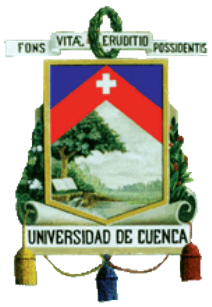


UNIVERSIDAD DE CUENCA

Facultad de Ciencias Químicas

Carrera de Bioquímica y Farmacia

"Relación estructura-actividad como estrategia para la selección de moléculas candidatas para el desarrollo de inhibidores de tirosinasas"

Tesis previa a la obtención
del título de Bioquímica Farmacéutica

Autoras:

Katherine Alexandra Mogrovejo Mata C.I.0302011879
katty7mm@hotmail.com

Doménica Victoria Muñoz Vázquez C.I.0301981510
dome.14@hotmail.com

Directora:

Dra. Maritza Raphaela Ochoa Castro C.I.0301843090

Asesor:

Dr. Cristian Xavier Rojas Villa C.I.0103596722

Cuenca-Ecuador
08/ 01/ 2020

RESUMEN

En este trabajo se ha desarrollado un modelo basado en las Relaciones Cuantitativas Estructura-Actividad (QSAR) para predecir la concentración inhibitoria media máxima (IC_{50}) de 581 moléculas sobre la enzima tirosinasa. Cada estructura molecular fue optimizada en el programa HyperChem mediante la mecánica molecular (MM+) y el método semiempírico PM3. Posteriormente, se calcularon 5274 descriptores moleculares y 166 huellas dactilares moleculares MACCS en el programa alvaDesc, los cuales fueron reducidos mediante el método no supervisado V-WSP. Así, 1692 descriptores se sometieron a un proceso de selección supervisada de variables mediante Algoritmos Genéticos (GAs) acoplados con el método de clasificación de los *k*-vecinos más cercanos (*k*NN). En esta etapa se aplicó el método simplex para optimizar el umbral para la separación entre las clases de alta y baja actividad. Se obtuvo un modelo óptimo con ocho descriptores moleculares y cuatro vecinos (NER_{cal} = 0.82). Este modelo se validó mediante validación cruzada de ventanas venecianas (NER_{cv} = 0.82) y un grupo externo de predicción constituido por 174 moléculas (NER_{pred} = 0.86). Adicionalmente, se definió el dominio de aplicabilidad del modelo y se brindó la interpretación mecánica de los descriptores moleculares. El modelo QSAR propuesto fue desarrollado usando los cinco principios definidos por la Organización para la Cooperación y el Desarrollo Económico (OECD) para garantizar su aplicabilidad.

Palabras Clave: Melanina. Inhibidores de tirosinasa. QSAR. *k*NN. Algoritmos genéticos (GAs). NER. Dominio de Aplicabilidad.

ABSTRACT

The purpose of this work was to calibrate a Quantitative Structure-Activity Relationship (QSAR) model to predict the half maximal inhibitory concentration (IC_{50}) of the tyrosinase activity of 581 molecules. For geometry optimization, the molecular mechanic force field (MM+) was used, followed by the PM3 semi-empirical method to refine the structures. Then, compounds were described by 5274 alvaDesc molecular descriptors and 166 MACCS structural keys, which were merged into a single dataset and subsequently reduced by the V-WSP unsupervised variable reduction. Thus, 1692 descriptors were submitted to the Genetic Algorithms (GAs) supervised variable selection process coupled with the *k*-nearest neighbors classifier. In this step, the simplex method was applied in order to optimize the threshold for the separation between the high and low activity classes. A model composed of eight molecular descriptors and four neighbors classifiers was retained as the optimal one (NER_{train} = 0.82). This model was validated by a cross-validation protocol based on venetian blind (NER_{cv} = 0.82) and an external test set of 174 molecules (NER_{test} = 0.86). In addition, the applicability domain of the model was defined, and the mechanistic interpretation of molecular descriptors was provided. The proposed QSAR model was developed using the five principles defined by the Organization for Economic Co-operation and Development (OECD) in order to guarantee its applicability.

Keywords: Melanin. Tyrosinase inhibitors. QSAR. *k*NN. Genetic Algorithms (GAs). NER. Applicability domain.

ÍNDICE

Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

Yo, *Katherine Alexandra Mogrovejo Mata,* en calidad de autora y titular de los derechos morales y patrimoniales del trabajo de titulación "**Relación estructura-actividad como estrategia para la selección de moléculas candidatas para el desarrollo de inhibidores de tirosinasas**", de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca, 08 de Enero de 2020

Katherine Alexandra Mogrovejo Mata
C.I: 0302011879

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

Yo, *Doménica Victoria Muñoz Vázquez*, en calidad de autora y titular de los derechos morales y patrimoniales del trabajo de titulación "**Relación estructura-actividad como estrategia para la selección de moléculas candidatas para el desarrollo de inhibidores de tirosinasas**", de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca, 08 de Enero de 2020

Doménica Victoria Muñoz Vázquez
C.I: 030198151-0

Cláusula de Propiedad Intelectual

Yo, *Katherine Alexandra Mogrovejo Mata*, autora del trabajo de titulación **"Relación estructura-actividad como estrategia para la selección de moléculas candidatas para el desarrollo de inhibidores de tirosinasas"**, certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autora.

Cuenca, 08 de Enero de 2020

Katherine Alexandra Mogrovejo Mata
C.I: 0302011879

Cláusula de Propiedad Intelectual

Yo, *Doménica Victoria Muñoz Vázquez,* autora del trabajo de titulación **"Relación estructura-actividad como estrategia para la selección de moléculas candidatas para el desarrollo de inhibidores de tirosinasas"**, certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autora.

Cuenca, 08 de Enero de 2020

Doménica Victoria Muñoz Vázquez
C.I: 030198151-0

AGRADECIMIENTO

En primer lugar, agradezco a Dios y a la Virgen Santísima por derramar abundantes bendiciones sobre mí y haber sido mis guías durante todos los años de carrera.

A la Universidad de Cuenca por permitir el desarrollarme tanto a nivel académico como personal. A la Universidad del Azuay por abrirme sus puertas y darme la oportunidad de desarrollar mi tesis.

Un extensivo agradecimiento a nuestras directoras, Dra. María Elena Cazar y Dra. Maritza Ochoa, por la ayuda recibida durante el desarrollo de esta tesis.

Agradezco infinitamente al Dr. Cristian Rojas, quien más que un asesor se convirtió en un amigo que con su infinita paciencia y ayuda nos supo guiar paso a paso en el desarrollo de este trabajo de investigación, gracias por compartirme todos sus conocimientos y enseñarme que nada es difícil si uno se lo propone y se tiene perseverancia. Gracias por todo el tiempo compartido y que a pesar de sus múltiples ocupaciones siempre me supo brindar su ayuda y tiempo sin importar su cansancio para finalmente dar por concluido con éxito este proyecto. De igual manera al Dr. Piercosimo Tripaldi quien supo ganarse mi cariño y respeto, gracias por todas sus enseñanzas y consejos y por su gran ayuda en el desarrollo de esta tesis.

Finalmente, mi gratitud infinita a mis padres, hermanos y amigos por todo el apoyo desinteresado, paciencia, consejos y palabras de aliento, por ayudarme a continuar pese a cualquier obstáculo o barrera que se me haya presentado a lo largo de este camino. De manera especial a mi amiga y compañera de tesis Dome, gracias por todo el esfuerzo y dedicación, por no rendirnos jamás y haber logrado juntas este sueño que hoy es ya una realidad. A Elisa Pacheco Jaramillo quién con su carisma supo hacer de mi paso por la Universidad del Azuay un lugar lleno de risas y momentos maravillosos, gracias por todo lo compartido y por tu ayuda incondicional.

Katherine Alexandra Mogrovejo Mata

AGRADECIMIENTO

Dicen que la universidad sirve para aprender ciencia. ¿Tiene sentido, no? Pues es todo mentira. La universidad no va de eso. Va de aprender a convivir con gente que no conoces. A encontrar puntos en común. A descubrir que la comida hace milagros para unir a la gente. Sirve para aprender a formar parte de algo, convertir un grupo en una piña, porque si es en pera no funciona igual. Es aprender a estar dispuesto a ayudar a los demás, tengas tiempo o no, y entender que a veces quien necesita que le ayuden eres tú. Es asumir que el concepto horario es una palabra difusa y sin sentido. Sirve para desarrollar la paciencia. Para aprender a hablar y a no hablar. Para descubrir lo mejor de ti e intentar empezar a eliminar lo malo. Para saber que a veces lo pequeños logros te hacen sentir más feliz que los grandes. Sirve para desarrollar tu animal interior y querer enseñarle al mundo lo que ves en tu cabeza. Y al final, sin saber cómo, cuándo, ni por qué, en algún momento de toda esa locura te da por sugerir algo que resulta que puede hasta tener sentido. Y ahí es cuando empiezas a hacer lo que todos pensaban que estabas haciendo desde el principio, ciencia.

Y como un barco no se puede gobernar solo, pues aquí toda la tripulación:

Mi más sincero agradecimiento a Cristian Rojas profesor investigador de la Universidad del Azuay, por su ayuda, dedicación, amabilidad, generosidad y paciencia demostrada en cada momento, nada de esto hubiera sido posible. Siempre recordare con una sonrisa los largos días que compartimos junto con el Doc. y la Eli en el laboratorio de Quimiometría y QSAR, son momentos que se quedan guardados para siempre en mi corazón.

A mis amigos, familiares y personas especiales les agradezco no solo por estar presentes aportando buenas vibras a mi vida, si no por los grandes lotes de felicidad y de diversas emociones que siempre me han causado.

Gracias a la Dra. Maritza Ochoa, quien nos brindó su ayuda en la recta final para culminar este proyecto, de igual manera a la Dra. María Elena Cazar, al habernos sugerido este tema; nuevo, complicado, pero fascinante.

Finalmente, a mi compañera de tesis Katty, parece como si nunca hubiésemos estado en paz, batallando por cualquier cuestión, sin embargo, siempre llegaron los buenos momentos en los que nuestra lucha cesaba para hacer una tregua y lograr esta meta juntas.

Doménica Victoria Muñoz Vázquez

DEDICATORIA

A Dios por todas sus bendiciones y acompañarme en cada paso durante mi carrera, a mis padres Fernando y Nila por su apoyo de manera incondicional y enseñarme a luchar por un sueño y no darme por vencida, a mis abuelitos Homero (+), Julita (+) y Blanquita (+) por sus consejos y palabras de aliento, a mis hermanos Fernando, Javier y Marcelo por su apoyo absoluto durante toda mi carrera universitaria, a mis tías Germania y Ruth por siempre estar conmigo dándome ánimos y su preocupación en todo este tiempo, a mis sobrinas Marcela y Valentina que han sido mi motivación para seguir adelante, a Jacob quién con su infinita paciencia y amor me ayudó en todo momento y siempre estuvo pendiente de mí. A todos Uds. gracias infinitas por ayudarme a hacer realidad este gran sueño y ser lo que soy hoy en día.

Katherine Alexandra Mogrovejo Mata

DEDICATORIA

A la memoria de Andrés Vázquez, quién me animó en este campo de estudio y, con sus locuras y ocurrencias hacía de mis días más difíciles una obra de comedia. Su ejemplo de vida me mantuvo soñando cuando quise rendirme.

A José Oswaldo y Marcia Victoria por confiar en mí, por ser uno de los pilares más importantes de mi vida y demostrarme siempre su cariño y apoyo incondicional a pesar de nuestras diferencias, gracias por todo mamá y papá.

Doménica Victoria Muñoz Vázquez

INTRODUCCIÓN

La enzima tirosinasa ha estado bajo la atención de la comunidad científica internacional por sus múltiples aplicaciones en diferentes áreas como la medicina, los cosméticos, los alimentos y la agricultura. Juega un rol fundamental en la síntesis de la melanina debido a que regula directamente la cantidad de melanina producida mientras que otras enzimas solo modifican el tipo de melanina sintetizada en la vía bioquímica de la pigmentación, así su sobreproducción provoca desórdenes en la melanogénesis y ciertos tipos de cáncer de piel (McQuarrie & Simon, 1997). Los inhibidores farmacológicos de la tirosinasa pueden servir como inhibidores tópicos *de novo* de la melanogénesis, teniendo efecto despigmentante o blanqueador de la piel, cobrando una gran importancia en los productos médicos y cosméticos ya que la mayoría de los inhibidores usados actualmente presentan ciertos inconvenientes como: inestabilidad (ejemplo, L-mimosina y ácido kójico), alta toxicidad o mutagenicidad (como la 1,4-dihidroquinona) limitando así el uso continuo (Y.-H. Chang et al., 2007). Además, esta enzima se encuentra implicada en el pardeamiento enzimático de los alimentos, siendo un problema muy serio en frutas, champiñones, patatas, algunos crustáceos, e incluso en la industria del vino, al producir alteraciones en el color que reducen el valor comercial de los productos, o incluso los hacen inaceptables para el consumidor. Estas pérdidas son muy importantes en el caso de las frutas tropicales y de los camarones, productos trascendentales para la economía de muchos países poco desarrollados (Xu et al., 2009). Razón por la cual el hallazgo de nuevos agentes inhibidores de la tirosinasa obtenidos por identificación *in silico* al ser introducidos en el mercado, tendrán gran impacto socio-económico en la comunidad, ya que los métodos clásicos basados en experimentación de “prueba y error” no permiten hallar moléculas que tengan un perfil de poder ser comercializadas, ahorrando tiempo, costos y recursos (Riley, 2000; Xu et al., 2009).

Las herramientas computacionales han evolucionado de tal forma que se han transformando en tecnologías cada vez más importantes para la búsqueda de moléculas candidatas a fármacos, dentro de estas herramientas se encuentra el acoplamiento molecular (docking), técnica de mecánica molecular ampliamente utilizada para predecir energías y modos de enlace entre ligandos y proteínas, información de gran utilidad en el estudio de nuevos compuestos con efectos terapéuticos. No obstante, los resultados obtenidos mediante esta técnica tienden a la subjetividad, debido a que los programas utilizados para llevarla a cabo proporcionan más de un criterio de selección de la mejor pose.

Un problema fundamental con el docking molecular es que el espacio de orientación es muy grande y crece combinatoriamente con el número de grados de libertad de las moléculas que interactúan (Velásquez, Drosos, Gueto, Márquez, & Vivas-Reyes, 2013)

Por todo lo anterior se plantea el siguiente problema científico: ¿Los métodos convencionales empleados para el descubrimiento y desarrollo de nuevos fármacos inhibidores de tirosinasa presentan desventajas en cuanto al proceso complejo y costoso en términos de tiempo y dinero, en relación con los métodos *in silico?*
Para dar respuesta al problema científico planteado, se propone el siguiente objetivo general:

Objetivo General

Aplicar métodos de química computacional para establecer la relación estructura – actividad de un conjunto de moléculas, como estrategia para la selección de inhibidores de tirosinasas.

Objetivos específicos

-Construir una base de datos con moléculas candidato para el desarrollo de inhibidores enzimáticos, y asociarlos a descriptores moleculares.

-Obtener a partir de descriptores moleculares y técnicas estadísticas, modelos matemáticos con el fin de poder clasificar a los inhibidores de la enzima tirosinasa.

-Estimar la potencia de los compuestos activos mediante los modelos matemáticos que han sido obtenidos por procesos estadísticos.

Los objetivos fueron planteados para probar la veracidad de la siguiente hipótesis:

- La química computacional es una estrategia válida para seleccionar compuestos candidatos al desarrollo de fármacos inhibidores de tirosinasas.

El presente trabajo de investigación demuestra la aplicabilidad de los métodos quimiométricos como estrategia en la búsqueda racional de blancos moleculares para el desarrollo de fármacos.

CAPÍTULO 1: CONTENIDO TEÓRICO

1.1 Tirosinasa

La tirosinasa conocida también como monofenol monooxigenasa o polifenol oxidasa (EC 1.14.18.1; número de registro CAS: 9002-10-2) es una enzima multifuncional y cuprífera, en la cual sus dos iones de cobre se ubican en el centro activo de la enzima y están individualmente conectados a tres residuos de histidina. Se encuentra altamente distribuida en la naturaleza en hongos, animales y plantas; además es la responsable del pardeamiento de frutas, vegetales y la coloración de la piel, cabello y ojos en los animales (Chen, Hung, Chen, Lai, & Chan, 2016).

Las primeras investigaciones bioquímicas se llevaron a cabo en 1895 en el hongo *Russula nigricans*, cuya carne cortada se vuelve roja y posteriormente negra al exponerse al aire, en este estudio se ha encontrado que la enzima está distribuida ampliamente en toda la escala filogenética desde bacterias hasta mamíferos. Las tirosinasas mejor caracterizadas se derivan de *Streptomyces glausescens,* los hongos *Neurospora crassa* y *Agaricus bisporus* (T.-S. Chang, 2009).

1.1.1 Melanogénesis

La tirosinasa cataliza dos reacciones distintas que son el paso inicial en la biosíntesis de la melanina: la hidroxilación de la tirosina a L- dopa (actividad monofenolasa o creolasa) y la oxidación de L-dopa a sus dopaquinonas correspondientes (actividad difenolasa o catecolasa). Estas quinonas al ser altamente reactivas tienden a polimerizarse continuando con una serie de reacciones: por un lado el dihidroxiindol (DHI) y el dihidroxiindol 2- acido carboxílico (DHICA) que son productos de la reacción del dopacromo forman la eumelanina y, por otro lado, en presencia de cisteína o glutatión la dopaquinona se convierte en cistenildopa o glutionildopa para posteriormente formar la feomelanina, pigmentos marrones de alto peso molecular que juntos conforman la melanina, como se muestra en la siguiente Figura 1 (Ashooriha et al., 2019; T.-S. Chang, 2009).

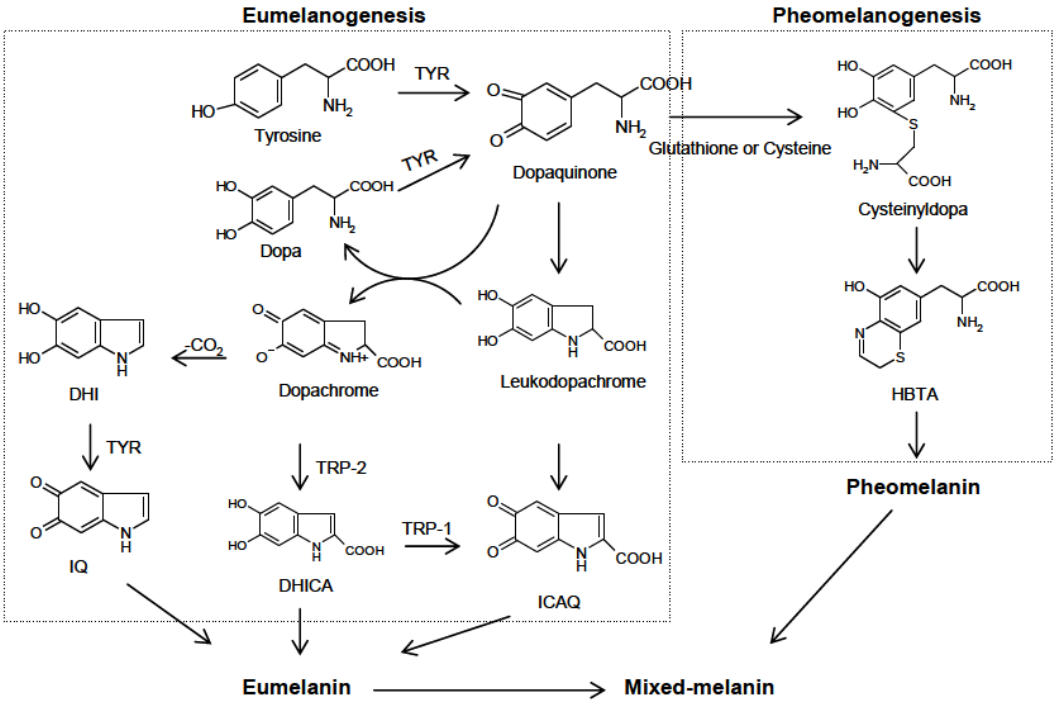


Figura 1. Ruta metabólica de la melanina

Fuente: (T.-S. Chang, 2009).

Las dos reacciones elementales de oxidación son pasos limitantes en la producción de melanina que cataliza la enzima tirosinasa. Por lo tanto, esta enzima puede ser un blanco adecuado para inhibir la formación de la melanina.

1.1.2 Melanina

La melanina es el nombre general para un rango amplio de pigmentos naturales que se encuentran en muchas especies de organismos vivos y microorganismos como animales, plantas, hongos y bacterias. En los mamíferos, la melanina es sintetizada en varias partes del cuerpo como la piel, ojos, cabello y cerebro. Varios roles son atribuidos a la melanina, pero realmente su función más importante es la protección de la piel humana de los efectos nocivos de la radiación UV del sol. Es sintetizada y secretada por los melanocitos que se encuentran en la capa basal de la dermis. A pesar del rol importante que cumple, su sobreproducción puede resultar en varios desórdenes de hiperpigmentación como pecas, léntigos seniles y melasma. En la industria agrícola una gran cantidad de frutas y vegetales se desperdician debido al pardeamiento. Este cambio de color se debe a la sobreproducción de la melanina. Los blanqueadores y agentes despigmentantes los cuales inhiben la formación de la melanina se utilizan

ampliamente en productos cosméticos por lo tanto tienen un alto valor económico en la industria cosmética (Chen et al., 2016).

En la formación de pigmentos de melanina están involucrados tres tipos de tirosinasa (oxi, mer y desoxitirosinasa) con diferentes estructuras binucleares de cobre del sitio activo. La forma oxigenada (oxitirosinasa, Eoxy) consta de dos átomos de cobre tetragonal II cada uno coordinado por dos ligandos NHis axiales ecuatoriales fuertes y uno más débil. La molécula de oxígeno está unida como peróxido y une los dos centros de cobre. Metirosinasa (Emet), similar a la forma oxi, contiene dos iones de cobre tetragonal II acoplados a través de un puente endógeno, aunque los ligandos exógenos de hidróxido distintos del peróxido están unidos al sitio de cobre. La desoxitirosinasa (Edeoxy) contiene dos iones de cobre I con una disposición de coordinación similar a la forma encontrada, pero sin el puente de hidróxido. La forma de reposo de la tirosina es decir la enzima obtenida después de la purificación se encuentra en una mezcla de 85% de emet y 15% de formas oxi (T.-S. Chang, 2009).

Muchos inhibidores putativos se examinan en presencia de tirosinasa o dopa como sustrato enzimático, y la actividad se evalúa en términos de formación de dopacromo, por lo tanto, la observación experimental de la inhibición de la actividad de la tirosina se puede lograr mediante uno de los siguientes métodos:

1. Reducción: agentes que causan la reducción química de dopaquinona como el ácido ascórbico, que se usa como inhibidor de la melanogénesis debido a su capacidad para reducir la o-dopaquinona a dopa, evitando así las formaciones de dopacromo y melanina.
2. Eliminador de o-dopaquinona: la mayoría de los compuestos que contienen un grupo funcional tio que son inhibidores de la melanogénesis bien conocidos y reacción con dopaquinona para formar productos incoloros, por lo tanto, el proceso melanogenético se ralentiza hasta que se consume todo el eliminador y luego va a su velocidad original.
3. Sustratos enzimáticos alternativos: algunos compuestos fenólicos cuyos productos de reacción quinoide absorben en un rango espectral diferente al del dopacromo. Cuando estos compuestos fenólicos muestran una buena afinidad por la enzima se evita la formación de dopacromo y podrían clasificarse erróneamente como inhibidores.

4. Inactivadores de enzimas inespecíficos: como los ácidos o bases que desnaturalizan la enzima de forma inespecífica, inhibiendo así su actividad.
5. Inactivadores de tirosinasa específicos: como los inhibidores basados en mecanismos que también se denominan sustratos suicidas. Estos inhibidores pueden ser catalizados por la tirosinasa y formar un enlace covalente con la enzima, inactivando irreversiblemente la enzima durante la reacción catalítica. Inhiben la actividad de la tirosinasa al inducir la enzima que cataliza la ¨reacción suicida¨.
6. Inhibidores: los compuestos se unen irreversiblemente a la tirosinasa y reducen su capacidad catalítica.

La actividad inhibitoria es uno de los criterios principales de eficiencia, la actividad de un inhibidor generalmente se expresa como la concentración de un inhibidor necesaria para inhibir la mitad de la actividad enzimática en condiciones probadas (CI_{50}) (T.-S. Chang, 2009).

1.1.3 Inhibidores de la tirosinasa

La oxidación enzimática de la L-tirosina a melanina es de considerable importancia debido a que la melanina cumple muchas funciones, y una alteración en la síntesis de la melanina puede desencadenar en muchas enfermedades, razón por la cual los inhibidores de tirosinasa son cada vez más importantes en la industria alimenticia, así como en productos medicinales y cosméticos (Cho, Roh, Sun, Kim, & Park, 2006).

Hoy en día existen muchos compuestos naturales y sintéticos como la arbutina, hidroquinona, ácido azelaico y ácido kójico han sido reportados como inhibidores de tirosinasas. Sin embargo, solo pocos de ellos poseen suficiente seguridad y potencia para su uso, por ejemplo: hay algunas evidencias que indican los efectos mutagénicos y citotóxicos de la hidroquinona contra los melanocitos, además la hidroquinona puede ocasionar sequedad o irritación local de la piel lo que lleva a la hiperpigmentación postinflamatoria. El ácido kójico es el compuesto más utilizado entre los blanqueadores de la piel en productos cosméticos, pero este compuesto tiene poca eficacia e insuficiente estabilidad y capacidad para penetrar a la piel. Para superar estos inconvenientes, se debe usar una concentración más alta de ácido kójico, lo que resulta en incidencia de efectos secundarios. El ácido kójico muestra un efecto inhibidor competitivo sobre la actividad monofenolasa y un efecto inhibidor mixto sobre la actividad difenolasa de la tirosinasa de hongos (Ashooriha et al., 2019).

1.2 Relaciones Cuantitativas Estructura-Actividad (QSAR)

Las relaciones cuantitativas estructura-actividad (QSAR: quantitative structure - activity relationship) son métodos *in silico* que se desarrollaron a inicios de los años 60 con los estudios pioneros de Corwin Hansch y Toshio Fujita. Estos investigadores, cuantificaron la relación entre efectos biológicos y la densidad en la posición orto del anillo aromático de los derivados del ácido fenoxiacético (Dearden, 2016). Con este trabajo abrieron un campo nuevo de investigación ampliamente usado para la predicción de actividades biológicas y propiedades fisicoquímicas para el diseño racional de fármacos. No obstante, la teoría QSAR a lo largo de los años ha ido evolucionando desde modelos de regresión lineal múltiple hasta ser una herramienta de importante aplicación en las ciencias biológicas, químicas, ambientales, medicinales, toxicológicas, farmacológicas y de los alimentos. Hoy en día la teoría QSAR sigue desarrollándose y prueba de ello es la gran variedad de métodos QSAR que han sido presentados para la predicción de las actividades de compuestos desconocidos (Galvez & Garcia-Domenech, 2010).

El objetivo de la teoría QSAR es desarrollar modelos matemáticos predictivos y capaces de ser aplicados para evaluar la actividad de nuevos compuestos químicos que no han sido previamente sintetizados ni evaluados experimentalmente. Además, la teoría QSAR complementa otros estudios, ya sean teóricos o experimentales, cuyo propósito es resolver las interrogantes de tipo químico de forma racional (Kubinyi, 2008). Actualmente diferentes organismos internacionales han propuesto como una herramienta útil el desarrollo de modelos QSAR predictivos para estudiar mediante técnicas racionales la información que se encuentra de forma implícita en la estructura química de las moléculas (Todeschini, Consonni, & Gramatica, 2009).

1.2.1 Definición y formalismo

El modelado QSAR se define como la relación matemática entre una respuesta (actividad, propiedad, toxicidad y otras) y características químicas definidas por los descriptores moleculares de las moléculas analizadas. El estudio puede tomar cualquier nombre específico dependiendo de la naturaleza de la respuesta que se va a modelar teniendo así, dos grandes clases denominadas relaciones cuantitativas estructura-actividad o QSAR y relaciones cuantitativas estructura-propiedad o QSPR que son modelos matemáticos asistidos por computadora que relacionan propiedades

fisicoquímicas de los compuestos con su estructura molecular (Kaliszan, 2007; Rojas, Duchowicz, Pis Diez, & Tripaldi, 2016; Roy, Kar, & Das, 2015a; Todeschini & Consonni, 2009).

Comúnmente para referirse a todos estos tipos específicos de estudios se utiliza el término QSAR, el mismo que permite predecir la actividad de las moléculas ya sean nuevas o hipotéticas en función de las características estructurales de cada molécula. Por lo tanto, matemáticamente el formalismo básico de la teoría QSAR se representa de la siguiente manera en la ecuación 1.1 (Hongmao, 2015; Roy et al., 2015a; Todeschini & Consonni, 2009)

$$Actividad\ /\ Propiedad\ biologica = f\left(Estructura\ química\right) \qquad (0.1)$$

Las propiedades fisicoquímicas son aquellas características que se obtienen de forma experimental o teórica (por ejemplo, coeficiente de partición octanol/agua, índices de retención cromatográfico, refractividad molar, umbral de olor, punto de ebullición y otros), en cambio las actividades biológicas son el efecto farmacológico que ejerce un compuesto sobre un blanco molecular (por ejemplo, antiinflamatoria, antipirética, toxicidad, antihistamínica, anticolinérgica, antimuscarínica entre otras). Por otro lado, los descriptores moleculares codifican información de la estructura química de los compuestos y se los obtiene de forma experimental o teórica. La interrelación entre la propiedad/actividad y la estructura química en los modelos QSAR se presentan en la Figura 2.

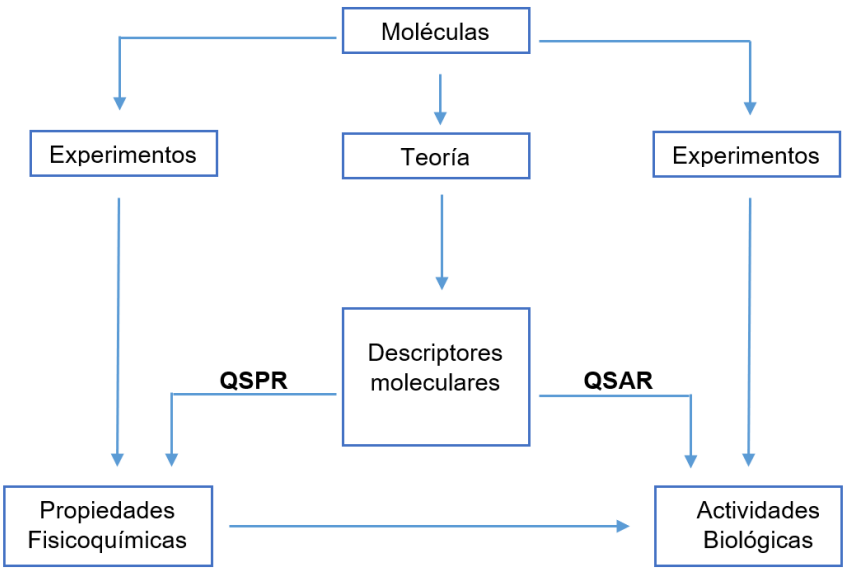


Figura 2. Esquema general de un estudio QSAR/QSPR

Fuente: (Todeschini, 2003).

El formalismo presentado anteriormente en la ecuación 1.1 puede ser aplicado a modelos discretos (clasificación), así como a modelos continuos (regresión) (Chaudhry et al., 2007; Hongmao, 2015; Liaw & Svetnik, 2015; Todeschini & Consonni, 2009). En los modelos de regresión la actividad/propiedad es una variable cuantitativa continua, es decir que toma cualquier valor dentro de la escala de medida que se utilice. En cambio, en clasificación la respuesta se toma como una variable cualitativa nominal, en el cual la actividad se representa en forma de categorías no ordenadas, por ejemplo, compuestos de alta inhibición (clase 1), compuestos de baja inhibición (clase 2) y compuestos de inhibición intermedia (clase 3). Es muy común en clasificación que el grupo de calibración se forme considerando la numerosidad de las clases, debido a que los modelos tienden a sesgarse hacia la clase más numerosa debido a que el balance entre las clases tiene gran influencia en la calidad del modelo (Hongmao, 2015).

1.2.2 Objetivos de un modelo QSAR

El objetivo de un estudio QSAR es desarrollar un modelo matemático predictivo y racional, junto con la interpretación de la información química involucrada. El modelo se desarrolla a partir de los compuestos para los cuales se disponga de una determinada respuesta de actividad (por ejemplo el IC_{50}) y permitirá también realizar la predicción de la actividad para un número mayor de moléculas (Hamzeh-Mivehroud, Sokouti, & Dastmalchi, 2015). Por este motivo, es de gran utilidad en procesos de investigación, así como en otros campos donde es importante la predicción de actividades de compuestos químicos. Otros objetivos que se pueden enlistar son (Cronin, 2010; Roy et al., 2015a; Roy, Kar, & Das, 2015b):

- Comprensión de los mecanismos de acción dentro de un grupo de sustancias químicas.
- Barrido virtual de bibliotecas químicas.
- Predicción de una actividad química de interés.
- Reducción y reemplazo de la experimentación de laboratorio usado en animales (la experimentación normalmente es larga y costosa).
- Optimización de la síntesis química de moléculas con actividades deseadas (minimizar la eliminación de desechos y reducir el costo).
- Aplicaciones con fines regulatorios por parte de agencias gubernamentales.

- Predicción de la toxicidad de compuestos en seres humanos por exposición ocasional u ocupacional, al igual que la predicción de la toxicidad sobre especies ambientales.
- Identificación de compuestos peligrosos en las etapas iniciales del diseño de los mismos.
- Refinamiento estructural de moléculas objetivo sintéticas.

1.2.3 Componentes y etapas de un modelo QSAR

En el modelado QSAR existen dos componentes fundamentales: los datos cuantitativos o cualitativos y el uso de técnicas quimiométricas adecuadas, los datos se consiguen por medición de la actividad/propiedad de interés y por información química contenida en los descriptores moleculares. Debido a la gran magnitud de información que se maneja, es necesario el uso de computadores con alto rendimiento de cálculo y procesamiento. Para llevar a cabo el desarrollo de un modelo QSAR es necesario realizar los siguientes cuatro pasos (Golbraikh, Wang, Zhu, & Tropsha, 2017; Roy et al., 2015b):

1. Preparación de los datos

   Para dar inicio al modelado de la información se debe tener un conjunto de compuestos químicos junto con su actividad de interés en este caso (IC_{50}) ordenados de manera conveniente y útil.

   Cuando la actividad tiene una amplia escala de medida, es necesario realizar una trasformación de la variable de tal forma que los valores sean lo más cercanos posibles; además es importante la representación de las estructuras químicas de tal forma que sea posible el cálculo de los descriptores moleculares.

2. Procesamiento de los datos

   Antes del desarrollo de un modelo QSAR es importante realizar un pretratamiento de la información. En esta etapa se busca eliminar moléculas duplicadas, excluir descriptores con valores faltantes o casi constantes, así como los que se encuentren correlacionados más arriba de un cierto umbral.

3. Validación y predicción de los datos

   Un aspecto importante a considerar durante el desarrollo de un modelo QSAR es la división de la matriz de datos en dos grupos: entrenamiento o calibración (training set) y predicción (test set). El grupo de calibración se usa para ajustar

el modelo, para lo cual se utilizan técnicas de regresión o clasificación acopladas con métodos de selección de variables. Por otra parte, el conjunto de predicción se emplea para determinar la capacidad predictiva y poder de generalización del modelo. También se utilizan técnicas de validación interna o cruzada (cross-validación) para medir la robustez del modelo. Además, se debe definir bien el dominio de aplicabilidad (AD).

4. Interpretación de los datos
   Después de que el modelo ha sido validado, la información interpretada a partir de los descriptores se utiliza para estudiar el mecanismo de acción de los compuestos químicos. Igualmente se puede utilizar el modelo para diseñar nuevas sustancias.

1.2.4 Principios del modelado QSAR

Para que los modelos QSAR sean confiables, hay que cumplir con ciertas condiciones que establece la Organización para la Cooperación Económica y el Desarrollo OECD (Organización para la Cooperación y el desarrollo Económico) (OECD, 2007)

- Actividad/propiedad definida
  El objetivo de este principio es el de garantizar la claridad en la definición de la actividad/propiedad que será utilizada para desarrollar el modelo debido a que las mismas pueden tener diferentes métodos o protocolos para la medición experimental.

- Algoritmo inequívoco
  Este objetivo busca asegurar transparencia en el algoritmo matemático que se utilizó para desarrollar el modelo, debido a que varios enfoques de modelización han sido propuestos en la literatura.

- Dominio de aplicabilidad definido
  Este principio hace referencia a que la aplicabilidad de un modelo QSAR está limitado a los componentes químicos que son estructuralmente similares a los que fueron utilizados para calibrar el modelo (principio de congenericidad), debido a que estos modelos son reduccionistas y están asociados con las limitaciones en términos de los tipos de estructuras químicas consideradas, de las propiedades fisicoquímicas y los mecanismos de acción.

- Medida apropiada de la bondad de ajuste, robustez y predictividad
  Es necesario saber si el modelo es robusto, si no está sobreajustado y si es capaz de predecir confiablemente la actividad/propiedad para moléculas externas para poder tener una mejor evaluación de calidad de un modelo.

- Interpretación del mecanismo de acción de los descriptores (de ser posible).
  Este último principio intenta dar una explicación de los descriptores moleculares, pero no siempre es posible obtener esta interpretación debido a que en algunos casos la definición propia del descriptor es compleja. Sin embargo, este aspecto no indica que el modelo pierda utilidad.

1.3 Aplicaciones de un Modelo QSAR

Las áreas de aplicación del modelado QSAR han ido creciendo a lo largo del tiempo. Son muy variadas y de gran utilidad para los diversos campos científicos, debido a que es una opción conveniente para el monitoreo de la actividad de los compuestos químicos. Así las aplicaciones de los modelos QSAR se agrupan de tres maneras distintas (Roy et al., 2015a).

1.3.1 Naturaleza de la respuesta

La respuesta experimental se puede categorizar en actividades biológicas, toxicidades biológicas. Como ejemplos de actividades biológicas se pueden citar: antibacteriales, antihistamínicos, antioxidantes, antidepresivos, antimalaria, antihipertensivas, antiepilépticos, anti-VIH, antidiuréticos, etc. Entre las propiedades fisicoquímicas asociadas a moléculas activas podemos citar: hidrólisis, biodegradación, coeficiente partición octanol/agua, bioacumulación, temperatura de transición vítrea, oxidación atmosférica, índices de retención cromatográfica, etc. Finalmente, las toxicidades pueden ser: aguda por inhalación, aguda oral, aguda en peces, acuática persistente, hepatotoxicidad, nefrotoxicidad, cardiotoxicidad, irritación de piel y ojos, carcinogenicidad (Roy et al., 2015a).

1.3.2 Naturaleza de los compuestos

Los compuestos químicos se pueden diferenciar entre aquellos usados en procesos industriales y de laboratorio (solventes, perfumería, surfactantes y reactivos), compuestos químicos beneficiosos para la salud (fármacos y aditivos alimentarios) y

compuestos químicos nocivos para la salud (pesticidas, agentes carcinogénicos, contaminantes orgánicos persistentes, toxinas) (Roy et al., 2015a)

1.3.3 Áreas de aplicación

Las principales áreas de aplicación de la teoría QSAR son la ciencia de los materiales, toxicología predictiva y diseño de fármacos. En ciencia de materiales se utilizan para el estudio de polímeros, líquidos iónicos, catálisis, nanomateriales, fullerenos, surfactantes, biomateriales, cerámicos, etc. En toxicología predictiva se puede valorar la toxicidad sistémica y el control de riesgos ecotoxicológicos. Finalmente, en el diseño de fármacos es importante el enfoque ADME (Absorción, Distribución, Metabolismo y Excreción), el cual permite monitorear el perfil farmacocinético de un fármaco antes de su síntesis, contribuyendo al diseño y eficacia de un compuesto dentro un sistema biológico (Roy et al., 2015a).

1.4 Descriptores moleculares

Los descriptores moleculares han estado en la mira de los científicos, quienes se han enfocado en la manera de capturar y transformar de forma teórica la información codificada dentro de una estructura química, para relacionarlos de forma cuantitativa con actividades/propiedades biológicas de interés. Por lo tanto, son de gran importancia ya que constituyen las variables independientes utilizadas para predecir dichas actividades. Para el cálculo de los descriptores es imprescindible representar adecuadamente la estructura molecular mediante herramientas quimioinformáticas apropiadas que permitan desarrollar modelos QSAR. (Gasteiger & Engel, 2006).

Los descriptores moleculares son considerados un grupo de parámetros que describen de forma cuantitativa a una estructura molecular y se los puede extraer a partir de diferentes formas de representación molecular. Se distinguen dos clases de descriptores moleculares: 1) experimentales que se obtienen a través de experimentos estandarizados como es el caso de la refractividad molar, polarizabilidad, y 2) teóricos que se obtienen aplicando algoritmos matemáticos bien establecidos a una representación inequívoca de la estructura molecular (Todeschini & Consonni, 2009).

Debido al avance en quimioinformática se ha hecho factible la relación entre diferentes formas de representación de la estructura molecular con las actividades/propiedades de los compuestos químicos, razón por la cual son muy variadas las áreas en las que se

puede aplicar los descriptores, por ejemplo, química medicinal, farmacología, toxicología, barrido virtual, entre otros (Mauri, Consonni, & Todeschini, 2017).

1.4.1 Definición

Se define a un descriptor molecular como "*el resultado final de un procedimiento lógico y matemático que transforma la información química codificada en una representación simbólica de una molécula en un número útil o el resultado de algún experimento estandarizado*" (Todeschini & Consonni, 2009). El campo de investigación en descriptores moleculares es amplio, por lo que se siguen proponiendo descriptores en la literatura, por tal motivo se han establecido ciertos criterios que se deben cumplir para ser reconocidos como tal (Guha & Willighagen, 2012; Randić, 1996):

1. Un descriptor debe ser importante para una amplia clase de compuestos.
2. Debe ser invariante al etiquetado, numerado de átomos, roto-traslación de la molécula, y ser calculado mediante un algoritmo bien definido.

Adicionalmente, para ser potencialmente útil debe cumplir con algunos requisitos:

3. Poseer una interpretación estructural y correlación adecuada por lo menos con una propiedad experimental.
4. No tener relación trivial con otros descriptores moleculares, ni estar basado en propiedades experimentales.
5. Debe ser simple, continuo, poseer mínima degeneración, ser capaz de separar isómeros y ser de preferencia aplicable a una amplia clase de moléculas.
6. Tener valores calculados en un rango numérico apropiado para el grupo de moléculas sobre las que se aplicará.

Básicamente, los dos primeros criterios permiten conocer si un descriptor molecular está bien definido, sin embargo, no permite identificar si éste será o no adecuado para predecir una propiedad/actividad determinada. A partir de la tercera regla se refiere al uso de un descriptor, es decir, debe ser interpretable y relacionarse con al menos una propiedad experimental pero no con los demás descriptores de forma estrecha (Guha & Willighagen, 2012; Randić, 1996). Por otra parte, la continuidad y baja degeneración se refiere a que deben ser capaces de considerar variaciones (incluso mínimas) en la estructura molecular (Leszczynski, 2012).

1.4.2 Representación de la estructura molecular

Los compuestos químicos pueden ser representados siguiendo distintos criterios y reglas. Dependiendo de la representación molecular que se use se encontrará información diferente que se ve reflejada en los descriptores moleculares que se calculen (Edwards, Anker, & Jurs, 1991; Testa & Kier, 1991). Las estructuras moleculares pueden ser representadas a través de esquemas gráficos que muestran el tipo de complejidad estructural (Figura 3) o mediante notaciones lineales de cadena como se indica en la Tabla 1.

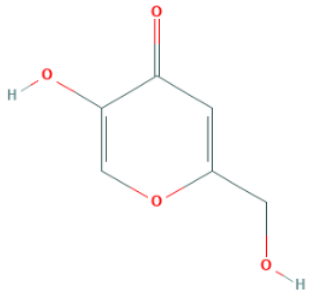


Figura 3. Representación de la estructura química del ácido kójico

Fuente: (Kim et al., 2015).

La forma más simple de representar a una molécula es mediante su fórmula química, que codifica los tipos de átomos y sus ocurrencias dentro de dicha molécula. La fórmula química no contiene ninguna información sobre conexión atómica, por lo tanto, los descriptores moleculares obtenidos por este tipo de representación son denominados descriptores moleculares 0D o constitucionales.

Tipo	Notación
Nombre comercial	Ácido kójico
Nombre IUPAC	5-Hydroxy-2-(hydroxymethyl)-4H-pyran-4-one
Fórmula molecular	$C_6H_6O_4$
Número de registro CAS (Chemical Abstracts Service)	501-30-4
SMILES canónico	C1=C(OC=C(C1=O)O)CO

Tabla 1. Diferentes tipos de notación lineal para la estructura química del ácido kójico
Fuente: (Kim et al., 2015).

La lista de fragmentos estructurales contenidos en el interior de una molécula constituye la lista de subestructuras, dichos fragmentos no simbolizan la topología de la estructura química, razón por la cual pueden ser calculados e interpretados con facilidad. Se denominan como descriptores moleculares 1D a todos aquellos que han sido obtenidos mediante la representación antes mencionada. Estos descriptores se usan sobre todo para análisis de similitud/diversidad molecular o en barrido virtual de bases de datos de gran tamaño.

La representación de las moléculas en dos dimensiones, conocida como representación topológica, es la más común y a diferencia de la 1D contiene información de la conectividad atómica que constituye la molécula. Los descriptores moleculares 2D son calculados en base a la representación topológica y dentro de este grupo se pueden mencionar los índices topológicos, índices de información, índices de conectividad, entre otras. Por otra parte, los descriptores moleculares 3D son aquellos que presentan información de conectividad atómica y que han sido obtenidos mediante representación geométrica de los átomos de una molécula en el espacio tridimensional, y cuyos valores pueden verse modificados debido a las diferentes conformaciones de equilibrio existentes en los compuestos químicos, a este grupo de descriptores pertenecen, por ejemplo, los descriptores GETAWAY, CATS 3D, WHIM, 3D-MoRSE, RDF, perfiles moleculares de Randić, descriptores de carga y cuánticos (Garcia, Duchowicz, & Castro, 2016).

1.4.3 Grafos moleculares

Los grafos moleculares *G = (V, E)* *(Janežič, Miličević, Nikolić, & Trinajstić, 2015; Polansky, 1991)* son la herramienta que se usa con mayor frecuencia para la representación 2D de la estructura química. Se definen como una representación matemática de un grupo de vértices (*V)* y un grupo de aristas que están localizadas entre los vértices (*E*) que reflejan los enlaces químicos entre átomos. Un grafo molecular es ponderado debido a que a cada vértice se le asigna un número que significa el orden de los enlaces, y es disperso cuando el número de aristas es menor al valor del cuadrado del número de vértices $(E) < (V)^2$.

A las propiedades estructurales de las moléculas se las puede estudiar por medio de algoritmos útiles y bien definidos que permiten el manejo de grafos moleculares con la aplicación de la teoría de los grafos sobre la estructura molecular (Balaban, 1985).

Un grafo puede ser representado de dos maneras: 1) como una colección de listas de adyacencia que son una colección de (V) listas, una para cada *i-ésimo* átomo, en donde cada lista *Adj [i]* contiene los átomos que están conectados al *i-ésimo* átomo y su correspondiente orden de enlace, o 2) como una matriz de adyacencia.

1.4.4 Curado de las estructuras moleculares

Antes de llevar a cabo el cálculo de los descriptores moleculares, es de gran importancia evidenciar que las estructuras moleculares sean las correctas, ya que de aquellas que se encuentran disponibles en bases de datos e incluso en publicaciones científicas pueden contener algún tipo de error. A este proceso se le conoce como curado y se aplica sobre el conjunto de datos completo con la aplicación de cinco pasos fundamentales (Fourches, Muratov, & Tropsha, 2010):

1. Eliminación de estructuras moleculares que presenten inconvenientes para ser procesadas por los programas de cálculo de descriptores.
2. Conversión y limpieza de las estructuras químicas.
3. Estandarización y normalización de quimiotipos específicos.
4. Eliminación de compuestos duplicados.
5. Control final manual.

Durante el curado de la información, existen algunos compuestos químicos que se eliminan, por ejemplo, mezclas de sustancias químicas, compuestos organometálicos e inorgánicos, debido a que ciertos programas que se utilizan para el cálculo de descriptores moleculares no reconocen o admite este tipo de compuestos. A continuación, se prepara el conjunto de datos, donde se codifican las estructuras químicas en ciertos formatos computacionales determinados, por ejemplo, HyperChem (hin) o SMILES (Simplified Molecular Input Line Entry Specification), entre los más importantes usados en este trabajo. Esta operación se realiza en programas editores específicos o a partir de alguna biblioteca de acceso público o comercial. Es importante realizar un control para evitar errores durante la construcción de las moléculas.

Debido a que una estructura química se puede representar de diversas maneras puede afectar los valores de los descriptores moleculares, por ello es importante y necesario escoger la misma forma de representación y que se aplique a todas las moléculas. Un ejemplo de esto es la representación en formato aromático (Figura 4a) y representación de anillos aromáticos representados en el formato de Kekulé (Figura 4b). Dependiendo del tamaño del conjunto de datos es recomendable hacer un control de las estructuras

moleculares codificadas de forma manual para identificar moléculas que presenten algún error y evitar problemas posteriores.

a b

Figura 4. Representación molecular en formato aromático (a) y del anillo aromático en formato de Kekulé (b) Fuente: Los autores.

1.4.5 Principales tipos de descriptores moleculares

A continuación, se describirán de entre los miles de descriptores existentes, aquellos que se han utilizado en el modelo QSAR de la presente tesis.

1.4.5.1 Índices topológicos

Los índices topológicos (TIs) (Bonchev, 2015; Roy et al., 2015a; Todeschini & Consonni, 2009) son descriptores que se calculan a partir de una representación topológica de la estructura química 2D (grafo molecular), por lo que no brindan ningún tipo de información sobre la distribución espacial de los átomos. A este tipo de índices se los conoce como descriptores moleculares topológicos.

1.4.5.1.1 Índices del átomo topoquímico ampliado

Los índices del átomo topoquímico ampliado (ETA) (Roy & Das, 2012; Roy & Ghosh, 2003; Roy et al., 2015b) son índices topológicos calculados a partir de un grafo molecular libre de hidrógenos, en el que cada vértice se considera compuesto por un núcleo y un ambiente electrónico de valencia. El recuento de núcleos (α_i) es un invariante de vértice local relacionado al abultamiento molecular. Los índices ETA contienen información relacionada con la naturaleza de los átomos, enlaces, entorno electrónico atómico, grupos funcionales, fragmentos moleculares y grados de ramificación molecular.

1.4.5.2 Fragmentos centrados en el átomo

Los fragmentos centrados en el átomo (Ghose, Viswanadhan, & Wendoloski, 1998; Viswanadhan, Ghose, Revankar, & Robins, 1989) llevan a cabo conteos de los distintos

fragmentos (tipos de átomos específicos) presentes en la estructura molecular, donde cada fragmento representa un átomo en la molécula descrito por sus átomos vecinos.

1.4.5.3 Descriptores geométricos

Este tipo de descriptores se derivan a partir de una representación 3D de la estructura molecular), es decir, a partir de un grafo 3D (índices topográficos) que considere las posiciones de los átomos y la conexión entre ellos. Para el cálculo de este tipo de descriptores es importante optimizar la geometría molecular a través de métodos computacionales mecano-cuánticos (Mauri et al., 2017). Los valores de estos descriptores dependen del tipo de optimización geométrica utilizada, la que presenta ciertos inconvenientes en cuanto a tiempo y costo computacional al trabajar con moléculas grandes y bases de datos extensas. Otro aspecto importante durante la optimización es la flexibilidad molecular, es decir, considerar los diferentes estados conformacionales en los cuales puede existir un compuesto químico (estudio conformacional).

1.4.5.3.1Descriptores de ensamblado de pesos de átomos, geometría y topología

Los descriptores GETAWAY (Geometry, Topology, and Atom-Weights Assembly) (Consonni, Todeschini, & Pavan, 2002; Consonni, Todeschini, Pavan, & Gramatica, 2002) se calculan a partir de la matriz de influencia molecular **H** , la cual contiene en sus elementos diagonales los valores de influencia de cada átomo de la molécula en la determinación de su forma; es decir, átomos cercanos al centro tendrán baja influencia y viceversa. Los valores de influencia son sensibles a cambios conformacionales significativos y a las longitudes de los enlaces, los cuales representan los tipos de átomos y la multiplicidad de enlaces.

1.4.5.3.2 Descriptores de carga y descriptores cuánticos

Los descriptores de carga o descriptores electrónicos (Roy et al., 2015; Todeschini & Consonni, 2009) describen la distribución de cargas en una molécula o regiones particulares de la misma, por ejemplo, átomos, enlaces, fragmentos moleculares. Es necesario que la molécula haya sido previamente optimizada mediante métodos semiempíricos. Las cargas eléctricas presentes en la molécula son la fuerza motriz de las interacciones electrostáticas, y es bien conocido que la densidad de electrones local o cargas tienen un rol importante en muchas propiedades fisicoquímicas y reacciones químicas (Roy et al., 2015a). Entre los descriptores cuánticos usados en estudios QSAR está (Doucet & Panaye, 2010; Karelson, Lobanov, & Katritzky, 1996):

1. Dipolo o momento dipolar: es un descriptor usado para describir la polaridad de una molécula; representando el comportamiento de resistencia y orientación de la misma en presencia de un campo electrostático. Se calcula usando las cargas atómicas parciales y las coordenadas atómicas.
2. Orbital molecular de más alta energía ocupado (HOMO): es el nivel de energía más alto en la molécula que contiene electrones y sirve para medir la nucleofilicidad de la misma. Cuando la molécula actúa como una base de Lewis (es decir, un donante de pares de electrones) en la formación de enlaces, los electrones son donados por ese orbital.
3. Orbital molecular no ocupado de más baja energía (LUMO): es el nivel de energía más bajo en la molécula que no contiene electrones y sirve para medir la electrofilicidad de la misma. Cuando la molécula actúa como un ácido de Lewis (aceptor de pares de electrones) en la formación de enlaces, los electrones entrantes se dirigen a este orbital.
4. Superdelocalizabilidad: es un descriptor de reactividad de los orbitales ocupados y desocupados. Indica la contribución realizada por un átomo especifico de la molécula a la energía de estabilización durante la formación de complejos de transferencia de carga con otra molécula; así como a la capacidad de un compuesto de formar enlaces a través de la transferencia de carga.
5. Polarizabilidad: es un descriptor que representa la habilidad de una molécula para formar dipolos instantáneos, y que se encuentra relacionado con la hidrofobicidad. Asimismo, la polarizabilidad electrónica de una molécula comparte características comunes con la superdelocalizabilidad electrofílica.

1.4.5.4 Descriptores topo-geométricos

Dentro de este grupo se encuentran aquellos descriptores que se calculan a partir de una representación topológica o tridimensional de la estructura molecular.

1.4.5.4.1 Pares de átomos

Corresponden a representaciones de cadena de la estructura química que consideran pares de átomos, a excepción de aquellos pares que contienen el átomo de hidrógeno, y la separación interatómica entre ellos. Dentro de este grupo existen los pares de átomos 2D que consideran la distancia topológica (por ejemplo, de 1 a 10) y los pares de átomos 3D que para medir la separación entre átomos usan la distancia Euclidiana (Carhart, Smith, & Venkataraghavan, 1985).

1.4.5.4.2 Descriptores de búsqueda de plantillas químicamente avanzadas

Los descriptores CATS (Chemically Advanced Template Search) se basan en la estructura bidimensional de una molécula. La definición del tipo de átomo en estos descriptores se relaciona con la presencia de potenciales farmacóforos (PPP), un tipo de átomo generalizado definido considerando ciertos aspectos fisicoquímicos. Para el cálculo de descriptores CATS se utilizan cinco PPP: donante de enlaces de hidrógeno (D), aceptor de enlaces de hidrógeno (A), cargados positivamente o ionizables (P), con carga negativa o ionizable (N) y lipofílicos (L). Por lo tanto, en los descriptores CATS 2D se le puede asignar cero, uno o dos tipos de PPP a cualquier átomo de una molécula y usar la distancia topológica entre 0 y 9 para medir la distancia entre ellos; en el caso de los descriptores CATS 3D no se puede asignar varios PPP y se usa la distancia euclidiana para medir la distancia en el espacio tridimensional (Fechner, Franke, Renner, Schneider, & Schneider, 2003).

1.5 Métodos Quimiométricos

La quimiometría se define como la ciencia de extraer información de las medidas realizadas en un sistema o proceso químico, mediante la aplicación de métodos matemáticos o estadísticos. Este término fue propuesto por Svante Wold en 1972 con el objetivo de describir la disciplina de obtener información importante de los experimentos químicos (Wold, 1995). Posteriormente se estableció una definición más precisa cuando se fundó la primera Sociedad de Quimiometría en 1974: disciplina química que usa la matemática, estadística y la lógica formal para diseñar o seleccionar procedimientos experimentales óptimos, maximizar la información química relevante de un análisis de datos químicos y obtener conocimientos de los sistemas químicos en estudio (Massart, Vandeginste, Buydens, Lewi, & Smeyers-Verbeke, 1997).

La quimiometría se fundamenta en el uso de un enfoque multivariado para explorar sistemas complejos químicos y diseñar racionalmente los experimentos. Debido a su naturaleza, los sistemas complejos necesitan varias variables para ser descritos y la quimiometría ofrece los métodos más adecuados para recopilar la mayor información de estos sistemas complejos. Las técnicas quimiométricas que se utilizan se caracterizan por analizar todas las variables simultáneamente, permitiendo así tener una visión completa del sistema en estudio (Wold, 2015).

1.5.1 Métodos no supervisados

1.5.1.1 Técnicas de reducción de variables

Cuando se tienen bases de datos con numerosas variables y en las que se puede encontrar correlación casual, ruido, redundancia o multicolinealidad los métodos de reducción no supervisados de variables son técnicas muy útiles, debido a que la presencia de estas variables insignificantes puede llegar a modificar el patrón de los datos e influenciar en los modelos finales. Es frecuente que este tipo de problemas van a estar presentes en un conjunto de datos QSAR, donde es de vital importancia que solamente los descriptores relevantes se retengan. Así se podrán obtener modelos de regresión o clasificación parsimoniosos que permiten predicciones confiables (Bagheri, Omidikia, & Kompany-Zareh, 2013). La reducción no supervisada de variables se realiza sin considerar la respuesta experimental (Consonni, Ballabio, Manganaro, Mauri, & Todeschini, 2009; Questier, Put, Coomans, Walczak, & Vander Heyden, 2005; Whitley, Ford, & Livingstone, 2000).

Estas técnicas funcionan como filtro para extraer moléculas según un criterio y de esta manera tener bases de datos depuradas ya solamente con variables relevantes, como consecuencia se reduce el costo computacional.

1.5.1.2 Método V-WSP

Este método es una modificación del algoritmo propuesto por Wootton, Sergente y Phan-Tan-Lu (WSP) para diseño de experimentos, que se basa en la selección de un subconjunto de variables de tal forma que se encuentren a una mínima correlación entre ellas en un espacio multidimensional (Ballabio et al., 2014). El algoritmo trabaja de la siguiente manera, dada una matriz de datos de $n \times p$, el método de reducción V-WSP realiza:

1. Se elige una variable inicial *i* como semilla y un valor umbral de correlación (*thr*).

2. Calcula el coeficiente lineal de correlación de Pearson (*R*) entre la variable *i* y todas las demás.

3. Elimina las variables *v* cuyo valor absoluto $R_{vi} \geq thr$

4. Se fija la variable *i* y se selecciona entre las variables restantes aquella que tenga la correlación absoluta más alta con *i*.

5. Repetir los pasos 2, 3 y 4 hasta que no existan variables para ser seleccionadas.

1.5.2 Técnicas de modelado de datos

1.5.3 Métodos supervisados

1.5.3.1 Métodos de regresión

Estos métodos se utilizan para procesar matrices de datos separando la parte más importante de la información para obtener modelos confiables pero complejos. El objetivo es predecir una variable continua ¨Y¨ a partir de una o varias variables explicativas (o covariables) ¨X¨, además de describir la estructura común subyacente de las variables (Rencher & Schaalje, 2008). Esta relación se expresa de la siguiente manera:

$$y = X\beta + e \tag{0.2}$$

donde:

y es la variable dependiente o el vector respuesta;

X es la matriz de las variables independientes del modelo;

β es el vector de los coeficientes de regresión;

e es el vector de errores que se comete en la predicción de los parámetros.

Un modelo matemático particular toma la siguiente forma:

$$y = X\,b \tag{0.3}$$

donde:

b es el vector de las estimaciones de los coeficientes verdaderos β;

y es el vector de las respuestas calculadas.

1.5.3.1.1 Mínimos cuadrados ordinarios

El método de mínimos cuadrados ordinarios (OLS: ordinary least squares) es una técnica de modelado lineal que minimiza la suma de los cuadrados entre las respuestas observadas en el conjunto de datos y las respuestas predichas por la aproximación lineal.(Rencher & Schaalje, 2008; Varmuza & Filzmoser, 2009)

1.5.3.1.2 Mínimos cuadrados parciales

El método de los mínimos cuadrados parciales (PLS: parcial least squares) es un método lineal (Pirouz, 2006) que presenta similitudes con el PCA (Principal Component

Analysis), el cual se utiliza como técnica de exploración de datos y no como técnica interpretativa. PCA trabaja con proyecciones de los objetos en un nuevo espacio definido por combinaciones lineales de las variables originales o variables latentes (LVs). Se considera un método más apropiado que el OLS cuando la relación observaciones/variables es menor a uno y cuando existe multicolinealidad en la matriz de diseño (Wold, Sjöström, & Eriksson, 2001).

1.5.3.2 Métodos de clasificación

1.5.3.2.1 *k*NN

El método de los *k*-vecinos más cercanos (*k*NN) es no lineal y no paramétrico, es decir, un objeto se clasifica en función de las clases a las que pertenecen la mayoría de los *k*-vecinos más cercanos en el espacio multidimensional de los datos. kNN usa la distancia Euclidiana existente entre el objeto a clasificar y los *k*-vecinos más cercanos. El valor de k varía normalmente entre 1 y 10. Estas distancias se ordenan de forma decreciente para luego clasificar un objeto en función de la clase a la que pertenecen la mayoría de los *k* vecinos. El valor óptimo de *k* se obtiene mediante validación cruzada, es decir, se evalúan los distintos valores de *k* y se selecciona aquel que brinde la mayor tasa de aciertos (NER: non-error rate) en validación cruzada (interna) (Cover & Hart, 1967). El algoritmo de clasificación *k*NN realiza las siguientes operaciones:

1. Escalado de datos

2. Selección de la distancia a usar

3. Optimización del número de vecinos *k*

4. Cálculo de la matriz de distancias

5. Cada objeto se clasifica según la clase más representativa de los *k* vecinos más cercanos.

1.5.4 Técnicas de selección de variables

1.5.4.1 Algoritmos genéticos

Los algoritmos genéticos (GAs) son una serie de pasos que describen el proceso para resolver problemas de búsqueda y optimización. Los GAs se aplican en la selección de variables en los modelos QSAR buscando optimizar un parámetro objetivo (R^2_{cv} en regresión y NERcv en clasificación). El proceso inicia con una población de "cromosomas" originada de forma casual, cada cromosoma se considera como un

Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

vector binario que cuenta con *p* bits (número total de variables) asociada a un modelo, donde un bit con valor 1 indica que dicha variable (descriptor) está presente en el modelo y 0 cuando no está presente en el modelo (Leardi, 2009). Durante el proceso evolutivo se dan dos etapas para generar nuevos cromosomas:

1. Reproducción o crossover: de toda la población se eligen dos padres que generan cromosomas hijos, estos comparten material genético de sus padres, así los bits con valores 0 o 1 se mantienen en los padres y aquellos bits con valores diferentes serán fijados en 0 o 1 de acuerdo a una regla de probabilidad.
2. Mutación: en este proceso los cromosomas se invierten generando mutantes, pero la posibilidad de que esto suceda es menor a la posibilidad de reproducción ya que se debe evitar alejar de la población que se está acercando al óptimo y con ello se evita que la población quedase atrapada en un mínimo local.

En el caso de bases de datos con gran número de variables se pueden presentar problemas como el sobreajuste de los modelos, para ello se ha propuesto realizar pocas corridas independientes a partir de varias poblaciones iniciales en lugar de una sola corrida simple (Leardi, 2009; Leardi & Gonzalez, 1998).

1.5.4.2 Parámetros de evaluación de los modelos de clasificación

Para realizar el cálculo de los parámetros de evaluación en los modelos de clasificación se utiliza una “matriz de confusión”, la cual se construye con las clases verdaderas y las clases predichas por el modelo. A partir de esta matriz se construyen los parámetros que permiten diagnosticar la calidad de los modelos de clasificación:

Precisión: se refiere a la capacidad del modelo de no introducir en la clase considerada objetos de otras clases. Se usa la siguiente ecuación:

$$\mathrm{Pr}_g = \frac{n_{gg}}{n'_g} \tag{0.4}$$

donde n_{gg} representa el número de elementos de la *g*-ésima clase perfectamente clasificados y n'_g es el número total de muestras que han sido asignadas a la *g*-ésima clase.

Sensibilidad: Es la capacidad del modelo para reconocer adecuadamente aquellos elementos que pertenecen a la *g*-ésima clase.

$$Sn_g = \frac{n_{gg}}{n_g} \tag{0.5}$$

donde n_g representa al número total de muestras que pertenecen a la *g*-ésima clase.

Especificidad: Se refiere a la capacidad de la *g*-ésima clase del modelo para refutar muestras diferentes a esta clase.

$$Sp_g = \frac{\sum_{k=1}^{G}\left(n_k^{'} - n_{gk}\right)}{n - n_g} \qquad para\, k \neq g \tag{0.6}$$

donde *n* es el número total de muestras y n'_k el número de muestras que han sido asignadas a la *k*-ésima clase, esta última se calcula aplicando la siguiente ecuación:

$$n^{'}_k = \sum_{g=1}^{G} n_{gk} \tag{0.7}$$

Si se consideran únicamente dos clases, la sensibilidad de la clase 1 pertenece a la especificidad de la clase 2 y viceversa.

Tasa de aciertos: Es el resultado del promedio de las sensibilidades de todas las clases, es de gran importancia ya que estima mejor la calidad de los modelos.

$$NER = \frac{\sum_{g=1}^{G} Sn_g}{G} \tag{0.8}$$

1.6 Técnicas de validación

Una vez que el modelo ha sido desarrollado se debe someter a diversas técnicas de validación, sea interna o externa. Durante la validación se somete al modelo a pequeñas perturbaciones para evaluar su estabilidad y verificar su capacidad predictiva.

Al aumentar la complejidad del modelo se corre el riesgo de sobreajustar los datos (aumenta la capacidad descriptiva), con lo que se reduce la capacidad de ser usado en predicción. Por este motivo, se necesitan de técnicas que permitan evaluar la presencia de sobreajuste y predictividad del mismo (Hawkins, 2004).

1.6.1 Validación cruzada o interna

Para la validación cruzada o interna se ha propuesto la técnica de *k*-grupos de validación cruzada (*k*-Fold Cross Validation), que consiste en dividir el grupo de entrenamiento en *k*-grupos de validación siguiendo una lógica (Hastie, Tibshirani, & Friedman, 2011). Normalmente, cada grupo *k* de validación se excluye una sola vez del modelo, se

recalibra el mismo y se usa para la predicción de las observaciones del grupo *k*. Las muestras se dividen en los *k*- grupos de validación según ventanas venecianas (venetian blinds), en el cual cada objeto del grupo de validación es seleccionado a partir del primer objeto en el grupo de entrenamiento y los siguientes cada *k*-ésimo objeto.

Cuando se trabaja en clasificación y las clases siguen un orden lógico, es conveniente trabajar con ventanas venecianas porque depende de la forma en que las clases están distribuidas en el vector respuesta (Ballabio & Consonni, 2013).

1.6.2 Validación externa

Para la validación externa, el conjunto de datos se divide de forma aleatoria y proporcional a la numerosidad de las clases en dos grupos, normalmente en el grupo de validación se coloca entre el 10% y 50% de los elementos.

1. Grupo de entrenamiento (training set): con este grupo se construye el modelo que se utilizará para predecir los objetos que constituyen el grupo de validación.
2. Grupo de validación (test set): Con el modelo desarrollado con el grupo de entrenamiento se predice la respuesta de los elementos del grupo de validación (Ballabio & Consonni, 2013).

1.7 Dominio de aplicabilidad

El dominio de aplicabilidad (AD) (Jaworska, Nikolova-Jeliazkova, & Aldenberg, 2005) se define como un espacio químico teórico definido por los descriptores moleculares y la actividad del grupo de calibración, dentro del cual se puede garantizar la predicción de nuevos compuestos. Es decir, el dominio de aplicabilidad define que tan similar es un compuesto del grupo de predicción con respecto a los del grupo de calibración que se usaron para construir el modelo (Dimitrov et al., 2005). En consecuencia, las predicciones serán confiables para únicamente los compuestos que caen dentro de este espacio teórico, caso contrario se consideran extrapolaciones sustanciales del modelo (predicciones no confiables).

Para los modelos de clasificación se ha definido el AD basado en la similitud *k*NN, en el cual se parte del cálculo de la distancia promedio que existe entre cada molécula del grupo de predicción con respecto a sus *k*-vecinos más cercanos del grupo de calibración y se compara dicho valor con el del umbral pre-definido (Sahigara, Ballabio, Todeschini, & Consonni, 2013; Sheridan, Feuston, Maiorov, & Kearsley, 2004).

CAPÍTULO 2. Metodología

2.1 Tipo de investigación

Este trabajo de investigación es de tipo teórico computacional.

2.2 Equipos y programas

2.2.1 Equipos

Computadoras

2.2.2 Programas

HyperChem versión 8.0.6

KNIME versión 3.7.2

alvaDesc versión 1.0.14

MarvinSketch versión 19.13

Open Babel GUI version 2.4.1

Matlab R2016

2.3 Métodos y técnicas de análisis

2.3.1 Generación del conjunto de datos

En este trabajo de titulación se desarrolló un conjunto de datos mediante una búsqueda exhaustiva de la literatura especializada, para identificar los diversos compuestos inhibidores de la enzima tirosinasa. Se ha consultado un total de 50 artículos científicos, que se detallan en el material anexo. Para cada compuesto químico se reporta el valor de concentración inhibitoria media máxima (IC_{50}) expresada en µmol/L, cabe mencionar que para llevar a cabo los procedimientos de la totalidad de artículos revisados para determinar la actividad difenolasa usan la enzima tirosinasa del hongo, pero bajo condiciones de temperatura diferente, así, aproximadamente el 90% de ellos realizan sus procedimientos a temperatura ambiente, el 8% a una temperatura de 37ºC y el 2% restante con diferentes temperaturas; para algunos compuestos se reporta también el nombre químico del compuesto, mientras que en los casos que no se reporta el nombre

se ha usado una notación de compuesto seguido de una numeración secuencial, en otros casos se ha indicado el nombre del compuesto derivado del que proviene seguido de un número secuencial. Para algunos compuestos se obtuvo el número de registro CAS y la notación de cadena SMILES canónico e isomérico. Esta notación permite describir sin ambigüedades la estructura de una molécula usando códigos SCII (American Standard Code for Information Interchange).

2.3.2 Representación de la estructura molecular

Los compuestos inhibidores de la enzima tirosinasa fueron representados en el programa HyperChem (Hypercube Inc.), donde se optimizaron las geometrías mediante los campos de fuerza de la mecánica molecular (MM+) y a continuación fueron refinadas sus coordenadas mediante el método semiempírico PM3. Para los dos métodos se usó el algoritmo de gradiente conjugado en la versión Polak-Ribiere y las coordenadas espaciales se optimizaron hasta que la desviación estándar del vector gradiente sea menor a 0.01kcal×(Å×mol)$^{-1}$.

2.3.3 Curado del conjunto de datos

Para el curado del conjunto de datos se utilizó el programa KNIME (Konstanz Information Miner), que es un programa quimioinformático muy útil cuando se trabaja con bases de datos de gran tamaño. En este programa se trabaja mediante la programación de diversos nodos, cada uno de los cuales realiza una o varias operaciones específicas (algoritmos). Estos nodos se ensamblan según las operaciones secuenciales que se requieren y forman un diagrama de flujo de trabajo que se detallan en la Figura 5.

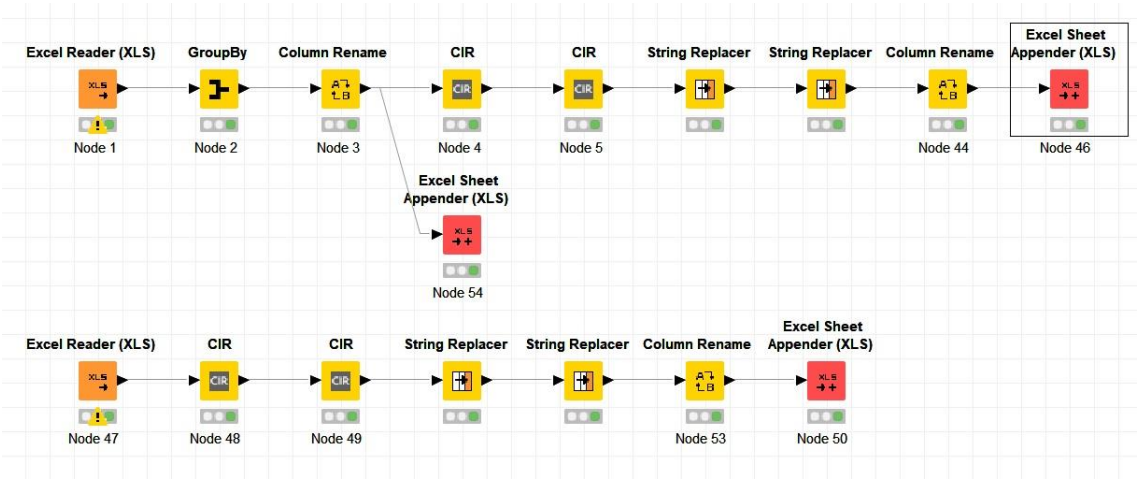


Figura 5. Diagrama de flujo KNIME para el filtrado y curado del conjunto de datos

El repositorio de nodos de KNIME se encuentra organizado por categorías y subcategorías. Para realizar un diagrama de flujo, los nodos requeridos se buscan y se importan en el editor del diagrama. A continuación, se conectan los nodos de forma secuencial de tal forma que se desarrollen las operaciones necesarias. De esta forma, se filtraron las moléculas que no presentaban un valor de (IC_{50}) pero que la bibliografía indicaba que tenían acción inhibitoria sobre la tirosinasa, las cuales posteriormente serán utilizadas para predecir su poder inhibitorio. Por otra parte, para las moléculas que presentaban el mismo nombre en distintas fuentes bibliográficas, pero con dos valores distintos de (IC_{50}) (por ejemplo, el oxiresveratrol, la hidroquinona, el resveratrol, el ácido ascórbico y la umbeliferona) se usó la media aritmética y para moléculas con 3 o más valores (por ejemplo, el ácido kójico y la arbutina) se aplicó el test de Dixon para identificar datos atípicos (outliers). Este test se basa en la relación de la diferencia entre: el valor que se sospecha es atípico y su vecino más próximo, con la diferencia de los valores más grandes y más pequeños en el grupo, para de esta manera obtener el valor atípico.

Para el test de Dixon se considera un conjunto de datos con n observaciones X_1(i= 1, 2, ……. n), arreglados en orden de magnitud. De acuerdo al tamaño de los datos se aplican las siguientes ecuaciones:

Para n ≥ 13

$$Q_{22} = \frac{(X_n - X_{n-2})}{(X_n - X_3)} \tag{2.1}$$

Para n = 3 a 10

$$Q_{10} = \frac{(X_n - X_{n-1})}{(X_n - X_1)} \tag{2.2}$$

El valor obtenido a partir de las ecuaciones se compara con un valor crítico *Q* reportado en la Tabla “Valores críticos de *Q* para la evaluación de valores atípicos” (Massart et al., 1997) y se consideran atípicos si este supera dicho valor.

2.3.4 Cálculo de descriptores moleculares

Con las moléculas optimizadas en HyperChem, se calcularon diversos tipos de descriptores moleculares en el programa alvaDesc (Alvascience Srl, 2019). Las familias

de descriptores incluidas en dicho programa son: índices constitucionales, descriptores de anillo, índices topológicos, número de trayectos moleculares, índices de conectividad molecular, índices de información, descriptores basados en la matriz 2D y 3D, autocorrelaciones 2D y 3D, descriptores de carga, descriptores tipo P_VSA, índices ETA, índices de Adyacencia de Arista, descriptores geométricos, descriptores RDF, descriptores 3D-MoRSE, descriptores WHIM, descriptores GETAWAY, perfiles moleculares de Randić, fragmentos centrados en el átomo y número de grupos funcionales, índices del estado Electrotopológico por tipo de átomo, descriptores farmacóforos, pares de átomos 2D y 3D, descriptores de carga, propiedades moleculares, descriptores CATS 3D.

Se excluyeron descriptores con valores constantes, casi constantes o con valores faltantes. Adicionalmente se analizaron mediante el método de reducción no supervisado de variables V- WSP con la finalidad de obtener una matriz con el menor número de descriptores que preserve la información sobre actividad inhibidora.

2.3.5 Métodos de modelamiento

Para el modelado de datos inicialmente se aplicó la técnica de regresión lineal múltiple con los métodos de OLS y PLS acoplados a los GAs para la selección supervisada de variables usando los valores de (IC_{50}) y el logaritmo del (log (IC_{50})). Posteriormente se utilizó el método de clasificación *k*-NN acoplado con los GAs. Para la separación de las clases se usó el percentil 33 y el percentil 50 del log (IC_{50}) como puntos iniciales, a partir de los cuales se optimizó mediante el método simplex. Este método se basa en el principio de movimiento de pasos lógicos realizados de forma secuencial establecido en el cambio simultáneo de variables. Así, el punto de partida se encuentra representado por dos puntos (P_1, P_2) llamados también vértices. A continuación, se aplica la ecuación 2.3 para obtener los puntos subsiguientes

$$X_i = X_{1c} + \alpha(X_{1c} - X_{1.1}) \qquad (2.3)$$

Donde:

X_i= Coordenada del umbral

X_{1c}= centroide de los dos mejores resultados

α = Factor de expansión (0,5)

$X_{1,1}$= Coordenada del umbral desechado

2.3.6 Validación del Modelo

Para la validación del modelo de clasificación *k*NN se realizó una validación interna o cruzada y una validación externa:

2.3.6.1 Validación cruzada

Aquí se definen *k*-grupos de validación cruzada a través de una secuencia lógica que permite la partición del grupo de calibración, es decir, ayuda a dividir dicho grupo en *k*-grupos de validación, los que se excluyen uno a la vez, recalibrar el modelo y posteriormente realizar la predicción de las moléculas excluidas. Un enfoque con ventanas venecianas, donde cada objeto del grupo de validación se selecciona a partir del primer objeto del grupo de calibración y los subsiguientes cada *k*-ésimo objeto.

2.3.6.2 Validación externa

Para la validación externa, el conjunto de datos se dividió de forma aleatoria y proporcional a la numerosidad de las clases en grupos de calibración y validación manteniendo una relación estructura-actividad en los dos grupos. Este algoritmo fue programado en MATLAB (The MathWorks Inc). Para esta validación, el conjunto de datos se divide en:

1. Grupo de calibración (training set): con este grupo se construye el modelo que se utilizará posteriormente para predecir los objetos que constituyen el grupo de validación.
2. Grupo de validación (test set): es un grupo externo que se utiliza para evaluar la capacidad predictiva del modelo, mediante predicción de la respuesta de sus elementos.

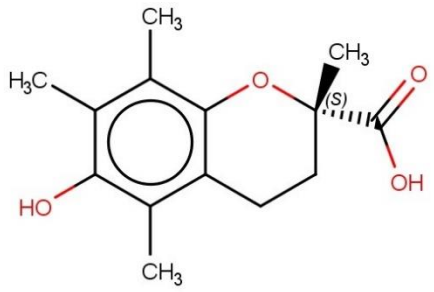
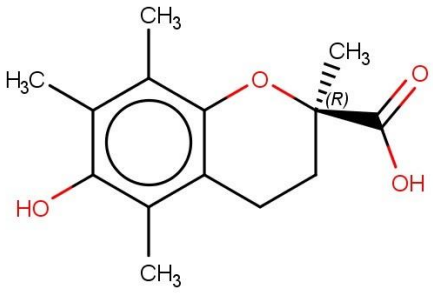
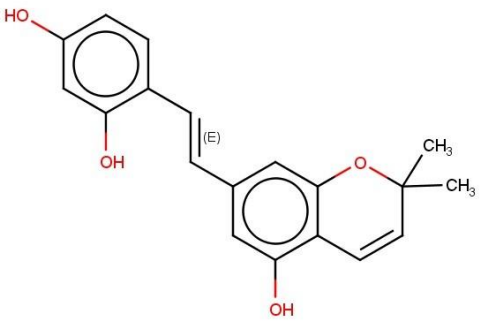
2.3.7 Dominio de aplicabilidad del modelo

El dominio de aplicabilidad (AD) (Jaworska, Nikolova-Jeliazkova, & Aldenberg, 2005) de un modelo QSAR se define como un espacio químico teórico dentro del cual el conjunto de calibración ha sido desarrollado y es aplicable con la finalidad de realizar la predicción de nuevos compuestos. Para los modelos de clasificación, un enfoque es basado en la similitud *k*NN entre los compuestos de calibración y predicción (Sahigara, Ballabio, Todeschini, & Consonni, 2013; Sheridan, Feuston, Maiorov, & Kearsley, 2004). Se parte del cálculo de la distancia promedio de cada molécula del conjunto de predicción con respecto a sus *k* vecinos más cercanos del conjunto de calibración, y se compara dicha

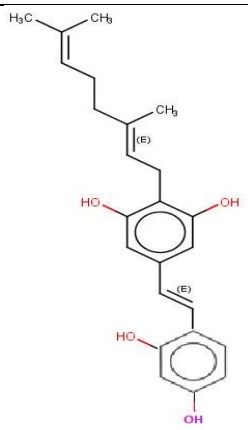
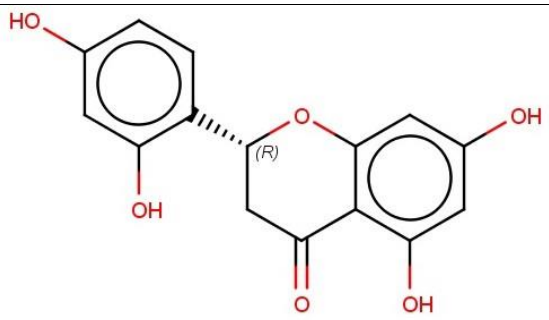
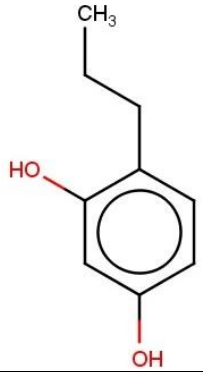
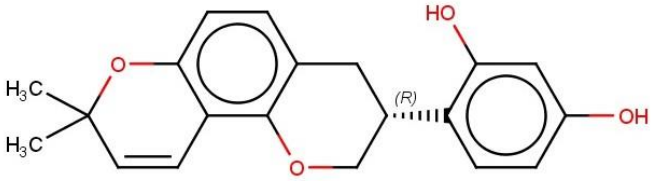
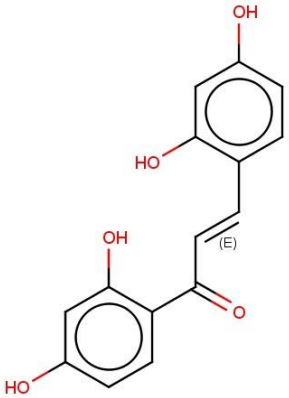
distancia promedio con un valor umbral pre-definido. Si la distancia promedio de una molécula del conjunto de predicción es menor a dicho umbral, la predicción de dicha molécula será confiable; caso contrario, su predicción será considerada una extrapolación del modelo. Para el cálculo del AD basado en similitud *k*NN se usó el AD toolbox programado en MATLAB (The MathWorks Inc).

2.3.8 Aplicación práctica del modelo

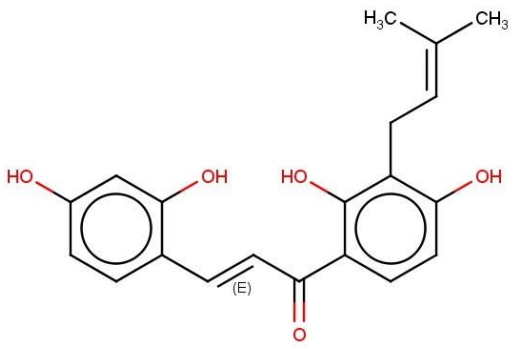
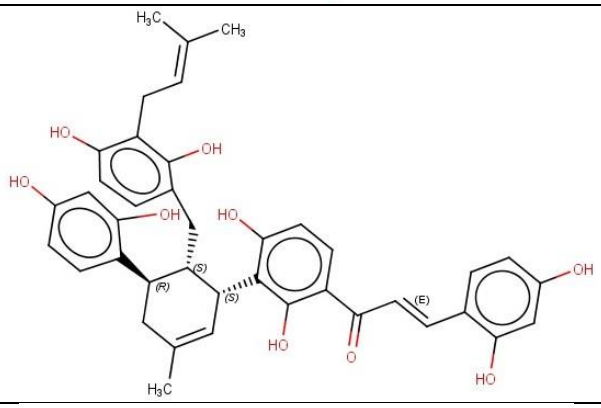
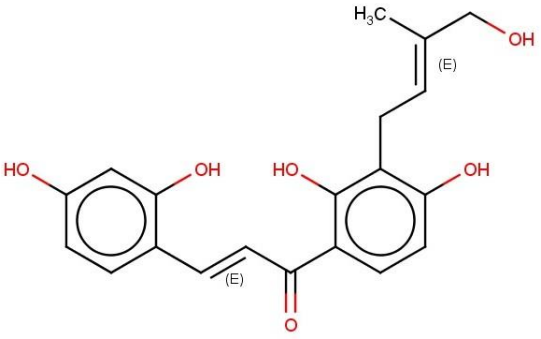
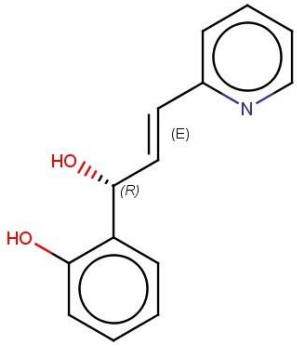
A partir del modelo desarrollado con las 581 moléculas fue posible realizar la estimación de la actividad para 19 moléculas del conjunto de datos inicial que no contaban con el valor reportado de la capacidad inhibitoria (IC_{50}). Las moléculas que se usaron para la predicción se presentan en la Tabla 2.

Ácido S-6-hidroxi-2,5,7,8-tetrametilcromo-2-carboxílico	
Ácido R-6-hidroxi-2,5,7,8-tetrametilcromo-2-carboxílico	
7-(2,4-Dihidroxiesteril)-5-hidroxi-2,2-metil-2H-1-benzopirano	

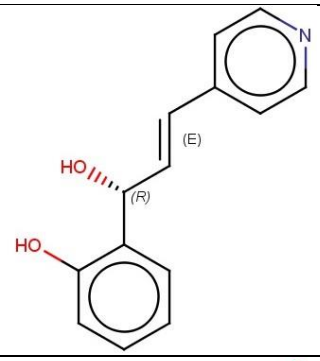
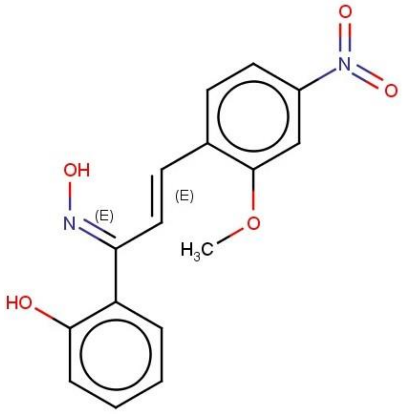
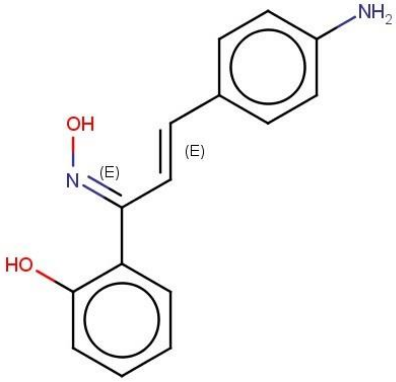
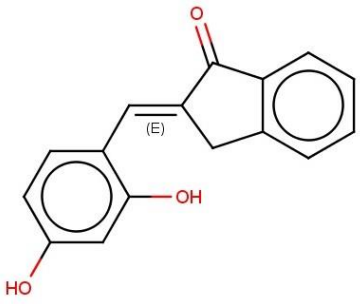
Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

4,4 '- [(1E) -1,2-Dietil-1,2-etanodiil] bisfenol	
2', 4', 5,7-Tetrahidroxiflavanona	
4-Propylresorcinol	
Glabridina	
Chalcona 9	

Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

Chalcona 10	
Chalcona 11	
Chalcona 12	
Azachalcona 13	

Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

Azachalcona 14	
Chalcona a base de oxima serie 15	
Chalcona a base de oxima serie 16	
Derivado chalconico 2,3-dihidro-1H-inden-1-onico 17	

Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

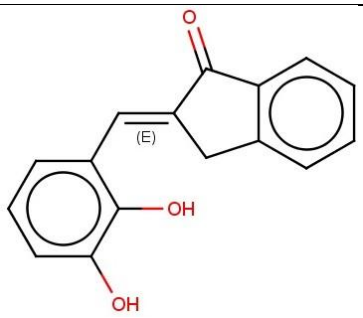
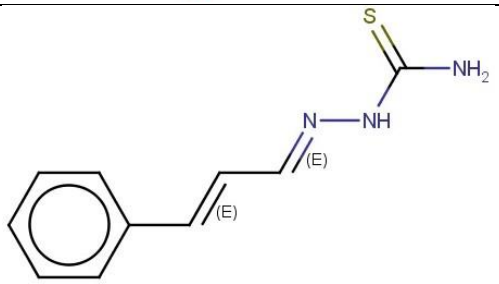
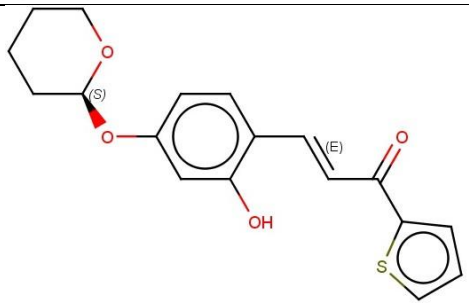
Derivado chalconico 2,3-dihidro-1H-inden-1-onico 18	
Trans-Cinnamaldheido thiosemicarbazona 17	
Cis-3-(2,4-Dihidroxifenil)-1-(tiofen-2-yl) prop-2-en-1-one	

Tabla 2. Moléculas usadas para la predicción de la actividad inhibitoria

Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

CAPÍTULO 3: Resultados y Discusiones

3.1 Resultados

3.1.1 Generación del conjunto de datos

El conjunto de datos inicial está constituido de 631 moléculas. Para las estructuras que fue posible, se verificó que la notación lineal de cadena SMILES coincida con el obtenido a partir del nombre químico o número del registro CAS en las diferentes bibliotecas químicas.

3.1.2 Representación de la estructura molecular

Las geometrías de las 631 moléculas se optimizaron mediante el método MM+, seguido del método semiempírico PM3, hasta que la desviación estándar del vector gradiente sea menor a 0.01 kcal×(Å×mol)$^{-1}$.

3.1.3 Curado del conjunto de datos

Se programaron diversos nodos en el programa de KNIME, así se filtraron 50 moléculas del conjunto de datos inicial de las cuales 19 fueron separadas debido a que no presentaban un valor de (IC_{50}); las 31 moléculas restantes fueron agrupadas mediante la aplicación de la media aritmética y con las ecuaciones 2.1 y 2.2 del test de Dixon se eliminaron los valores atípicos tanto para el ácido kójico como para la arbutina como se indica en la Tabla 3 y 4, respectivamente.

Ácido Kójico			
IC_{50}	**Q_{22}**	**Q_{crit}**	**$Q_{22} \geq Q_{crit}$**
6	0.083	0.489	-
54	0.015	0.489	-
318	0.857	0.478	Se elimina
934.3	0.951	0.468	Se elimina
1800	0.827	0.459	Se elimina

Tabla 3. Valores atípicos eliminados mediante el test de Dixon para el Ácido Kójico

Arbutina			
IC_{50}	**Q_{10}**	**Q_{crit}**	**$Q_{10} \geq Q_{crit}$**
10400	0.986	0.97	Se elimina

Tabla 4. Valores atípicos eliminados mediante el test de Dixon para la Arbutina

Como resultado, se obtuvo un conjunto de datos curada de 581 moléculas.

3.1.4 Cálculo de descriptores moleculares

En el programa alvaDesc se calcularon 5274 descriptores moleculares para cada molécula. En una primera etapa de filtración se eliminaron descriptores no informativos: 1541 descriptores constantes, 1646 casi constantes y 339 con al menos un valor faltante, cuyo criterio de clasificación se basa en que para filtrar los descriptores previo al análisis se define un umbral de desviación estándar 0.0001, por lo tanto, todos los descriptores con una desviación estándar igual o inferior al valor del umbral predeterminado serán excluidos automáticamente. Adicionalmente, se calcularon 166 huellas dactilares moleculares MACCS. En consecuencia, se fusionaron los descriptores filtrados y las huellas dactilares para generar 3559 descriptores.

3.1.5 Reducción no supervisada de descriptores moleculares

Los 3559 descriptores se analizaron mediante el método V-WSP definiendo un umbral de correlación de 0.95 (thr=0.95). De esta manera fue posible excluir 1867 descriptores moleculares correlacionados por encima de dicho umbral.

3.1.6 Selección Supervisada de Descriptores Moleculares

3.1.6.1 Método de regresión

Con los 1692 descriptores finales, en una primera etapa de modelado, se buscó desarrollar modelos de regresión lineal múltiple de mínimos cuadrados ordinarios y parciales acoplados con los algoritmos genéticos para la selección supervisada de descriptores. En estos modelos se usó el (IC_{50}) y el log (IC_{50}), posteriormente se dividieron las moléculas de manera aleatoria y proporcional a la numerosidad de las clases en grupos de calibración (407) y predicción (174) y durante la selección basada en los GAs se optimizó el coeficiente de determinación en validación cruzada de ventanas venecianas, para luego evaluar la calidad predictiva de cada modelo a través de la estimación de la actividad para el conjunto de predicción. Los modelos de regresión se presentan en la Tabla 5.

Modelo	d	LVs	R^2_{cal}	RMSEC	R^2_{cv}	$RMSEC_{cv}$	R^2_{pred}	RMSEP
IC_{50}-OLS	11	**	0,231	426,968	0,193	437,308	-5,965	1284,783

Katherine Alexandra Mogrovejo Mata
Doménica Victoria Muñoz Vázquez

IC_{50}-PLS	7	2	0,198	436,033	0,174	442,526	-5,978	1285,934
log (IC_{50})-OLS	7	**	0,479	0,861	0,463	0,875	0,341	0,969
log (IC_{50})-PLS	3	2	0,443	0,89	0,435	0,897	0,328	0,979

d: número de descriptores en el modelo**; LVs:** número de variables latentes**;** $\mathbf{R^2_{cal}}$**:** coeficiente de determinación en calibración**; RMSEC:** error cuadrático medio de calibración**;** $\mathbf{R^2_{cv}}$**:** coeficiente de determinación en validación cruzada**; RMSECcv:** error cuadrático medio de validación cruzada**;** $\mathbf{R^2_{pred}}$**: c**oeficiente de determinación en predicción**; RMSEP:** error cuadrático medio de predicción

Tabla 5. Resultados de los modelos de regresión QSAR para inhibidores de la enzima tirosinasa.

3.1.6.2 Métodos de Clasificación

Para el desarrollo del método de clasificación, se usó el valor de log IC_{50} para dividir las moléculas en compuestos de alta actividad (Clase 1) y baja actividad (Clase 2). Posteriormente, las moléculas se dividieron de manera aleatoria y proporcional a la numerosidad de las clases en grupos de calibración (285), validación (122) y predicción (174). Para las particiones iniciales se consideraron los valores de percentil 33 y el percentil 50 del log (IC_{50}), a partir de los cuales, y usando los conjuntos de calibración y validación, se usó el método Simplex para la separación de las clases obteniendo los mejores resultados para el punto tres como se indica en la Tabla 6. De esta forma se obtuvo un umbral óptimo de 0.4159 para separación de las dos clases.

Modelo	d	k	NER_{cal}	NER_{cv}	NERval
Opt. 1	2	1	0,72	0,73	0,77
Opt. 2	4	1	0,76	0,75	0,78
Opt. 3	**8**	**4**	**0,85**	**0,84**	**0,81**
Opt. 4	4	1	0,75	0,77	0,77

Tabla 6. Resultados del método simplex
Fuente: Los autores

Posteriormente, para el desarrollo del modelo *k*NN, se fusionó el grupo de calibración y validación para generar un nuevo conjunto de calibración de 407 moléculas. Con este grupo se aplicó nuevamente el método *k*NN acoplado con los GAs para obtener un modelo óptimo constituido por 4 vecinos (usando la distancia euclidiana) y 8 descriptores

moleculares. La calidad del modelo final fue evaluada considerando la tasa de aciertos (NER), la sensibilidad (*Sn*) y la especificidad (*Sp*) de las clases, tal como se indica en la Tabla 7 y, en la Tabla 8 se detallan los descriptores moleculares que forman parte del modelo.

	NER	**Sn**		**Sp**	
		Clase 1	Clase 2	Clase 1	Clase 2
Calibración	0.82	0.80	0.84	0.84	0.80
Validación	0.82	0.79	0.86	0.87	0.79
Predicción	0.86	0.85	0.87	0.87	0.85

Tabla 7. Parámetros de calidad del modelo QSAR basado en clasificación kNN

Nombre	**Descripción**	**Bloque**
F07[C-N]	Frecuencia de pares de átomos de C - N a una distancia topológica 7	Pares de átomos 2D
F02[N-S]	Frecuencia de pares de átomos de N - S a una distancia topológica 2	Pares de átomos 2D
N-069	Ar-NH2 / X-NH2	Fragmentos centrados en el átomo
R3v	Autocorrelación R a desplazamiento 3 / ponderado por el volumen de van der Waals	Descriptores GETAWAY
R8e	Autocorrelación R a desplazamiento 8 / ponderado por la electronegatividad de Sanderson	Descriptores GETAWAY
HATS4p	Autocorrelación a desplazamiento 4 ponderada por la matriz de influencia y la polarizabilidad	Descriptores GETAWAY
CATS2D_05_PL	CATS2D Positivo-Lipofílico a distancia topológica 5	Descriptores Farmacóforos
Eta_sh_y	Índice de forma	Indices ETA

Tabla 8. Descriptores moleculares incluidos en el modelo *k*NN

3.1.7 Validación del Modelo

La validación interna del modelo muestra buena estabilidad del mismo con un (NER_{cv}=0.82). Por otra parte, el modelo presenta buena capacidad predictiva (NER_{pred}=0.86), es decir, presenta 86% de probabilidad de predecir apropiadamente la actividad de nuevas moléculas con potencial actividad inhibitoria sobre la enzima tirosinasa.

3.1.8 Dominio de aplicabilidad del modelo

El dominio de aplicabilidad basado en similitud *k*NN definió un umbral de 1.5620. De esta manera se estable la región teórica definida por los descriptores moleculares dentro del cual las predicciones son confiables.

3.1.9 Predicción de moléculas

Una vez que el modelo ha sido validado, se lo utilizó para predecir la clase de las 19 moléculas para las cuales no existía reportada la actividad biológica. Se identificó que únicamente 2 moléculas caen fuera del AD (outlier): la Chalcona 10 y Chalcona a base de Oxima serie 15. Por otra parte, solo una molécula para la cual su predicción es confiable, es clasificada en la clase de alta actividad inhibitoria: trans-Cinnamaldehído thiosemicarbazona 17. Los resultados de las predicciones y del dominio de aplicabilidad se presentan en la Tabla 9.

Nombre	Clase
Ácido S-6-hidroxi-2,5,7,8-tetrametilcromo-2-carboxílico	2
Ácido R-6-hidroxi-2,5,7,8-tetrametilcromo-2-carboxílico	2
7-(2,4-Dihidroxiesteril)-5-hidroxi-2,2-metil-2H-1-benzopirano	2
4,4 '- [(1E) -1,2-Dietil-1,2-etanodiil] bisfenol	2
2', 4', 5,7- Tetrahidroxiflavanona	2
4-Propylresorcinol	2
Glabridina	2
Chalcona 9	2
Chalcona 10	2
Chalcona 11	2
Chalcona 12	2

Azachalcona 13	2
Azachalcona 14	2
Chalcona a base de oxima serie 15	2
Chalcona a base de oxima serie 16	2
Derivado chalconico 2,3-dihidro-1H-inden-1-onico 17	2
Derivado chalconico 2,3-dihidro-1H-inden-1-onico 18	2
Trans-Cinnamaldheido thiosemicarbazona 17	**1**
Cis-3-(2,4-Dihidroxifenil)-1-(tiofen-2-yl) prop-2-en-1-one	2

[a] moléculas fuera el AD del modelo

Tabla 9. Clase predicha para el conjunto externo de ITs

3.2 Discusiones

El curado del conjunto de datos inicial permitió excluir moléculas duplicadas, así como moléculas para las cuales se reportaban valores dispersos de (IC_{50}). De esta manera, al proveer un conjunto de datos filtrado de 581 estructuras moleculares permitió generar un modelo QSAR confiable (Golbraikh et al., 2017; Roy et al., 2015b). En este sentido, el programa KNIME resultó ser una herramienta apropiada para automatizar el curado de la información y de esta manera evitar el incluir errores que se hubiesen cometido al realizar el curado de forma manual (Berthold et al., 2008).

La reducción no supervisada de descriptores basado en el método V-WSP permitió excluir 1867 descriptores moleculares correlacionados por encima del 95% y de esta manera, al usar 1692 descriptores, se redujo el costo computacional durante la selección supervisada mediante los algoritmos genéticos acoplados con los modelos de regresión OLS y PLS, así como el método de clasificación *k*NN (Ballabio et al., 2014).

En una primera etapa se realizaron cuatro modelos de regresión mediante la combinación de las respuestas (IC_{50}) y log (IC_{50}) con los métodos OLS y PLS. Los resultados de la Tabla 5 indican que en los modelos el coeficiente de determinación R^2 (bondad de ajuste) es bajo. De hecho, los modelos de (IC_{50}) OLS e (IC_{50}) PLS presentan R^2_{pred} negativos. En consecuencia, los modelos obtenidos no resultan representativos de la realidad. Adicionalmente, se observa que los valores de error cuadrático medio RMSEC (mide la cantidad de error que existe entre los valores experimentales y predichos) (Balzarini et al., 2016; Hubert & Verboven, 2002) son altos, lo que confirma

la falta de ajuste y predictividad de los modelos de regresión para la actividad inhibitoria de nuevas dianas moleculares.

La estrategia adoptada en este trabajo, de dicotomizar la respuesta continua, en este caso el log (IC_{50}), para evaluarla mediante modelos locales de clasificación es una estrategia válida dentro del modelado *in silico* (Tripaldi et al., 2018). Por otra parte, los modelos de clasificación basados en similitudes locales son una buena estrategia cuando no existe una discriminación lineal de las clases en el espacio multidimensional de los datos (Rojas et al., 2016; Rojas, Duchowicz, Tripaldi, & Pis Diez, 2017). Debido a que no se conoce a priori el umbral más adecuado de separación, se utilizó el método de optimización simplex, el cual permitió identificar el umbral óptimo para discriminar las dos clases de compuestos (Tabla 6). Los resultados indican que el paso de optimización 3 fue la mejor debido a que genera la mayor tasa de aciertos en validación (NER_{val} = 0.81) (Turina, 1986). Este enfoqué permitió definir un umbral óptimo de log (IC_{50}) = 0.4159.

El umbral óptimo de separación se usó para dividir entre compuestos de alta actividad y baja actividad inhibitoria. Los resultados presentados en la Tabla 7 indican que el modelo que se obtuvo mediante algoritmos genéticos (GAs) acoplados con el método de clasificación de los *k*-vecinos más cercanos (*k*NN) son satisfactorios. De hecho, el modelo tiene buena capacidad de descripción de los datos (NER_{cal} = 0.82), estabilidad en validación interna de dejar-varios-fuera de ventanas venecianas (NER_{val} = 0.82) y principalmente buena capacidad predictiva (NER_{pred} = 0.86) (Ballabio & Consonni, 2013; Varmuza & Filzmoser, 2009). Estos resultados indican que el modelo no presenta sobreajuste y puede ser aplicado para predecir la actividad biológica de nuevas dianas moleculares, es decir, de compuestos que aún no han sido evaluados (barrido virtual de bibliotecas químicas) o sintetizados con una probabilidad de 86% de acierto.

El mecanismo de acción de los descriptores moleculares que forman parte del modelo indica que la separación entre las dos clases modeladas se encuentra descrita por tres descriptores GETAWAY: **R3v** indica que cada 3 enlaces existe una autocorrelación ponderada por el volumen de van der Waals, **R8e** muestra que por cada 8 enlaces existe una autocorrelación ponderada por la atracción que existe entre un electrón en la superficie de un átomo y su propio núcleo (electronegatividad de Sanderson) y **HATS4p** indica que cada 4 enlaces existe una autocorrelación ponderada por la polarizabilidad combinada con los elementos diagonales de la matriz de influencia. Los índices R consideran los elementos no diagonales de la matriz de influencia/distancia y en general,

los descriptores GETAWAY brindan información de la influencia de cada átomo para determinar la forma de la molécula y evalúa las interacciones entre los mismos con respecto a su posición geométrica en el espacio 3D. Por otra parte, se incluyen dos pares de átomos 2D: frecuencia de átomos C-N separados por una distancia topológica 7 (**F07[C-N]**) y frecuencia de átomos N-S a una distancia topológica 2 (**F02[N-S]**). Por otra parte, el descriptor **N-069** indica la presencia de grupos NH_2 (amina) unidos a un grupo aromático (Ar) o a cualquier heteroátomo (O, N, S, P, Se y halógenos), mientas que el descriptor **CATS2D_05_PL** indica la presencia de un átomo cargado positivamente (P) y un átomo lipofílico (L) separados por cinco enlaces. Finalmente, el índice de forma **Eta_sh_y** se deriva del recuento de núcleos centrales α_i, donde los vértices están unidos a tres átomos distintos al hidrógeno.

En lo que respecta al dominio de aplicabilidad del modelo, existen 17 moléculas que presentan un valor de distancia promedio menor al valor del umbral definido y por tanto se encuentran dentro del AD, lo que indica que las clases predichas son interpolaciones del modelo, es decir, existe suficiente similitud con los compuestos del grupo de calibración. Contrariamente, las 2 moléculas restantes tienen un valor de distancia promedio mayor al umbral definido encontrándose fuera del dominio de aplicabilidad y por tanto constituyen extrapolaciones del modelo y como consecuencia su predicción no es confiable. De los compuestos que están dentro del AD, únicamente el compuesto Trans-Cinnamaldehído Thiosemicarnazona 17 pertenece a la clase de alta actividad; no obstante, con este modelo, no es posible estimar el valor numérico del (IC_{50}).

De esta manera fue posible utilizar este modelo *in silico* para predecir la actividad de los 19 compuestos para los cuales no se reportó actividad inhibitoria (IC_{50}) en la literatura. Sin embargo, al ser un modelo de clasificación, únicamente permite conocer que tan activo es un compuesto, es decir, si tiene alto o bajo poder inhibitorio, al contrario de un modelo de regresión el cual permite obtener un valor numérico (Rojas, Duchowicz, & Castro, 2019; Rojas, Tripaldi, Pérez-González, Duchowicz, & Diez, 2018).

Otro aspecto importante durante el modelado QSAR es contrastar los resultados con modelos similares desarrollados para la misma actividad biológica, en este caso la capacidad inhibitoria sobre la enzima tirosinasa. Luego de una revisión exhaustiva de la literatura especializada, se han identificado cuatro estudios relacionados a estudios QSAR sobre la enzima tirosinasa (Tabla 10).

En el año 2007 Casañola-Martin y col realizaron siete modelos QSAR para 653 inhibidores de tirosinasas basado en análisis discriminante lineal (LDA), de los cuales 245 presentaban actividad inhibitoria. El conjunto de datos se dividió en grupos de calibración y predicción mediante el análisis de agrupamientos de *k*-medias. El modelo óptimo tiene una precisión del 0.99, una especificidad del 0.98 y una sensibilidad de 1. Más adelante, en el año 2010, estos mismos autores usaron una base de datos de 658 compuestos (246 activas y 412 inactivas) para desarrollar doce modelos QSAR basados en LDA, usando una combinación de descriptores Dragon y TOMOCOM-CARDD. Al igual que en su estudio previo, la base de datos se dividió en grupos de calibración y predicción usando el análisis de conglomerados (CA). Los índices cuadráticos basados en enlaces no estocásticos se usaron para construir seis modelos, mientras que las huellas dactilares moleculares estocásticas se utilizaron para los restantes. Finalmente, estos autores seleccionaron dos modelos, uno para cada enfoque, con una precisión de 0.93; 0.91, especificidad de 0.91; 0.87 y sensibilidad de 0.91; 0.90, respectivamente. En el mismo año, Marrero-Ponce y col. usaron una base de datos de 658 compuestos (183 activos y 295 inactivos) para desarrollar doce modelos QSAR basados en LDA, usando descriptores TOMOCOM-CARD. La base de datos se dividió en grupos de calibración (478 compuestos) y predicción (180 compuestos). Para desarrollar los primeros seis modelos usaron índices bilineales no estocásticos, mientras que para los seis modelos restantes utilizaron índices lineales estocásticos. Por último, estos autores seleccionaron dos modelos como los mejores, uno para cada enfoque, así para el primer modelo se obtuvo una precisión de 0.92 y 0.91, especificidad de 0.89 y 0.90, y sensibilidad de 0.91 y 0.86, respectivamente para los grupos de calibración y predicción; por el contrario, para el segundo modelo se obtuvo valores de precisión de 0.89; 0.90, especificidad de 0.85; 0.88, y sensibilidad de 0.85; 0.83 respectivamente, como se indica en la Tabla 10.

Por otra parte, Le Thi Thu y col. desarrollaron cuatro modelos QSAR aplicando diferentes métodos quimiométricos: análisis discriminante lineal (LDA), análisis discriminante cuadrático (QDA), regresión logística binaria (BLR) y árboles de clasificación (CART) utilizando el análisis de agrupamientos (CA) para dividir la base de datos de 1429 compuestos representados por descriptores 2D TOMOCOMD-CARDD. Estos autores concluyen que los mejores modelos son los que se basan en LDA y QDA, siendo este último el mejor de todos.

Referencia	Modelo	d	Calibración			Predicción		
			Pr	Sp	Sn	Pr	Sp	Sn
(Casañola-Martín et al., 2007)	LDA	3	0.99	0.99.	1	0.99	0.98	1
(Casañola-Martin et al., 2010)	LDA índices cuadráticos no estocásticos	8	0.93	0.91	0.91	0.90	0.81	0.92
	LDA huellas moleculares estocásticas	11	0.91	0.87	0.90	0.89	0.80	0.92
(Le-Thi-Thu et al., 2010)	LDA	11	0.91	0.92	0.89	0.91	0.92	0.89
	QDA	11	0.92	0.93	0.90	0.92	0.93	0.90
	BLR	11	0.91	0.92	0.90	0.88	0.88	0.88
	CART	15	0.91	0.94	0.90	0.89	0.89	0.89
(Marrero-Ponce et al., 2010)	LDA índices bilineales	-	0.92	0.89	0.91	0.91	0.90	0.86
	LDA índices lineales	-	0.89	0.85	0.85	0.90	0.88	0.83
Este trabajo	*k*NN	8	0,80	0,84	0,79	0,85	0,87	0,83

Tabla 10. Modelos de clasificación QSAR para la predicción de la capacidad inhibitoria de la enzima tirosinasa.

De los modelos presentados en la Tabla 10 se observa que los diferentes autores en sus investigaciones desarrollan modelos de clasificación utilizando métodos quimiométricos basados en LDA, QDA, BLR y CART para discriminar compuestos identificados como activos e inactivos. Además, utilizan programas como DRAGON y TOMOCOMD-CARD para calcular los descriptores moleculares y STATISTICA como software para desarrollar los modelos *in silico*. A diferencia de estas investigaciones, en este trabajo se han utilizado únicamente moléculas activas que han sido divididas en clases de alta y baja actividad para ser modeladas mediante el método no paramétrico de clasificación *k*NN. Otro aspecto fundamental que se debe indicar es que las estructuras moleculares fueron representadas por diversos descriptores moleculares del programa alvaDesc, los que permiten obtener un modelo con una capacidad predictiva aceptable (86%). Finalmente, este estudio puede complementar los modelos QSAR previamente realizados para que una vez que se identifique un potencial blanco molecular como activo, pueda ser predicha su clase dentro de los de alta capacidad inhibitoria de la enzima tirosinasa.

CAPÍTULO 4. Conclusiones

En esta investigación se determinó una relación cuantitativa entre la estructura molecular y la actividad de compuestos inhibidores de la enzima tirosinasa de 581 moléculas. El modelo se centró en el método de clasificación *k*-vecinos más cercanos (*k*NN) acoplado a los algoritmos genéticos (GAs) y se usó el método Simplex para optimizar la separación de las clases. El modelo final fue desarrollado de acuerdo con los principios establecidos por la Organización para la Cooperación y el desarrollo Económico (OECD) mostró un buen resultado en calibración, validación interna y validación externa (predicción). La relación QSAR sigue una regla simple de clasificación y brinda una explicación de los descriptores moleculares que sirven para predecir los compuestos de alta actividad inhibitoria. Así, el modelo permitió predecir la actividad de 19 compuestos para los cuales no se reportó el valor del (IC_{50}), lo cual indica su potencial uso para el diseño racional de nuevos compuestos inhibidores de la enzima tirosinasa con alta actividad inhibitoria. Finalmente, el mérito de esta investigación radica en que los resultados fueron presentados en el VII Congreso Latinoamericano de Plantas Medicinales (COLAPLAMED) “Plutarco Naranjo” desarrollado el mes de septiembre del 2019.

CAPÍTULO 5. Recomendaciones

- Se recomienda trabajar con bases de datos de gran tamaño, que permitan proporcionar una estimación de probable potencia de nuevos productos químicos como inhibidores de la tirosinasa.
- Utilizar moléculas externas con las cuales se podrá comprobar la validez del modelo, que pertenezcan a la misma clase que las del grupo de entrenamiento.
- En el presente trabajo de titulación se utilizó para la clasificación el método de *k*-NN o *k*- vecinos más cercanos, pero se recomienda el uso de otros métodos de clasificación con el objetivo de que la capacidad predictiva del modelo mejore.
- Impulsar el uso de QSAR y las herramientas quimiométricas en el área de la Farmacología con el fin de encontrar potenciales sustancias con acción biológica para el desarrollo racional de nuevos fármacos.

CAPÍTULO 6. Bibliografía

Alvascience Srl. (2019). alvaDesc (software for molecular descriptors calculation) version 1.0.12, https://www.alvascience.com.

Ashooriha, M., Khoshneviszadeh, M., Khoshneviszadeh, M., Moradi, S. E., Rafiei, A., Kardan, M., & Emami, S. (2019). 1, 2, 3-Triazole-based kojic acid analogs as potent tyrosinase inhibitors: Design, synthesis and biological evaluation. *Bioorganic chemistry, 82*, 414-422.

Bagheri, S., Omidikia, N., & Kompany-Zareh, M. (2013). Unsupervised Selection of Informative Descriptors in QSAR Study of Anti-HIV Activities of HEPT Derivatives. *Chemometrics and Intelligent Laboratory Systems, 128*, 135-143.

Balaban, A. T. (1985). Applications of Graph Theory in Chemistry. *Journal of chemical information and computer sciences, 25*(3), 334-343.

Ballabio, D., & Consonni, V. (2013). Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Analytical Methods, 5*(16), 3790-3798.

Ballabio, D., Consonni, V., Mauri, A., Claeys-Bruno, M., Sergent, M., & Todeschini, R. (2014). A Novel Variable Reduction Method Adapted from Space-Filling Designs. *Chemometrics and Intelligent Laboratory Systems, 136*, 147-154.

Balzarini, M. G., Di Rienzo, J., Tablada, M., Gonzalez, L. A., Bruno, C., Córdoba, M., . . . Casanoves, F. (2016). *Estadística y biometría: ilustraciones del uso de InfoStat en problemas de agronomía*: : Brujas

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., . . . Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 319-326). Germany: Springer.

Bonchev, D. (2015). On the Concept for Overall Topological Representation of Molecular Structure. In S. C. Basak, G. Restrepo, & J. L. Villaveces (Eds.), *Advances in Mathematical Chemistry and Applications* (pp. 42-75). Netherlands: Elsevier.

Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *Journal of chemical information and computer sciences, 25*(2), 64-73.

Casañola-Martín, G. M., Marrero-Ponce, Y., Khan, M. T. H., Ather, A., Khan, K. M., Torrens, F., & Rotondo, R. (2007). Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental in vitro assays. *European journal of medicinal chemistry, 42*(11-12), 1370-1381.

Casañola-Martin, G. M., Marrero-Ponce, Y., Khan, M. T., Khan, S. B., Torrens, F., Pérez-Jiménez, F., . . . Abad, C. (2010). Bond-Based 2D Quadratic Fingerprints in QSAR Studies: Virtual and In vitro Tyrosinase Inhibitory Activity Elucidation. *Chemical biology & drug design, 76*(6), 538-545.

Chang, T.-S. (2009). An updated review of tyrosinase inhibitors. *International journal of molecular sciences, 10*(6), 2440-2475.

Chang, Y.-H., Kim, C., Jung, M., Lim, Y.-H., Lee, S., & Kang, S. (2007). Inhibition of melanogenesis by selina-4 (14), 7 (11)-dien-8-one isolated from Atractylodis Rhizoma Alba. *Biological and Pharmaceutical Bulletin, 30*(4), 719-723.

Chaudhry, Q., Chrétien, J., Craciun, M., Guo, G., Lemke, F., Müller, J.-A., . . . Trundle, P. (2007). Algorithms for (Q)SAR Model Building. In E. Benfenati (Ed.), *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes* (pp. 111-147). The Netherlands: Elsevier B.V.

Chen, M.-J., Hung, C.-C., Chen, Y.-R., Lai, S.-T., & Chan, C.-F. (2016). Novel synthetic kojic acid-methimazole derivatives inhibit mushroom tyrosinase and melanogenesis. *Journal of bioscience and bioengineering, 122*(6), 666-672.

Cho, S. J., Roh, J. S., Sun, W. S., Kim, S. H., & Park, K. D. (2006). N-Benzylbenzamides: a new class of potent tyrosinase inhibitors. *Bioorganic & medicinal chemistry letters, 16*(10), 2682-2684.

Consonni, V., Ballabio, D., Manganaro, A., Mauri, A., & Todeschini, R. (2009). Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between Sets of Data: Part 2. Variable Reduction. *Analytica Chimica Acta, 648*(1), 52-59.

Consonni, V., Todeschini, R., & Pavan, M. (2002). Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *Journal of chemical information and computer sciences, 42*(3), 682-692.

Consonni, V., Todeschini, R., Pavan, M., & Gramatica, P. (2002). Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *Journal of chemical information and computer sciences, 42*(3), 693-705.

Cover, T., & Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE transactions on information theory, 13*(1), 21-27.

Cronin, M. T. (2010). Quantitative Structure-Activity Relationships (QSARs)-Applications and Methodology. In T. Puzyn, J. Leszczynski, & M. T. Cronin (Eds.), *Recent*

Advances in QSAR Studies: Methods and Applications (pp. 3-11). Germany: Springer Science+Business Media B.V.

Dearden, J. C. (2016). The History and Development of Quantitative Structure-Activity Relationships (QSARs). *International Journal of Quantitative Structure-Property Relationships, 1*(1), 1-44.

Dimitrov, S., Dimitrova, G., Pavlov, T., Dimitrova, N., Patlewicz, G., Niemela, J., & Mekenyan, O. (2005). A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *Journal of chemical information and modeling, 45*(4), 839-849.

Doucet, J. P., & Panaye, A. (2010). *Three Dimensional QSAR: Applications in Pharmacology and Toxicology*. USA: CRC Press.

Edwards, P. A., Anker, L. S., & Jurs, P. C. (1991). Quantitative structure-property relationship studies of the odor threshold of odor active compounds. *Chemical senses, 16*(5), 447-465.

Fechner, U., Franke, L., Renner, S., Schneider, P., & Schneider, G. (2003). Comparison of Correlation Vector Methods for Ligand-Based Similarity Searching. *Journal of Computer-Aided Molecular Design, 17*(10), 687-698.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but Verify: on the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of chemical information and modeling, 50*(7), 1189-1204.

Galvez, J., & Garcia-Domenech, R. (2010). Molecular Topology in QSAR and Drug Design Studies. In E. Castro (Ed.), *QSPR-QSAR Studies on Desired Properties for Drug Design* (pp. 63-94): Research Signpost.

Garcia, J., Duchowicz, P. R., & Castro, E. A. (2016). Considering the Molecular Conformational Flexibility in QSAR Studies. In A. G. Mercader, P. R. Duchowicz, & P. M. Sivakumar (Eds.), *Chemometrics Applications and Research: QSAR in Medicinal Chemistry* (pp. 129-158). USA: Apple Academic Press.

Gasteiger, J., & Engel, T. (2006). *Chemoinformatics: a textbook*: John Wiley & Sons.

Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (1998). Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *The Journal of Physical Chemistry A, 102*(21), 3762-3772.

Golbraikh, A., Wang, X. S., Zhu, H., & Tropsha, A. (2017). Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment. In J. Leszczynski, A. Kaczmarek-Kedziera, T. Puzyn, M. G. Papadopoulos, H. Reis,

& M. K. Shukla (Eds.), *Handbook of computational chemistry* (Second ed., Vol. 3, pp. 2303-2340). Switzerland: Springer International Publishing.

Guha, R., & Willighagen, E. (2012). A Survey of Quantitative Descriptions of Molecular Structure. *Current topics in medicinal chemistry, 12*(18), 1946-1956.

Hamzeh-Mivehroud, M., Sokouti, B., & Dastmalchi, S. (2015). An Introduction to the Basic Concepts in QSAR-Aided Drug Design. In K. Roy (Ed.), *Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment*: IGI Global.

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.). Germany: Springer-Verlag.

Hypercube Inc. HyperChem version 8, http://www.hyper.com.

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of chemical information and computer sciences, 44*(1), 1-12.

Hongmao, S. (2015). *A Practical Guide to Rational Drug Design*. UK: Woodhead Publishing/Elsevier.

Hubert, M., & Verboven, S. (2002). *A robust PCR method for high-dimensional regressors and several response variables.* Paper presented at the ICRM 2002: International Chemometrics Research Meeting, Location: Velthoven.

Janežič, D., Miličević, A., Nikolić, S., & Trinajstić, N. (2015). *Graph-Theoretical Matrices in Chemistry*. USA: CRC Press.

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR Applicability Domain Estimation by Projection of the Training Set Descriptor Space: A Review. *ATLA, 33*(5), 445–459.

Kaliszan, R. (2007). QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. *Chemical reviews, 107*, 3212-3246.

Karelson, M., Lobanov, V. S., & Katritzky, A. R. (1996). Quantum-chemical Descriptors in QSAR/QSPR Studies. *Chemical reviews, 96*(3), 1027-1044.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Shoemaker, B. A. (2015). PubChem substance and compound databases. *Nucleic acids research, 44*(D1), D1202-D1213.

Kubinyi, H. (2008). *QSAR: Hansch Analysis and Related Approaches* (Vol. 1). Germany: VCH.

Leardi, R. (2009). Genetic Algorithms. In R. Tauler, B. Walczak, & S. D. Brown (Eds.), *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Vol. 1, pp. 631-653). The Netherlands: Elsevier B.V.

Leardi, R., & Gonzalez, A. L. (1998). Genetic Algorithms Applied to Feature Selection in PLS Regression: How and When to Use Them. *Chemometrics and Intelligent Laboratory Systems, 41*(2), 195-207.

Le-Thi-Thu, H., Cardoso, G. C., Casañola-Martin, G. M., Marrero-Ponce, Y., Puris, A., Torrens, F., . . . Abad, C. (2010). QSAR models for tyrosinase inhibitory activity description applying modern statistical classification techniques: A comparative study. *Chemometrics and Intelligent Laboratory Systems, 104*(2), 249-259.

Leszczynski, J. (2012). *Handbook of computational chemistry*: Springer Science & Business Media.

Liaw, A., & Svetnik, V. (2015). QSAR Modeling: Prediction of Biological Activity from Chemical Structure. In A. L. Gould (Ed.), *Statistical Methods for Evaluating Safety in Medical Product Development* (pp. 66-83): Wiley.

Marrero-Ponce, Y., M Casanola-Martin, G., Tareq Hassan Khan, M., Torrens, F., Rescigno, A., & Abad, C. (2010). Ligand-based computer-aided discovery of tyrosinase inhibitors. Applications of the TOMOCOMD-CARDD method to the elucidation of new compounds. *Current pharmaceutical design, 16*(24), 2601-2624.

Massart, D. L., Vandeginste, B. G., Buydens, L., Lewi, P., & Smeyers-Verbeke, J. (1997). *Handbook of Chemometrics and Qualimetrics: Part A*. The Netherlands: Elsevier Science Inc.

Mauri, A., Consonni, V., & Todeschini, R. (2017). Molecular Descriptors. In J. Leszczynski, A. Kaczmarek-Kedziera, T. Puzyn, M. G. Papadopoulos, H. Reis, & M. K. Shukla (Eds.), *Handbook of computational chemistry* (Second ed., Vol. 3, pp. 2065-2093). Switzerland: Springer International Publishing.

McQuarrie, D. A., & Simon, J. D. (1997). *Physical chemistry: a molecular approach* (Vol. 1): University science books Sausalito, CA.

OECD, D. (2007). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q) SAR] Models. *Organisation for Economic Co-operation and Development, Paris, France*.

Pirouz, D. M. (2006). An overview of partial least squares. *Available at SSRN 1631359*.

Polansky, O. E. (1991). Elements of Graph Theory for Chemists. In D. Bonchev & D. H. Rouvray (Eds.), *Chemical Graph Theory: Introduction and Fundamentals* (Vol. 1, pp. 41-96). The Netherlands: Abacus Press.

Questier, F., Put, R., Coomans, D., Walczak, B., & Vander Heyden, Y. (2005). The Use of CART and Multivariate Regression Trees for Supervised and Unsupervised

Feature Selection. *Chemometrics and Intelligent Laboratory Systems, 76*(1), 45-54.

Randić, M. (1996). Molecular Bonding Profiles. *Journal of Mathematical Chemistry, 19*(3), 375-392.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear Models in Statistics*. USA: John Wiley & Sons, Inc.

Riley, P. A. (2000). Tyrosinase kinetics: a semi-quantitative model of the mechanism of oxidation of monohydric and dihydric phenolic substrates. *Journal of theoretical biology, 203*(1), 1-12.

Rojas, C., Duchowicz, P. R., Pis Diez, R., & Tripaldi, P. (2016). Applications of Quantitative Structure-Relative Sweetness Relationships in Food Chemistry. In A. G. Mercader, P. R. Duchowicz, & P. M. Sivakumar (Eds.), *Chemometrics Applications and Research: QSAR in Medicinal Chemistry* (pp. 317-339). USA: Apple Academic Press.

Rojas, C., Ballabio, D., Consonni, V., Tripaldi, P., Mauri, A., & Todeschini, R. (2016). Quantitative Structure–Activity Relationships to Predict Sweet and Non-Sweet Tastes. *Theoretical Chemistry Accounts, 135:66*, 1-13.

Rojas, C., Duchowicz, P. R., & Castro, E. A. (2019). Foodinformatics: Quantitative Structure-Property Relationship Modeling of Volatile Organic Compounds in Peppers. *Journal of Food Science, 84*(4), 770-781.

Rojas, C., Duchowicz, P. R., Tripaldi, P., & Pis Diez, R. (2017). Quantitative Structure–Property Relationships for Predicting the Retention Indices of Fragrances on Stationary Phases of Different Polarity. *Journal of the Argentine Chemical Society, 104*(2), 173-193.

Rojas, C., Tripaldi, P., Pérez-González, A., Duchowicz, P. R., & Diez, R. P. (2018). A Retention Index-Based QSPR Model for the Quality Control of Rice. *Journal of Cereal Science, 79*, 303-310.

Roy, K., & Das, R. N. (2012). On Extended Topochemical Atom (ETA) Indices for QSPR Studies. In E. A. Castro & A. K. Haghi (Eds.), *Advanced Methods and Applications in Chemoinformatics: Research Progress and New Applications* (pp. 380-412). USA: IGI Global.

Roy, K., & Ghosh, G. (2003). Introduction of Extended Topochemical Atom (ETA) Indices in the Valence Electron Mobile (VEM) Environment as Tools for QSAR/QSPR Studies. *Internet Electronic Journal of Molecular Design, 2*(9), 599-620.

Roy, K., Kar, S., & Das, R. N. (2015a). *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*. USA: Springer.

Roy, K., Kar, S., & Das, R. N. (2015b). *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. USA: Academic Press.

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Defining a Novel k-Nearest Neighbours Approach to Assess the Applicability Domain of a QSAR Model for Reliable Predictions. *Journal of cheminformatics, 5*(1), 27.

Sheridan, R. P., Feuston, B. P., Maiorov, V. N., & Kearsley, S. K. (2004). Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *Journal of chemical information and computer sciences, 44*(6), 1912-1928.

Testa, B., & Kier, L. B. (1991). The Concept of Molecular Structure in Structure-Activity Relationship Studies and Drug Design. *Medicinal research reviews, 11*(1), 35-48.

The MathWorks Inc. MatLab, http://www.mathworks.com.

Todeschini, R. (2003). Introduzione alla chemiometria.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics* (Second ed.). Germany: Wiley-VCH Verlag GmbH & Co.

Todeschini, R., Consonni, V., & Gramatica, P. (2009). Chemometrics in QSAR. In R. Tauler, B. Walczak, & S. D. Brown (Eds.), *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Vol. 4, pp. 129-170). The Netherlands: Elsevier B.V.

Tripaldi, P., Pérez-González, A., Rojas, C., Radax, J., Ballabio, D., & Todeschini, R. (2018). Classification-based QSAR Models for the Prediction of the Bioactivity of ACE-inhibitor Peptides. *Protein and peptide letters, 25*(11), 1015-1023.

Turina, S. (1986). 2 Optimization in Chromatographic Analysis. *Planar chromatography, 1*, 15.

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. USA: CRC press.

Velásquez, M., Drosos, J., Gueto, C., Márquez, J., & Vivas-Reyes, R. (2013). Método acoplado Autodock-PM6 para seleccionar la mejor pose en estudios de Acoplamiento Molecular. *Revista Colombiana de Química, 42*(1)

Viswanadhan, V. N., Ghose, A. K., Revankar, G. R., & Robins, R. K. (1989). Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for

Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *Journal of chemical information and computer sciences, 29*(3), 163-172.

Whitley, D. C., Ford, M. G., & Livingstone, D. J. (2000). Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *Journal of chemical information and computer sciences, 40*(5), 1160-1168.

Wold, S. (1995). Chemometrics; What do We Mean with It, and What do We Want from It? *Chemometrics and Intelligent Laboratory Systems, 30*(1), 109-115.

Wold, S. (2015). Chemometrics and Bruce: Some Fond Memories. In B. K. Lavine, S. D. Brown, & K. S. Booksh (Eds.), *40 Years of Chemometrics - From Bruce Kowalski to the Future* (Vol. 1199, pp. 1-13). USA: American Chemical Society.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109-130.

Xu, Y., Stokes, A. H., Freeman, W. M., Kumer, S. C., Vogt, B. A., & Vrana, K. E. (2009). Tyrosine mRNA is expressed in human substantia nigra. *Molecular brain research, 45*(1), 159-162.

ANEXOS

Lista de los artículos científicos revisados para la generación del conjunto de datos:

1. Ashooriha, M., Khoshneviszadeh, M., Khoshneviszadeh, M., Moradi, S. E., Rafiei, A., Kardan, M., & Emami, S. (2019). 1, 2, 3-Triazole-based kojic acid analogs as potent tyrosinase inhibitors: Design, synthesis and biological evaluation. *Bioorganic chemistry*, *82*, 414-422.
2. Casañola-Martín, G. M., Marrero-Ponce, Y., Khan, M. T. H., Ather, A., Khan, K. M., Torrens, F., & Rotondo, R. (2007). Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental in vitro assays. *European journal of medicinal chemistry*, *42*(11-12), 1370-1381.
3. Casañola-Martin, G. M., Marrero-Ponce, Y., Khan, M. T., Khan, S. B., Torrens, F., Pérez-Jiménez, F., ... & Abad, C. (2010). Bond-Based 2D Quadratic Fingerprints in QSAR Studies: Virtual and In vitro Tyrosinase Inhibitory Activity Elucidation. *Chemical biology & drug design*, *76*(6), 538-545.
4. Chang, T. S. (2009). An updated review of tyrosinase inhibitors. *International journal of molecular sciences*, *10*(6), 2440-2475.
5. Chang, T. S., Ding, H. Y., & Lin, H. C. (2005). Identifying 6, 7, 4′-trihydroxyisoflavone as a potent tyrosinase inhibitor. *Bioscience, biotechnology, and biochemistry*, *69*(10), 1999-2001.
6. Chen, M. J., Hung, C. C., Chen, Y. R., Lai, S. T., & Chan, C. F. (2016). Novel synthetic kojic acid-methimazole derivatives inhibit mushroom tyrosinase and melanogénesis. *Journal of bioscience and bioengineering*, *122*(6), 666-672.
7. Cho, S. J., Roh, J. S., Sun, W. S., Kim, S. H., & Park, K. D. (2006). N-Benzylbenzamides: a new class of potent tyrosinase inhibitors. *Bioorganic & medicinal chemistry letters*, *16*(10), 2682-2684.
8. Cui, H. X., Duan, F. F., Jia, S. S., Cheng, F. R., & Yuan, K. (2018). Antioxidant and Tyrosinase Inhibitory Activities of Seed Oils from Torreya grandis Fort. ex Lindl. *BioMed research international*, *2018*.
9. da Silva, A. P., Silva, N. D. F., Andrade, E. H. A., Gratieri, T., Setzer, W. N., Maia, J. G. S., & da Silva, J. K. R. (2017). Tyrosinase inhibitory activity, molecular docking studies and antioxidant potential of chemotypes of Lippia origanoides (Verbenaceae) essential oils. *PloS one*, *12*(5), e0175598.

10. Hałdys, K., Goldeman, W., Jewgiński, M., Wolińska, E., Anger, N., Rossowska, J., & Latajka, R. (2018). Inhibitory properties of aromatic thiosemicarbazones on mushroom tyrosinase: Synthesis, kinetic studies, molecular docking and effectiveness in melanogenesis inhibition. *Bioorganic chemistry*, *81*, 577-586.
11. Ingle, S. S., & Khobragade, C. N. (2013). In silico drug docking and screening for the drug discovery of new tyrosinase inhibitors. *Journal of Pharmacy Research*, *6*(7), 704-708.
12. JARA, J. R., SOLANO, F., & LOZANO, J. A. (1988). Assays for mammalian tyrosinase: a comparative study. *Pigment cell research*, *1*(5), 332-339.
13. Khan, M. T. H., Choudhary, M. I., Khan, K. M., & Rani, M. (2005). Structure–activity relationships of tyrosinase inhibitory combinatorial library of 2, 5-disubstituted-1, 3, 4-oxadiazole analogues. *Bioorganic & medicinal chemistry*, *13*(10), 3385-3395.
14. Khatib, S., Nerya, O., Musa, R., Shmuel, M., Tamir, S., & Vaya, J. (2005). Chalcones as potent tyrosinase inhibitors: the importance of a 2, 4-substituted resorcinol moiety. *Bioorganic & medicinal chemistry*, *13*(2), 433-441.
15. Kim, C., Noh, S., Park, Y., Kang, D., Chun, P., Chung, H., ... & Moon, H. (2018). A Potent Tyrosinase Inhibitor, (E)-3-(2, 4-Dihydroxyphenyl)-1-(thiophen-2-yl) prop-2-en-1-one, with Anti-Melanogenesis Properties in α-MSH and IBMX-Induced B16F10 Melanoma Cells. *Molecules*, *23*(10), 2725.
16. Kim, Y. J., No, J. K., Lee, J. H., & Chung, H. Y. (2005). 4, 4′-Dihydroxybiphenyl as a new potent tyrosinase inhibitor. *Biological and Pharmaceutical Bulletin*, *28*(2), 323-327.
17. Kubo, I., & Kinst-Hori, I. (1999). 2-Hydroxy-4-methoxybenzaldehyde: a potent tyrosinase inhibitor from African medicinal plants. *Planta Medica*, *65*(01), 019-022.
18. Kubo, I., Kinst-Hori, I., Chaudhuri, S. K., Kubo, Y., Sánchez, Y., & Ogura, T. (2000). Flavonols from Heterotheca inuloides: tyrosinase inhibitory activity and structural criteria. *Bioorganic & medicinal chemistry*, *8*(7), 1749-1755.
19. Lai, J. S., Lin, C., & Chiang, T. M. (2014). Tyrosinase inhibitory activity and thermostability of the flavonoid complex from Sophora japonica L (Fabaceae). *Tropical Journal of Pharmaceutical Research*, *13*(2), 243-247.
20. Le-Thi-Thu, H., Cardoso, G. C., Casañola-Martin, G. M., Marrero-Ponce, Y., Puris, A., Torrens, F., ... & Abad, C. (2010). QSAR models for tyrosinase inhibitory activity description applying modern statistical classification techniques:

A comparative study. *Chemometrics and Intelligent Laboratory Systems*, *104*(2), 249-259.

21. Li, W., & Kubo, I. (2004). QSAR and kinetics of the inhibition of benzaldehyde derivatives against Sacrophaga neobelliaria phenoloxidase. *Bioorganic & medicinal chemistry*, *12*(4), 701-713.
22. Lin, Y. S., Chen, H. J., Huang, J. P., Lee, P. C., Tsai, C. R., Hsu, T. F., & Huang, W. Y. (2017). Kinetics of tyrosinase inhibitory activity using Vitis vinifera leaf extracts. *BioMed research international*, *2017*.
23. Liu, J., Cao, R., Yi, W., Ma, C., Wan, Y., Zhou, B., ... & Song, H. (2009). A class of potent tyrosinase inhibitors: alkylidenethiosemicarbazide compounds. *European journal of medicinal chemistry*, *44*(4), 1773-1778.
24. Liu, J., Yi, W., Wan, Y., Ma, L., & Song, H. (2008). 1-(1-Arylethylidene) thiosemicarbazide derivatives: A new class of tyrosinase inhibitors. *Bioorganic & medicinal chemistry*, *16*(3), 1096-1102.
25. Maisuthisakul, P., & Gordon, M. H. (2009). Antioxidant and tyrosinase inhibitory activity of mango seed kernel by product. *Food chemistry*, *117*(2), 332-341.
26. Mann, T., Gerwat, W., Batzer, J., Eggers, K., Scherner, C., Wenck, H., ... & Kolbe, L. (2018). Inhibition of human tyrosinase requires molecular motifs distinctively different from mushroom tyrosinase. *Journal of Investigative Dermatology*, *138*(7), 1601-1608.
27. Marrero-Ponce, Y., M Casanola-Martin, G., Tareq Hassan Khan, M., Torrens, F., Rescigno, A., & Abad, C. (2010). Ligand-based computer-aided discovery of tyrosinase inhibitors. Applications of the TOMOCOMD-CARDD method to the elucidation of new compounds. *Current pharmaceutical design*, *16*(24), 2601-2624.
28. Masamoto, Y., Ando, H., Murata, Y., Shimoishi, Y., Tada, M., & Takahata, K. (2003). Mushroom tyrosinase inhibitory activity of esculetin isolated from seeds of Euphorbia lathyris L. *Bioscience, biotechnology, and biochemistry*, *67*(3), 631-634.
29. Matos, MJ, Santana, L., Uriarte, E., Delogu, G., Corda, M., Fadda, MB, ... y Fais, A. (2011). Nuevas fenilcumarinas halogenadas como inhibidores de la tirosinasa. *Cartas de química bioorgánica y medicinal*, *21* (11), 3342-3345.
30. Matsuura, R., Ukeda, H., & Sawamura, M. (2006). Tyrosinase inhibitory activity of citrus essential oils. *Journal of agricultural and food chemistry*, *54*(6), 2309-2313.

31. Nerya, O., Musa, R., Khatib, S., Tamir, S., & Vaya, J. (2004). Chalcones as potent tyrosinase inhibitors: the effect of hydroxyl positions and numbers. *Phytochemistry*, *65*(10), 1389-1395.
32. Nguyen, H. X., Nguyen, N. T., Nguyen, M. H. K., Le, T. H., Van Do, T. N., Hung, T. M., & Nguyen, M. T. T. (2016). Tyrosinase inhibitory activity of flavonoids from Artocarpus heterophyllous. *Chemistry Central Journal*, *10*(1), 2.
33. Park, JW, Ha, YM, Moon, KM, Kim, SR, Jeong, HO, Park, YJ, ... & Byun, Y. (2013). Inhibidor de la tirosinasa de novo: 4- (6, 7-dihidro-5H-indeno [5, 6-d] tiazol-2-il) benceno-1, 3-diol (MHY1556). *Cartas de química bioorgánica y medicinal*, *23* (14), 4172-4176.
34. Rescigno, A., Casañola-Martin, G. M., Sanjust, E., Zucca, P., & Marrero-Ponce, Y. (2011). Vanilloid Derivatives as Tyrosinase Inhibitors Driven by Virtual Screening-Based QSAR Models. *Drug testing and analysis*, *3*(3), 176-181.
35. Shiino, M., Watanabe, Y., & Umezawa, K. (2001). Synthesis of N-substituted N-nitrosohydroxylamines as inhibitors of mushroom tyrosinase. *Bioorganic & medicinal chemistry*, *9*(5), 1233-1240.
36. Shiino, M., Watanabe, Y., & Umezawa, K. (2003). Synthesis and tyrosinase inhibitory activity of novel N-hydroxybenzyl-N-nitrosohydroxylamines. *Bioorganic Chemistry*, *31*(2), 129-135.
37. Shin, N. H., Ryu, S. Y., Choi, E. J., Kang, S. H., Chang, I. M., Min, K. R., & Kim, Y. (1998). Oxyresveratrol as the potent inhibitor on dopa oxidase activity of mushroom tyrosinase. *Biochemical and biophysical research communications*, *243*(3), 801-803.
38. Song, Y. M., Ha, Y. M., Kim, J. A., Chung, K. W., Uehara, Y., Lee, K. J., ... & Moon, H. R. (2012). Synthesis of novel azo-resveratrol, azo-oxyresveratrol and their derivatives as potent tyrosinase inhibitors. *Bioorganic & medicinal chemistry letters*, *22*(24), 7451-7455.
39. Tan, X., Song, Y. H., Park, C., Lee, K. W., Kim, J. Y., Kim, D. W., ... & Park, K. H. (2016). Highly potent tyrosinase inhibitor, neorauflavane from Campylotropis hirtella and inhibitory mechanism with molecular docking. *Bioorganic & medicinal chemistry*, *24*(2), 153-159.
40. TH Khan, M. (2012). Novel tyrosinase inhibitors from natural resources–their computational studies. *Current medicinal chemistry*, *19*(14), 2262-2272.
41. Therdphapiyanak, N., Jaturanpinyo, M., Waranuch, N., Kongkaneramit, L., & Sarisuta, N. (2013). Development and assessment of tyrosinase inhibitory activity

of liposomes of Asparagus racemosus extracts. *asian journal of pharmaceutical sciences*, *8*(2), 134-142.

42. Uchida, R., Ishikawa, S., & Tomoda, H. (2014). Inhibition of tyrosinase activity and melanine pigmentation by 2-hydroxytyrosol. *Acta Pharmaceutica Sinica B*, *4*(2), 141-145.
43. Vontzalidou, A., Zoidis, G., Chaita, E., Makropoulou, M., Aligiannis, N., Lambrinidis, G., ... & Skaltsounis, A. L. (2012). Design, synthesis and molecular simulation studies of dihydrostilbene derivatives as potent tyrosinase inhibitors. *Bioorganic & medicinal chemistry letters*, *22*(17), 5523-5526.
44. Xie, W., Zhang, H., He, J., Zhang, J., Yu, Q., Luo, C., & Li, S. (2017). Synthesis and biological evaluation of novel hydroxybenzaldehyde-based kojic acid analogues as inhibitors of mushroom tyrosinase. *Bioorganic & medicinal chemistry letters*, *27*(3), 530-532.
45. Yoshimori, A., Oyama, T., Takahashi, S., Abe, H., Kamiya, T., Abe, T., & Tanuma, S. I. (2014). Structure–activity relationships of the thujaplicins for inhibition of human tyrosinase. *Bioorganic & medicinal chemistry*, *22*(21), 6193-6200.
46. You, A., Zhou, J., Song, S., Zhu, G., Song, H., & Yi, W. (2015). Rational design, synthesis and structure–activity relationships of 4-alkoxy-and 4-acyloxy-phenylethylenethiosemicarbazone analogues as novel tyrosinase inhibitors. *Bioorganic & medicinal chemistry*, *23*(5), 924-931.