compuestos (i,j) que cumplan con: $d_{i,j} \leq 0{,}45$ y que tengan clase opuesta. Agregar los pares identificados a la lista de candidatos de generadores de *activity cliffs* (LC)

Paso 5: Inicializar la lista de compuestos a excluir (LE) como una lista vacía.

Paso 6: Mientras LC no esté vacía, para cada candidato en LC calcular: el número de compuestos con los que forma *activity cliffs*.

Paso 7: Agregar a la LE el compuesto que forme la mayor cantidad de *activity cliffs*, actualizar LC quitando todos los pares en los que aparece el compuesto movido a LE y comenzar de nuevo a partir del Paso 6. De existir empate, definir una LC* que contenga solamente estos compuestos, e ir al paso siguiente.

Paso 8: Para cada compuesto en LC*, determinar cuáles pertenecen a la clase mayoritaria, si existe uno solo, agregarlo a la LE, actualizar LC y comenzar de nuevo a partir

del Paso 6. De haber empate, actualizar la LC* dejando solamente estos compuestos, e ir al paso siguiente.

Paso 9: Para cada compuesto en LC*, buscar el más cercano a la frontera de clasificación, agregarlo a la LE, actualizar la LC* y comenzar de nuevo a partir del Paso 6. Si existe empate, agregar uno de los compuestos en LC* aleatoriamente a LE, actualizar LC y comenzar de nuevo a partir del Paso 6.

Paso 10: Eliminar de conjunto de datos los compuestos en LE.

Para ejecutar los pasos 1 y 2 existe gran diversidad de herramientas disponibles. En esta investigación se escoge el paquete *Mayachemtools* (50), el cual se emplea para calcular los valores de los *fingerprints* de tipo ECFP y de las matrices de similitud.

2.1.5 División del conjunto de datos

Para la obtención de modelos QSAR predictivos el conjunto de datos se dividió en subconjuntos de modelación y validación externa. Este primer subconjunto se dividió a su vez en subconjuntos de entrenamiento y prueba. El subconjunto de entrenamiento se utiliza para la selección de variables y para el entrenamiento de los clasificadores, el de prueba para la validación y selección de los modelos y finalmente el subconjunto de validación externa solamente se emplea para la validación externa de los modelos.

Para la formación del subconjunto de validación externa se seleccionó aleatoriamente el 20% del total del conjunto de datos iniciales curado como se explicó en el epígrafe 2.1.2. El resto de los compuestos se reservó como subconjunto de modelación. El subconjunto de modelación fue dividido en subconjuntos de entrenamiento y prueba empleando tres algoritmos de exclusión de esferas (SE) según lo propuesto en (51) y aplicados en nuestro laboratorio. Los algoritmos SE aseguran la cercanía en el espacio químico de los subconjuntos de entrenamiento y prueba.

Para la división del conjunto de modelación se emplearon las variantes 1M, 2M y 3M del algoritmo de SE implementadas en MATLAB (52). A diferencia de los algoritmos originales, para la partición de los datos basada en SE las estructuras de los compuestos se codifican como

Fingerprints topológicos de 1024 bits del tipo ECFP del *ChemAxon* con el programa *GenerateMD* (53). El coeficiente de Tanimoto se selecciona como la medida de distancia o criterio de similitud.

El algoritmo comienza con la selección de un compuesto a partir del conjunto de datos y la construcción de una hiperesfera alrededor de él. A continuación, el compuesto más cerca del centro de la hiperesfera es asignado al conjunto de selección mientras que el centro y el resto de los compuestos dentro de ella se asignan al conjunto de entrenamiento. A continuación los compuestos dentro de la hiperesfera se eliminan del grupo inicial de casos y el proceso se repite hasta que el grupo inicial de compuestos este vacío.
Para evaluar la calidad de las particiones de los conjuntos de datos obtenidos para cada radio de la esfera se evalúan tres parámetros: el índice de diversidad del subconjunto de prueba con respecto al subconjunto de entrenamiento $M_{test,train}$ (c), el índice de diversidad del subconjunto de entrenamiento con respecto al subconjunto de prueba $M_{train,test}$ (c) y el índice de diversidad del subconjunto de entrenamiento I_{train} (c). Estos parámetros caracterizan la cercanía de los compuestos en el subconjunto de prueba a los compuestos del subconjunto de entrenamiento, la cercanía de los compuestos en el entrenamiento a los del subconjunto de prueba y la diversidad del subconjunto de datos de entrenamiento, respectivamente(54). Una partición ideal del conjunto de datos debería producir valores de $M_{test,train}$(c) = 0, $M_{train,test}$ (c) = 0 y Itrain (c) = 1 (54).

2.1.6 Codificación estructural

Para la codificación de la información estructural se emplearon descriptores 2D obtenidos a partir del software Fragmentor2013 del proyecto ISIDA(47). Para el cálculo se introduce al programa la base de datos curada en formato .sdf. Este software ofrece tres tipos fundamentales de fragmentaciones: “secuencias” (I), “fragmentos centrados en átomos” (II) y “tripletes” (III), para el cálculo de nuestros descriptores se emplean los siguientes subtipos de fragmentos:

-t3: Secuencias de átomos y enlaces

-t6: Fragmentos centrados en átomos basados en secuencias de átomos y enlaces

-t9: Fragmentos centrados en átomos basados en secuencias de átomos y enlaces de longitud fija

-t10: Tripletes

Como longitud mínima de los fragmentos se emplea una secuencia de dos átomos y como longitud máxima ocho átomos. Se emplea como formato de salida el formato SVM que contiene el valor de los fragmentos (descriptores) para cada compuesto. Además el programa genera un archivo de cabecera que contiene la lista de todos los fragmentos descubierto en el conjunto de datos. El cálculo inicial de los descriptores se realiza para el conjunto de datos de entrenamiento.

Una vez calculados los descriptores, se eliminan lo que tienen 99% o más de sus valores constantes. Este proceder constituye el primer filtro aplicado en la reducción de dimensionalidad y busca eliminar los descriptores con poca varianza cercana a 0. Luego se aplica otro filtro que busca obtener los 250 descriptores más relevantes y con el mínimo de redundancia empleando el algoritmo MRMR. A partir de los 250 descriptores obtenidos para el subconjunto de entrenamiento se generan las plantillas con las cuales se calculan los descriptores para el resto de los subconjuntos existentes.

2.2 Modelación QSAR

Para la modelación QSAR se toma como entrada un conjunto de datos que contiene *n* muestras y *m* variables $Z=\{ x_1, ..., x_n\}$ así como un vector de clase observado **y**. Este conjunto de datos se divide en tres subconjuntos diferentes: entrenamiento, prueba y validación externa.

En base a esto, el algoritmo de modelación propuesto comienza con el entrenamiento de un conjunto de modelos base que, en general, por si solos carecen de capacidad predictiva. Posteriormente se seleccionan un subconjunto de estos modelos para ser parte del multiclasificador y finalmente estos modelos seleccionados son agregados en el multiclasificador final. Para el entrenamiento de los modelos base se emplea un muestreo aleatorio de las variables considerando todos los compuestos en el conjunto de entrenamiento.

Por otra parte, en la selección de los modelos base para el multiclasificador se exploran todos los modelos, subconjuntos de estos seleccionados empleando análisis de conglomerados y algoritmos genéticos como técnicas de muestreo. En la etapa de agregación de los modelos base se emplean técnicas de Voto Mayoritario, Voto de Borda y *AdaBoost*. En la Figura 2.1 se resume la estrategia de modelación empleada y en los epígrafes siguientes se explica cada uno de los bloques en esta Figura. La metodología de modelación QSAR propuesta fue implementada en MATLAB(52).

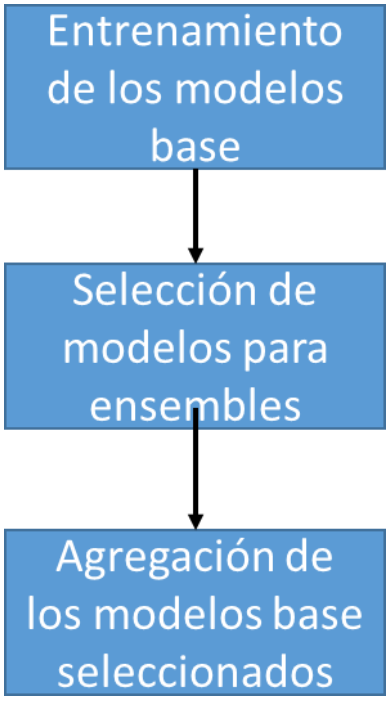


Figura 2.1 Esquema general de la metodología QSAR implementada

Al emplear tres métodos de agregación diferentes para construir los multiclasificadores se tienen tres metodologías de modelación por cada método de selección de los modelos que formarán parte del multiclasificador. En el caso de la selección basada en conglomerados se emplean dos medidas diferentes de distancia entre modelos y además se emplean dos variantes de función de ajuste para los Algoritmos Genéticos. Teniendo estos aspectos en cuenta se tienen 12 metodologías de modelación diferentes que se emplean en la búsqueda del mejor multiclasificador.

2.2.1 Entrenamiento de los modelos base

Obtener los modelos base es el primer paso para la modelación basada en multiclasificadores. Estos modelos no necesariamente poseen poder predictivo pero si permiten su combinación para mejorar el rendimiento respecto a un clasificador único.

Para generar los modelos bases se empleó un algoritmo de clasificación basado en Máquinas de Soporte Vectorial de Mínimos Cuadrados (LS-SVM) que es un enfoque de modelado no lineal. Para la obtención de estos modelos bases se emplea el conjunto de datos de entrenamiento. Para garantizar la diversidad de los modelos base se hace un muestreo aleatorio de las variables tomando una cantidad de variables también aleatoria entre 5 y 25.
Los modelos LS-SVM se desarrollan mediante *LS-SVMlab Toolbox* para MATLAB(55). Cada variable se escala al intervalo {0-1} y se seleccionó como núcleo (kernel) la función de base radial (RBF) que se define por la transformación $K(x, x_k) = e^{\frac{-\|x-x_k\|^2}{2\sigma^2}}$. El parámetro de regularización γ, y el parámetro del núcleo σ^2 de los modelos LS-SVM se optimizaron utilizando una combinación de Recocido Simulado (Simulated Annealing, SA) y búsqueda por malla como se propone la documentación de la herramienta LS-SVMlab Toolbox para MATLAB.
La etapa de optimización de los parámetros γ y σ^2 se guía por la tasa de errores de clasificación de los datos de formación con validación cruzada de 10 veces. El algoritmo SA se utiliza primero para encontrar buenos valores iniciales de los parámetros del kernel y luego la estrategia de búsqueda por mallas se aplica para el ajuste de los parámetros.
Durante el proceso de obtención de los clasificadores base se genera una gran cantidad de estos con diferente desempeño. A pesar de que no es importante que estos modelos bases sean robustos y predictivos, si es necesario que tengan un desempeño mejor que el de un modelo aleatorio. Por esta razón es necesario adicionar un filtro durante la generación de los modelos base para descartar aquellos que tengan un desempeño muy pobre. En base a esto se establecen los siguientes criterios de aceptabilidad de los clasificadores base:

a) Al menos el 65% de los casos predichos correctamente en el conjunto de entrenamiento (exactitud >65%)
b) Al menos el 65% de los casos predichos correctamente en el conjunto de prueba (exactitud >65%)
c) Al menos un 65% de exactitud en una validación cruzada de 5 veces.

Los modelos base que cumplan con estos tres criterios son considerados como modelos base válidos y son elegidos hasta obtener un número de modelos base ya previamente fijado. En este caso se estable 501 como el máximo de clasificadores base a generar, los que no cumplan con los criterios de aceptabilidad son desechados durante el proceso de búsqueda.

2.2.2 Selección de modelos para multiclasificadores

Como se mencionó anteriormente unas de las características que debe poseer un multiclasificador es que los modelos base que lo componen sean diversos. Por esta razón se implementan dos estrategias para seleccionar los modelos base que formarán parte de los multiclasificadores. La primera estrategia consiste realizar un análisis de conglomerados (*Clustering*) de las salidas de los modelos base. Este método de selección es un método de selección de tipo filtro pues primero se selecciona un subconjunto de modelos representativos y posteriormente se evalúa el desempeño del multiclasificador formado por dichos modelos.

La segunda estrategia es una de tipo “contenedor” (*wrapper*). En este tipo de estrategia se emplea alguna medida de desempeño del multiclasificador que se está buscando como función objetivo de un algoritmo computacional de optimización. En este caso se empleó una búsqueda basada en algoritmos genéticos (AG).

2.2.2.1 Selección de modelos basada en conglomerados

El análisis de agrupamiento jerárquico (HCA) es una técnica que agrupa los datos en segmentos que tienen atributos semejantes. Su propósito es presentarlos de una manera que acentúe los grupos naturales más comunes, procurando definir esas categorías en primer lugar.

La selección de los modelos base basada en conglomerados se realizó en base a dos medidas diferentes de distancia. Estas medidas de distancia son el estadígrafo Q y la correlación entre la salida de los modelos (45, 56).

En vez fijada una cantidad de modelos representativos para el análisis de conglomerados se diseñó un procedimiento de aglomeración iterativo. Este proceso consistió en, para cada

medida de distancia entre modelos, comenzar con la selección de 2 modelos representativos. Para estos dos modelos se construye un multiclasificador y se evalúa su calidad (ver más abajo el criterio para comparar el desempeño de los multiclasificadores). A continuación se repite el proceso de selección de los modelos representativos aumentando la cantidad de modelos representativos en 1 en cada ciclo hasta completar 501 aglomerados que se corresponde con todos los modelos base.

Después de cada uno de estos ciclos de selección de los modelos base representativos, estos se agregan en un multiclasificador empleando los métodos de agregación descritos a continuación y se evalúa su desempeño.

2.2.2.2 Selección de modelos basada en algoritmos genéticos

A diferencia de los métodos de filtro, los métodos del tipo contenedor emplean un criterio de desempeño de los multiclasificadores para realizar la selección de los modelos que forman el mejor multiclasificador. En esta investigación se emplean algoritmos genéticos (AG) como algoritmo de optimización del desempeño de los multiclasificadores.

El AG se corre 10 veces, cada vez con una población inicial máxima de 5,10 y 15 individuos diferentes (modelos base) y para 500 generaciones (Ngen= 500). Dado que el AG se implementa como un enfoque de selección de modelos base, se utiliza una codificación de cadena de bits de los cromosomas. Un 1 en la posición i significa que el modelo LS-SVM i es considerado por la función de ajuste, mientras que 0 significa que este modelo es excluido del cómputo del ajuste. La población inicial se crea aleatoriamente de tal manera que cada individuo contiene a lo sumo 5, 10 o 15 genes codificados como 1 y la longitud de cada individuo se establece igual al tamaño del conjunto de modelos base de entrada (501).

La función de selección se establece como un torneo de tamaño 2, mientras que el operador de cruce combina aleatoriamente cada posición de dos padres para producir una descendencia. En resumen, la selección del torneo consiste en la selección aleatoria de n individuos de la población, comparar sus valores de ajuste y seleccionar el que tiene el mejor ajuste (minimizamos nuestra función objetivo) como el ganador. El proceso de mutación selecciona al

azar una posición codificada como 1 en el individuo y la cambia a 0 y de la misma manera cambia a 1 una de las posiciones excluida del multiclasificador.

El tamaño de la población se establece en 100 individuos, las tasas de cruce y mutación se establecen en 0,7 y 0,3, respectivamente, y dos individuos de élite sobreviven y pasan a la siguiente generación de forma automática.
Se emplearon dos funciones de ajuste en los experimentos con AG. La primera función de ajuste minimiza $\sqrt{Se * Sp}$ para el conjunto de prueba Se y Sp la sensibilidad y la especificidad del multiclasificador respectivamente. La segunda función de ajuste para el AG está basada en el índice de Akaike (57) y minimiza $\log((1-\sqrt{Se*Sp})^2)+2\frac{m}{n}$ siendo m la cantidad de modelos en el multiclasificador y n el número de casos en el conjunto de entrenamiento.

2.2.3 Agregación de los modelos base

La manera en que se agregan los resultados de los clasificadores individuales es una cuestión fundamental para obtener un buen multiclasificador. Este paso consiste en optimizar la decisión que se va a tomar a partir de las decisiones individuales de los modelos base seleccionados. Existen varios métodos de agregación de modelos base(45). En este caso en particular se empleó la Votación Mayoritaria, el Voto de Borda y el algoritmo *AdaBoost*.

2.2.3.1 Voto Mayoritario

En el Voto Mayoritario se trata de llegar un consenso a partir de todos los votos emitidos. El consenso por simple mayoría exige que más de la mitad de los clasificadores concuerden en una clase(45).

La salida de cada clasificador para cada compuesto en particular es un voto y el voto Mayoritario no es más que contar para ese compuesto cuantos votos positivos tiene y cuantos negativos. Finalmente, la salida del multiclasificador estará de acuerdo con la clase que haya recibido más de la mitad de los votos.

2.2.3.2 Voto de Borda

El método del conteo de Borda es un método de decisión donde los votantes se expresan a través de un ordenamiento de sus opciones y tiene en cuenta el consenso entre los clasificadores(45). Para esto cada clasificador asigna un ordenamiento del grupo de casos a predecir. En esta investigación el ordenamiento para cada modelo se obtiene de las puntuaciones (*scores*) asignados por los modelos base de tipo LS-SVM.

Una vez ordenados se le asigna un orden (ranking) a cada caso en el conjunto de datos. Para un conjunto de datos de n casos al primer caso en el ordenamiento se le asigna una cantidad de votos n-1, al segundo n-2 y así sucesivamente hasta asignar 0 votos al caso ordenado en último lugar. Este ordenamiento se realiza para cada clase y cada caso se predice de acuerdo a la clase que mayor cantidad de votos reciba.

2.2.3.3 *AdaBoost*

El ultimo método para la agregación de los modelos base es el algoritmo *AdaBoost*(45, 58, 59). El algoritmo *AdaBoost* toma como entrada la predicción hecha por un grupo de clasificadores base. Este algoritmo se basa en un proceso de agregación iterativo en el cual en cada paso se adiciona al multiclasificador el modelo base que minimice una función de error pesado. Esta función de error pesado se calcula a partir la asignación de un peso a cada caso a predecir.
Una de las características de este algoritmo es que inicialmente se le asigna el mismo peso a cada caso a predecir y en cada ciclo se aumenta el peso de los casos que son consistentemente mal clasificados por los modelos incluidos en el multiclasificador. De esta forma durante el proceso de selección del próximo modelo a incluir en el multiclasificador se les asignará mayor peso a los modelos que sean capaces de predecir correctamente casos que son incorrectamente predichos por la mayoría de los modelos base en el multiclasificador. Una vez que se determina el peso de cada modelo en el multiclasificador se realiza la suma pesada de las salidas de estos modelos y se asigna la clase predicha en base a un valor de corte pre-calculado.

2.2.4 Evaluación del desempeño de los modelos.

El desempeño de cada modelo, ya sea de los modelos base o de los multiclasificadores se evalúa empleando las siguientes medidas:

$$\text{Exactitud} = \frac{\text{Número de muestras correctamente clasificadas}}{\text{Número total de muestras}}$$

$$\text{Sensibilidad} = \frac{\text{Número de muestras positivas correctamente clasificadas}}{\text{Número total de muestras positivas}}$$

$$\text{Especificidad} = \frac{\text{Número de muestras negativas correctamente clasificadas}}{\text{Número total de muestras negativas}}$$

Como se presentó previamente, se emplean diferentes algoritmos y metodologías para construir los multiclasificadores. Por esta razón es necesario emplear una medida de desempeño única que permita comparar el desempeño de diferentes métodos de modelación. Es este caso seleccionamos como criterio de comparación de los diferentes modelos obtenidos $\sqrt{Se * Sp}$. Esta medida permite para modelos con similar error de predicción priorizar aquellos que tengan un balance entre Sensibilidad y Especificidad. En otras palabras, esta medida priorizará la selección de modelos que tengan una exactitud similar en la predicción de las dos clases modeladas.

CAPÍTULO 3. RESULTADOS Y DISCUSIÓN

3.1 Preparación de los datos

3.1.1 Proceso de curación de los conjuntos de datos.

El conjunto de datos iniciales formado por 5697 compuestos fue sometido al proceso de curación descrito en los Materiales y Métodos. Fueron eliminados los compuestos que no podían ser asignados a alguna de las clases por presentar valores de medición ambiguos. Además se eliminaron compuestos que presentaban elementos traza o poco representados, elementos raros o poco comunes en sistemas biológicos. Siguiendo estos criterios se eliminaron 579 compuestos del conjunto de datos inicial.

Una vez de estandarizadas y protonadas las estructuras, se realizó la comprobación de duplicados estructurales, encontrándose 255 pares de compuestos duplicados. Se encontró que algunos se correspondían con isómeros cis-trans. La representación empleada (2D) es incapaz de manejar adecuadamente este tipo de isómeros y además el protocolo de determinación de la actividad antimalárica no especificaba si las mediciones fueron tomadas para ambos, por lo que todos los pares de compuestos duplicados fueron eliminados. El conjunto de datos fue sometido al proceso de inspección visual en el cual no se detectaron irregularidades, por lo que finalmente el conjunto de datos curado quedo conformado por 4608 compuestos.

3.1.2 División del conjunto de datos en subconjuntos de modelación y validación externa.

El conjunto de 4608 moléculas que forma la base de datos curada se dividió en subconjuntos de modelación y validación externa siguiendo el protocolo descrito en los Materiales y Métodos. Estos quedaron formados por 3686 y 922 compuestos respectivamente. La serie de validación externa se seleccionó al azar y no se empleó para la modelación o selección de los modelos, lo que permite poner a prueba la capacidad predictiva del mejor modelo obtenido.

3.1.3 Detección y eliminación de *activity cliffs* del conjunto de modelación.

La detección y eliminación de activity cliffs se realizó para el subconjunto de modelación. Los *activity cliffs* son pares de compuestos que presentan elevada similitud estructural pero grandes diferencias en los valores de actividad biológica o propiedad modelada. Su eliminación permite optimizar el conjunto de entrenamiento para la modelación basada en aprendizaje automatizado previamente a cualquier esfuerzo de modelación (60). La esencia de este proceso es remover del proceso de entrenamiento compuestos responsables de la discontinuidad de las relaciones estructura-actividad (SAR), y consecuentemente restaurar la continuidad de las SAR requerida para derivar modelos QSAR más confiables y predictivos.

Es importante señalar que para la detección y eliminación de activity cliffs es necesario generar la matriz de distancia para cada una de las representaciones que se emplee, motivo por el cual se decidió solamente como forma de codificación los fingerprints del tipo ECFP. Aunque emplear varios tipos de representación mejoraría la descripción de los datos pero al mismo tiempo aumentaría mucho el tiempo de trabajo computacional dado la gran cantidad de compuestos a modelar.

Como resultado de este proceso se detectaron 359 generadores de cliffs que fueron removidos del conjunto de modelación. Por lo que en este punto la serie de aprendizaje quedó formada por 3327 compuestos.

3.1.4 División del conjunto de modelación en subconjunto de entrenamiento y prueba.

Los 3327 compuestos restantes considerados como el conjunto de aprendizaje se dividieron en subconjuntos de entrenamiento y prueba utilizando los tres algoritmos de exclusión esfera 1M, 2M y 3M descritos en el Capítulo 2. La selección de la partición óptima se guió por un valor de corte de $M_{test,\ train}$ (c) < 0.5, $M_{train,\ test}$ (c) < 0.5, I_{train} (c) > 0.5 y el número de muestras de cada clase seleccionados para el subconjunto de prueba. La condición de que el subconjunto de prueba contenga al menos 15% de los compuestos de cada clase se impone para garantizar

que este contenga suficientes muestras para una evaluación fiable de las capacidades de generalización potencial de los modelos durante el proceso de selección del mejor modelo. Las series de entrenamiento y prueba quedaron compuestas por 2814 y 513 compuestos respectivamente.

3.1.5 Codificación estructural

Para el cálculo de los descriptores moleculares fue necesario ajustar las configuraciones de los comandos empleados. La longitud de los fragmentos calculados se estableció entre 2 y 8. Para distancias mayores de 8 enlaces las interacciones comienzan a disminuir considerablemente además de que por encima de este valor el número de descriptores generados seria computacionalmente prohibitivo durante el proceso de modelación.

Los descriptores obtenidos fueron filtrados según lo descrito en el Capítulo 2 y finalmente se obtuvo la plantilla de 250 descriptores con la cual se calcularon los descriptores para el resto de los subconjuntos.

3.2 Modelación QSAR

La base de datos que se modeló posee una gran cantidad de moléculas con elevada diversidad estructural y esto dificulta encontrar un patrón único. Por esta razón se decide implementar un método basado en multiclasificadores.

Como se mencionó en el Capítulo 1, cualquier estrategia de modelación basada en algoritmos multiclasificadores tiene dos componentes básicos: 1) la selección de los modelos base (previamente entrenados) y 2) la agregación de las salidas de los modelos seleccionados.

3.2.1Entrenamiento de los modelos base

Los modelos base se entrenaron siguiendo el procedimiento descrito en los Materiales y Métodos. Se entrenaron 501 modelos de clasificación de tipo LS-SVM que cumplieran con los criterios de aceptabilidad previamente establecidos. Las estadísticas de los modelos base para el subconjunto de prueba se muestran en el Gráfico 3.1.

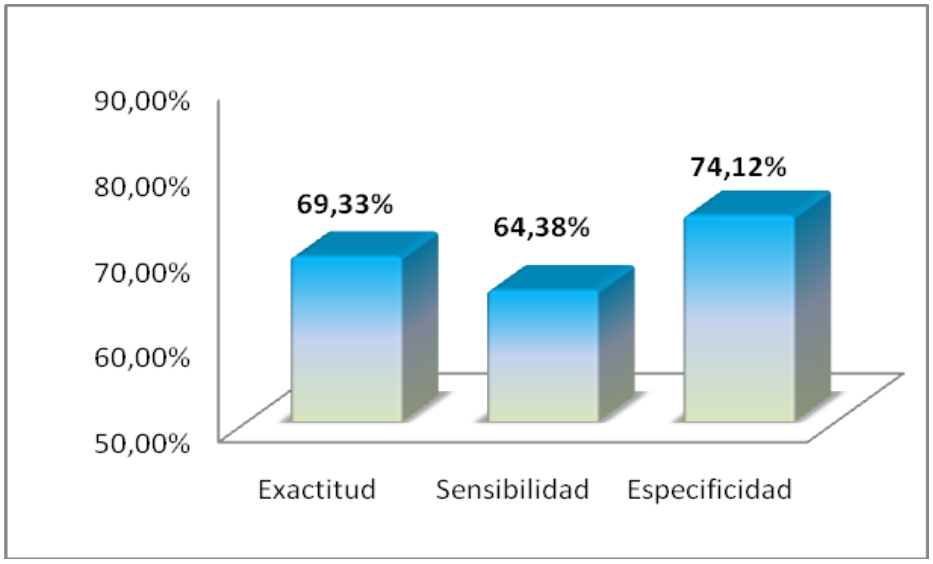


Grafico 3.1. Exactitud, Sensibilidad y Especificidad media de los modelos base para el conjunto de prueba.

Adicionalmente, se analizó el desempeño del mejor modelo base obtenido. Como se presentó en los Materiales y Métodos, como métrica de desempeño para seleccionar el mejor modelo en todos los experimentos presentados es esta tesis se empleó $\sqrt{Se * Sp}$ para el conjunto de prueba. El desempeño de este mejor modelo base se presenta en el Gráfico 3.2.

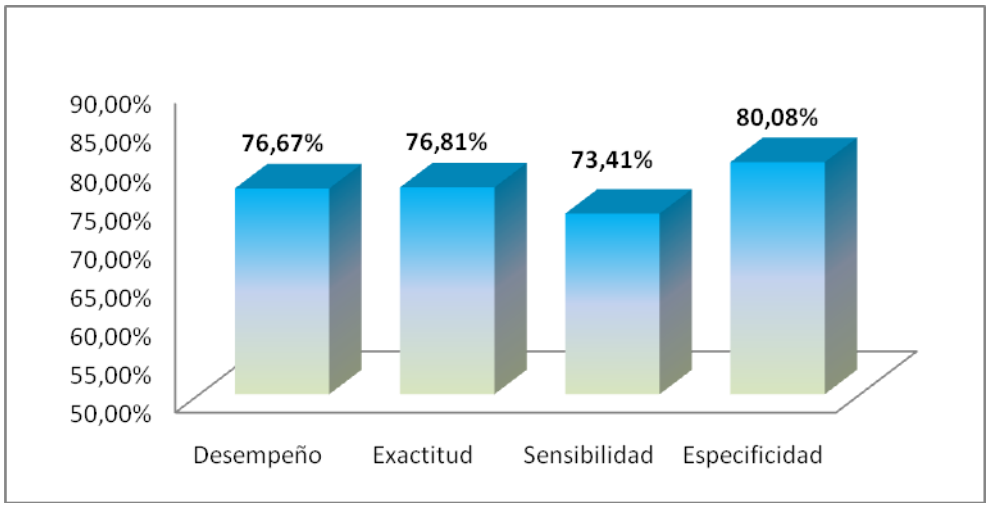


Gráfico 3.2. Desempeño, Exactitud, Sensibilidad y Especificidad del mejor modelo base para el subconjunto de prueba.

3.2.2 Formación de un multiclasificador a partir de los 501 modelos base.

Inicialmente, a modo de control, se decidió antes de emplear cualquiera de los métodos de selección de modelos base descritos en el Capítulo 2, formar un multiclasificador compuesto por los 501 modelos base y combinar sus salidas empleando los diferentes métodos de agregación. Es importante señalar que además de los tres métodos propuestos en el Capítulo 2: Voto mayoritario, Voto de Borda y AdaBoost, en este experimento las salidas también fueron combinadas a través del Voto Pesado por Exactitud y pesado por Sensibilidad y Especificidad. Esta estrategia se sigue con el objetivo dar un peso mayor en la votación a aquellos modelos de los cuales se espera un mejor desempeño. En el Gráfico 3.3 se muestra el desempeño del mejor modelo base y del multiclasificador obtenido a partir de la agregación de los 501 modelos base para cada uno de los métodos de agregación.

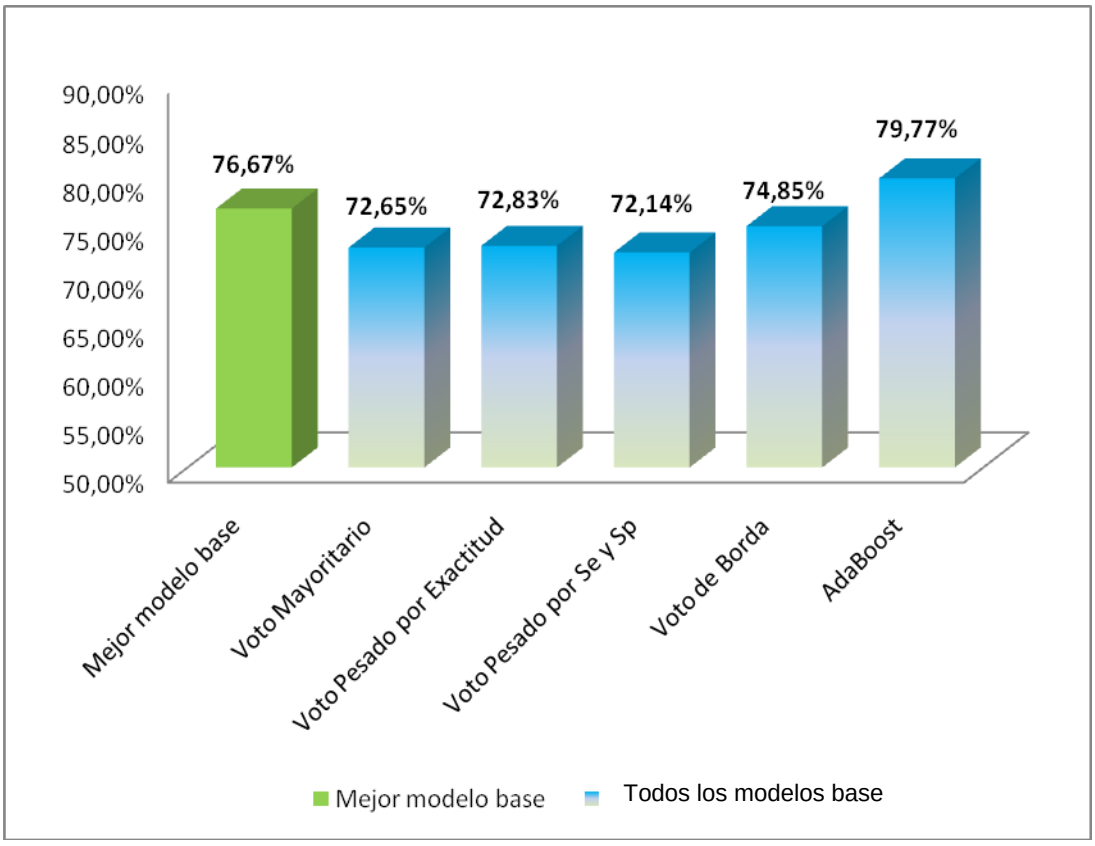


Gráfico 3.3. Desempeño del mejor modelo base y del multiclasificador de 501 modelos base agregados por Voto Mayoritario, Voto Pesado por Exactitud, Voto Pesado por Sensibilidad y Especificidad, Voto de Borda y AdaBoost respectivamente.

Como puede observarse en el Gráfico 3.3 al agregar la salida de todos los modelos base a través de las estrategias propuestas y evaluar su desempeño este es menor que el del mejor modelo base en todos los casos exceptuando cuando el método de agregación es AdaBoost. Este resultado tiene sentido ya que aunque los modelos base fueron generados procurando su diversidad, aun así, la salida de muchos de estos modelos es redundante y no podemos pasar por alto que la diversidad es uno de los aspectos fundamentales de un buen multiclasificador.

El Gráfico 3.3 también muestra que no existen diferencias entre los desempeños de los 501 modelos cuando se combinan las salidas por Voto Mayoritario, Voto Pesado por Exactitud y Voto Pesado por Sensibilidad y Especificidad razón por la cual estas dos últimas formas de combinación de salidas no se emplean de aquí en lo adelante.

3.2.3 Formación de multiclasificadores

Luego de la obtención de los 501 modelos base y la evaluación de su desempeño como el de un único clasificador. El siguiente paso en esta investigación fue la formación de los multiclasificadores combinando los dos métodos de selección de modelos con cada una de sus variantes con los tres métodos de agregación de las salidas. Como resultado de este proceso los multiclasificadores formados responden a un total de 12 metodologías, estas se muestran en la Tabla 3.1. Es necesario que para la minimización de la función objetivo de los métodos basados en algoritmos genéticos se emplearon las predicciones realizadas en la serie de prueba.

Tabla 3.1. Metodologías empleadas en la obtención de los multiclasificadores

#	Definición	Representación
1	Agrupamiento empleando como medida la matriz de correlación y agregación con Voto Mayoritario	**A-VM-MatrizCor**
2	Agrupamiento empleando como medida la matriz de correlación y agregación con Voto de Borda	**A-VB-MatrizCor**
3	Agrupamiento empleando como medida la matriz de correlación y agregación con AdaBoost	**A-AB-MatrizCor**
4	Agrupamiento empleando como medida la matriz Q y agregación con Voto Mayoritario	**A-VM-MatrizQ**
5	Agrupamiento empleando como medida la matriz Q y agregación con Voto de Borda	**A-VB-MatrizQ**
6	Agrupamiento empleando como medida la matriz Q y agregación con AdaBoost	**A-AB-MatrizQ**
7	Algoritmo Genético minimizando $1-\sqrt{Se*Sp}$ y agregación con Voto Mayoritario	**AG-VM-MinError**
8	Algoritmo Genético minimizando $1-\sqrt{Se*Sp}$ y agregación con Voto de Borda	**AG-VB-MinError**
9	Algoritmo Genético minimizando $1-\sqrt{Se*Sp}$ y agregación con AdaBoost	**AG-AB-MinError**

10	Algoritmo Genético minimizando AIC y agregación con Voto Mayoritario	**AG-VM-MinAIC**
11	Algoritmo Genético minimizando AIC y agregación con Voto de Borda	**AG-VB-MinAIC**
12	Algoritmo Genético minimizando AIC y agregación con Adaboost	**AG-AB-MinAIC**

Una vez generados los multiclasificadores empleando las 12 metodologías propuestas fueron elegidos de estos los de mejor desempeño para cada una de las metodologías. Los análisis que se muestran de aquí en adelante se basan en estos 12 mejores multiclasificadores representando cada una de las metodologías implementadas. El desempeño de los multiclasificadores seleccionados como los mejores para cada metodología se muestra en la Tabla 3.1 del Anexo 3.

Para determinar la mejor estrategia de selección de modelos base se promediaron los desempeños de los multiclasificadores de la Tabla 3.1 del Anexo 3 por cada método de selección. Los resultados se muestran en el Gráfico 3.4.

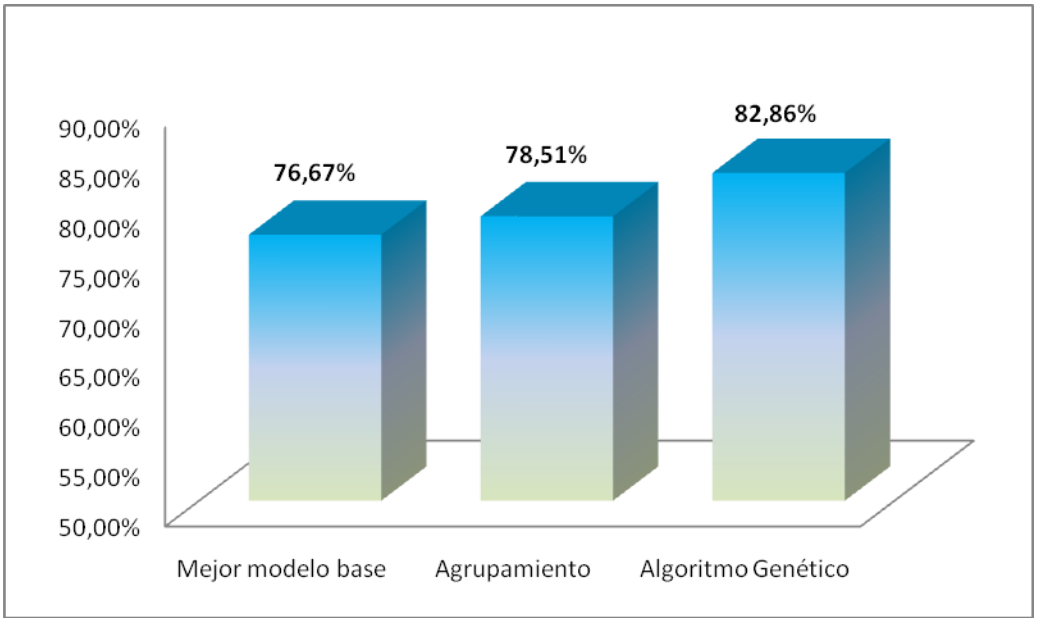
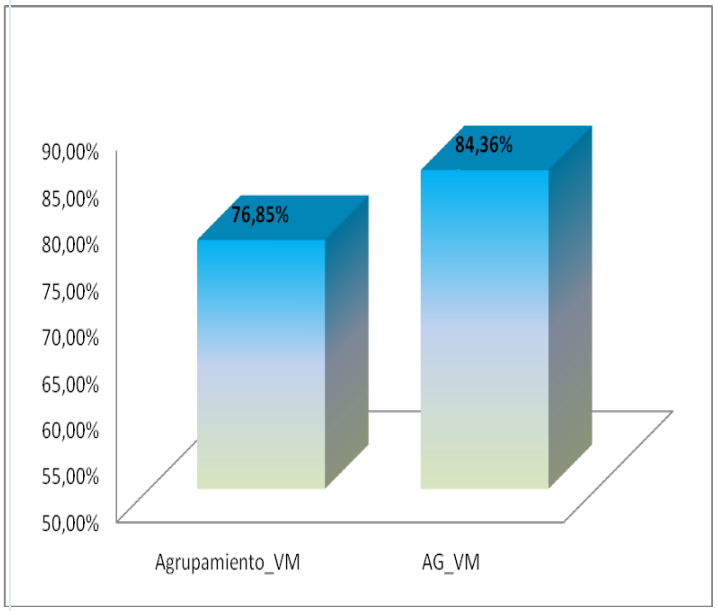
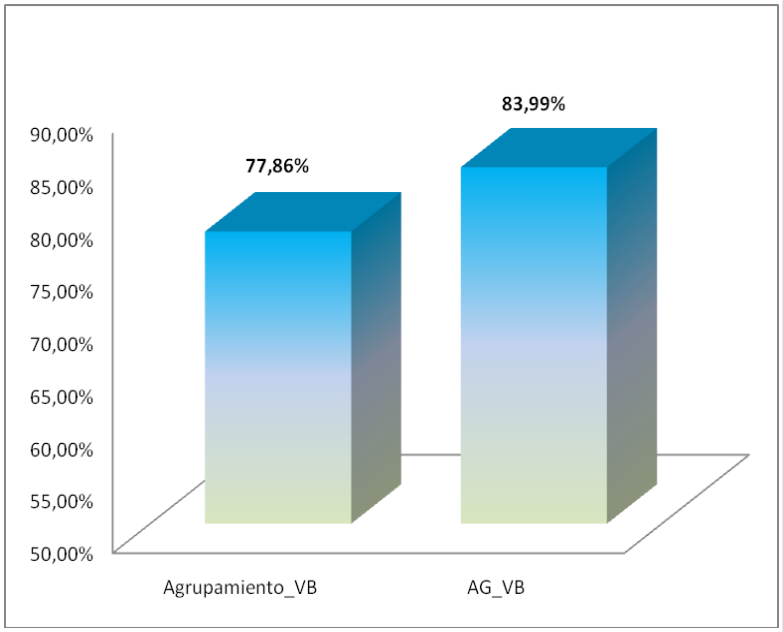


Gráfico 3.4. Desempeño del mejor modelo base y promedio de desempeño para los multiclasificadores formados por técnicas de Agrupamiento y Algoritmo Genético respectivamente.

En el Gráfico 3.4 se evidencia la superioridad del empleo de Algoritmos Genéticos como método de selección respecto al Agrupamiento. Puede observarse que cuando se utiliza una técnica de selección de modelos base para la formación de multiclasificadores, estos superan el desempeño del mejor modelo base.

A demás, si comparamos el promedio del desempeño de los multiclasificadores formados por técnicas de Agrupamiento y Algoritmo Genético para cada uno de los métodos de agregación como se muestra en el Gráfico 3.5, tenemos que cuando los métodos de agregación empleados son Voto Mayoritario y Voto de Borda, el Algoritmo Genético demuestra un mejor rendimiento. En el caso de la agregación empleando AdaBoost se observa que no existen diferencias cuando se emplea uno u otro método de selección. Aun así podemos afirmar que desempeño de los multiclasificadores está estrechamente relacionado con el método de selección que se emplee y que empleando Algoritmos Genéticos se obtiene un mejor desempeño de los multiclasificadores.





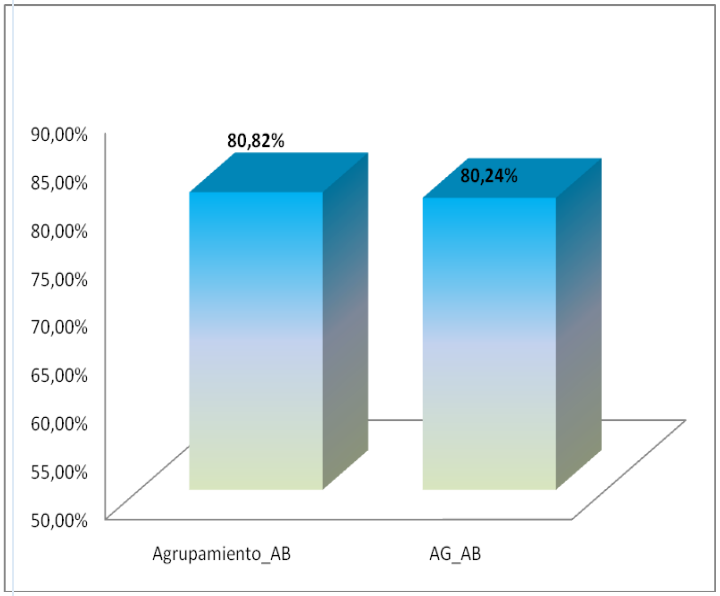


Grafico 3.5 Comparación entre el promedio de los desempeños de los multiclasificadores formados por técnicas de Agrupamiento y AG para la agregación por Voto Mayoritario(VM), Voto de Borda(VB) y AdaBoost(AB) respectivamente.

Finalmente, si comparamos los dos mejores multiclasificadores de cada uno de los métodos de selección empleados en relación al desempeño y número de modelos base que lo conforman,

obtenemos mejor desempeño cuando se emplea AG (84%), que agrupamiento (80%). En el gráfico 3.6 se muestra que el número de modelos base que conforman el multiclasificador es mucho menor para el AG, lo que disminuye notablemente la complejidad de los multiclasificadores. Este resultado está en concordancia con el hecho de que durante la búsqueda de multiclasificadores empleando AGs la selección de los miembros del multiclasificador es guiada por la optimización del desempeño del multiclasificador.

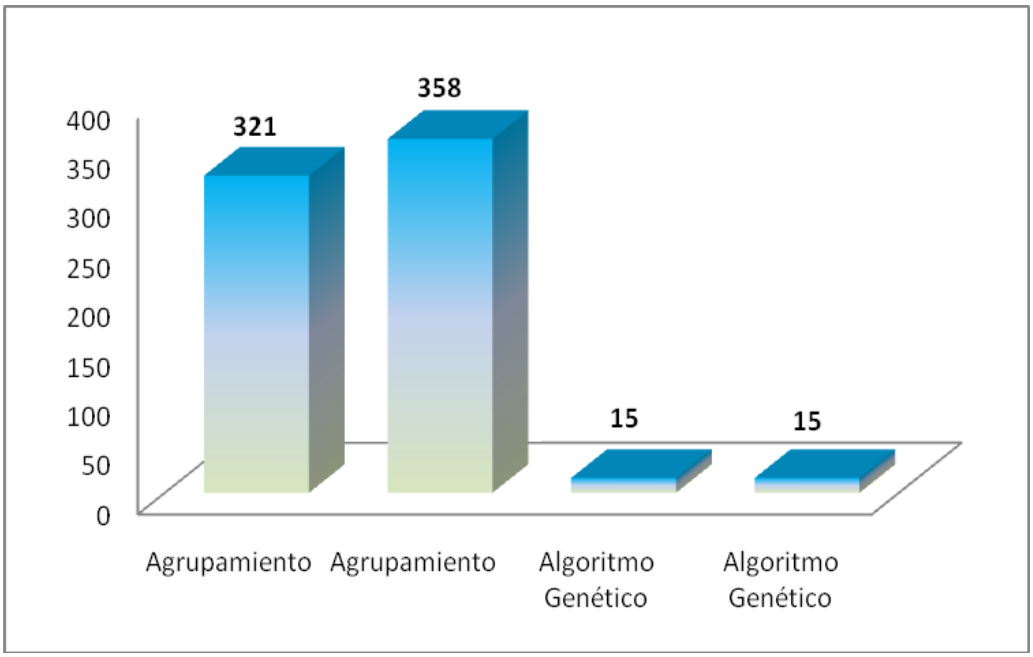


Gráfico 3.7. Número de modelos en los dos mejores multiclasificadores para cada método de selección

Además se investigó si alguno de los métodos de agregación reporta un desempeño superior para los multiclasificadores independientemente de si el método de selección es AG o Agrupamiento. Para esto se estudió el promedio del desempeño obtenido con Voto Mayoritario, Voto de Borda y Adaboost respectivamente. Estos resultados se muestran en el **Gráfico 3.8**.

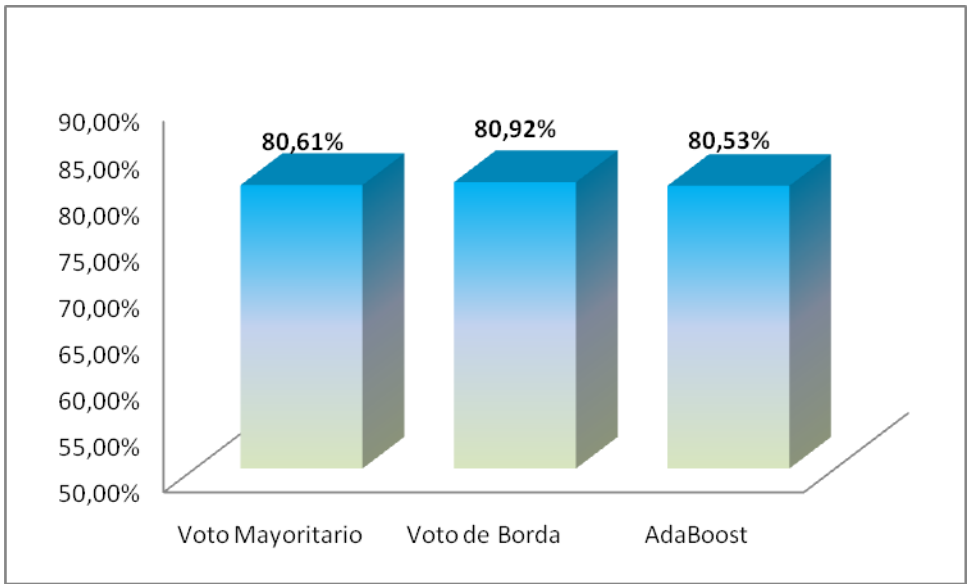


Gráfico 3.8. Desempeño del mejor modelo base, Promedio de desempeño para los multiclasificadores formados por Voto Mayoritario, Voto de Borda y Adaboost respectivamente

Según lo observado en este gráfico no parecen existir diferencias si se emplea una forma de agregación u otra. Sin embargo si separamos los resultados para cada uno de los métodos de selección de modelos base, como se observa en los Gráficos 3.9 y 3.10 respectivamente, cuando la técnica de selección empleada es AG el mejor desempeño se obtiene cuando se agregan las salidas por Voto Mayoritario, por otra parte, para los casos en que se utiliza Agrupamiento el mejor desempeño se logra para la combinación de las salidas por AdaBoost.

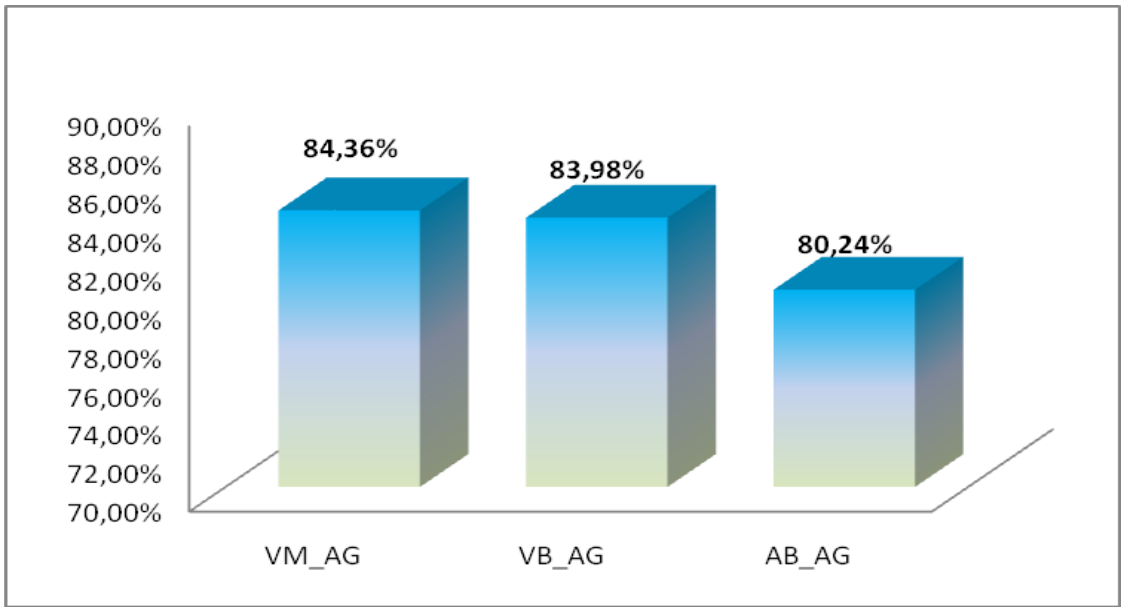


Gráfico 3.9. Promedio del desempeño de los multiclasificadores cuando se emplea Voto Mayoritario, Voto de Borda y AdaBoost respectivamente para la técnica de selección de AG

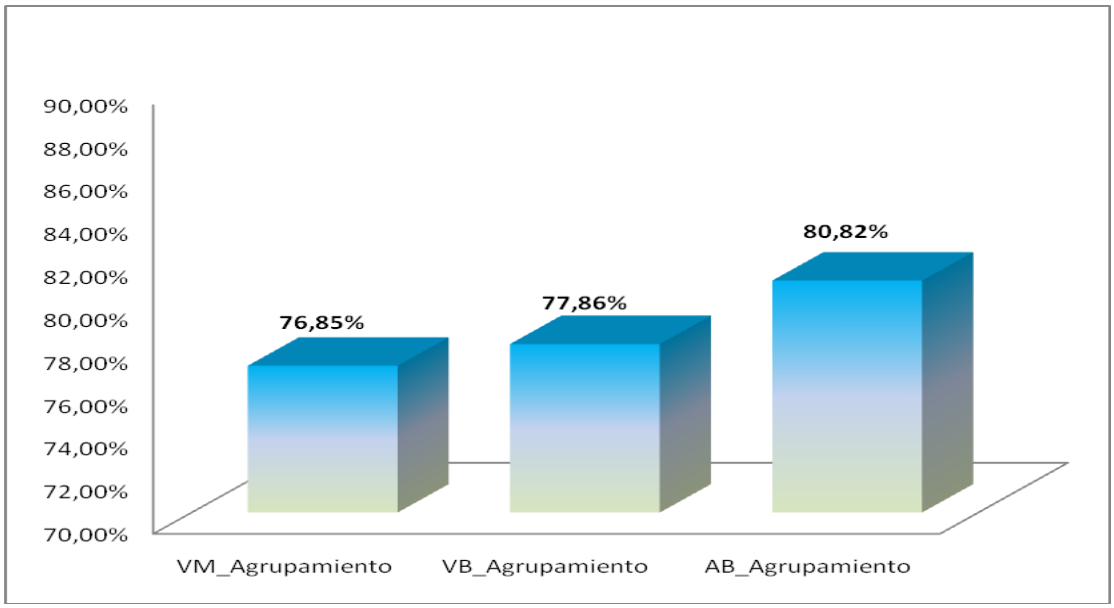


Gráfico 3.10. Promedio del desempeño de los multiclasificadores cuando se emplea Voto Mayoritario, Voto de Borda y AdaBoost respectivamente para la técnica de selección de Agrupamiento

Aun así, si se compara el rendimiento de los multiclasificadores que se obtienen para cada una de estas metodologías, cuando se emplea AG para la selección de modelos y Voto Mayoritario para la agregación de salidas el desempeño que se obtiene es más prometedor. Este comportamiento se evidencia en el Gráfico 3.11 donde también se muestra el desempeño de los mejores multiclasificadores que se obtuvieron por cada una de las metodologías propuestas, este gráfico nos permite realizar una comparación final mucho más general y seleccionar el multiclasificador de mejor desempeño.

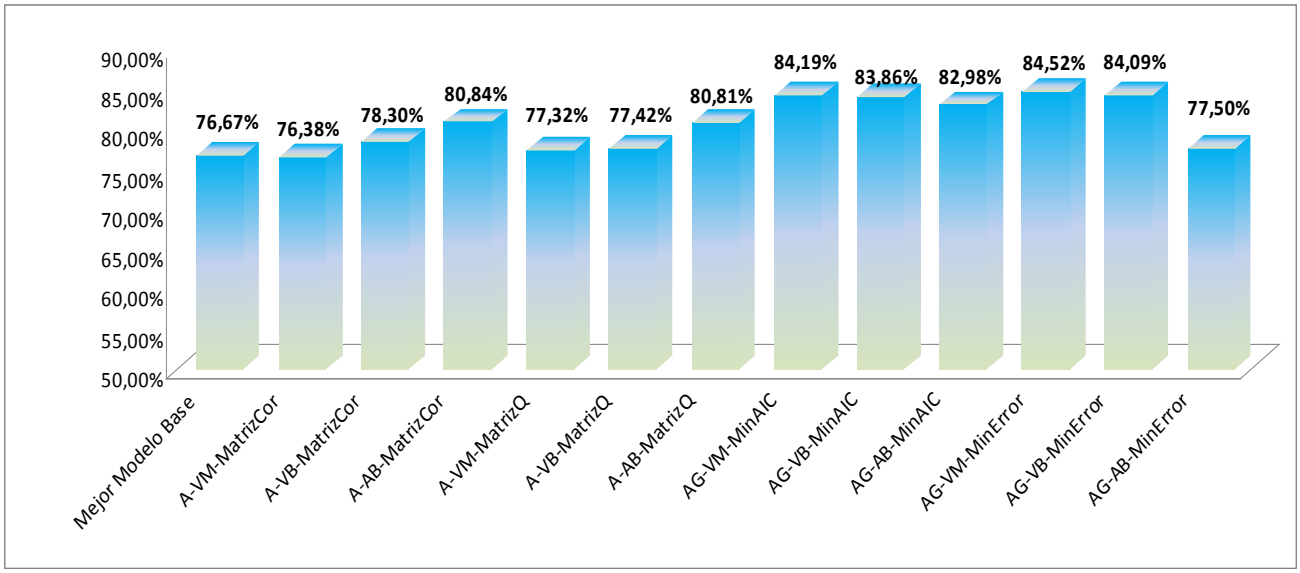


Gráfico 3.11. Desempeño de los mejores multiclasificadores para cada una de las metodologías propuestas

En base a estos resultados se determinó que la mejor metodología para la modelación de la actividad antimalárica a partir del conjunto de datos utilizado se corresponde con la que emplea como método de selección AG minimizando el error del multiclasificador que se forma con la combinación con Voto Mayoritario de los modelo base, en la Tabla 3.2 se muestra el desempeño y las estadísticas detalladas del mejor multiclasificador sobre el conjunto de validación externa para la metodología escogida como la mejor.

Tabla 3.2. Desempeño y estadísticas para la serie de validación externa del multiclasificador que emplean la mejor metodología

Parámetro	Validación Externa (AG-VM-MinError) (%)
Desempeño	84,52352043
Exactitud	70,60737527
Sensitividad	75,91836735
Especificidad	64,58333333

Es importante señalar que para cualquiera de las funciones objetivo que se emplearon para la obtención del AG la complejidad de los multiclasificadores fue muy similar, por lo que el uso indistinto de uno u otro clasificadores puede ser óptimo en la solución de nuestro problema, aunque consideramos que el más exacto y que ofrece la mejor solución es el que se muestra en la Tabla 3.2

CONCLUSIONES

1) Se recopiló un conjunto de datos obtenidos de un único laboratorio y bajo el mismo protocolo, apropiados para la modelación de la actividad antimalárica de compuestos químicos.

2) Se curó rigurosamente el conjunto de datos de partida, quedando estos libres de errores y ambigüedades que afecten el proceso de modelación QSAR.

3) Se estudió el desempeño de métodos basados en multiclasificadores para la modelación QSAR de hits antimaláricos, encontrándose que los multiclasificadores de mejor rendimiento son los que emplean como método de selección Algoritmo Genético y como forma de agregación de salidas el Voto Mayoritario

4) Se determinó que la mejor metodología para la clasificación de hits antimaláricos se obtiene cuando se combina Algoritmo Genético minimizando el error y Voto Mayoritario

RECOMENDACIONES

- Considerar el dominio de aplicación durante el proceso de entrenamiento de los multiclasificadores

- Incluir métodos de manipulación del conjunto de datos de entrenamiento

- Explorar otros métodos de agregación

REFERENCIAS

1. Organización Mundial de la Salud Informe Mundial sobre el Paludismo 2015. 2015 94.
2. Flores González JR, Heredia E, Losana Ferrer P, Martínez López L, Manuela Pérez B, Pérez Moreno M, et al. Género Plasmodium. Microbiología Clínica. Madrid2014. p. 7-16.
3. World Health Organization Signos clínicos y características epidemiológicas de la enfermedad. Guidelines for the treatment of malaria. 2 ed2011. p. 5-6.
4. Avendaño C. Una breve revisión actualizada de la batalla contra la malaria. Anales de la Real Academia Nacional de Farmacia. 2015;81:145-57.
5. Historia de la malaria. [cited 2016 www.rph.wa.gov.au/labs/haem/malaria/spain/history.htm].
6. Champin G. Pesticide use and malaria resurgence in Central America and India. Sociedades Médicas ed1983. p. 273-90.
7. Pereira Á, Pérez M. Epidemiología y tratamiento del paludismo. . Parasitología [Internet]. http://www.doymafarma.com [cited 2016; 21.
8. Jiménez J, Muskus C, Vélez I. Diversidad genética de Plasmodium falciparum y sus implicaciones en la epidemiología de la Malaria. Biomédica. 2005;25:588-602.
9. Guevara R. Gaceta Médica. Biología de los parásitos del género Plasmodium Cracas1997. p. 24-6.
10. Cowman A, Crabbs B. Invasion of red blood cells by malaria parasites. 2006. p. 755-66.
11. Castro R, Rodríguez G. Proteomic analysis of Plasmodium, the causal agent of Malaria. Salud Pública de México 2009:395-402.
12. Méndez-Cuadro D. Proteómica redox de membrana de eritrocito humano en Malaria y polimorfismos de grupos sanguíneos y G6PD. [Doctorado]. Madrid: Universidad Complutense de Madrid; 2011.
13. Chen Q, Schlichtherle M, Wahlgren M. Molecular aspects of severe malaria. Clinical Microbiology 2000:439-50.
14. Kwiatkowski D. Tumour necrosis factor, fever and fatality in falciparum malaria. Immunology 1990. p. 213-6.
15. López H, Fumadó P, González T. Actualización en el diagnóstico y tratamiento de la malaria. . Anales de Pediatría 2012.

16. Ramasamy M. Interactions of human malaria parasites, Plasmodium vivax and P. falciparum, with the midgut of Anopheles mosquitoes. 1997. p. 290-6.
17. Beier J. Ookinete rates in Afrotropical anopheline mosquitoes as a measure of human malaria infectiousness. . Journal of Tropical Medicine and Hygiene 1992:41'6.
18. Florez J. Fármacos antiparasitarios.Protozoos. Farmacología humana 3ra edición. 6 ed. p. 1230-1.
19. Avendaño C. Una breve revisión actualizada de la batalla contra la malaria. ANALES DE LA REAL ACADEMIA NACIONAL DE FARMACIA, . 2015;81:145-57.
20. Antimalarials. Medicinal Chemistry 2007. p. 613-48.
21. Adams C, Brantner V. Estimating The Cost Of New Drug Development: Is It Really $802 Million? . Health Affairs 2006;25:420-8
22. Katsuno K, Burrows JN, Duncan K, Hooft van Huijsduijnen R, Kaneko T, Kita K, et al. Hit and lead criteria in drug discovery for infectious diseases of the developing World. Nature Reviews| Drug Discovery. 2015;14:753.
23. Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. J Chem Inf Model. 2010;50:1189-204.
24. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics. 2010;29:476 – 88.
25. Golbraikh A, Wang X, Zhu H, Tropsha A. Predictive QSAR Modeling: Methods and Applications in Drug Discovery andChemical Risk Assessment. In: Leszczynski J, editor. Handbook of Computational Chemistry: Springer Netherlands; 2012. p. 1309-42.
26. Gramatica P. Chemometric Methods and Theoretical Molecular Descriptors in Predictive QSAR Modeling of the Environmental Behavior of Organic Pollutants. In: Puzyn T, Leszczynski J, Cronin MT, editors. Recent Advances in QSAR Studies. Challenges and Advances in Computational Chemistry and Physics. 8: Springer Netherlands; 2010. p. 327-66.
27. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Molecular diversity. 2000;5(4):231-43.
28. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des. 2003;17(2-4):241-53.
29. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods. 2013;5(16):3790-8.

30. Bajot F. The Use of Qsar and Computational Methods in Drug Design. Recent Advances in QSAR Studies: Springer Netherlands; 2010. p. 261-82.
31. Cramer R. The inevitable QSAR renaissance. Journal of Computer-Aided Molecular Design2012. p. 35-8.
32. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics. 2010;29:476 – 88.
33. Maggiora GM. On Outliers and Activity Cliffss Why QSAR Often Disappoints. J Chem Inf Model. 2006;46.
34. Medina-Franco JL, Yongye AB, López-Vallejo F. Consensus Models of Activity Landscapes. Statistical Modelling of Molecular Descriptors in QSAR/QSPR2012. p. 307-26.
35. Todechini R, Consonni V. Hand Book of Molecular Descriptors. 2000.
36. Dearden JC. Phisico-Chemical Descriptor. In Practical Applications of QSAR in Enviromental Chemestry and Toxicology. 1990. p. 25-9.
37. Newman MS. .J. Am. Chem. Soc. 1950. p. 4783-6.
38. Taft R. Organic Chemestry. In: Newman MS, editor. Seoaration Polar, Steric and Resonance Effects in Reactivity 1956. p. 556-675.
39. Tosato MS, Piazza R, Chiorboli C, Passerini L, Pino A, Cruciani G, et al. Chemom. Intell. Lab. Syst. . 1992. p. 155-67.
40. Rios-santamarina I, Garcia R, Galves J, Cortijo J, Santamarina P, Marcillo E. Biorg.Med. Chem. 1998. p. 477-82.
41. Todeshini R, Consonni V, Maiocchi A. Lab. Syst. 1998. p. 13-29.
42. Randic M. J. Math. Chem. 1996. p. 375-92.
43. Syswerda G. Handbook of Genetic Algorithms. In: Davis L, editor. Schedule optimization using genetic algorithms. New York1991. p. 332-49.
44. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least Squares Support Vector Machines. World Scientific. 2002.
45. Polikar R. Ensemble Based Systems in Decision Making. Feature. 2006;34:21-42.
46. Standardizer. 6.1.0.110 ed: ChemAxon; 2012.
47. Marcou G, Solov'ev V, Horvath D, Varnek A. ISIDA Fragmentor2013. 2013.
48. Velázquez Libera JL. Evaluación de la influencia de los acantilados de actividad (activity cliffs) en la modelación QSAR. [Maestría]. Cuba: Universidad Central "Marta Abreu" de Las Villas 2015.
49. Hu Y, Stumpfe D, Bajorath J. Advancing the activity cliff concept F1000Research. 2013.

50. Sud M, editor MayaChemTools: An open source package for computational discovery. 243rd ACS National Meeting & Exposition, March 25-29 2012, San Diego, CA; 2012.
51. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Molecular diversity. 2000;5:231-43.
52. MATLAB. The MathWorks Inc. 8.1.0.604 (R2013a) ed. Massachusetts2009.
53. ChemAxon. JChem. 6.3.1 ed. Budapest, Hungary2013.
54. Cruz-Monteagudo M, Morales Helguera A, Perez-Castillo Y, Tejera E, Paz C, Sánchez-Rodrígueze A, et al. Ligand-Based Virtual Screening Using Tailored Ensembles: A Prioritization Tool for Dual A2A Adenosine Receptor Antagonists / Monoamine Oxidase B Inhibitors Current Pharmaceutical Design. 2016;22:5.
55. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least Squares Support Vector Machines. World Scientific: Singapore. 2002.
56. Kuncheva LI. Combining Pattern Classifiers Methods and Algorithms: John Wiley & Sons; 2004.
57. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory 1973. p. 267-81.
58. Freund Y, Schapire RE. “Decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 1977;55:119-39.
59. Pérez-Castillo Y, CruzMonteagudo M, Lazar C, Taminau J, Froeyen M, Cabrera-Pérez MÁ. Toward the computer-aided discovery of FabH inhibitors. Do predictive QSAR models ensure high quality virtual screening performance. Molecular Diversity 2014;18:638-54.
60. Cruz-Monteagudo M, Medina-Franco JL, Pérez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discovery Today. 2014;19(8):1069-80.

ANEXOS

ANEXO 1. Malaria

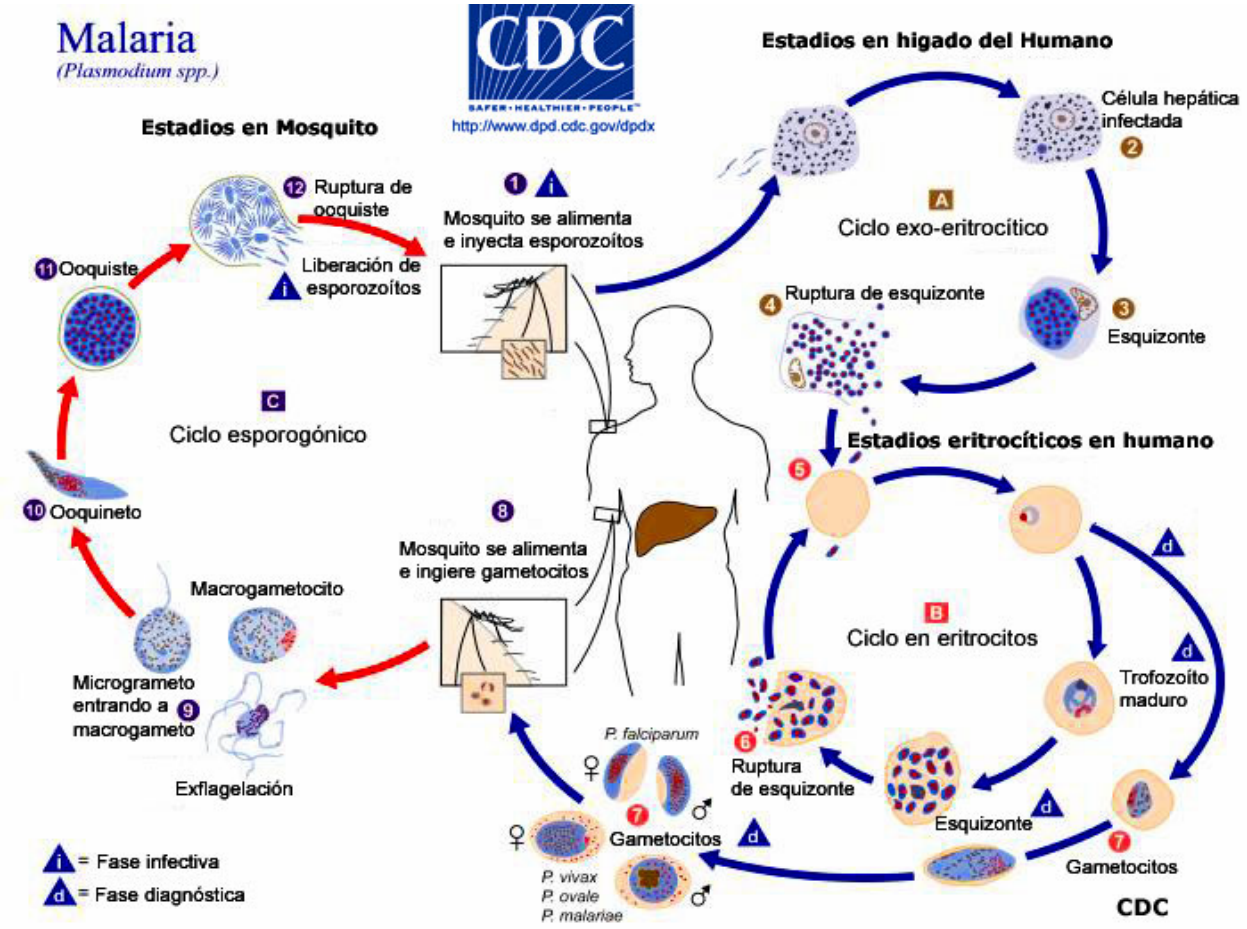


Figura 1.1 Ciclo de vida de *Plasmodium* spp. Tomado de:

ftp://www.facmed.unam.mx/deptos/microbiologia/parasitologia/paludismo.htmL

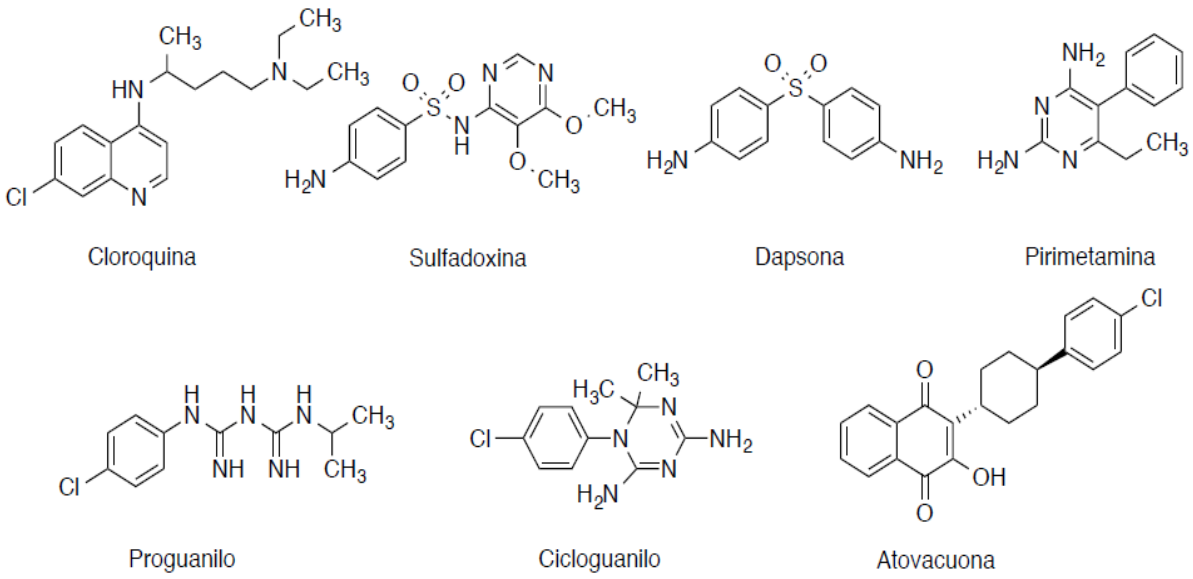


Figura 1.2 Fármacos clásicos para el tratamiento de la malaria

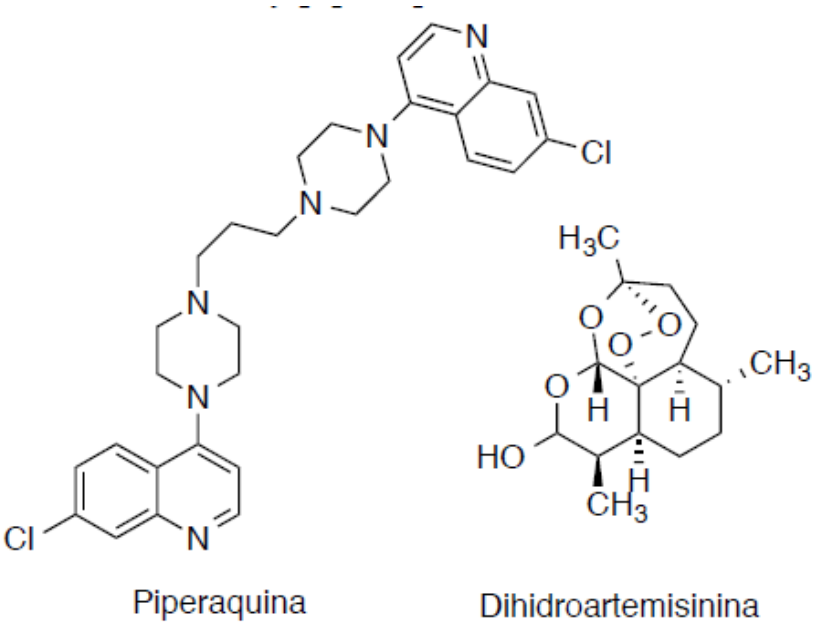


Figura 1.3 Combinación de piperaquina y dihidroartemisina para el tratamiento de la malaria por *P. falciparum.*

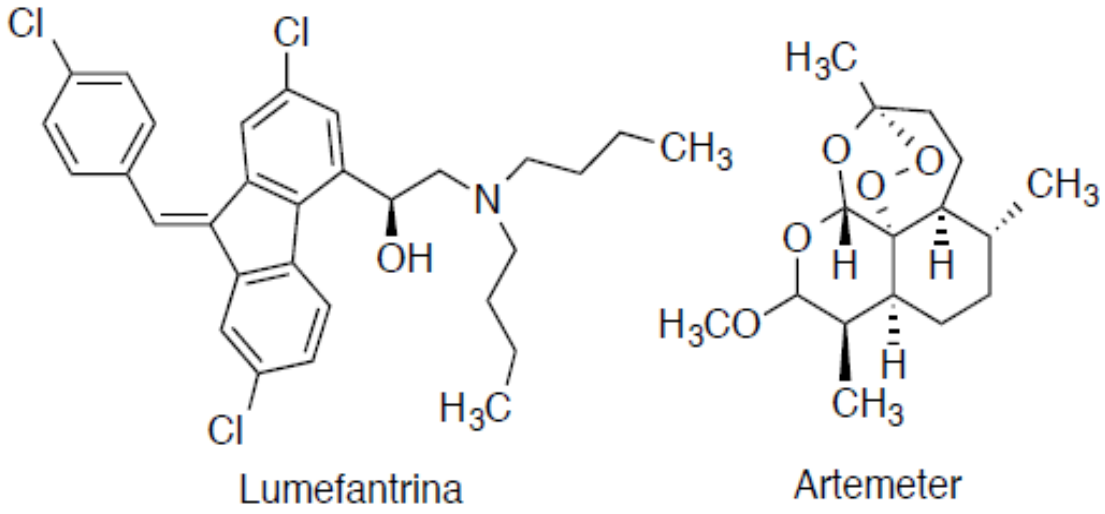
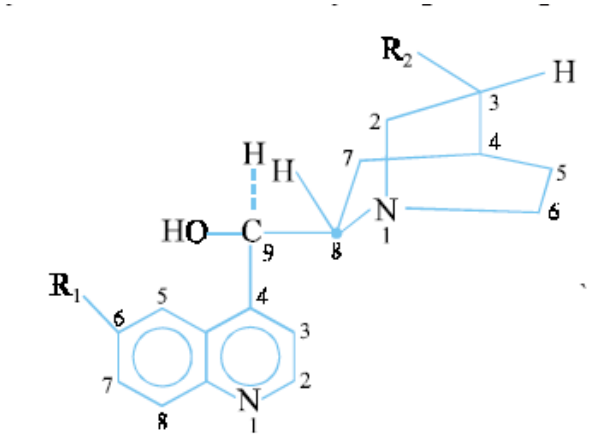


Figura 1.4 Combinación de Lumefantrina y Artemeter para el tratamiento de para el tratamiento de la malaria por *P. falciparum.*

ANEXO 2. Estructuras de algunos de los compuestos que conforman los principales grupos de fármacos antimaláricos



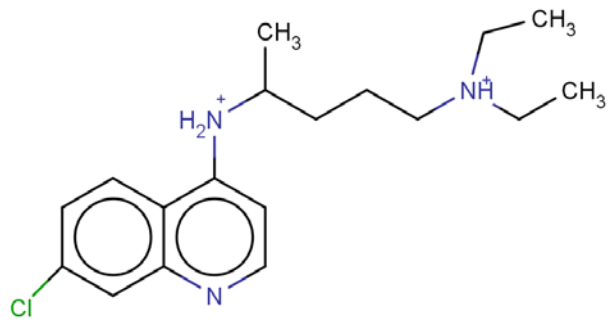
Quinine : $R_1 = OCH_3$; $R_2 = -CH = CH_2$; (–) 8S : 9R isomer

Quinidine : $R_1 = OCH_3$: $R_2 = -CH = CH_2$; (+) 8R : 9S isomer

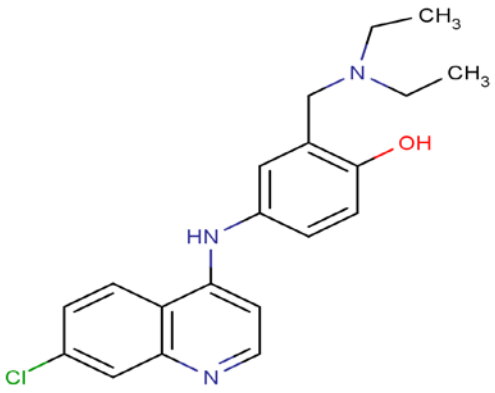
Cinchonine : $R_1 = H$; $R_2 = -CH = CH_2$; (+) 8R : 9S isomer

Cinchonidine : $R_1 = H$; $R_2 = -CH = CH_2$; (–) 8S : 9R isomer

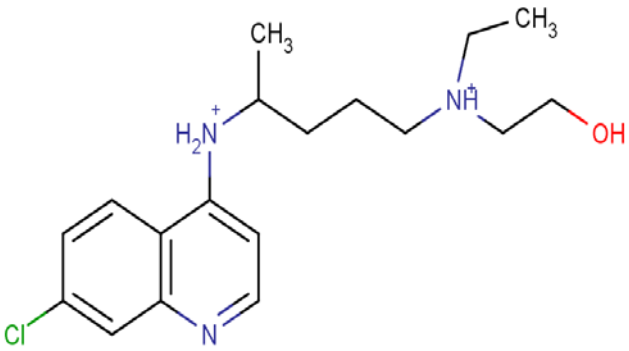
Figura 2.1 Derivados de la quinina



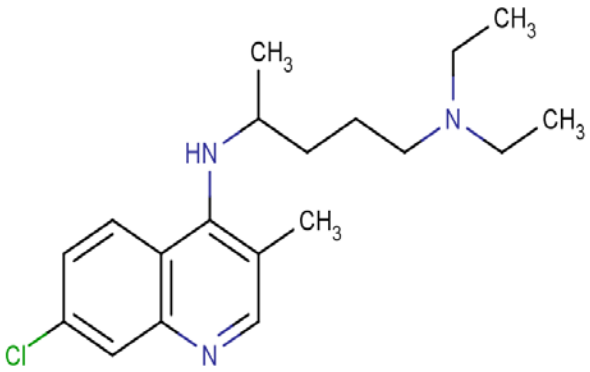
Cloroquina



Amodiaquina

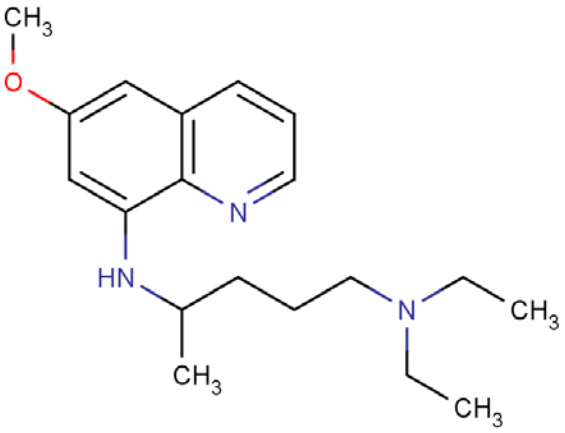


Hidroxicloroquina

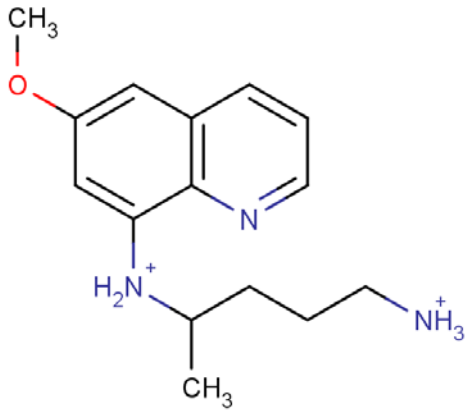


Santoquina

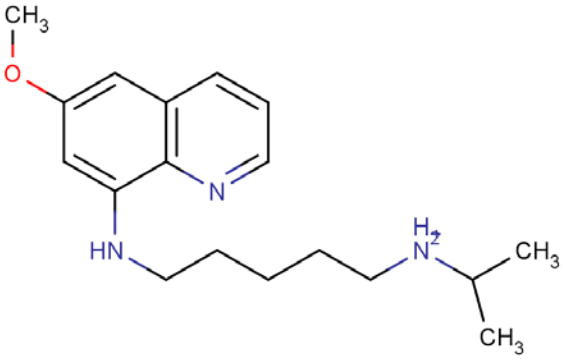
Figura 2.2 Análogos de 4-aminquinolina



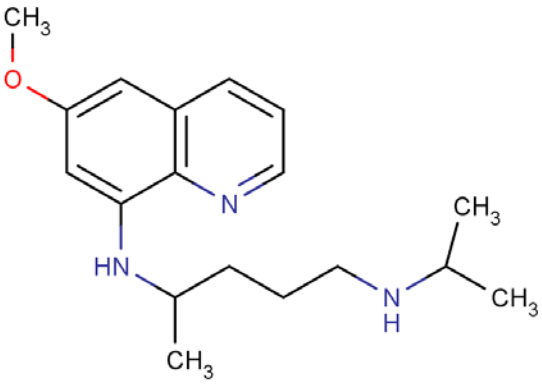
Pamaquina



Primaquina

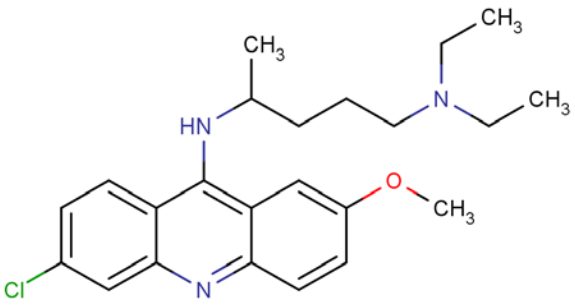


Pentaquina

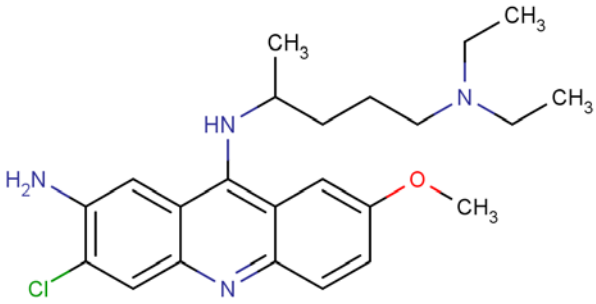


Isopentaquina

Figura 2.3 Análogos de 8-aminoquinolina

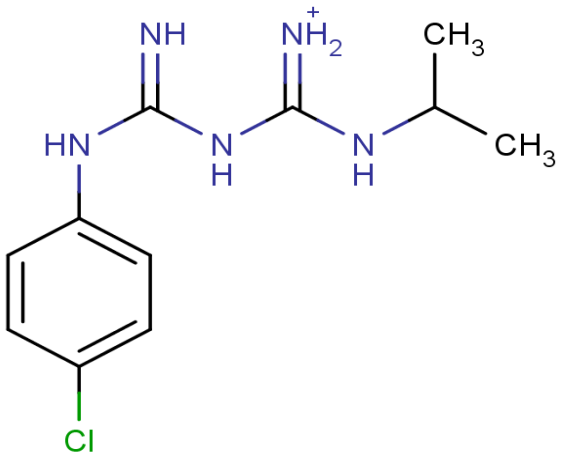


Mecaprina

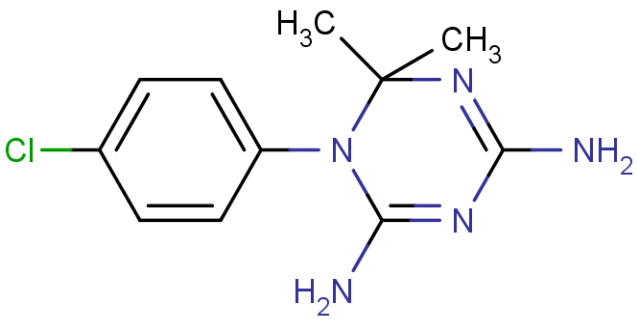


Aminoacridina

Figura 2.4 Análogos de 9-aminoacridina

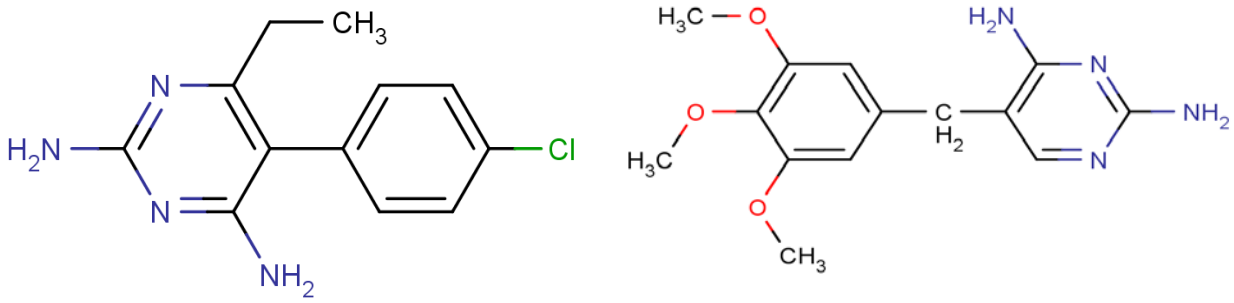


Proguanil



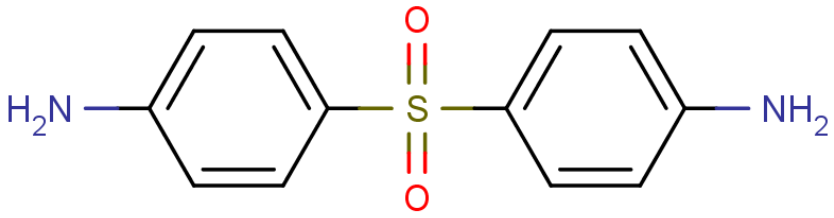
Cicloguanil

Figura 2.5 Análogos de guanidina (biguanidas)



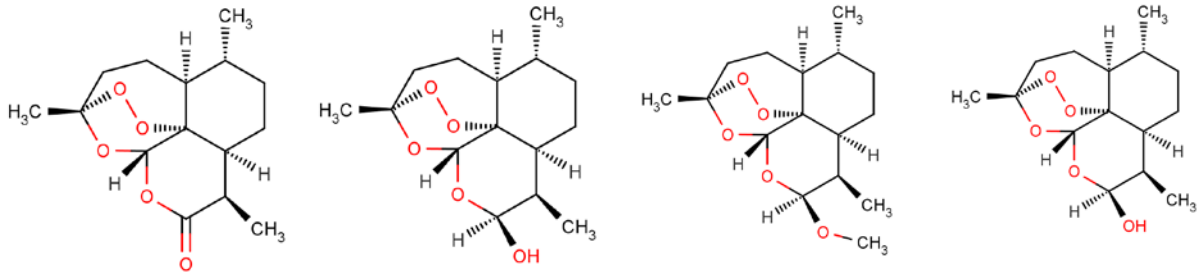
Trimetoprima Pirimetamina

Figura 2.6 Análogos de pirimidina (diaminopirimidinas)



Dapsona

Figura 2.7 Sulfonas



Artemisina Dihidroartemisina Artemeter Artesunato

Figura 2.8 Artemisina y sus derivados

ANEXO 3. QSAR

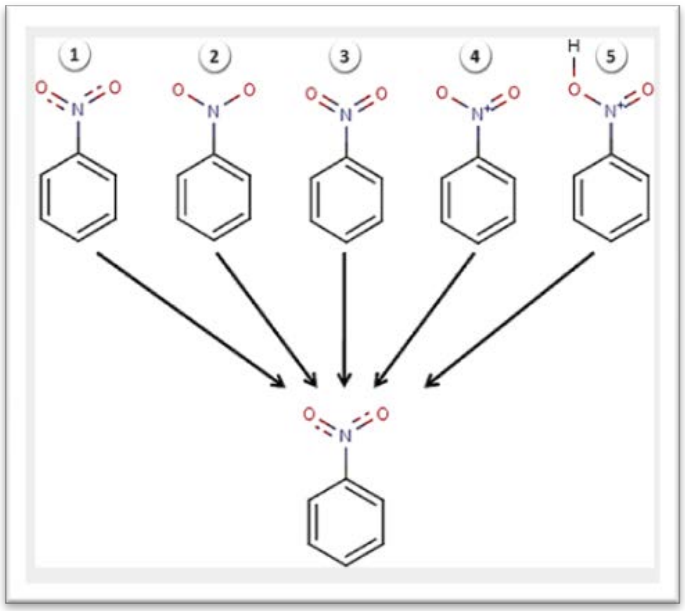


Figura 3.1. Cinco formas diferentes de representar el grupo nitro

ANEXO 4. Resultados

Tabla 3.1 Desempeño del mejor multiclasificador seleccionado para cada metodología

Metodología	Desempeño (%)
All Base Models Major Vote	72,64831573
All Base Models Borda Vote	74,8500904
All Base Models AB	79,77344311
Iterative Clustering Major Vote Corr Matrix	76,37725691
Iterative Clustering Borda Vote Corr Matrix	78,29602927
Iterative Clustering AB Corr Matrix	80,84122572
Iterative Clustering Major Vote Q Matrix	77,32293307
Iterative Clustering Borda Vote Q Matrix	77,42118584
Iterative Clustering AB Q Matrix	80,80548399
GA Major Vote Min AIC 10 init models	84,19098886
GA Borda Vote Min AIC 15 init models	83,86348423
GA AB Vote Min AIC 10 init models	82,97954457
GA Major Vote Min Error 15 init models	84,52352043
GA Borda Vote Min Error 15 init models	84,0934138
GA AB Vote Min Error 15 init models	77,49675559