



UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
VERITATE SOLA NOBIS IMPONETUR VIRILISTOGA. 1948

Facultad de Química-Farmacia
Departamento de Licenciatura en Química

Trabajo de Diploma



Modelación de la actividad antimalárica de compuestos orgánicos frente al parásito protozoario Plasmodium falciparum

Autor: Adilen Isabel Acosta González

Tutor: Dr.C. Oscar Martínez Santiago, PA

Consultante: Dr.C. Yovani Marrero Ponce, PT

SANTA CLARA

Curso 2015-2016



RESUMEN

La caracterización estructural de las sustancias es uno de los campos de mayor interés dentro de las ciencias químicas. El objetivo fundamental de este trabajo consiste en desarrollar regresiones sobre un conjunto químico estructuralmente diverso, conocido como Malaria Box, para encontrar relaciones lineales que correlacionen las estructuras moleculares con sus actividades correspondientes. En este estudio se modeló la actividad antimalárica (expresada como el EC_{50}) de 317 entidades químicas, frente al parásito *Plasmodium falciparum*. Las estructuras químicas fueron codificadas usando índices de derivada discreta de grafos moleculares. Estos fueron calculados utilizando el módulo DIVATI del software TOMOCOMD-CARDD. Posteriormente, para seleccionar los descriptores con mayor variabilidad, fue utilizado el software IMMAN, que permite desarrollar análisis de variabilidad basado en el cálculo de la entropía de Shannon. Los modelos de regresión lineal múltiple (RLM) fueron encontrados en el software MobyDigs, este último usa algoritmo genético para la selección de parámetros. El modelo de RLM obtenido luego de eliminar 60 estructuras outliers demostró mejor ajuste y predictibilidad que el obtenido en un inicio sin la exclusión de las mismas.

ABSTRACT

Structural characterization of the substances is one of the areas of greatest interest in the chemical sciences. The fundamental objective of this work consists in developing regressions of one dataset with high chemical diversity, known as Malaria Box, to find linear relations between the molecular structures and their corresponding activities. In this study was modeling the antimalarial activity (expressed like the EC_{50}) of 317 chemical entities, in front of the parasite *Plasmodium falciparum*. Chemical structures were encoded using indexes of discrete derivative of molecular graphs. These were calculated using the module DIVATI of the software TOMOCOMD CARDD. At a later time, in order to select the descriptors with bigger variability, was used the software IMMAN, that it allows developing analysis of variability based on the calculation of the Shannon entropy. The best MLR equations were found with MObyDigs software. These programs obtain the optimum parameters using genetic algorithm. Last to separate 60 molecular structures of data set, the model show better statistic values and mildly robustness.

PENSAMIENTO
*“La curiosidad es hija de la ignorancia
y madre de la ciencia. ”*
Vico

DEDICATORIA

Mis Padres.

Mi madre Marja Estela.

Por todo el amor y confianza que depositaste en mí para dejarme realizar mis sueños lejos de casa. Solo tú y yo conocemos desde mis primeros pasos en la vida hasta mis mayores sacrificios. Por ser mi amiga del alma y la fuente de inspiración de mis mejores obras.

Mi padre Luis Alberto.

Por todos tus sacrificios que siempre han valido la pena. Por tus enseñanzas, por los valores que aprendí de ti y por la felicidad que pones a mis días.

Mis segundos padres Iliana y Alexis.

Por el amor y cariño con los que han cuidado de mí. Por el apoyo que siempre tuve de ustedes y por darme el calor de un hogar.

Will.

Por el cariño y la responsabilidad con la que cuidaste de mí desde que yo era muy pequeñita. Siempre serás como un padre para mí.

Jorge.

Por tu apoyo incondicional y todos tus mimos. A todos ustedes mis padres les debo mi vida.

Mi novio.

Por ser mi más grande amor y sostén en la vida. Gracias por ser en los momentos necesarios mi padre, mi hermano, mi amigo, mi confidente. Gracias por ser mi paño de lágrimas y por soportarme cuando ni yo misma lo logro. Sin ti mi vida no sería tan linda.

Mis hermanos.

Mis amores, a Lili, la mayor, por librarme de ser la más vieja y a los que me siguen Adriana, Addel, Glauser, Luis Alberto y Miguel Ángel por la responsabilidad y fuerza con las que he tenido que crecer por ustedes. A todos los adoro y son la alegría de mis días.

Mis abuelos.

Por los consentimientos y los mimos que me permitieron sentirme segura. Por el amor y el cuidado que recibí siempre de ustedes.

A ti abuelo que Dios te tenga en la gloria, espero que donde estés te sientas orgulloso de mí. Siempre seguiré tus consejos.

Mis tíos.

Belkis, Isabel, Maribel, Irabel, Dilía, Doralina, Tío Pito (que descanses en paz), Arle y Yanetsy, por el apoyo que he recibido de todos ustedes, el amor y los abrazos cariñosos.

Mis primos.

Por ser muchos no los menciono pero a todos los quiero mucho por ser mis compañeros de andanzas en la escuela y en la vida.

Mis amigos.

Ana, Disle, Ruji, Juan, Liesly, Rosi, Caco, Lía, Ivan, Elaine, Leidy Laura, Taimí, Carli, Pepe, Orlando, Campal, mis compañeros de aula y las muchachitas de cuarto año de farmacia, por los consejos, compañía y apoyo incondicional. Nunca los olvidaré.

Mi tutor Oscar.

Por llenarme de esperanza y confianza al darme su apoyo y amistad. Gracias por tenderme tu mano amiga en esta tarea tan importante como es terminar mi carrera universitaria. Te estaré siempre agradecida por los conocimientos que me has brindado.

Agradecimientos

Agradezco a:

Mis profesores.

A todos aquellos que me enseñaron con amor y dedicación. Les agradezco de todo corazón por los conocimientos y la preparación que recibí de todos ustedes para mi vida como profesional y como ser social.

Raúl.

Por su apoyo en mis primeros pasos. Gracias compañero siempre te tendré presente.

Viviana.

Por ser una buena compañera de cálculos. Gracias por ayudarme con mi trabajo.

Mis compañeros de aula.

Por el apoyo y las enseñanzas que recibí de ustedes. Siempre recordaré los buenos momentos que vivimos.

A todos les estaré eternamente agradecida.

Tabla de Contenido

INTRODUCCIÓN	10
1 REVISIÓN BIBLIOGRÁFICA.....	15
1.1 Malaria	15
1.1.1 Historia de la malaria	15
1.1.2 Impacto Económico de la Malaria	16
1.1.3 Historia de los Fármacos Antimaláricos	17
1.1.4 Resistencia a los Fármacos Antimaláricos	19
1.2 Métodos in silico en la búsqueda de nuevos compuestos activos	21
1.2.1 Descriptores Moleculares	21
1.2.2 Estudios de Relación Cuantitativa Estructura-Actividad (QSAR)	25
1.2.3 Métodos Estadísticos	27
1.2.4 Regresión	28
1.2.5 Regresión Lineal Múltiple	28
1.2.6 Calidad Estadística de un Modelo	31
1.2.7 Datos Atípicos u Outliers	32
2 MATERIALES Y MÉTODOS.....	34
2.1 Base de Dato Química	34
2.2 Herramientas Computacionales	35
2.2.1 Software DIVATI. Cálculo de los nuevos Índices de Derivada del Grafo	35
2.2.2 Software IMMAN. Análisis de Variabilidad (AV)	37
2.2.3 Software Moby Digs. Regresión Lineal Múltiple	38
3 RESULTADOS Y DISCUSIÓN.....	40
3.1 Modelación de actividad antimalárica de sustancias orgánicas	40
3.1.1 Cálculos de los Índices de Derivada del Grafo	40
3.1.2 Selección de los mejores descriptores moleculares	40
3.2.1 Modelos de RLM	41
3.2.2 Identificación de outliers	45
3.2.3 Discusión de los resultados de la regresión lineal múltiple:	48
CONCLUSIONES:	50
RECOMENDACIONES:	51
REFERENCIAS BIBLIOGRÁFICAS	52
ANEXOS	56

Glosario

2D	Bidimensional
3D	Tridimensional
DMs	Descriptores Moleculares
ADL	Análisis Discriminante Lineal
RLM	Regresión Lineal Múltiple
F _{exp}	Razón de Fisher experimental
ITs	Índices Topológicos
LOO	Procedimiento de validación interna cruzada 'leave-one-out'
q ²	Coeficiente de correlación al cuadrado del procedimiento de validación cruzada LOO
QSAR	Quantitative Structure Activity Relationships
CV	Cribado Virtual
R ²	Coeficiente de determinación o coeficiente de correlación al cuadrado
GDI	Índices de Derivada del Grafo
AG	Algoritmo Genético
LOVIs	Invariantes Locales del Vértice
PCA	Análisis de Componentes Principales

INTRODUCCIÓN

“La ciencia es el fundamento de todo progreso, que mejora la vida humana y alivia del sufrimiento.”

Irene Joliot-Curie

La malaria o paludismo se puede considerar la enfermedad parasitaria tropical que más afecta al mundo. (Laguna, 2011) El alto riesgo de contraer esta enfermedad potencialmente mortal en el 40% de la población mundial, los más de 500 millones de personas afectadas anualmente, y el impacto económico que causa, son algunas de las razones que sitúan a la malaria entre las mayores preocupaciones a nivel de Salud Pública (Peydró, 2008). La malaria es una enfermedad infecciosa, producida por un protozoo intracelular del género *Plasmodium* (Tugores, 2014).

El paludismo humano está producido por cuatro especies diferentes del parásito protozoario *Plasmodium*: *Plasmodium falciparum*, *P. vivax*, *P. ovale* y *P. malariae*. Los humanos, ocasionalmente, pueden llegar a infectarse con una especie de *Plasmodium* que normalmente infecta a animales, tal como el *P. Knowlesi*, sin embargo hasta el momento no hay informes de transmisión humano-mosquito-humano de dicha forma “zoonótica” de paludismo. (Isea, 2010) La forma más frecuente y grave de la enfermedad se debe a *P. falciparum* y es transmitido por la especie hembra del mosquito *Anopheles*, que pica principalmente entre el anochecer y el amanecer (Geneva, 2009).

La infección plasmodial puede cursar desde asintomática, pasando por un cuadro leve de fiebre, escalofrío, sudoración y malestar general, y llegar hasta un síndrome grave o complicado, con hiperpirexia, anemia progresiva, confusión, somnolencia o debilidad extremas y luego presentarse deterioro de la consciencia hasta un coma profundo (malaria cerebral), disturbios hidroelectrolíticos, septicemia (malaria álgida) y disfunción multiorgánica que puede llevar a la muerte (Castaño, 2006).

La transmisión de la malaria varía entre regiones, incluso en un mismo país; depende de la especie del parásito, sus propiedades y las de los vectores, las condiciones ecológicas que intervienen en la transmisión de la parasitosis y factores socioeconómicos. La diversidad genética le confiere a *Plasmodium* la capacidad para evadir la respuesta inmune del hospedador y producir variantes resistentes a medicamentos y vacunas, siendo esto en gran parte, causa del éxito de la supervivencia de este parásito en la historia evolutiva, así como del fracaso de las medidas empleadas con el objetivo de erradicarlo.

El tratamiento ha sido posible durante muchos años gracias a la existencia de un número restringido de fármacos presentando cada uno de ellos una serie de limitaciones de tipo farmacológico, aunque el mayor problema es la aparición de resistencias. La resistencia a los medicamentos está documentada para todas las clases de antimaláricos, incluidos los derivados de la artemisinina, y es una seria amenaza para el control de la malaria.(Zúñiga, 2002)

Durante la pasada década, las inversiones realizadas en la prevención y el control de esta enfermedad han tenido un impulso sin precedentes que han salvado más de un centenar de vidas. Sin embargo la enfermedad sigue causando millones de muertes principalmente en niños menores de cinco años de edad.(Tugores, 2014).

En la actualidad se han realizado grandes esfuerzos para lograr más eficiencia en la selección de compuestos antimaláricos. Tal es el caso de la utilización de cribados automatizados de alto rendimiento (HTS, por sus siglas en inglés) sobre la colección de productos de GlaxoSmithKline (GSK), a través de los cuales se han identificado más de 13000 compuestos activos en cepas de *P. falciparum*. Estos resultados son solamente la punta del "iceberg" y requieren inversiones sustanciales para que esto consolide (Ekins, Williams, Krasowski, & Freundlich, 2011), ya que muchos de los compuestos analizados pueden no tener las condiciones necesarias para llegar a ser fármacos. Hay pocos ejemplos de la utilización previa de los métodos *in silico* con este propósito. Dentro de estas técnicas, el cribado virtual (CV) tiene la ventaja de ser más económica (ahorro en compra de reactivos y robotización), rápida, y permite tener en cuenta una cantidad de compuestos del orden de billones, cifra impensable experimentalmente. Todo esto mejora la eficiencia y aporta racionalidad a la búsqueda de nuevos antimaláricos. Por otra parte los estudios de Relaciones Cuantitativas Estructura-Actividad (QSAR, por sus siglas en inglés) se han utilizado ampliamente en la modelización de disímiles propiedades moleculares de naturaleza física, química y biológica, siendo actualmente, el enfoque más utilizado en el diseño de fármacos (Lazarou et al., 2009). Este tipo de análisis es muy útil y generalmente es utilizado como principal herramienta en la selección de compuestos durante el protocolo de CV.

En la Universidad Central "Marta Abreu" de las Villas existen grupos de investigación [Modelación Molecular del Centro de Bioactivos Químicos (CBQ), Bioinformática e

Inteligencia Artificial del Centro de Estudios de Informática (CEI) y el CAMD-BIR de la Facultad de Química-Farmacia fundamentalmente] que dirigen sus estudios al desarrollo de nuevos principios activos para la industria farmacéutica en las ramas de la medicina veterinaria y humana. (Santiago, 2015b). En específico el Grupo de Descubrimiento Molecular Asistido por Computadora e Investigaciones Bioinformáticas (CAMD-BIR Unit, por sus siglas en inglés) ha orientado sus investigaciones hacia la utilización de diferentes métodos para los estudios QSAR, priorizando el desarrollo y aplicación de nuevos descriptores grafo-teóricos para el diseño molecular de compuestos orgánicos potencialmente bioactivos. En el CAMD-BIR existe una tradición en la utilización de ITs (Índices Topológicos) y descriptores 3D marcada por la definición de los índices algebraicos (basados en formas bilineales, trilineales y cuatrilineales) moleculares por el Dr. Yovani Marrero Ponce y cols., los cuales han sido aplicados con éxito a diversos estudios de relación estructura-actividad/propiedad.(Schultz, 2003)

Los Descriptores Moleculares o Índices agrupan una serie de parámetros que a través de valores numéricos representan información estructural, electrónica, estérica, así como una determinada propiedad químico-física de la molécula en estudio. (Escalona, 2010)

Los ITs han comenzado a ocupar un lugar importante dentro del conjunto de DMs utilizados en estudios donde se aplican métodos de Química-Matemática y Quimiinformática, siendo probablemente el diseño y descubrimiento de nuevos compuestos bioactivos, una de las áreas más activas de investigación donde se aplican estos DMs a problemas biológicos.(Santiago, 2015b)

Existen muchas técnicas quimiométricas que pueden ser utilizadas para describir las actividades biológicas y propiedades en general de sustancias orgánicas, pero la regresión lineal se destaca por su simplicidad, y además su uso ha sido ampliamente descrito y aplicado en quimiometría. El criterio de selección de la Regresión Lineal Múltiple (RLM) empleado en el presente trabajo fue, fundamentalmente, la simplicidad de la misma, la amplia descripción bibliográfica al respecto y la robustez de la técnica. Sin embargo no se han realizado estudios de (RLM) para determinar la actividad antimalárica debido a la dificultad de cuantificación de esta actividad y a la garantía de la homogeneidad de condiciones en la determinación de los valores. Recientemente se ha recopilado información cuantitativa de 400 moléculas sobre su actividad frente al parásito *Plasmodium*

Falciparum y han sido organizadas en un conjunto químico denominado MalariaBox,(Spangenberg, 2013) a pesar de ello no se han reportado estudios de regresión usando esta base de datos química. Además no se conoce una estrategia de cribado farmacológico que integre métodos *in silico* e *in vitro* fiables, rápidos y económicos para el descubrimiento de nuevos compuestos que contribuyan al tratamiento de la malaria.

Toda la información anterior tributa a que se plantee el siguiente problema científico:

Problema científico:

¿Cómo encontrar modelos *in silico* de RLM que permitan describir adecuadamente la acción de moléculas de estructuras diversas frente al parásito *Plasmodium falciparum*?

Como vía para solucionar el problema científico se formula la siguiente hipótesis:

Hipótesis:

Es posible encontrar relaciones lineales entre la actividad frente al parásito *Plasmodium falciparum* (causante de la malaria) y DMs basados en el uso del concepto de Derivada Discreta de Grafos Moleculares.

Para demostrar la anterior hipótesis y dar respuesta al problema científico planteado, se proponen los siguientes objetivos:

Objetivo General:

Encontrar modelos matemáticos que describan satisfactoriamente la actividad de moléculas orgánicas contra el parásito *Plasmodium falciparum* y que puedan ser usados en posteriores estudios de CV y desarrollo de nuevas entidades químicas como candidatos prometedores para combatir la malaria.

Objetivos específicos:

- Organizar y limpiar un conjunto químico de 400 moléculas diversas estructuralmente para las cuales se ha reportado actividad frente al parásito *Plasmodium falciparum*, las que han sido agrupadas bajo el nombre de Malaria Box.
- Calcular índices de derivada discreta de grafos moleculares (GDI), para el conjunto anteriormente mencionado.

- Encontrar modelos robustos de RLM que relacionen las estructuras del conjunto químico con sus respectivas actividades biológicas.

La **novedad científica** de este trabajo se fundamenta en la obtención de nuevos modelos matemáticos lineales para el descubrimiento de potenciales fármacos antimaláricos, usando GDI como estrategia de codificación de las estructuras químicas para su posterior aplicación al cribado virtual y el diseño molecular de nuevos candidatos a fármacos.

1 REVISIÓN BIBLIOGRÁFICA

*“La ciencia es el alma de la prosperidad de las naciones
y la fuente de todo progreso.”*

Louis Pasteur

En este capítulo se abordan los resultados de una extensa y exhaustiva búsqueda bibliográfica en la literatura internacional acerca de los aspectos teóricos que justifican y sustentan la presente investigación.

1.1 Malaria

1.1.1 Historia de la malaria

La malaria ha infectado a los humanos por más de 50,000 años, y puede que haya sido un patógeno humano durante la historia entera de nuestra especie (Rojas, 2008). Esta enfermedad se originó en África (Catalina, 2012). De cierto, especies cercanas a los parásitos humanos de la malaria se han encontrado en los chimpancé, pariente ancestral de los humanos. Se encuentran referencias de las peculiares fiebres periódicas de la malaria a lo largo de la historia, comenzando desde 2700 a. C. en China (Rojas, 2008). En el pasado, la malaria era común en las zonas de ciénagas en Roma y por esto su nombre se origina del italiano “mal – aria” o “mal aire”, también se le llamó “Fiebre Romana” o paludismo que se deriva del latín “palud” que significa pantano, asimismo se le denominó “calentura o fiebre de los pantanos”, por las asociaciones que se le daba con estas tierras. La malaria fue común en gran parte de Europa y América del Norte.(Catalina, 2012)

Hipócrates fue el primero en describir los síntomas de la enfermedad, vinculándolos con la época del año y el lugar donde vivían los afectados. La relación del agua reposada con el gran número de casos de malaria, hizo que en Roma se construyeran sistemas de desagüe.

Los primeros avances sobre la malaria, los realizó el médico francés Charles Louis Alphonse Laveran, en 1880; quien observó por primera vez la presencia de los parásitos en los glóbulos rojos de personas con malaria (Catalina, 2012).

Charles propuso por ello que la malaria era causada por un protozoario, la primera vez que se identificó a un protozoario como causante de una enfermedad.(Gamenara, 2005) Al protozoario en cuestión se le llamó Plasmodium, por los científicos italianos Ettore Marchiafava y Angelo Celli. Un año después, Carlos Finlay, un médico cubano tratando pacientes con fiebre amarilla en la Habana, sugirió que también (al igual que en la fiebre

amarilla) en el paludismo eran los mosquitos quienes transmitían la enfermedad de un humano a otro (Rojas, 2008).

En 1886, Camillo Golgi, estableció que la enfermedad se presentaba en más de una variante. En 1890, investigadores italianos, introdujeron por primera vez los términos *Plasmodium vivax* y *malariae*, luego en 1897 se descubrió una tercera clase de parásito el *P. falciparum* y no fue hasta 1922 que se descubrió al parásito *P. ovale* (Catalina, 2012).

Posteriormente, fue el británico Sir Ronald Ross, trabajando en la India, quien finalmente demostró en 1898 que la malaria es de hecho transmitida por mosquitos. Lo probó al mostrar que ciertas especies del mosquito transmitían malaria a pájaros, aislando los parásitos de las glándulas salivales de mosquitos que se alimentaban de aves infectadas. Por su aporte investigativo, Ross recibió el premio Nobel de Medicina en 1902 (Rojas, 2008).

Fue en 1948 cuando se identificó el ciclo vital de este mosquito.(Todeschini, 2009) Sin embargo no fue sino hasta 1980 que se observó la forma latente hepática del parásito. El descubrimiento de esta forma latente del parásito, finalmente explicó por qué daba la impresión de que algunas personas se curaban de la enfermedad, para luego recaer años después de que el parásito había desaparecido de su circulación sanguínea. Además en 1980 se encontró que el parásito al comienzo de su ciclo en el cuerpo (ver en anexos figura 1) , se alojaba en el hígado.(Piñeros, 2010)

1.1.2 Impacto Económico de la Malaria

El paludismo produce pérdidas económicas considerables, que han llevado a diferencias entre los valores del PIB de los países con y sin paludismo. Los costos sanitarios de esta enfermedad incluyen gastos tanto personales como públicos en prevención y tratamiento.(Tugores, 2014)

En los últimos años la industria farmacéutica está sumergida en una crisis importante, únicamente, el 1% de los *hits* descubiertos en las etapas preliminares del proceso de desarrollo de fármacos llegan al mercado, y de estos, solamente 3 de cada 10 recuperan la inversión realizada.(Carrero, 2003)

La Organización Mundial de la Salud (OMS) describe que se necesita un estimado de 5.100 millones de dólares por año entre 2011 y 2020 para alcanzar el acceso universal a las intervenciones en malaria y actualmente, solo hay disponibles 2.300 millones de dólares.(Quiñones, 2015)

En la actualidad el gasto en el tratamiento de la enfermedad es muy superior al de las pruebas de diagnóstico rápido (PDR), pero se espera que disminuya por la estrategia de ampliar la prueba parasitológica a todos los casos sospechosos de malaria antes de usar el tratamiento. Con los precios actuales de las PDR y las terapias combinadas con artemisina (0,50 dólares para las PDR y 1,40 dólares para la combinación de arteméter-lumefantrina), y el estricto cumplimiento del protocolo de tratamiento, el ahorro en materias primas podría llegar a ser considerable.(Meléndez, 2014)

Si no se puede garantizar un apoyo financiero y político sostenido para el control y eliminación de la malaria, la enfermedad reaparecerá, lo que mermará los beneficios obtenidos de las inversiones realizadas hasta la fecha y generará costos todavía mayores para los países y la comunidad internacional.(Hoz, 2014)

Desde 1998, los fondos para la lucha contra la malaria han aumentado sustancialmente, han alcanzado los 2700 millones de dólares en 2013 y se espera que lleguen a los 3200 millones de dólares en 2016. Los recursos externos (procedentes de donantes) financiaron la mayor parte del gasto creciente, que pasó de menos de 100 millones de dólares en 1998 a 2100 millones de dólares en 2013, con un aumento anual del 43% entre 2005 y 2009. Los fondos nacionales de los programas de malaria también aumentaron, representando una quinta parte de todo el financiamiento (527 millones de dólares, es decir, un 20%) en 2013.

Conseguir los objetivos contra la malaria para 2030 dependerá de la movilización de un mayor financiamiento predecible y mantenido y requerirá un esfuerzo exponencial en un mundo cada vez más complejo.(Pínel, 2011) Muchos de los principales donantes tendrán problemas para mantener los niveles actuales de apoyo a menos que la economía mundial mejore. Aunque todavía no se han visto los resultados en términos de desembolso, los compromisos de ayuda al desarrollo están disminuyendo, y los países donde la malaria es endémica tienen múltiples problemas con los que lidiar. Nunca ha sido tan grande la necesidad de estrategias de movilización de recursos para la malaria a todos los niveles y la amenaza omnipresente de reaparición de la enfermedad hace que sea peligroso disminuir el financiamiento de la malaria incluso en contextos de baja tasa de transmisión. (Salud, 2015)

1.1.3 Historia de los Fármacos Antimaláricos

La historia del medicamento se remonta a los orígenes de la sociedad humana.(Escalona, 2010) El primer tratamiento eficaz para la malaria fue la corteza del árbol Cinchona, el cual

contiene el alcaloide quinina. Este árbol crece en las colinas de los Andes, en particular en Perú. El producto natural se usaba por habitantes del Perú para controlar la malaria, y los Jesuitas introdujeron esta práctica en Europa durante los años 1640 donde fue aceptada rápidamente.(Monge, 2003)

A comienzos del siglo XX, antes de los antibióticos, los pacientes con sífilis eran intencionalmente infectados con malaria para crear una fiebre. Al controlar la fiebre con quinina, los efectos tanto de la sífilis como la malaria podían ser minimizados. Algunos de los pacientes murieron por la malaria, aunque ello era preferido por encima de la casi segura muerte por sífilis. (Rojas, 2008).

Entre 1934 y 1946 se hizo el primer medicamento químico antimalárico, la cloroquina, por un alemán que trabajaba para los laboratorios Bayer. En 1939, un suizo llamado Paul Müller, hizo el primer insecticida, el DDT, el cual fue usado para fumigar las áreas más endémicas de la enfermedad; pero con el paso del tiempo los mosquitos crearon inmunidad contra este insecticida.

Posteriormente, en 1955, la Organización Mundial de la Salud se dio a la tarea de erradicar la enfermedad de países con climas tropicales. Esta campaña consistía en cuatro pasos: preparación, ataque, consolidación y mantenimiento; logrando que en varios de los países participantes disminuyera el número de muertos a causa de esta enfermedad.

Sin embargo, muchos países de África fueron excluidos del programa de erradicación de la malaria.

Hoy, los gobiernos de África y de otros países y diferentes fundaciones están centrando sus esfuerzos en encontrar la manera de controlar la enfermedad de manera definitiva y no provisional. (Catalina, 2012).

En la lucha contra esta enfermedad el trabajo en asociación y patneriado público/privado es fundamental e imprescindible. Esta enfermedad es una de las primeras causas de muerte en los países en los que es endémica, pero los avances en la investigación y el acceso de la sociedad a los tratamientos contribuirán a que en 2020 pueda llegar a erradicarse.(Barrio, 2014)

La historia de los fármacos antimaláricos siempre ha estado marcada por la competición entre la evolución de parásitos resistentes al medicamento y la formulación de nuevos fármacos. En la actualidad, los expertos en medicamentos contra la malaria se están

centrando en terapias que combinan derivados de artemisina con otros compuestos acompañantes, con el fin de evitar estas resistencias.(Berganza, 2009-2010)

1.1.4 Resistencia a los Fármacos Antimaláricos

La resistencia a los fármacos antimaláricos ha sido definida como la habilidad de una cepa del parásito para sobrevivir y/o multiplicarse a pesar de la administración y absorción de un fármaco dado en dosis igual o más alta que aquella recomendada usualmente pero entre lo tolerado por el sujeto. Esta definición fue después modificada para especificar que el fármaco en cuestión debería tener acceso al parásito o al glóbulo rojo infectado con la duración del tiempo necesario para su acción normal.

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.(Rodríguez, 2006)

Al comienzo de la resistencia desaparecen los síntomas iniciales de la infección y el paciente parece mejorar durante un período corto; sin embargo, los síntomas reaparecen (generalmente entre tres a seis semanas después del tratamiento), la anemia puede empeorar y hay una mayor probabilidad de que el paciente sea portador de gametocitos (formas parasitarias que a su vez portan los genes que confieren resistencia). El paciente y el proveedor de tratamiento interpretan principalmente estas características tempranas de resistencia, como una infección recién contraída. A medida que esta empeora se acorta el intervalo entre infección primaria y recrudescencia; tiempo después los síntomas no desaparecen con el tratamiento y la incidencia de malaria probablemente aumente en áreas de baja transmisión, mientras que la mortalidad probablemente aumente en todas las áreas.

Aunque el problema de la resistencia a los fármacos está presente en todas las zonas endémicas de paludismo, las áreas reconocidas actualmente como más afectadas son la península de Indochina y la región del Amazonas en Suramérica. (Tugores, 2014). *Plasmodium falciparum* es la cepa con el mayor índice de incidencia, llegando a registrar 500 a 800 millones de casos anuales aproximadamente en todo el mundo. En el mapa mundial que está disponible en los anexos (figura 2) se muestran las zonas que están afectadas por Malaria y aquellas zonas que están en riesgo de ser afectadas por la cepa de

Plasmodium falciparum de acuerdo con datos obtenidos de la Organización Mundial de la Salud.(Isea, 2010)

Mecanismos de Resistencia Antimalárica

En general, la resistencia parece ocurrir a través de mutaciones espontáneas que confieren una sensibilidad reducida a un fármaco o a una clase de fármacos dado. Para algunos fármacos, se requiere solamente un punto único de mutación para conferir resistencia, mientras que para otros, parece ser requerido múltiples mutaciones. Las mutaciones presentes no son perjudiciales para la supervivencia o reproducción de los parásitos, la exposición a los fármacos puede remover los parásitos susceptibles mientras que los resistentes sobrevivirán. Se ha encontrado que un aislado de malaria único puede ser originado de una población heterogénea de parásitos que puede tener una amplia variedad de respuestas características a los fármacos, desde una alta resistencia a una sensibilidad completa (Heydra, 2006).

Alcance de la Resistencia a los Antimaláricos

La resistencia a los antimaláricos está documentada para *P. falciparum*, *P. malariae* y *P. vivax*. Se ha observado resistencia de *P. falciparum* a todos los antimaláricos actualmente en uso (amodiaquina, cloroquina, mefloquina, quinina y sulfadoxina-pirimetamina) y, más recientemente, a derivados de la artemisinina. *P. vivax* ha desarrollado rápidamente resistencia a la sulfadoxina-pirimetamina en muchas áreas, mientras que su resistencia a la cloroquina se limita en gran parte a Indonesia, Papua Nueva Guinea, Timor-Leste y otras partes de Oceanía.(Fonseca, 2015) Existen también informes sobre resistencia en Brasil y Perú. *P. vivax* sigue siendo susceptible a la cloroquina en la mayor parte de Asia Sudoriental, el subcontinente indio, la península coreana, el Oriente Medio, el África nororiental y la mayor parte de América del Sur y Central. (Salud., 2010).

En las últimas décadas, la frustración por el fracaso de la esperada erradicación de la enfermedad en los años 60 y las crisis económicas y financieras de los años 70 y 80, produjeron serias reducciones de los recursos dedicados a la lucha antimalárica, dificultando el progreso y permitiendo el desarrollo de resistencias en el parásito. Al mismo tiempo la explotación abusiva de recursos naturales, los conflictos bélicos y disturbios sociopolíticos y el constante aumento de poblaciones desplazadas, han producido graves resurgimientos de la enfermedad y su extensión a áreas previamente libres.

Aunque la malaria es una enfermedad completamente prevenible y tratable (Salud, 2012) el deterioro de la situación global afecta también a las poblaciones de los países industrializados, no sólo por el creciente número de casos en viajeros internacionales sino por el riesgo de un posible resurgimiento debido al cambio climático.

Los indicadores sociales y de salud han señalado que la pobreza familiar es clave en la génesis de la enfermedad y que la morbilidad se concentra, generalmente, en los individuos, familias y grupos o clases sociales más pobres. (Griffith, 2015)

Sin embargo, el desarrollo de nuevos métodos de lucha y el progreso de la investigación científica abren nuevas esperanzas, que se ven respaldadas por el resurgir de la solidaridad internacional, reflejada en aportaciones financieras sin precedentes al desarrollo de las campañas de lucha, a la subvención de los instrumentos básicos (medicamentos y mosquiteros tratados) y a la investigación científica. (Nájera, 2009).

1.2 Métodos *in silico* en la búsqueda de nuevos compuestos activos

El costo del descubrimiento de nuevos fármacos excede los 750 millones por cada nueva entidad química, por lo que terapéuticas novedosas para enfermedades olvidadas pueden estar fuera del alcance de países del tercer mundo. (Pérez, 2011) Universidades e instituciones sin ánimo de lucro pueden jugar un papel importante en la búsqueda de nuevas estrategias terapéuticas, y tienen el potencial para crear nuevos paradigmas de descubrimiento de fármacos antipalúdicos que aumenten la efectividad y eficiencia de los métodos tradicionales de experimentación de “prueba y error”. (Tugores, 2012)

Los métodos *in silico* son una de las pocas técnicas que han tenido y tienen potencial para mejorar significativamente el descubrimiento y posterior desarrollo de fármacos (Tugores, 2014).

Los procedimientos *in silico* evitan los procesos actuales de síntesis y bioensayos, complementan y/o reemplazan los estudios *in vitro/in vivo*, resultan la única opción para el estudio y tratamiento de compuestos virtuales, y además el costo de su implementación es lo suficientemente inferior a los métodos experimentales. (Pastor, 2003)

1.2.1 Descriptores Moleculares

Los Descriptores Moleculares (DMs) juegan un rol fundamental actualmente en el desarrollo de las Ciencias Químicas, las Ciencias Farmacéuticas, las políticas de protección ambiental e investigaciones de la salud. Un DM es: el resultado final de un procedimiento

lógico y matemático en el cual se transforma la información química codificada en un número de utilidad (o conjunto de ellos) como representación simbólica de la molécula.

El término utilidad, posee aquí un doble sentido, dado que los resultados numéricos finales pueden brindar información sobre la naturaleza físico-química y estructural de las moléculas así como ser relacionados con sus propiedades. Aun, si la interpretación en términos estructurales y químico-físicos del descriptor es débil, temporal o ausente, el mismo puede estar correlacionado con al menos una propiedad.

Los descriptores pueden ser agrupados inicialmente en dos clases generales:

1. Los derivados de medidas experimentales como: logP, refractividad molar, momento dipolo, polarizabilidad y otras propiedades químico-físicas en general.
2. Descriptores Moleculares Teóricos, los cuales son derivados de representaciones simbólicas de las moléculas y estos a su vez pueden ser clasificados acorde con diferentes formas de representación molecular.(Jurs, 1995)

La naturaleza de los DMs depende de cuál haya sido el proceder utilizado para la definición de los mismos. En ese sentido si se tiene en cuenta la fórmula química (la representación molecular más simple), los descriptores derivados de este tipo de representación son conocidos como **descriptores-0D**. Ejemplos de algunos descriptores de este tipo pueden ser, la masa molecular, conteo de átomos, en general descriptores constitucionales y/o alguna función de las propiedades atómicas.(Todeschini, 2009)

Una representación unidimensional de estructuras químicas consiste en listar fragmentos moleculares, grupos funcionales o sustituyentes de interés presentes en una determinada molécula sin ser requerido un conocimiento completo de la estructura molecular. Los descriptores derivados de este tipo de representación son conocidos como **descriptores-1D** y son mayormente usados en investigación y análisis de subestructuras.(Todeschini, 2009)

La representación bidimensional de las moléculas considera la conexión de los átomos y esta conectividad se define en términos de la presencia y naturaleza de los enlaces químicos presentes. Los grafos moleculares permiten desarrollar representaciones moleculares en dos dimensiones de las moléculas y estas representaciones son conocidas como representaciones topológicas. Los DMs derivados de aplicaciones de algoritmos matemáticos a representaciones topológicas se conocen como **descriptores-2D**; también son denominados invariantes grafo-teóricas o índices topológicos.(Kier, 1999)

La representación tridimensional de una molécula muestra a la misma como un objeto geométrico rígido en el espacio y permite representar tanto la naturaleza y conectividad de los átomos como la conformación espacial de la estructura molecular. Este tipo de representación se conoce como geométrica y los descriptores derivados de este tipo de representación constituyen **descriptores-3D**.(Todeschini, 2009)

La representación estéreo-electrónica de una molécula es una descripción molecular relacionada con propiedades moleculares que emanan de la distribución electrónica y la interacción con el espacio que rodea a la estructura. Los descriptores de este nivel pueden ser considerados **descriptores-4D** y son caracterizados por un campo escalar y una geometría molecular.(Todeschini, 2009)

Los descriptores pueden describir a la molécula como un todo (descriptores globales) o solo representar un fragmento presente en ella (descriptores fragmentos). Generalmente, un gran número de descriptores moleculares surgen de diferentes teorías, tales como la Teoría de Orbitales Moleculares, la Teoría de Grafos, La Mecánica Cuántica, entre otras.(Castro, 2011)

Además, los descriptores se pueden clasificar en función de:

1. La invariabilidad de sus propiedades, es decir, su capacidad para rendir un valor independiente de características particulares de la representación del compuesto. Estas propiedades son la invariabilidad química (tipos de átomos o enlaces), invariabilidad translacional y rotacional (en función del marco de referencia espacial) y la conformación de la representación geométrica. Los descriptores 3D que presentan invariabilidad translacional y rotacional son particularmente útiles, ya que no requieren el alineamiento previo de las moléculas, por lo que se ahorra tiempo de cálculo y se evitan problemas asociados con el alineamiento.
2. Su degeneración o capacidad de evitar asignar valores idénticos a compuestos distintos.
3. El tipo de propiedad que describen (estéricas, electrónicas, lipofílicas, de forma, descriptores farmacofóricos...).

En el cálculo y selección de descriptores existe básicamente un compromiso entre su eficacia y la eficiencia. La eficacia se entiende como la bondad de un descriptor en

términos de diferenciar entre moléculas diferentes, mientras que la eficiencia hace referencia a la velocidad de cálculo asociada al descriptor.(Rodríguez, 2006)

Los descriptores se utilizan en una amplia variedad de tareas, destacando las técnicas QSAR y predicción de propiedades, clasificación de compuestos, diseño de quimiotecas diversas, búsquedas de estructuras en bases de datos e interpretación de reactividad química y bioquímica.(Gracia, 2010)

Los GDI (Índices de Derivada del Grafo) son DMs que recogen información sobre la simetría molecular, el tamaño, cantidad y posición de ramificaciones, regiones de elevada densidad electrónica y su posición en la estructura, además de conjugación, en caso de existir. Estos TIs tienen en cuenta la presencia de heteroátomos y cómo se afecta el resto molecular por su presencia. Los GDI codifican de forma local y global características de la distribución electrónica de átomos y enlaces y pueden ser expresados como reactividad electrónica y estérica de cada átomo en un entorno molecular determinado y de moléculas en general, además los GDI cuantifican información novedosa de la estructura química, y correlacionan adecuadamente con propiedades/actividades de moléculas orgánicas.(Santiago, 2015a)

Índices de Derivada de Grafos Moleculares (GDI).

Dentro de las diferentes formas modernas de representar las posibles estructuras de las moléculas, los más comunes esquemas desarrollados y semidesarrollados (fundamentalmente moléculas orgánicas) no son más que grafos químicos. Este hecho permite pensar en un posible tratamiento matemático basado en la aplicación de conceptos de la Matemática Discreta y el Álgebra Lineal a estos grafos químicos, surgiendo así toda una disciplina conocida como química grafo-teórica.(Harary, 1971)

La estructura orgánica se fragmenta convenientemente según un suceso determinado. De acuerdo a la teoría de información, la representación de la molécula como un grafo representa la fuente de información, mientras que el suceso escogido para fragmentarla se considera: la forma de caracterización estructural de dicha fuente.(Pareja, 2016)

El suceso prefijado determina la matriz de incidencia (Q), teniendo en cuenta la frecuencia de aparición de cada vértice. Por manipulación algebraica de la matriz Q se obtiene la matriz de frecuencia de relaciones F ($F = Q^T \times Q$).

Para obtener una descripción local se determina la derivada sobre los pares, ternas y cuaternas de los átomos sobre las derivadas:

$$\text{Joint: } \frac{f_i + f_j - f_{ij}}{f_{ij}} \quad (1.1)$$

$$\text{Mutual: } f_{ij} \quad (1.2)$$

$$\text{Conditional: } \frac{f_i - f_{ij}}{f_{ij}} \quad (1.3)$$

$$\text{Jenssent: } \frac{f_i - 2 f_{ij} + f_j}{f_{ij}} \quad (1.4)$$

que son la base para hallar los valores de las Invariantes Locales del Vértice [LOVIs (Δ_i)] para cada uno de los elementos del grafo (o sea para cada átomo de la estructura). Luego es necesaria la obtención de índices totales (involucra a todos los átomos de la estructura) y locales sobre agrupaciones atómicas (involucra a determinados átomos de interés particular, no necesariamente conectados) lo que aumenta las posibilidades de una codificación más completa y adecuada descripción de las propiedades/actividades de las moléculas.(Santiago, 2015b)

1.2.2 Estudios de Relación Cuantitativa Estructura-Actividad (QSAR)

En el área de estudio de nuevos fármacos de la industria farmacéutica, institutos de investigación privados y públicos se han desarrollado herramientas computacionales (in silico *drug design*) y se han establecido nuevas disciplinas entorno de la Química Farmacéutica. Una de ellas es la metodología que relaciona numéricamente estructuras químicas con sus actividades biológicas, denominada QSAR (acrónimo de *Quantitative Structure-Activity Relationships*).(Aponte, 2012)

Los métodos para correlacionar cuantitativamente características estructurales con actividad biológica fueron introducidos por Hansch y Fujita en la década de 1960(Contreras, 2011), con el transcurso del tiempo los métodos QSAR han emergido no sólo como una vía para el descubrimiento de compuestos con una actividad deseada, sino además como un método con aplicaciones al estudio del mecanismo de acción de los mismos (Tugores, 2014).

Un estudio QSAR reúne un conjunto de técnicas computacionales relacionadas con diseño y visualización espacial virtual de moléculas, cálculo de propiedades fisicoquímicas moleculares (descriptores), bioinformática y estadística. Todo esto con el fin de hacer una predicción teórica de la actividad biológica que permita el diseño teórico de posibles

nuevos fármacos, evitando pasar por el proceso a prueba y error de síntesis orgánica. Al ser una ciencia que existe sólo en un entorno virtual, desmaterializado de necesidades de infraestructura (tales como equipo, instrumentos, materiales y personal de laboratorio), con enfoque en las relaciones estructura (química) - actividad (biológica), el diseño de candidatos a nuevos fármacos es mucho más económico y rápido (Aponte, 2012).

En estos métodos, se utilizan múltiples descriptores moleculares (DMs) de la estructura química combinados con la aplicación de técnicas quimiométricas para derivar modelos, por ello se requiere disponer de datos biológicos de calidad, definir descriptores químicos relevantes y elegir un modelo adecuado. (Tugores, 2014)

Dado que los bioensayos son más costosos, es común que los experimentos no se repitan de forma debida para asegurar un valor estadísticamente validado, por su parte al ser un método estadístico, el estudio QSAR incluye este componente de incertidumbre de la variable dependiente por lo que la predicción puede fallar. En contraste, los valores de las variables independientes o descriptores, no tienen dicha naturaleza, ya que cada valor es reproducible e invariable pues descansa en un algoritmo de cálculo definido en el *software*. Eso no quiere decir que los descriptores no fallan, al contrario, hay que discriminar entre descriptores inadecuados (mala selección en ciertos casos), redundantes (covarianza o colinealidad) u otros mal programados (mal diseño del algoritmo).

Los pasos comunes de un estudio QSAR constan de:

1. Formulación del problema
2. Parametrización cuantitativa de la estructura molecular
3. Medición de la propiedad de interés (efectos biológicos u otros)
4. Elección del tipo de modelo QSAR
5. Selección de los compuestos (diseño estadístico de las series)
6. Análisis matemático de los datos y validación (interna y externa) de los modelos obtenidos
7. Interpretación de los resultados y aplicación de los modelos desarrollados a los procedimientos de CV (Cribado virtual).

Sin embargo, el desarrollo de cualquier estudio QSAR, es un ciclo iterativo (Tugores, 2014).

Las ventajas de QSAR son el bajo costo (programas gratuitos, información relativamente accesible, no se utiliza instrumental ni reactivos químicos, etc.), el uso de interfaces que facilitan el manejo y diseño, además de que la construcción de las moléculas y el cálculo de descriptores pueden ser sumamente rápidos. Sus desventajas son la familiarización con metodologías computacionales (diferentes sistemas operativos e interfaces gráficas, manejo de bases de datos, desarrollo del software) y en este sentido, la resolución de diferentes problemas de computo (compatibilidad, actualizaciones, registros, formatos de datos) así como el hecho de tener que disponer de datos de actividad biológica de las moléculas en estudio. (Aponte, 2012)

Tipos de estudios QSAR.

En la actualidad existen diferentes técnicas por las que se puede desarrollar un estudio QSAR en dependencia del tipo de tratamiento matemático que se le dé a la muestra y se pueden clasificar en:

1. QSAR tradicional
2. QSAR por redes de neuronas
3. QSAR tridimensional

QSAR tradicional.

El método tradicional incluye el tratamiento estadístico de los datos por métodos multivariados, que incluye el análisis de regresión, análisis cluster y análisis por componentes principales. En ellos se valora la actividad biológica como variable dependiente del conjunto de descriptores moleculares que constituyen las variables independientes.

Como resultado de esto se obtienen modelos que describen la actividad biológica como una determinada función matemática de los descriptores moleculares, bien sean estos estructurales o químico-físicos.

El método más familiar es la regresión, en la que se obtendrá la ecuación de una recta, un plano, o de un hiperplano, según sea el número de variables independientes incluida en la expresión. (Escalona, 2012)

1.2.3 Métodos Estadísticos

El término *quimiometría* surgió en la década del 70 y se define como la disciplina química que combina herramientas matemáticas y estadísticas con procedimientos para el

análisis e interpretación de los datos químicos. Estas técnicas son utilizadas para la recopilación, elaboración, análisis, reducción y la caracterización de conjuntos grandes de datos, como los que se utilizan en los estudios quimionfórmicos y bioinformáticos.(Waterbeemd, 1995) Actualmente las técnicas de análisis quimiométrico se interceptan no solo con varios campos de la Matemática y la Estadística clásica sino también de la Inteligencia Artificial y otras ramas de la ciencia de la computación.(Vapnik, 1995) A continuación se explicará en qué consiste la regresión línea y serán descritos algunos aspectos importantes del método de regresión lineal múltiple (RLM) que ayudarán al entendimiento de los resultados del presente trabajo.

1.2.4 Regresión

El análisis de regresión lineal es una técnica estadística utilizada para estudiar la relación entre variables. Se adapta a una amplia variedad de situaciones.(Kizys, 2003) El análisis regresivo es una técnica utilizada para inter y extrapolar las observaciones, las cuales pueden clasificarse como regresión lineal o no lineal. Hablamos de modelo de regresión cuando la variable de respuesta y las variables explicativas son todas ellas cuantitativas.(Márquez, 2013) Si sólo disponemos de una variable explicativa hablamos de regresión simple, mientras que si disponemos de varias variables explicativas se trata de un problema de regresión múltiple.(Gómez, 2010)

1.2.5 Regresión Lineal Múltiple

Las metodologías usadas para ver qué parámetros se relacionan mejor con la actividad biológica son muy diversas. La más popular es el análisis de RLM.(Salazar, 2016)

El Análisis de Regresión Lineal Múltiple (RLM) nos permite establecer la relación que se produce entre una variable dependiente (criterio, explicada, respuesta) Y y un conjunto de variables independientes (predictoras o explicativas) (X1, X2,... XK).(López, 2008) El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que, directa e indirectamente, participan en su concreción.

La anotación matemática del modelo o ecuación de regresión lineal múltiple es la que sigue:

$$Y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n + e \quad (1.5)$$

en donde:

Y es la variable a predecir;

$a, b_1x_1, b_2x_2 \dots b_nx_n$, son parámetros desconocidos a estimar;

y e es el error que cometemos en la predicción de los parámetros.(Gravano, 2014)

Algoritmo Genético (AG) como método de selección de parámetros:

Según D. Goldberg los AGs son: *Algoritmos de búsqueda basados en los mecanismos de selección natural y genética natural. Combinan la supervivencia de los más compatibles entre las estructuras de cadenas, con una estructura de información ya aleatorizada, intercambiada para construir un algoritmo de búsqueda con algunas de las capacidades de innovación de la búsqueda humana.*(Goldberg, 1989)

En la naturaleza todo el proceso de evolución biológica se hace de forma natural pero para aplicar el AG al campo de la resolución de problemas habrá que seguir una serie de pasos. Como premisa se debe conseguir que el tamaño de la población sea lo suficientemente grande para garantizar la diversidad de soluciones. Se aconseja que la población sea generada de forma aleatoria para obtener dicha diversidad.

Los pasos básicos de un algoritmo genético son:

1. Evaluar la puntuación de cada uno de los cromosomas generados.
2. Permitir la reproducción de los cromosomas siendo los más aptos los que tengan más probabilidad de reproducirse.
3. Con cierta probabilidad de mutación, mutar un gen del nuevo individuo generado.
4. Organizar la nueva población.

Estos pasos se repetirán hasta que se dé una condición de terminación. Se puede fijar un número máximo de iteraciones antes de finalizar el algoritmo genético o detenerlo cuando no se produzcan más cambios en la población (convergencia del algoritmo). Esta última opción suele ser la más habitual.(Pareja, 2016)

Selección de variables.

En la ecuación de regresión se pueden incluir tantas variables predictoras como se desee. Si el número de predictores es muy elevado, la cantidad de parámetros a determinar también lo será. Por otro lado, si muy pocos predictores se incluyen en el modelo la función de regresión corre el riesgo de generar una pobre explicación de los datos. La noción de que características hacen que una variable sea importante no es aún muy claro, pero una

interpretación es que una variable se vuelve importante si su exclusión del modelo afecta seriamente la precisión de las predicciones.(García, 2016)

Una de las principales dificultades a la hora de ajustar un modelo de regresión múltiple surge cuando es necesario identificar entre el conjunto de variables disponibles aquellas que están relacionadas con la respuesta y que la predicen de la mejor forma posible. Una primera alternativa es construir un modelo por inclusión o hacia delante ("forward"), considerando en primer lugar la relación de cada variable con la respuesta e ignorando todas las demás variables, valorándola por medio del coeficiente de correlación lineal de Pearson. Aquella que muestra una correlación más alta con la variable dependiente se introduce en un modelo inicial.(Izenman, 2008)

El segundo paso consiste en seleccionar entre las variables restantes aquella que al introducirla en el modelo permite explicar una mayor parte de la variabilidad residual. La comparación entre distintos modelos debe hacerse en términos del valor relativo de los coeficientes de determinación.(Díaz, 2001)

Cuando el objetivo es seleccionar aquellas variables que sin ser redundantes proporcionen la mejor explicación de la variable dependiente existen métodos secuenciales en SPSS: hacia atrás, hacia delante, pasos sucesivos.(Miranda, 2012)

Al combinar la regresión lineal múltiple con la técnica de selección de variables se obtienen una serie de modelos con distinta cantidad de variables predictoras. Se determina la cantidad óptima de variables a retener en el modelo a través de los errores que los distintos modelos generan. El modelo que presente el menor error será designado como óptimo.(Talento, 2011)

Los procedimientos de selección de variables por pasos son usualmente criticados porque no existen garantías de que los modelos seleccionados con uno u otro tipo de esquema (backward o forward) conduzcan al mismo conjunto de variables, ni al “mejor conjunto de variables” en algún sentido. Mas en general, los métodos de selección de variables son criticados porque utilizan los propios datos para agregar o eliminar variables y, por lo tanto, cambian el modelo que se asumía tenía variables predictoras determinadas a priori. Esto podría implicar que si los datos cambian levemente, las variables seleccionadas podrían también cambiar, haciendo a estos procedimientos muy inestables.(Izenman, 2008)

1.2.6 Calidad Estadística de un Modelo

El análisis de regresión múltiple, como se planteó anteriormente, es el más utilizado dentro del QSAR tradicional. En él, una vez de establecidos los conjuntos de valores de las variables independientes X_i y la actividad biológica A , se obtiene un modelo en forma de ecuación de una recta (ecuación 1.2), la cual describe la dependencia de la actividad (A) en función del conjunto de descriptores (X_i) así como la magnitud de las contribuciones de cada uno de ellos.

$$A = f(X_i) + cte \quad (1.6)$$

Para poder evaluar la capacidad de predicción del modelo ha de ser analizado en función de su calidad estadística y cuanto mayor sea la misma, más confiables y exactas serán las predicciones a realizar.(Vallejo, 2012)

La calidad estadística de un modelo se evalúa por diferentes estadígrafos. Los más comúnmente aceptados son el coeficiente de regresión r (el método de mínimos cuadrados se usa de manera típica para estimar los coeficientes de regresión de un modelo de regresión lineal múltiple)(Neria, 2010), el valor F de Fischer y la desviación estándar s . Cuanto más tienda a la unidad el valor de R mejor ajuste tendrán los datos al modelo. El valor F correlaciona la varianza explicada (R^2) por el número de grados de libertad, con la varianza no explicada ($1 - R^2$) por el número de variables del modelo. Cuanto más alto es el porcentaje de varianza explicada por el modelo mayor será el valor de F , mientras que la existencia de variables con baja contribución a la explicación de la varianza, tenderán a disminuir dicho valor, que tiende a infinito en los mejores modelos. La desviación estándar s depende de la varianza no explicada y de los grados de libertad del modelo y es una medida de cuánto se alejan los valores predichos por el modelo de la línea, plano o hiperplano. La tendencia a cero de este valor pudiera presuponer mayor calidad en la predicción. Sin embargo, esto puede conducir a modelos sobrepredictivos en los cuales se refleje exactamente el comportamiento de la muestra pero que no sea posible utilizarlo para la extrapolación de valores en el caso de la predicción de nuevos compuestos. Un valor de s es válido cuando está en el mismo orden de magnitud del error experimental de las mediciones de la variable dependiente.(Escalona, 2012)

Con el objetivo de lograr agrupar el criterio de varios autores se tomó como referencia artículos de revisión muy citados que plantean:

R^2 : Es el parámetro utilizado por lo general para estimar el ajuste del modelo al comportamiento estudiado (en este caso RLM). Sus valores, se plantea, deben estar lo más cercanos posibles a 1.0, pero no debe considerarse a R^2 como parámetro único, debido a que existen muchas posibilidades de sobreajustes arbitrarios.(Eriksson, 2003) Son aceptados aquellos resultados en los cuales se exhiban valores de R^2 por encima de 0.6.(Tropsha, 2010)

Q^2 : Expresa el poder predictivo del modelo. Es imposible obtener altos valores de Q^2 sin haber obtenido altos valores de R^2 . Generalmente se acepta un $Q^2 > 0.5$ y $Q^2 > 0.9$ se considera excelente (pero estos mínimos dependen de la aplicación del estudio).(Tropsha, 2010)

$R^2 - Q^2$: La diferencia entre ellos no debe exceder a 0.3. Valores mayores pueden indicar: un sobreajuste del modelo, la presencia de variables irrelevantes o de *outliers* en la Data.(Eriksson, 2013)

1.2.7 Datos Atípicos u Outliers

Se denominan casos atípicos u outliers a aquellas observaciones con características diferentes de las demás.(Rubin, 1987) Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar.(Peinado, 2011)

Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.(Hair, 1999)

Tipos de outliers

Los casos atípicos pueden clasificarse en 4 categorías:

1. Casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberán subsanarse en el filtrado de los datos, y si no se puede, deberán eliminarse del análisis o recodificarse como datos ausentes.

2. Observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.
3. Observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberán ser retenidas en el análisis pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.
4. Datos extraordinarios para los que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista deberá reportarlo en sus conclusiones.(Rial, 2001)

2 MATERIALES Y MÉTODOS

“La ciencia es la expresión de una necesidad inherente al ser humano y, en todo caso, está ligada a la función superior de su naturaleza inteligente: la capacidad de crear.”

René Gerónimo Favaloro

En el presente capítulo se detallan los materiales, es decir, las herramientas o recursos computacionales utilizados para llevar a cabo la investigación bioinformática; así como los métodos o procedimientos que permitieron: la generación de la Data, el cálculo de descriptores moleculares y la búsqueda de significación estadística utilizando RLM.

2.1 Base de Dato Química

Preparación de la Base de Datos:

La base de datos que se utiliza en esta tesis es Malaria Box, la misma fue creada a partir de una recopilación de información brindada por tres bases de datos diferentes (St Jude's, Novartis y GSK), las cuales fueron sometidas a varios análisis que se exponen a continuación. La actividad antimalárica es reportada en EC50 (la concentración que aniquila el 50 % de la población del parásito) expresada en nM (10^{-9} mol/L).

Previo a cualquier análisis, los compuestos en cada base de datos fueron procesados para 1- devastar sales, 2- quitar pequeños fragmentos, 3- desprotonar bases/ protonar ácidos, 4- generar tautómeros canónicos, y 5- quitar duplicados. Esta preparación de la base de datos fue realizada usando Pipeline Pilot 8.5. En este punto cualquier compuesto con peso molecular >1000 o con más de 20 enlaces rotativos no fue tomado en consideración. Siguiendo estos pasos, la base de datos St Jude's incluyó 1523 estructuras únicas; la base de datos Novartis incluyó 5661 estructuras únicas; y la base de datos GSK incluyó 13257 estructuras únicas.

Las representaciones canónicas (SMILES) de moléculas en las bases de datos St Jude's, GSK y Novartis fueron comparados para identificar duplicados de estructuras entre bases de datos, quedando 1315 estructuras únicas para la base de datos St Jude's; 5173 estructuras únicas para la base de datos Novartis; y 12867 estructuras únicas para la base de datos GSK. La St Jude's y la Novartis comparten 158 estructuras, en común; La Novartis y GSK comparten 340 estructuras en común; y la GSK y St Jude's comparten 77 estructuras en común. Hay en total 27 estructuras comunes para las tres bases de datos y 19876 compuestos estructuralmente únicos presentados en las mismas. Estos compuestos

estructuralmente únicos constituyen la fuente compuesta para la selección de un subconjunto representativo de antimaláricos. (i.e., Malaria Box). (Spangenberg, 2013)

Análisis de la base de datos:

Cada base de datos St Jude's, Novartis y GSK fue perfilada con relación al peso molecular, el número donantes de hidrógeno, A LogP y N+O (la cuenta de nitrógeno y la cuenta de oxígeno). Éstas son las cuatro propiedades fisicoquímicas que se usaron para perfilar moléculas con relación a la probabilidad de convertirse en fármacos orales exitosos. (Spangenberg, 2013)

La diversidad química de la base de datos:

Dadas las limitaciones en el número de compuestos que pueden ser puestos a prueba en detalle, para maximizar el potencial impacto de Malaria Box, fue importante maximizar la diversidad estructural en los compuestos seleccionados. Un Análisis de Componentes Principales bidimensional (2D-PCA, por sus siglas en inglés) fue usado para evaluar la diversidad química. Las tres colecciones de estructuras seleccionadas ocupan espacio similar al de la propiedad, y la base GSK muestra la máxima diversidad, aunque ésta sea probablemente una consecuencia de ser la base de datos más grande. Para evitar el costo de re-síntesis y para hacer los compuestos rápidamente disponibles para el seguimiento experimental, el siguiente paso consistió en seleccionar compuestos comercialmente disponibles. Alrededor de una cuarta parte (5034) de los seleccionados fueron accesibles por vendedores on-line (ambulantes en línea directa). Sometiendo estos 5034 compuestos al mismo análisis de componentes principales, dando como resultado un buen solapamiento con la base de datos inicial. Esto sugiere que los compuestos comercialmente disponibles son descriptivos de la diversidad del conjunto determinado. Finalmente 200 compuestos como drogas y 200 como sondas, que cubren la diversidad química del conjunto comercial, fueron seleccionados para la Malaria Box. (Spangenberg, 2013)

2.2 Herramientas Computacionales

Programas Informáticos desarrollados en el CAMD-BIR

2.2.1 Software DIVATI. Cálculo de los nuevos Índices de Derivada del Grafo

La definición de los nuevos DMs tiene como propósito final su uso como una herramienta para el diseño molecular. Con el fin de facilitar, desde un punto de vista computacional, el cálculo de los DMs se desarrolló una aplicación visual, escrita en el lenguaje de

programación JAVA denominado DIVATI (acrónimo de DIscrete DeriVAtive Type Indices), un nuevo módulo del programa TOMOCOMDCARDD 1.0 (acrónimo de Topological Molecular Computational Design Computed-Aided ‘Rational’ Drug Design). Este programa carga las estructuras en ficheros de formato .MDL, .mol y .sdf. y permite salvar los descriptores calculados en ficheros de tipo .txt, .xls, .xlsx y .csv.

Este *software* facilita el cálculo de todas las familias de Nuevos índices basados en la derivada discreta de un grafo: derivadas, derivadas de orden superior y derivadas mixtas, sobre pares (duplas), tríos (ternas) y cuartetos (cuaternas) de átomos en cualquier tipo de estructura química orgánica. Para la diferenciación atómica se dispone de una serie de 21 ponderaciones (etiquetas) atómicas, agrupadas en tres grandes grupos: Químicas [número atómico (Z), volumen de Van der Waals (W), polarizabilidad (P), masa atómica (A), radio covalente (R), electronegatividad de Pauling (E)], Físicas [área superficial polar total (T), contribución atómica al logP (G), contribución atómica a la refractividad molar (M), carga (C)] y Topológicas basadas en diferentes criterios de Grados del Vértice [grado de valencia (N), conectividad excéntrica (Y), estado electrotopológico (S), grado del vértice de Kupchik (K), estado intrínseco (I), grado del vértice de enlace (B), grado del vértice de Li (D), grado del vértice de Hu-Xu (H), grado del vértice de Alikhanidi (L), grado del vértice de Ivanciuc (V), conteo de distancias (O)]. El programa permite además realizar cálculos locales atómicos individuales y para grupos de átomos. Para los cálculos locales de grupos de átomos se aplican operadores matemáticos. Si al evaluar estos operadores se tienen en cuenta todos los átomos de la molécula el resultado será un cálculo global (total) de la molécula. En el *software* DIVATI estos operadores han sido denominados colectivamente como invariantes y agrupados en normas, medias, estadísticos y algoritmos clásicos y constituyen una generalización del uso de la combinación lineal de las contribuciones atómicas para la descripción global de la molécula.(Santiago, 2015b). En la Figura 1 se muestra la interfaz gráfica del *software*.

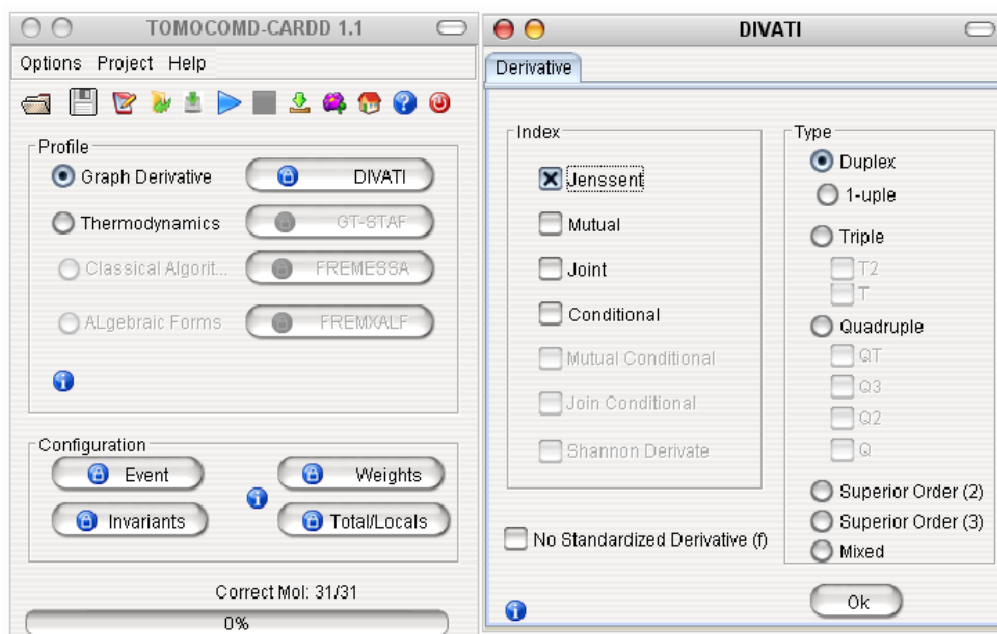


Figura 1. Interfaz gráfica del programa TOMOCOMD-CARDD y su módulo DIVATI.

2.2.2 Software IMMAN. Análisis de Variabilidad (AV)

Con el motivo de realizar el análisis de variabilidad de los parámetros moleculares propuestos en el presente trabajo, se ha implementado una herramienta computacional fundamentada en los conceptos de la teoría de información la cual se denomina IMMAN (acrónimo de Information Theory based CheMoMetric ANalysis). Esta aplicación permite el cálculo de la SE a los DMs, además de otros parámetros derivados de transformaciones matemáticas de la SE, tales como: la entropía de Shannon estandarizada (sSE), la Negentropía (nSE), el índice de redundancia de Brillouin (rSE), el índice de Gini (gSE) y el contenido de la energía de información (iSE). Otros parámetros basados en la teoría de información implementados en este programa incluyen la entropía diferencial generalizada (DSE) y la información de Jeffrey (JI). (Barigye, 2013) Los formatos de ficheros de entrada para el IMMAN incluyen TXT (tabbed ASCII files) y CSV (comma-separated value files). (Santiago, 2015b) En la Figura 2 se muestra la interfaz gráfica del software (aplicación visual).

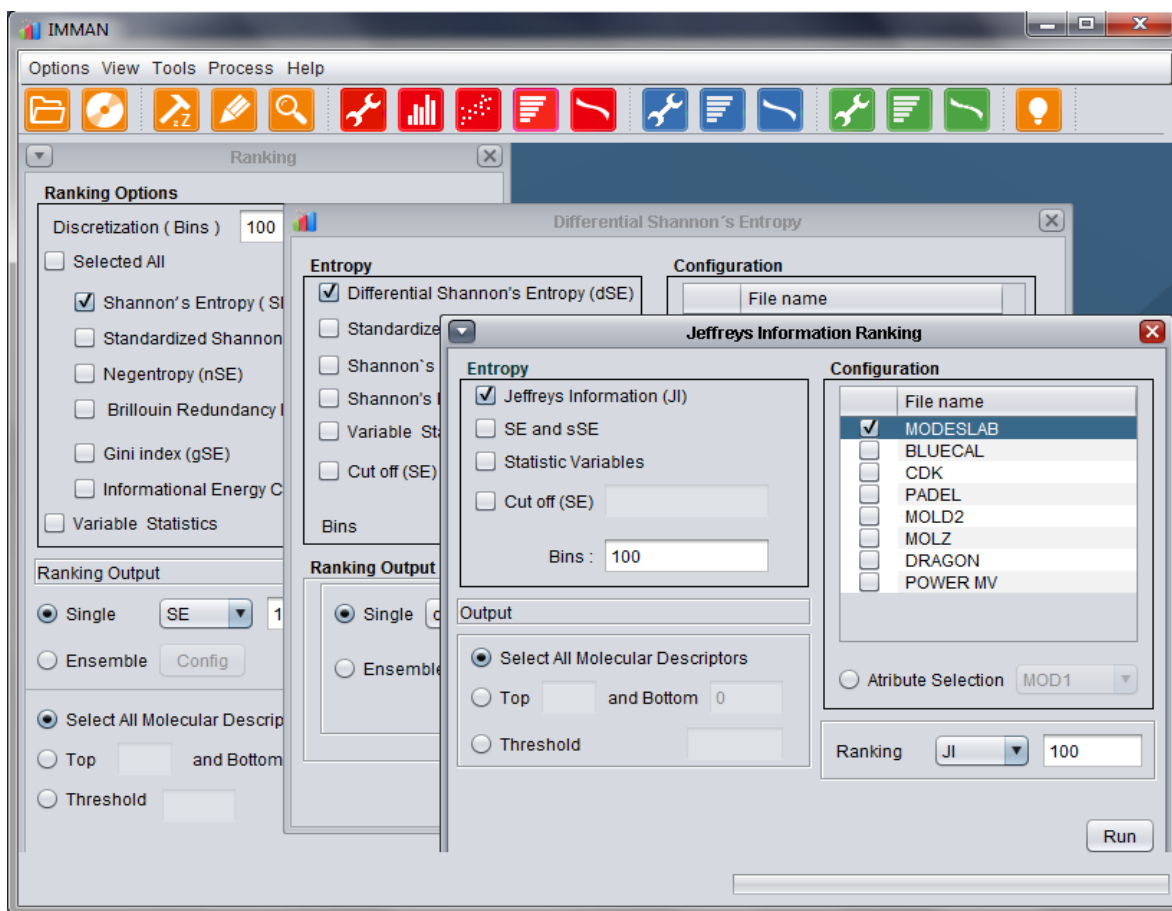


Figura 2. Interfaz gráfica del programa IMMAN.

2.2.3 Software Moby Digs. Regresión Lineal Múltiple

Los modelos QSAR (acrónimo de Quantitative Structure-Activity Relationship) se obtuvieron con el programa MOBYDIGS (versión 1.0 – 2004).¹⁷⁷ Una de las dificultades de la modelación QSAR aparece cuando existe un gran número de DMs y es necesario decidir cuáles son los que propician las mejores regresiones, considerando tanto el valor del ajuste como el sentido químico de la regresión. Además, puesto que la evaluación de un número elevado de todas las posibles combinaciones de las variables constituye una tarea tediosa y un procedimiento de elevado costo computacional, se emplea un algoritmo genético (GA) como método de selección de variables (optimización de los subconjuntos de variables). Los GA son una clase de algoritmos inspirados en el proceso de evolución natural en la que las especies de elevada adaptación bajo algunas condiciones pueden prevalecer y sobrevivir a la próxima generación; las mejores especies pueden adaptarse por cruzamiento y/ o mutación en búsqueda de mejores individuos. (Barigye, 2013) Este

programa utiliza un Algoritmo Genético como método de selección de parámetros, lo que le permite evaluar un número elevado de variables. Para cada estudio QSPR/QSAR, se determinó el tamaño (grado de libertad) deseado para los modelos a generar. El tamaño de la población para todos los estudios fue fijado a 100. El AG con un tamaño poblacional inicial de 100 rápidamente converge (200 generaciones) y alcanza un modelo QSAR en un número razonable de generaciones y por tanto un tiempo prudencialmente aceptable. Se optimizaron los modelos usando como función objetivo (función de optimización) del algoritmo genético el estadístico q^2_{LOO} y se validaron empleando las técnicas de remuestreo [*bootstrapping* (q^2_{boot})] y revuelto [*Y-scrambling*: $a(R^2)$, $a(Q^2)$]. La selección del mejor modelo fue desarrollada en términos del mayor coeficiente de correlación al cuadrado [coeficiente de determinación, (R^2)] o de la F-test (razón de Fisher a nivel- p [$p(F)$]) y la(s) ecuación(es) de la desviación estándar más baja. Se analizaron los parámetros estadísticos q^2_{LOO} (validación cruzada “dejando uno afuera”) y el q^2_{boot} para evaluar la calidad de los modelos. Muchos autores consideran elevados valores de Q^2 (por ejemplo, $Q^2 > 0.5$) como un indicador de elevado poder predictivo de un modelo QSAR. Sin embargo, es conocido que esta afirmación es solo cierta para datas pequeñas (< de 100 casos), y que en datas de gran dimensionalidad solo es una condición necesaria pero no suficiente para afirmar que un modelo posee un adecuado poder predictivo. (A. Golbraikh, 2002) En la Figura 3 se muestra la interfaz gráfica del *software*.



Figura 3. Interfaz Gráfica del Software Moby Digs 1.0.

3 RESULTADOS Y DISCUSIÓN

*“El verdadero logro en los descubrimientos...
es ver una analogía donde nadie jamás la ha visto”*

Arthur Koestler

En este capítulo se explican los principales resultados del trabajo realizado con el fin de exponer los aportes más relevantes durante el proceso de experimentación, que se desarrollará para la confección de esta tesis.

3.1 Modelación de actividad antimalárica de sustancias orgánicas

3.1.1 Cálculos de los Índices de Derivada del Grafo

En el programa interactivo DIVATI, nuevo módulo del programa TOMOCOMD-CARDD fueron desarrollados los cálculos de los Índices de Derivada del Grafo (GDI), donde están implementadas todas las familias de índices basadas en el concepto de Derivada Discreta de un Grafo. Los cálculos se desarrollaron inspirados en la entropía de Jentsen, Conditional, Joint y Mutual para pares, ternas y cuaternas de átomos respecto a 6 eventos diferentes que generaron modelos diversos para describir la estructura molecular. Los átomos en cada molécula fueron diferenciados usando diversas ponderaciones químicas, físicas y topológicas basadas en grados del vértice como estado intrínseco (I), electronegatividad de Pauling (E), polarizabilidad (P), volumen de Vander Waals (W), área de superficie total polar (T), contribución atómica al log P (G), grado de valencia (N), grado del vértice de Ivanic (V), grado del vértice de Hu-Xu (H), grado del vértice de Li (D) y el estado electrotopológico (S). Se cuantificaron además, índices globales y locales sobre heteroátomos (HT), total (T), donantes de protones (DH), halógenos (HL), carbonos metilos (MC) e insaturaciones (IS). Para los cálculos se tuvo en cuenta todas las invariantes y se emplearon los sucesos como: subestructuras (B), subgrafos de Sachs (H), sub-grafos conexos (S), multiplicidad (M), AlogP (A), y cuantización (Q).

3.1.2 Selección de los mejores descriptores moleculares

La degeneración es la propiedad de un DM para evitar la obtención de valores idénticos para moléculas diferentes. Es posible entonces que, los descriptores no posean degeneración (N) o presenten algún tipo de degeneración que puede ser baja (L), intermedia (I) o alta (H). El grado de degeneración de un descriptor puede ser medido por la *Entropía*

de Shannon (ecuación 3.1).(Santiago, 2014) De esta forma se obtiene la entropía de cada variable $h(\text{DMs})$, siendo elevada para variables de alta variabilidad y mínima para las de poca variabilidad en la data. Es válido decir que la degeneración es un atributo no deseable para un DM usado en la caracterización de estructuras químicas.(Santiago, 2014)

$$H(X) = -K \sum_{i=1}^N p_i \log_2 p_i \quad (3.1)$$

Donde, X es un suceso, constituido por N elementos o símbolos discretos con probabilidades $p_i (i = 1 \dots N)$ y K es una constante positiva.

El contenido de información y la entropía de Shannon de los DMs han sido extensivamente estudiados por Godden y colaboradores.(Godden, 2001) Basándose en estas ideas, un análisis de variabilidad (AV) cuantifica el contenido de información y, por lo tanto, la variabilidad de los DMs, basado en el concepto de entropía de Shannon. Para esto se introduce un procedimiento de discretización empleando los histogramas de distribución de frecuencias. Sea $p(a)_i$ la probabilidad de que el caso a esté en intervalo i , para un número de intervalos N se construye una función de distribución de probabilidades, $P(A)$, a la cual se le aplica la ecuación 3.1, obteniéndose la entropía de cada variable (DMs). Por tanto, esta técnica permite evaluar la calidad de los DMs como entidades independientes y se ha utilizado en la literatura para comparar el desempeño de conjuntos de DMs implementados de diferentes paquetes computacionales, así como en estudios de diversidad molecular.(Godden, 2000)

Es de suma importancia encontrar los mejores DMs usando análisis de variabilidad debido a que como se emplea una data de estructuras diversas es de esperar que los DMs que la describan deben ser altamente variables. Con el objetivo de encontrar los DMs que mayor variabilidad poseían, los anteriormente calculados en el programa DIVATI fueron sometidos a un análisis con el software IMMAN, el cual reportó el valor de la entropía de Shannon (SE) y permitió seleccionar los mejores atributos.

3.2.1 Modelos de RLM

El criterio de selección de la RLM empleado en el presente trabajo fue, fundamentalmente, la simplicidad de la misma, la amplia descripción bibliográfica al respecto y la robustez de la técnica. Muchas propiedades biológicas han sido

satisfactoriamente modeladas con el empleo de la RLM; de hecho, es aceptado que representa, en el caso de las relaciones cuantitativas estructura-actividad; la más empleada de las técnicas de regresión, y por tanto se consideró adecuado comenzar la descripción de la actividad antimalárica usando RLM.

Como se detalló en el Capítulo anterior, fue utilizado para la obtención y significación de los modelos de RLM; el software MobyDigs, que posee como método de selección de variables el Algoritmo Genético. Se predefinió que solo los 10 modelos que produjeran la capacidad predictiva más alta fueran retenidos finalmente y analizados en lo adelante.

De las 400 moléculas de la base de datos Malaria Box realmente se realizó la modelación sobre 317 ya que las 83 restantes no tenían definido el valor de la actividad antimalárica.(Spangenberg, 2013) De los modelos de RLM encontrados para los descriptores moleculares calculados se hizo una selección de las mejores variables.

De la combinación de las mejores variables de las cuatro derivadas trabajadas (conditional, joint, mutual y jenssent) se obtuvo tres primeros modelos para describir la actividad antimalárica (expresada como EC50_nM) de las moléculas de la base de datos Malaria Box. La ecuación del mejor modelo de regresión, los parámetros estadísticos y los correspondientes gráficos de regresión y predicción se muestran a continuación:

Ecuación del mejor modelo de regresión:

$$\begin{aligned}
 EC_{50} = & 459.36 (\pm 821.41) + 10.55 (\pm 2.86) \left[{}^{N/In}GI_1(MX) \right]_C^{C-T_2/f} - 265.66 (\pm 91.48) \\
 & \left[{}^{E/In}AC_4(Q_2) \right]_S^{J-QT/f} - 315.57 (\pm 120.96) \left[{}^{T/In}CN_5(M) \right]_M^{J-D/f} - 121267.18 \\
 & (\pm 53046.97) \left[{}^{V_3/In}AC_1(K) \right]_M^{J-D/f} - 38.00 (\pm 14.16) \left[{}^{V_3/In}CN_5(Q_2) \right]_Q^{J-D/f} + 5060.07 (\pm 1121.21) \\
 & \left[{}^{E/In}GI_7(Q_2) \right]_S^{J-D/f} - 5080.17 (\pm 1118.23) \left[{}^{T/In}GI_3(MN) \right]_S^{J-D/f} + 97.15 (\pm 40.75) \\
 & \left[{}^{E/In}IB(N_2) \right]_Q^{J-T_2/f} - 161.76 (\pm 43.09) \left[{}^{P/In}AC_4(MX) \right]_Q^{J-T_2/f} + 64.42 (\pm 18.60) \\
 & \left[{}^{E/In}CN_3(A) \right]_A^{M-QT/f} + 1872.83 (\pm 763.84) \left[{}^{E/In}CN_2(P_3) \right]_A^{M-D/f} - 759.00 (\pm 304.56) \\
 & \left[{}^{P/In}GI_2(N_2) \right]_Q^{M-D/f} - 0.69 (\pm 0.19) \left[{}^{D/In}TS_4(A) \right]_S^{M-D/f} + 44.37 (\pm 48.63) \\
 & \left[{}^{D/In}CN_5(K) \right]_Q^{M-T_2/f} + 1.08 (\pm 0.02) \left[{}^{C/In}AC_5(CV) \right]_S^{JE-D/f}
 \end{aligned} \tag{3.2}$$

Se aprecia en el modelo la presencia de índices totales solamente, lo cual es lógico, pues la propiedad que se modela depende de las características espaciales de la estructura molecular en todo su conjunto.

Tabla 1. Parámetros estadísticos del modelo 3.2:

N	R ²	R	Q ²	Q ² _{boot}	F	Scv	y-sc	s
317.000	38.030	0.616	30.720	27.430	11.130	525.776	0.042	511.711

Donde, N es el número de compuestos, R² es el coeficiente de determinación, R es el coeficiente de correlación, s es la desviación estándar de la regresión, Q² es el coeficiente de determinación obtenido a partir del método de validación cruzada (LOO), y-sc es la prueba de aleatoriedad (y-scrambling), Q²_{boot} es el coeficiente de determinación de la validación por Bootstrap, Scv es la desviación estándar de la validación cruzada y F es el radio de Fisher.

Como puede verse, el mejor modelo de RLM obtenido explica pobremente la variable dependiente; ya que el valor de ajuste (R²) se encuentra por debajo de 0.5, al igual (como es de esperar) que el valor de Q². La validación interna evaluada a través de los métodos de Validación Cruzada, *Bootstrapping*, y-aleatorización brindaron, igualmente, resultados desfavorables.

Una de las razones que pueden justificar los bajos valores de robustez, bondad de ajuste y predictibilidad del modelo obtenido por la técnica de RLM es la cantidad y diversidad de compuestos empleados en la conformación del modelo, difícil de ajustar a un comportamiento lineal.

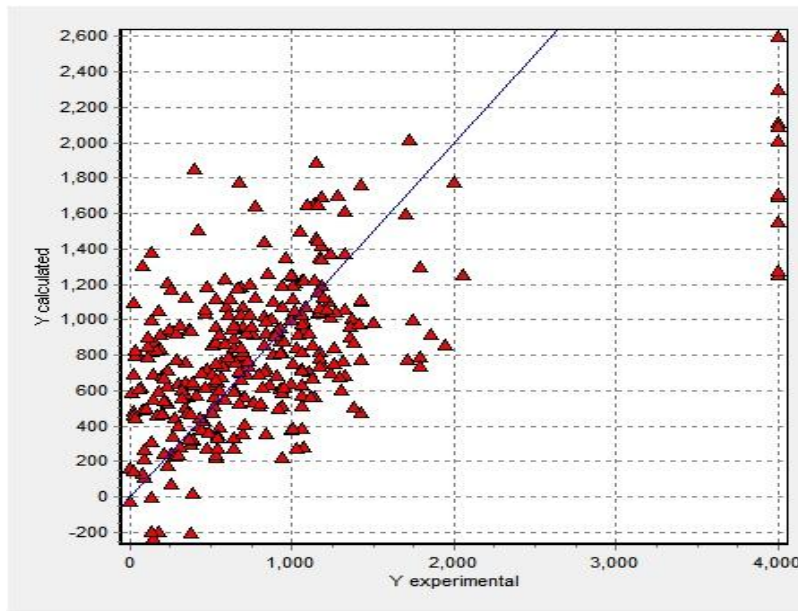


Figura 4. Gráfico de Regresión para el mejor modelo obtenido (Ec. 3.2).

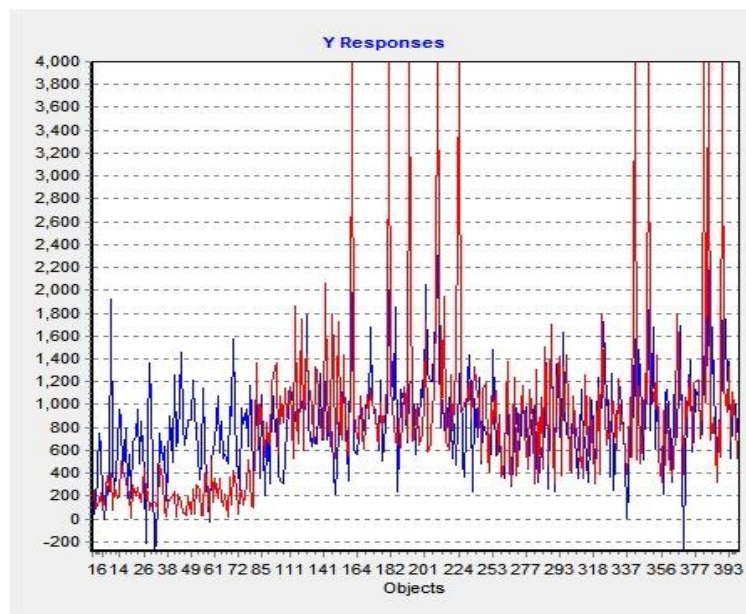


Figura 5. Gráfico de predicción para el mejor modelo obtenido (Ec. 3.2).

Al realizar un análisis de los parámetros estadísticos del modelo puede observarse que no posee una buena capacidad de ajuste a los datos experimentales. Analizando además los gráficos de regresión y predicción (Figuras 4 y 5) se puede apreciar que existían compuestos con un comportamiento completamente anómalo, por lo cual resultó necesaria su exclusión inmediata. Según el criterio de descartar solamente aquellos compuestos que se alejaran en gran medida del dominio de aplicación (DA), y que por tanto exhibieran un

comportamiento atípico muy marcado o un elevado *leverage*; se seleccionaron un total de 60 compuestos para su posterior exclusión.

3.2.2 Identificación de outliers

Un paso crucial en la construcción de modelos, resulta la detección de compuestos atípicos (“*outliers*”), que se definen como puntos que no se ajustan o son pobremente predichos; afectando así los parámetros estadísticos. Es decir, la identificación de los *outliers* busca un mejoramiento cuantitativo del modelo, y aunque no es necesario justificar la extracción de estos puntos, se recomienda determinar la razón para su peculiaridad en aquellos casos en que sea posible. (Verma, 2005) Existen varias técnicas para detectar la presencia de *outliers*, tales como: los análisis de los residuales estandarizados, los residuales studentizados, el método de *Leverage*, la estadística DFITS, la distancia de Cook y el método de “dejar varios fuera”.

En este trabajo se eliminaron 60 outliers en busca de un mejoramiento cuantitativo del modelo, para esto se utilizó el método de “análisis de los residuales estandarizados”.

Se muestran a continuación, para el mejor modelo de RLM obtenido luego de la exclusión de los 60 *outliers* y seleccionando solo las 14 mejores variables (DMs), con el objetivo de describir la actividad antimalárica expresada como EC50_nM de las moléculas de la base de datos Malaria Box, la ecuación, los parámetros estadísticos y sus correspondientes gráficos de regresión y predicción:

Ecuación del mejor modelo de regresión sin los outliers:

$$\begin{aligned}
 EC_{50} = & 438.24 (\pm 387.45) + 42.21 (\pm 18.81) [^{T/In}IB \ (PN)]_Q^{C-D/f} + 5.52 (\pm 1.53) \\
 & [^{N/In}GI_1(MX)]_S^{C-T_2/f} - 109.46 (\pm 17.08) [^{W/In}ACT \ (N_1)]_Q^{C-T_2/f} + 57.32 \\
 & (\pm 26.70) [^{E/In}CN_3(PN)]_A^{J-QT/f} - 917.22 (\pm 243.57) [^{E/In}CN_4(P_2)]_B^{J-QT/f} - 23.29 (\pm 7.63) \\
 & [^{V_3/In}CN_5(Q_2)]_Q^{J-D/f} + 58.25 (\pm 11.06) [^{E/In}CN_3(A)]_A^{M-QT/f} + 213.24 (\pm 403.30) \\
 & [^{E/In}CN_2(P_3)]_A^{M-D/f} + 55.91 (\pm 15.50) [^{V_3/In}TS_1(N_3)]_M^{M-D/f} + 184.85 (\pm 26.12) \\
 & [^{E/In}TS_3(N_1)]_M^{M-D/f} - 548.13 (\pm 153.82) [^{P/In}GI_2(N_2)]_Q^{M-D/f} - 169.69 (\pm 56.01) \\
 & [^{G/In}CN_7(P_2)]_S^{M-D/f} + 15.47 (\pm 2.96) [^{V_{10}/In}GI_3(MN)]_S^{J-E-D/f} - 19.38 (\pm 5.93) \\
 & [^{T/In}AC_5(MN)]_M^{J-E-D/f}
 \end{aligned} \tag{3.3}$$

Tabla 2. Parámetros estadísticos del modelo 3.3.

N	R ²	R	Q ²	Q ² _{boot}	F	Scv	y-sc	s
257.000	51.340	0.716	45.300	42.810	18.160	298.130	0.036	289.803

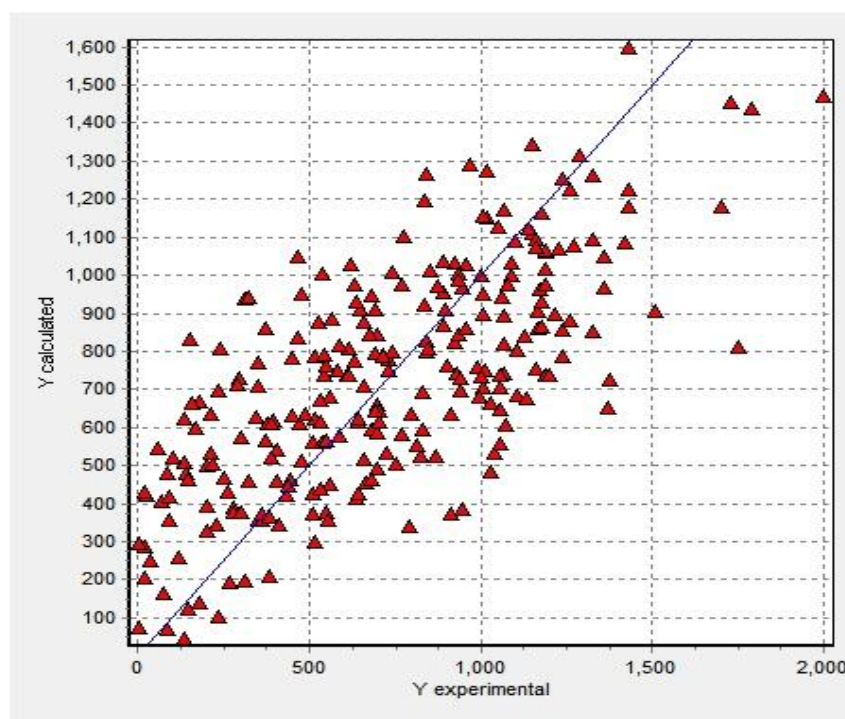


Figura 6. Gráfico de Regresión para el modelo obtenido (Ec. 3.3).

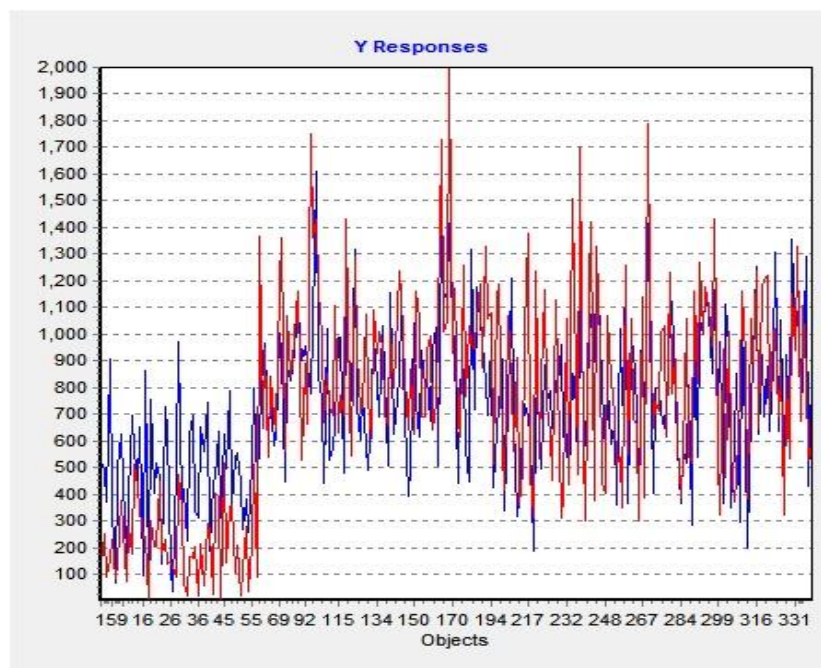


Figura 7. Gráfico de predicción para el modelo obtenido (Ec. 3.3).

Luego de la exclusión; se decidió no extraer ningún otro compuesto debido a que los parámetros estadísticos definitivamente no variarían, y se estaría perdiendo información valiosa al descartar casos solo por tratar de que se ajusten a un comportamiento lineal.

Como puede observarse en los gráficos anteriores (Figuras 6 y 7) y en los parámetro estadísticos el modelo, luego de eliminar los compuestos *outliers* mejora un tanto la capacidad de ajuste a los datos experimentales con respecto al primer modelo expuesto. El valor de R^2 es más cercano a uno, Q^2 presenta un valor aceptable, cercano a 0.5 y el valor de la diferencia: $R^2 - Q^2$ no excede a 0.3.

Este modelo indicó solamente 14 variables, por lo que se pudo descartar entonces una de las 15 con las que se había trabajado; resultado significativo debido a que, como plantea el *Principio de Parsimonia*, es conveniente explicar el comportamiento con la menor cantidad de variables posible (nunca excluyendo aquellas importantes solo por dejar un número menor).

Con respecto a la robustez del modelo, los valores del coeficiente de correlación difieren en un 10%, por lo que el modelo puede considerarse moderadamente robusto.

En la tabla 3 se compara la capacidad de ajuste a los datos experimentales entre los mejores modelos con y sin los compuestos outliers.

Tabla 3. Comparación entre los mejores modelos con y sin compuestos *outliers*

Modelo	R	R ²	s
Con todos los compuestos	0.616	38.030	511.711
Sin los 60 compuestos outliers	0.716	51.340	289.803

3.2.3 Discusión de los resultados de la regresión lineal múltiple:

Teniendo en cuenta los resultados de los gráficos de regresión y predicción y los parámetros estadísticos del primer modelo obtenido puede observarse que no posee una buena capacidad de ajuste a los datos experimentales. Esto puede deberse, entre otros factores, a que se está modelando una actividad biológica, modelación que puede resultar muy compleja ya que depende, además de la estructura, de factores no controlables por el experimentador, como por ejemplo, el efecto de la matriz biológica sobre la capacidad individual de acción de cada una de las moléculas, sin embargo se asume que la estructura química posee el protagonismo en cuanto a influenciar el valor de la propiedad.

Es posible que se dé el caso, además, de que la estructura de las moléculas no tenga una relación lineal con la propiedad evaluada, al menos cuando están descritas con los descriptores moleculares utilizados, influyendo así en la capacidad de ajuste y de predicción del mejor modelo obtenido.

Se podrían calcular otros DMs basados en derivadas del grafo y teoría de la información que sean ortogonales a los utilizados en este trabajo para abundar en un universo más vasto de los mismos que codifiquen la estructura de otra forma, esto podría mejorar los resultados de la RLM. Igualmente como la relación que existe entre la estructura de las moléculas y la propiedad de las mismas puede no ser lineal, se deben desarrollar técnicas no lineales para evaluar el comportamiento de las mismas.

El MobyDigs es un software que tiene implementado el AG como método de selección de variables, y fue utilizado en este trabajo para tales fines.

El método aleatorio de AG tiende a caer en máximos locales y por ende no evaluar otras combinaciones donde sería posible encontrar el máximo global, esto puede influir en los resultados de los modelos y consecuentemente ser otro factor que interfiera en la capacidad de ajuste y predicción del modelo. Se trató de atenuar esta situación en todo lo posible construyendo varias poblaciones y migrando cada determinado tiempo hacia la que

contuviera los mejores modelos, esto modificaba y enriquecía constantemente las poblaciones y se disminuyó así esta posibilidad.

La presencia de compuestos *outliers* también puede influenciar la capacidad de ajuste y de predicción del modelo, lo que ha sido demostrado en este trabajo al obtener un modelo cuantitativamente superior al eliminar estos compuestos atípicos.

A partir del análisis de los parámetros estadísticos del modelo obtenido luego de excluir los 60 compuestos *outliers* se observa que muestra un moderado poder predictivo ($q^2_{\text{loo}} \approx 0.5$), moderado valor del coeficiente de determinación ($R^2 > 0.5$), lo que demuestra que existe correlación con la propiedad modelada, es capaz de explicar más del 50 % de la varianza de la propiedad experimental. Además los valores de q^2_{boot} están cercanos al 50 %. Todos estos resultados dan una medida de la moderada robustez estadística del modelo de RLM obtenido. El error estándar resulta menor, lo que demuestra menos variación entre el valor experimental y el valor calculado que en el caso del modelo obtenido con los compuestos atípicos.

Se aprecia en la ecuación del modelo que todas las variables que penetran son totales, esto significa que debido a la alta variabilidad de la data es lógico que las descripciones no se enfoquen en aspectos específicos de algunas moléculas porque otras no tendrían esos aspectos sino otros, por eso las descripciones que mejor correlacionan son de tipo total. Además se recoge información diversa de las estructuras pues se enfoca la descripción con cuatro familias de GDI ortogonales entre sí (conditional, jenssen, joint y mutual).

A partir del análisis del gráfico de predicción se aprecia claramente la moderada capacidad predictiva del modelo, coincidiendo las predicciones para una parte considerable de las moléculas, (ambas curvas se solapan en varias estructuras, la roja representa los valores del EC50 para cada molécula y la azul representa los valores predichos por el modelo), el gráfico de regresión, muestra una moderada correlación entre la propiedad modelada y la predicha por el modelo, lo cual da una medida de la posibilidad de aplicación de estas en la predicción de propiedades.

Hasta el momento no se han reportado estudios de regresión para esta base de datos, que permitan la comparación del desempeño de los descriptores moleculares trabajados. Así mismo no pudo conocerse hasta qué punto es válido o no el conjunto de métodos utilizados con respecto a otros frente a las mismas condiciones.

CONCLUSIONES:

1. El resultado de la limpieza del conjunto de datos químicos conocido como Malaria Box permitió eliminar residuos moleculares, sales inorgánicas y eliminó los hidrógenos dejando las estructuras moleculares como grafos con hidrógenos suprimidos listos para el posterior cálculo de los GDI.
2. Se calcularon cuatro familias de GDI correspondientes a evaluar derivadas sobre pares, tríos y cuaternas de átomos lográndose en todos los casos una adecuada codificación de las estructuras químicas y los mejores índices fueron seleccionados basados en sus correspondientes valores de entropía de Shannon.
3. Las mejores regresiones lineales encontradas entre las estructuras y sus respectivas actividades antimaláricas, posterior a la separación de estructuras *outliers*, sugieren que entre dichos parámetros existe una relación lineal moderada.

RECOMENDACIONES:

- Calcular otras familias de GDI y otras familias de DMs con el objetivo de explorar todo lo posible el universo de nuevas codificaciones con información ortogonal entre sí.
- Usar otras técnicas de regresión usando máquinas de soporte vectorial que permitan obtener relaciones no lineales entre los índices y las actividades moleculares.

REFERENCIAS BIBLIOGRÁFICAS

1. A.Golbraikh. (2002). Mol Graph Model. 20, 269.
2. Aponte, J. L. (2012). ¿Qué sabe Ud. acerca de...QSAR? *Rev. mex. cienc. farm*, 43.
3. Barigye, S. J. (2013). *Teoría de Información en la Codificación de la Estructura Química*.
4. Barrio, L. S. R. d. (2014). Desarrollo de nuevos fármacos: desde la invención a la Farmacia.
5. Berganza, P. A. (2009-2010). La malaria. *Ambiociencias*.
6. Carrero, M. P. (2003). *Revolución Combinatoria Capítulo 1*.
7. Castaño, A. T. (2006). Clínica de la malaria complicada debida a *P. falciparum*. *Iatreia*, 19.
8. Castro, E. A. (2011). Diseño de Conjuntos moleculares balanceados para su aplicación en la teoría QSPR-QSAR
9. Catalina, R. (2012). Historia de la malaria.
10. Contreras, L. K. C. (2011). ¿Qué sabe Ud. acerca de Cribado Virtual? *Rev. mex. cienc. farm*, vol.42 no.2
11. Cronin. (2002). Structure-based classification of antibacterial activity. *J Chem Inf Comput Sci*, 869-878.
12. Díaz, S. P. (2001). Técnicas de regresión: Regresión Lineal Múltiple.
13. Ekins, S., Williams, A. J., Krasowski, M. D., & Freundlich, J. S. (2011). In silico repositioning of approved drugs for rare and neglected diseases. *Drug discovery today*, 16(7-8), 298-310.
14. Eriksson, L. (2003). Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs.
15. Eriksson, L. (2013). Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environmental Health Perspectives.*, 1361-1375.
16. Escalona, J. C. (2010). Introducción al diseño de Fármacos.
17. Escalona, J. C. (2012). Introducción al diseño de Fármacos.
18. Fonseca, J. C. (2015). Primaquina y recurrencias de malaria por *Plasmodium vivax*. Metanálisis de estudios clínicos controlados. *Rev Bras Epimideol*.
19. Gamenara, D. (2005). *Síntesis de potenciales antimaláricos a través de reacciones de Diels-Alder*. Universidad de la República Montevideo.
20. García, A. D. (2016). *Rendimiento académico en alumnado de educación primaria. Análisis de regresión lineal múltiple*.
21. Geneva. (2009). *Paludismo*.
22. Godden, J. (2000). Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *JChem InfComput Sci*, 40, 796–800.
23. Godden, J. (2001). Shannon entropy: a novel concept in molecular descriptor and diversity analysis.
24. Goldberg, D. (1989). *Genetics Algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

25. Gómez, A. J. C. (2010). *Minear de datos con Weka para la predicción del precio de automóviles de segunda mano* Universidad Politécnica de Valencia
26. Gracia, O. R. (2010). *Herramientas de cribado virtual aplicadas a inhibidores de tirosina quinasas. Contribución al desarrollo del programa Pralins para el diseño de quimiotecas combinatorias.*, Ramon Llull.
27. Gravano, A. (2014). *Aprendizaje Automático (Machine Learning)*.
28. Griffith, M. (2015). Conocimientos, actitudes y prácticas sobre la malaria en la población indígena guna de la comarca de Madungandí. *Biomédica*.
29. Hair, J. (1999). *Análisis Multivariante*. Prentice Hall.
30. Harary, F. (1971). *Graph Theory*. Addison-Wesley, Reading: MA.
31. Heydra, F. M. C. M. (2006). *Síntesis y evaluación preliminar de la posible actividad antimalárica y antineoplásica de derivados 7-cloroquinolina-4-sustituidos*. Unpublished Postgrado de Química de Medicamentos, Universidad Central de Venezuela.
32. Hoz, F. d. I. (2014). *Protocolo de Vigilancia en Salud Pública*. Instituto Nacional de Salud Pública.
33. Isea, R. (2010). Identificación de nuevos medicamentos a través de métodos computacionales. *Revista del Instituto Nacional de Higiene*.
34. Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*.
35. Jurs, P. C. (1995). Representations of molecules, in *Chemometrics Methods in Molecular Design*. New York, Vol. 2.
36. Kier, L. B. (1999). *Molecular Structure Description. The Electrotopological State: Academic Press: San Diego*.
37. Kizys, R. (2003). *Modelo de Regresión Lineal Múltiple*.
38. Laguna, F. (2011). *Predicción de incidencia de malaria mediante un modelo de retardo distribuido autorregresivo con componente espacial*.
39. Lazarou, M., Guevara Patino, J. A., Jennings, R. M., McIntosh, R. S., Shi, J., Howell, S., et al. (2009). Inhibition of erythrocyte invasion and Plasmodium falciparum Merozoite Surface Protein 1 processing by human IgG1 and IgG3 antibodies. *Infect Immun*.
40. López, F. J. B. (2008). *Regresión Múltiple. Capítulo 6*.
41. Márquez, F. J. B. (2013). *Regresión Lineal*.
42. Meléndez, J. L. C. (2014). *Determinación de metabolitos secundarios en tres pteridofitos, plantas con interés medicinal*
43. Miranda, M. D. M. (2012). *El Análisis de la Regresión a través de SPSS*, Universidad de Granada.
44. Monge, A. (2003). El Descubrimiento de Fármacos a partir de plantas medicinales. *Ciencia e Investigación*.
45. Nájera, J. A. (2009). *Malaria*
46. Neria, J. A. (2010). *Análisis de Incertidumbre para Resultados de Códigos Termohidráulicos de mejor estimación*. . Escuela Superior de Física y Matemática, México.
47. Pareja, R. T. (2016). *Nuevas ecuaciones en Diferencia Finita para codificar matemáticamente estructuras orgánicas*. Universidad Central de Las Villas, Santa Clara.

48. Pastor, M. (2003). Técnicas QSAR en el diseño de fármacos.
49. Peinado, F. M. O. n. (2011). Técnicas estadísticas en Nutrición y Salud. Tratamiento estadístico de outliers y datos faltantes.
50. Pérez, D. M. Á. C. (2011). *Métodos in-silico para la determinación de la permeabilidad y la solubilidad de fármacos. Aproximaciones al SCB.*
51. Peydró, R. B. M. y. R. J. (2008). Malaria en España: aspectos entomológicos y perspectivas de futuro. *Rev Esp Salud Pública* 82.
52. Pínel, A. B. (2011). Avances en Salud.
53. Piñeros, J. G. (2010). Malaria y determinantes sociales de la salud: un nuevo marco heurístico desde la medicina social latinoamericana. *Biomédica*.
54. Quiñones, M. L. (2015). Producción Científica de Investigadores del INS. *Bol - Inst Nac Salud*.
55. Rial, A. (2001). Depuración y Análisis Preliminares de Datos en SPSS. Sistemas Informatizados para la Investigación del Comportamiento. RA-MA.
56. Rodríguez, P. (2006). Evaluación de la actividad antimalárica de algunas plantas utilizadas en la medicina tradicional cubana. *Revista de Ciências Farmacêuticas Básica e Aplicada*.
57. Rojas, R. M. (2008). Historia de la malaria o el paludismo. *Medicinam*.
58. Rubin, D. (1987). Statistical Analysis with Missing Data.
59. Salazar, J. A. G. (2016). *Isoimida monosustituidas como inhibidores de la Acetilcolinesterasa*.
60. Salud, O. M. d. I. (2012). Ampliando el diagnóstico, tratamiento y vigilancia de la malaria.
61. Salud, O. M. d. I. (2015). Acción e Inversión para vencer a la Malaria 2016-2030: por un mundo libre de malaria.
62. Salud., O. P. d. I. (2010). Directrices de la OMS para el tratamiento de la malaria.
63. Santiago, O. M. (2014). Discrete Derivatives for Atom-Pairs as a Novel Graph-Theoretical Invariant for Generating New Molecular Descriptors: Orthogonality, Interpretation and QSARs/QSPRs on Benchmark Databases. *J Molecular Informatics*.
64. Santiago, O. M. (2015a). Índices de Derivada Discreta de Grafos Moleculares.
65. Santiago, O. M. (2015b). *Índices de Derivada Discreta de Grafos Moleculares. Descripción de Estructuras Químicas*. Unpublished Tesis presentada en opción al grado científico de Doctor en Ciencias Químicas, Universidad Central “Marta Abreu” de las Villas, Santa Clara.
66. Schultz, T. W. (2003). Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. *J Mol Struct* 1-22.
67. Spangenberg, T. (2013). The Open Access Malaria Box: A Drug Discovery Catalyst for Neglected Diseases.
68. Talento, S. (2011). *Bases para un sistema de predicción de caudales de aporte a Rincón del Bonete y Salto Grande*
69. Universidad de la República
70. Todeschini. (2009). Molecular Descriptors for Chemoinformatics. *wiley-VCH, Vol. I & II*.

71. Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation.
72. Tugores, Y. M. (2012). Descubrimiento de nuevos antimaláricos a partir de fármacos conocidos mediante cribado in silico e in vitro.
73. Tugores, Y. M. (2014). *Tamizaje farmacológico en la búsqueda de potenciales fármacos antimaláricos integrando nuevos modelos in silico y corroboración experimental*. Universidad Complutense de Madrid, Madrid.
74. Vallejo, P. D. (2012). Minería de Datos Aplicada
75. Vapnik, V. (1995). The Nature of Statistical Learning Theory. *Springer*.
76. Verma, R. (2005). An approach toward the problem of outliers in QSAR. *Bioorganic & Medicinal Chemistry*.
77. Waterbeemd, H. V. d. (1995). Chemometric Methods in Molecular Design (Methods and Principles in Medicinal Chemistry).
78. Zúñiga, A. J. P. (2002). Simulación del reconocimiento entre proteínas y moléculas orgánicas o docking. Aplicación al diseño de fármacos.

ANEXOS

Anexo1:



Figura A.1. Ciclo de vida de *Plasmodium*. El parásito se desarrolla tanto en el hospedero vertebrado como en el mosquito. La fase asexual se lleva a cabo en el hígado y en los eritrocitos, con la consecuente producción de los gametocitos; éstos son tomados por el mosquito *Anopheles*, en cuyo interior se desarrollan en esporozoítos. Los esporozoítos son transportados hacia las glándulas salivales del mosquito, de donde son inoculados al hospedero vertebrado, para dar inicio a un nuevo ciclo.

Anexo2:

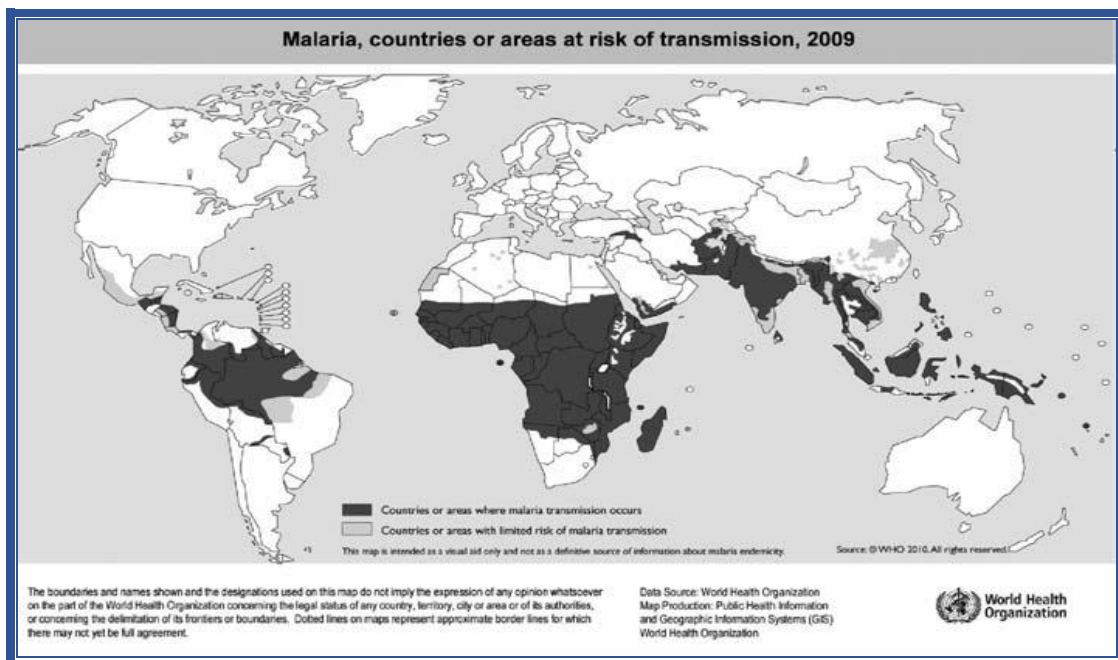


Figura A.2. Distribución mundial de la malaria. OMS, 2010.

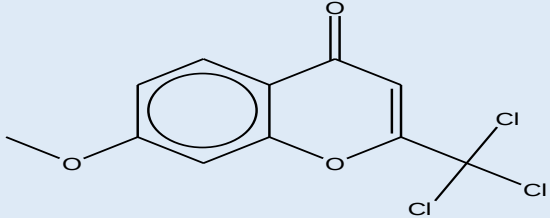
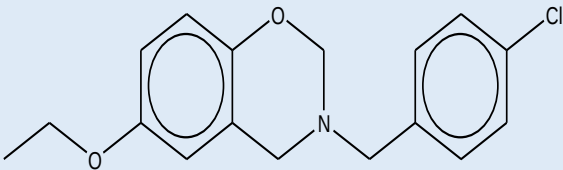
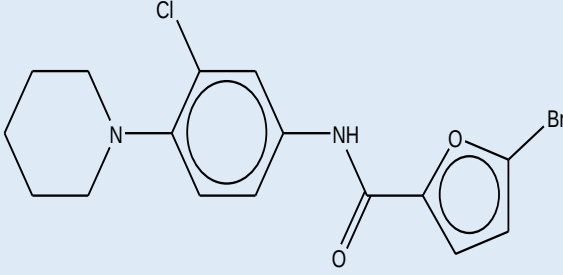
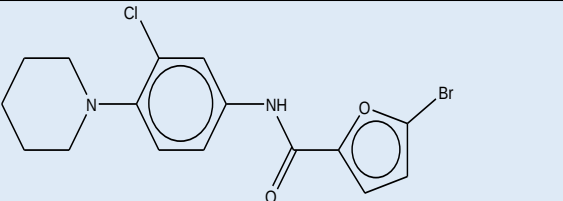
Anexo3:

Tabla A.3. Representación de 10 de los compuestos outliers:

Representación de los compuestos outliers		
Códigos	Smiles	Estructuras
MMV666023	<chem>C(c1ccccc1)n2c(N=Nc3c(Nc4ccccc4)ccc5ccccc35)nc6ccccc26</chem>	
MMV007384	<chem>COc1ccc(cc1)c2nc3cc(Cc4ccc5[nH]c(nc5c4)c6ccc(OC)cc6)ccc3[nH]2</chem>	
MMV007577	<chem>Cc1ccccc1OCCSc2nc3ccc(NC(=O)c4ccccc4)cc3s2</chem>	
MMV666079	<chem>Cc1ccc2cc3c(N)c4c(C)ccc(C)c4nc3nc2c1</chem>	
MMV008173	<chem>COc1cc(OC)cc(c1)N2C(=O)N(Cc3ccc(cc3)C(C)(C)C)c4ccccc4S2(=O)=O</chem>	
MMV007041	<chem>Cc1ccc2c(C)nc(Nc3nc(C)cc(n3)c4ccccc4)nc2c1</chem>	
MMV666687	<chem>COc1cc(cc(OC)c1OC)C(=O)ON=C(N)Cc2cccc3ccccc23</chem>	
MMV000839	<chem>OC(CNCc1ccccc1)Cn2c3ccccc3c4ccccc24</chem>	
MMV396705	<chem>c12c(C(=O)N(C)C(=O)N1C)n(Cc3cccc(OC)c3)c(Oc4ccccc4)n2</chem>	
MMV000963	<chem>n1c(NCc2ccccc2)c3c(cccc3)nc1Nc(cc4)ccc4OC</chem>	

Anexo4:

Tabla A.4. Representación de los 5 compuestos de menores residuales:

Compuestos de menores residuales		
Series	Smiles	Estructuras
MMV665820	<chem>COc1ccc2C(=O)C=C(Oc2c1)C(Cl)(Cl)Cl</chem>	
MMV000443	<chem>CCOc1ccc2OCN(Cc3ccc(Cl)cc3)Cc2c1</chem>	
MMV019700	<chem>Clc1cc(NC(=O)c2oc(Br)cc2)ccc1N3CCCCC3</chem>	
MMV019700	<chem>Clc1cc(NC(=O)c2oc(Br)cc2)ccc1N3CCCCC3</chem>	
MMV007977	<chem>Clc1ccc(C=CC2=NC(=O)c3ccccc3O2)cc1</chem>	