UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

“Avances en los estudios de relación estructura-actividad cuantitativa de compuestos con relevancia epigenética”

T E S I S

QUE PARA OBTENER EL TÍTULO DE

QUÍMICO FARMACÉUTICO BIÓLOGO

PRESENTA:

Mario Omar García Sánchez

Ciudad Universitaria, CDMX **2017**

JURADO ASIGNADO:

PRESIDENTE: **Dr. Francisco Hernández Luis**

VOCAL: **Dra. Elena Guadalupe Ramírez López**

SECRETARIO: **Dr. José Luis Medina Franco**

SUPLENTE 1: **Dr. Alberto Ortega Vázquez**

SUPLENTE 2: **Dra. Sara Margarita Garza Aguilar**

SITIO DONDE SE DESARROLLÓ EL TEMA:

El presente trabajo se realizó en los cubículos 305 y 108 DIFACQUIM, Diseño de Fármacos Asistido por Computadora, Edificio F, Departamento de Farmacia, Facultad de Química de la Universidad Nacional Autónoma de México (UNAM).

ASESOR DEL TEMA:

_____

Dr. José Luis Medina Franco

SUPERVISOR TÉCNICO:

_____

Dr. Eli Antonio Alonso Fernández de Gortari

SUSTENTANTE:

_____

Mario Omar García Sánchez

AGRADECIMIENTOS

A la Dirección General de Asuntos del Personal Académico (DGAPA) de la Universidad Nacional Autónoma de México (UNAM) por la beca para titulación otorgada dentro del Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) con clave IA204016: “Reposicionamiento de fármacos asistido por computadora en la identificación de moduladores de DNA metiltransferasas” y a la Facultad de Química a través del Programa de Apoyo a la Investigación y al Posgrado (PAIP 5000-9163).

PUBLICACIÓN

Parte de los resultados de investigación de este proyecto están publicados en el capítulo de libro:

> **García-Sánchez, MO**, Cruz-Monteagudo M, Medina-Franco JL. Quantitative Structure-Epigenetic Activity Relationships. K. Roy, Editor. Volume *Advances in QSAR modeling with Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*". Book series "Challenges and Advances in Computational Chemistry and Physics", ISSN: 2542-4491. Springer, (**2017**) en prensa.

La carta de aceptación y las primeras páginas del capítulo se encuentran en el Apéndice A.1.

ÍNDICE GENERAL

ABREVIATURAS GENERALES i

ÍNDICE DE FIGURAS Y TABLAS ii

RESUMEN iii

INTRODUCCIÓN 1

MARCO TEÓRICO 2

2.1 Quimioinformática 2

2.1.1 Estudio de las relaciones estructura-actividad 2

2.1.2 Modelado de panoramas de actividad (a*ctivity landscape modeling*) 3

2.1.3 Mapas de Similitud Estructura-Actividad (*Structure-activity similiarity maps*) 4

2.1.3.1 Similitud estructural y de actividades 5

2.1.3.2 Acantilados de actividad (a*ctivity cliffs*) 6

2.2 Epigenética 7

2.2.1 Tipos de modificaciones epigenéticas 7

2.2.2 Bromodominios (BRDs) 8

2.2.3 Inhibidores de Bromodominios (iBRDs) 9

OBJETIVOS 12

3.1 Objetivo general 12

3.2 Objetivos particulares 12

METODOLOGÍA 13

4.1 Curado de la base de datos 13

4.2 Modelado de los panoramas de actividad (*SAS maps*) 13

4.3 Análisis de las relaciones estructura-actividad de iBRDs 14

4.3.1 *Scaffolds* moleculares 14

4.3.2 Análisis del SAR global 15

4.3.3 Análisis de *activity cliffs* 15

4.3.3.1 Análisis individuales, hacia un solo BRD: BRD2, BRD3 y BRD4 16

4.3.3.2 Análisis pareado: BRD2-BRD3, BRD2-BRD4 y BRD3-BRD4 17

4.3.3.3 Análisis simultáneo hacía los tres receptores BRD2-BRD3-BRD4 18
4.4 Elaboración del *script* para la automatización del análisis estructura-actividad 18
RESULTADOS Y DISCUSIÓN 20
5.1 Base de datos 20
5.2 Modelado de los panoramas de actividad (*SAS maps*) 20
5.3 Análisis de las relaciones estructura-actividad de iBRDs 23
5.3.1 *Scaffolds* moleculares 23
5.3.2 Análisis del SAR global 24
5.3.3 Análisis de *activity cliffs* 25
5.3.3.1 Análisis individuales, hacia un solo BRD: BRD2, BRD3 y BRD4 25
5.3.3.2 Análisis pareado: BRD2-BRD3, BRD2-BRD4 y BRD3-BRD4 32
5.3.3.3 Análisis simultáneo hacía los tres receptores BRD2-BRD3-BRD4 34
CONCLUSIONES 35
PERSPECTIVAS 36
REFERENCIAS 37
APÉNDICE 41
A.1 Carta de aceptación y primeras páginas del capítulo de libro: *Quantitative Structure-Epigenetic Activity Relationships* 41
A.2 Bases de datos (IC_{50} para BRD2, BRD3 y BRD4) y referencia para SARI 44
A.3 *Script* elaborado en Python 3.5.2 47

ABREVIATURAS GENERALES

AC	*Activity cliffs*
ACG	*Activity cliffs generators*
ALM	*Activity landscape modeling*
Cmpd(s)	Compuesto(s)
Ct	Coeficiente de Tanimoto
|DifAct|	Valor absoluto de las diferencias entre actividades
DAD	*Dual-activity difference*
BET	*Bromodomain and Extra-Terminal*
BRD(s)	Bromodominio(s)
ECFP	*Extended Connectivity Fingerprints*
iBRDs	Inhibidores de los Bromodominios
IC_{50}	*Inhibitory concentration 50*
Kac	Lisina(K) acetilada
QSAR	*Quantitative structure-activity relationship*
SALI	*Structure-Activity Landscape Index*
SAR	*Structure-activity relationship*
SARI	*Structure-Activity Relationship Index.*
SAS	*Structure-activity similarity*
TAD	*Triple activity-difference*

ÍNDICE DE FIGURAS Y TABLAS

Figura 1. Modelado de panoramas de actividad

Figura 2. Representación gráfica general del mapa de similitud estructura-actividad

Figura 3. Representación gráfica de la distancia de enlace calculada por ECFP

Figura 4. Estructura general de los BRDs y modo de unión con la Kac

Figura 5. DAD *map* y las zonas que lo constituyen

Figura 6. TAD *map* y las zonas que lo constituyen

Figura 7. Diagrama de flujo de la automatización del análisis

Figura 8. SAS *maps* de los iBRDs (BRD2, BRD3 y BRD4)

Figura 9. Frecuencia de scaffolds moleculares de los 88 iBRDs

Figura 10. Frecuencia de los *activity cliffs generators* (ACG)

Figura 11. *Deep* AC de los iBRD2

Figura 12. *Deep* AC de los iBRD3

Figura 13. *Deep* AC de los iBRD4 con **Cmpd_87**

Figura 14. *Deep* AC de los iBRD4 con **Cmpd_37**

Figura 15. *Dual-Activity Difference maps* (DAD *maps*)

Figura 16. *Triple-Activity Difference map* (TAD *map*)

Tabla 1. Tipos de activity cliffs

Tabla 2. Modificaciones epigenéticas más comunes

Tabla 3. iBRDs de la familia BET en ensayos clínicos

Tabla 4. Criterios de clasificación de los activity cliffs

Tabla 5. Resumen general de las bases de datos utilizadas

Tabla 6. Porcentajes de las combinaciones en las diferentes regiones del SAS *map*

Tabla 7. Valores de SARI obtenidos para cada conjunto de compuestos iBRDs

Tabla 8. Cantidad de *activity cliffs* y categoría a la que pertenecen

RESUMEN

Se realizó el modelado de panoramas de actividad de un conjunto de compuestos inhibidores de Bromodominio de la familia BET (BRD2, BRD3 y BRD4). El estudio se hizo mediante la representación gráfica de mapas de similitud estructura-actividad (SAS *maps*) con el objetivo de identificar en forma sistemática acantilados de actividad (*activity cliffs*). Los compuestos fueron obtenidos de una base de datos pública (ChEMBL). Para automatizar el análisis se desarrolló e implementó un *script* en Python 3.5.2. Se analizó y cuantificó el tipo de núcleos base presentes en el grupo de datos, el tipo de SAR global y las diferentes regiones del SAS *map*. De las diferentes regiones del SAS *map* se realizó énfasis en el análisis de la región de *activity cliffs* (AC), debido a que dicha zona contiene pares de compuestos con alta similitud estructural y gran diferencia entre las actividades. Con la identificación de los AC se determinó la frecuencia de los generadores de AC (ACG) los cuales pueden ser utilizados en la optimización de modelos QSAR. Los pares de compuestos identificados como AC fueron clasificados en dos categorías, medianas y pronunciadas (*shallow* y *deep* AC), de acuerdo a las diferencias de actividad. Además, se realizó otra clasificación, basada en la concentración de los inhibidores para identificar a los compuestos como: inactivos, de actividad intermedia e inactivos. Para aquellos AC con modificaciones en la actividad muy pronunciadas (*deep AC)* y en dónde al menos un compuesto era activo, se analizaron los cambios subestructurales asociados a los de actividad. Finalmente se realizó un breve análisis con dos y tres dianas simultáneamente, con la finalidad de encontrar compuestos selectivos a una de ellas. Esta tesis es una aportación a la *Epi-informática*, especialmente para analizar cuantitativamente relaciones estructura actividad-epigenética.

1. INTRODUCCIÓN

En los últimos años las herramientas informáticas han tomado gran valor en la ciencia ya que dan soporte a la investigación científica y disminuyen tiempo y costos. Una de las primeras etapas en el descubrimiento de nuevos fármacos es la de seleccionar aquellos compuestos, entre cientos o miles, que cuenten con las características químico-farmacéuticas deseadas (*e.g.*, buena eficacia, potencia, afinidad, especificidad, etc.) hacia algún receptor relacionado con uno o varios problemas de salud. Con ayuda de programas computacionales y datos experimentales es posible discriminar y optimizar aquellas estructuras de utilidad terapéutica, y que con ello puedan pasar a las siguientes fases de desarrollo [1].

El presente trabajo tiene como objeto de estudio a los receptores bromodominios (BRDs) BRD2, BRD3 y BRD4; dichas proteínas están constituidas por aproximadamente 110 aminoácidos y se caracterizan por ser "lectores" epignéticos ya que identifican modificaciones post-traduccionales, específicamente la acetilación en los residuos de lisina en las colas de las histonas. Estas modificaciones alteran el estado de condensación de la cromatina y por lo tanto la expresión de genes [2]. La importancia terapéutica de los bromodominios mencionados anteriormente radica en que su inhibición permite la generación de nuevos tratamientos contra ciertos tipos de cáncer, enfermedades cardíacas e incluso en el proceso inflamatorio [3].

La búsqueda de compuestos que permiten inhibir a estos receptores ha aumentado en los últimos años, al igual que la cantidad de estructuras sintetizadas e información de pruebas biológicas. En este trabajo, a partir de información previamente publicada en bases de datos, se realizaron estudios de relación estructura-actividad cuantitativos y se identificaron subestructuras que permitan la optimización de los iBRD (isoformas 2,3 y 4). Para este fin también se desarrolló un procedimiento quimioinformático automatizado de relaciones estructura-actividad utilizando el concepto de panoramas de actividad.

2. MARCO TEÓRICO

2.1 Quimioinformática

La Quimioinformática es una disciplina que surgió recientemente y se basa en la trasformación de un gran número de datos en información y dicha información en conocimiento, esto con la intención de tomar mejores y rápidas decisiones en el área de optimización e identificación de fármacos [4]. Para poder llevar a cabo lo anterior los químicos realizan una serie de experimentos, los analizan, y buscan patrones comunes; esto con el objetivo de desarrollar modelos y colocar dichas observaciones en un esquema sistemático que permita llevar a cabo predicciones y corroborar los resultados de nuevos experimentos [5]. Una aplicación que tienen las herramientas quimioinformáticas es en el análisis automatizado de las relaciones-estructura actividad.

2.1.1 Estudio de las relaciones estructura-actividad

El estudio de las relaciones estructura-actividad o también conocida por sus siglas en inglés SAR (*Structure-activity relationship*), se basa en el análisis de la relación entre la estructura química y la actividad biológica. Para llevar a cabo la exploración y obtener información del "SAR" se pueden utilizar diferentes modelos que se basan en el ligando, tales como: QSAR (*Quantitative structure-activity relationship*), modelo del farmacóforo o bien modelado de panoramas de actividad (a*ctivity landscape modeling*).

El análisis cuantitativo de las relaciones estructura-actividad (QSAR) se enfoca en encontrar un modelo estadístico que proporcione una ecuación matemática la cual permita correlacionar las propiedades biológicas (descriptores) con una estructura dentro de una familia de compuestos; la finalidad es predecir propiedades como bioactividad, logP, toxicidad, entre otras [6].

Respecto al farmacóforo, en 1998, la IUPAC (*International Union of Pure and Applied Chemistry*) lo definió como: "El conjunto de características electrónicas y estéricas que son necesarias para asegurar la interacción supramolecular óptima con una diana específica y desatar (o bloquear) la respuesta biológica". El modelo es una herramienta que extrae información de las interacciones que tienen las conformaciones de las moléculas, más predichas, del ligando con el receptor.

La tercera herramienta que nos permite obtener información del SAR es el modelo de panoramas de actividad, el cual se caracteriza por ser cualquier representación gráfica que integre el análisis de la similitud estructural y la diferencia de actividad entre compuestos que comparten el mismo efecto biológico [7].

2.1.2 Modelado de panoramas de actividad (*activity landscape modeling*)

El primer tipo de representación gráfica generado para el modelado de los panoramas de actividad fueron los mapas de Similitud Estructura-Actividad (*Structure-activity similarity*, SAS *maps*), que son gráficos en dos dimensiones (2D) de los cuales se hablará en la sección 2.1.3. El segundo tipo está basado en panoramas en tres dimensiones (3D), el cual integra un descriptor molecular como un nuevo grado de libertad. A partir de estos modelos es posible identificar distintos tipos de SAR. Los principales tipos son (Figura 1): 1) SAR discontinuo, pequeños cambios estructurales en los compuestos que producen alteraciones dramáticas en la actividad (regiones con acantilados y muy irregulares). 2) SAR continuo, cambios estructurales en los compuestos que generan modificaciones moderadas en las actividades (regiones continuas y sin grandes alteraciones) y 3) Heterogéneo, que representa tanto al SAR continuo como discontinuo. Otra representación gráfica está basada en redes que se forman entre moléculas, los nodos (intersecciones) representan compuestos individuales y las aristas son relaciones de similitud, una aplicación de esto son los NSG (*Network-like similarity graphs*). En los cuales se coloca como límite de similitud estructural 80% u 85% (Figura 1) [8].

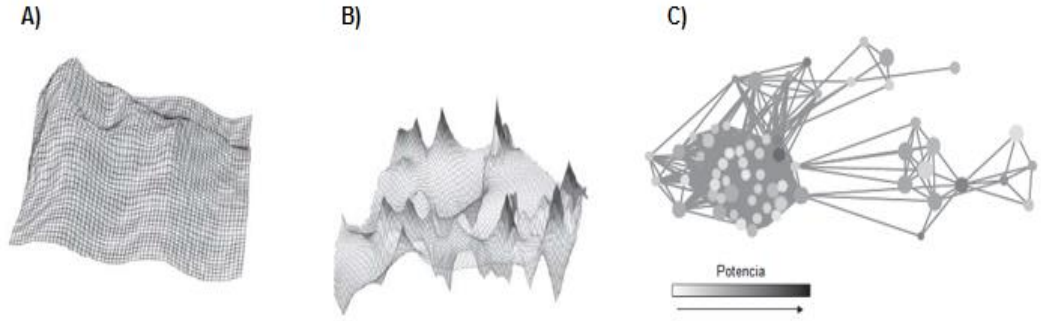


Figura 1. Modelado de panoramas de actividad. (A) Representación gráfica 3D del SAR continuo. (B) Gráfico en 3D del SAR discontinuo de un conjunto de inhibidores de acetilcolinesterasa y (C) representación de los gráficos NSG de una base de datos de inhibidores de tirosincinasa (los nodos indican compuestos y las uniones relaciones de similitud estructural). Tomado de Bajorath, J. Expert Opin. Drug Discov. 2012, 7:463-73.

2.1.3 Mapas de Similitud Estructura-Actividad (*structure-activity similiarity maps*)

En 2001, Shanmugasundaram y Maggiora introdujeron un nuevo concepto para la representación gráfica de los panoramas de actividad, al que denominaron mapa de relación estructura-actividad [9], el cual compara la similitud estructural y de actividad que existe entre pares de compuestos (Figura 2). El mapa se puede dividir en cuatro regiones principales: I) *Scaffold* o *R-hopping*, en donde predominan pares de compuestos tanto con baja similitud estructural como baja diferencia entre actividades. II) SAR continuo, que son pares con alta similitud estructural y baja diferencia entre sus actividades. III) Acantilados de actividad (*activity cliffs*), comparaciones de compuestos con alta similitud estructural y alta diferencia entre actividades. La región IV) No descriptiva, la cual contiene compuestos con baja similitud estructural pero grandes diferencias de actividad [10].

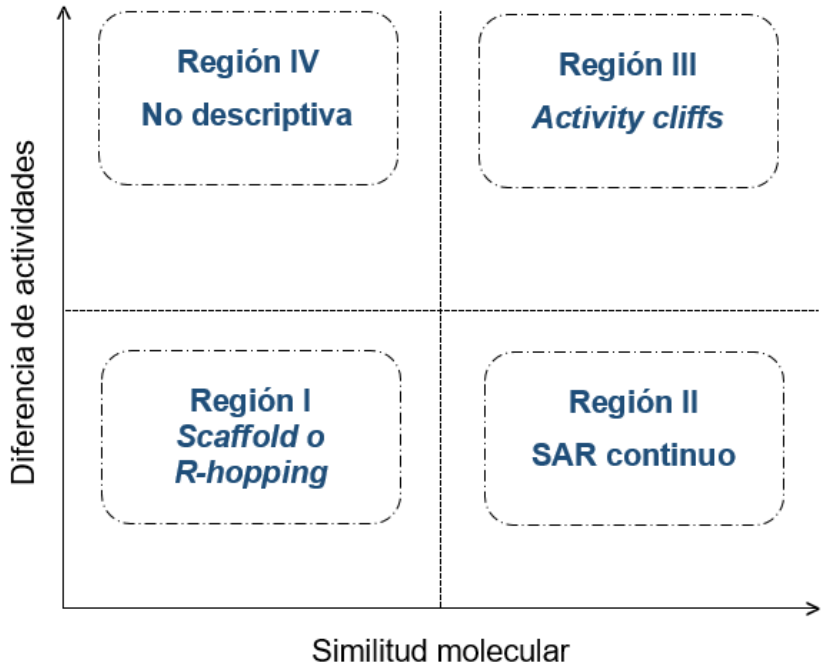


Figura 2. Representación gráfica general del mapa de similitud estructura-actividad y las regiones de las que está formado. Modificado de Medina-Franco, J.L. J. Chem. Inf. Model. 2012, 52:2485-2493.

Para medir la similitud estructural y de actividad se pueden utilizar varios métodos. De la manera en la que las similitudes son calculadas, depende el valor de los límites para definir las diferentes regiones del *SAS map*.

2.1.3.1 Similitud estructural y de actividades

La similitud estructural puede ser obtenida a partir de la representación de la estructura (en 2D o 3D) con ayuda de huellas digitales moleculares (*fingerprints*), descriptores obtenidos de modelos farmacofóricos o de propiedades fisicoquímicas. Las huellas digitales se pueden calcular de dos maneras: 1) Basadas en diccionario y 2) Basadas en topología. Las basadas en diccionario utilizan colecciones de fragmentos estructurales o características de los mismos y mediante un código binario indican "1" presencia o "0" ausencia; un ejemplo de ello son los diccionarios (de 166 o 322 características) de MACCS (*Molecular ACCess System*) *keys* [11]. En las basadas en topología se identifican los fragmentos de acuerdo a la distancia de enlace elegida (Figura 3), un ejemplo de esto son los ECFP (*Extended Connectivity FingerPrints*), que de manera similar a los basados en diccionario es producido un vector con valores binarios [12].

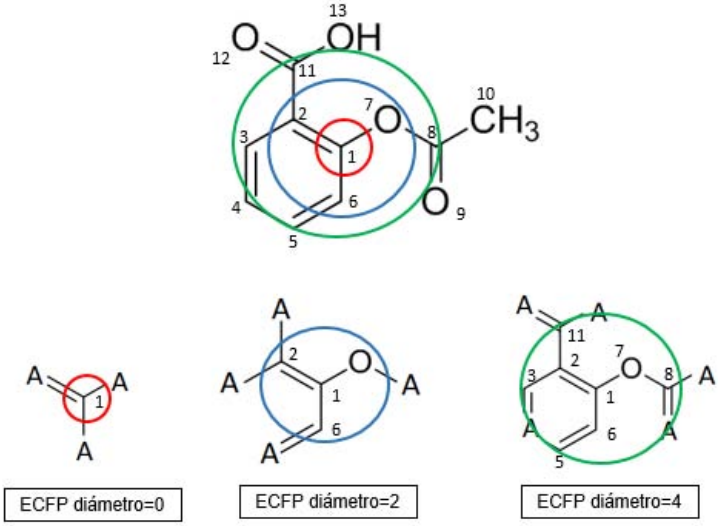


Figura 3. Representación gráfica de la distancia de enlace entre átomos calculada por ECFP. Un diámetro de dos indica que el radio elegido es uno y por lo tanto se contabiliza un átomo a partir del átomo central. Lo mismo sucede con un diámetro de cuatro, se considera un radio de dos y por lo tanto dos átomos a partir del átomo central. Modificado de Rogers, D. and M. Hahn. J. Chem. Inf. Model. 2010, 50: 742-754.

Para cuantificar la similitud que existe entre dos estructuras se utilizan los valores generados de las huellas digitales, frecuentemente es usado el coeficiente de Tanimoto (Ct, también llamado de Jaccard), que de manera general es calculado con la siguiente ecuación [13].

$$Tc_{(a,b)} = \frac{c}{a+b-c} \qquad (1)$$

Donde a y b corresponden al número de fragmentos únicos presentes (bits =1) de i-ésimo y j-ésimo compuesto, respectivamente, mientras que c representa el número de fragmentos en común.

Respecto a la similitud entre actividades es posible utilizar la siguiente expresión:

$$Sim_{(a,b)} = 1 - \frac{|Act_a - Act_b|}{\max - min} \qquad (2)$$

donde max es el valor más grande entre las actividades y min el menor,

o bien la diferencia entre el cologaritmo de las actividades [10]:

$$|\Delta pIC_{50}(T)_{a,b}| = pIC_{50}(T)_a - pIC_{50}(T)_b \qquad (3)$$

donde pIC_{50} es el cologaritmo de la concentración inhibitoria 50 de cada uno de los compuestos (a y b), y (T) hace referencia a un receptor molecular o *target*.

2.1.3.2 Acantilados de actividad (*activity cliffs*)

Las regiones del SAR discontinuo son generadas por los AC, los cuales están definidos como un par de compuestos estructuralmente similares o análogos que tienen una diferencia grande entre sus actividades [14]. Algunos puntos para considerar a los AC como tal son: 1) que cumplan con algún criterio de similitud previamente establecido, 2) que un compuesto en el par tenga actividad en el rango nM y 3) que exista por lo menos una diferencia en la actividad de dos órdenes de magnitud entre ambos compuestos [15]. Partiendo de este término han surgido diferentes clasificaciones. En la Tabla 1 se colocan algunos de los tipos más conocidos de AC:

Tabla 1. Tipos de *activity cliffs*

Término	Explicación	Referencia
Consensus activity cliffs	*Activity cliffs* en común que se identifican con dos o más metodologías distintas.	[10]
Shallow activity cliffs	Pares de compuestos que muestran grandes, pero no extremas, diferencias entre sus actividades.	[10] [16]
Deep activity cliffs	Pares de compuestos con alta similitud estructural y muy poca similitud entre actividades.	[10] [16]

Tabla 1 (continuación)		
Selectivity switch	Cambios estructurales que invierten la selectividad de dos compuestos similares, regularmente se estudian en dos dianas moleculares	[10]
Selectivity cliffs	A*ctivity cliffs* específicos a un receptor (en ocasiones dos), mientras que los multidianas (*Multitarget activity cliifs)* son relacionados con una serie de receptores	[17]
Activity cliff generators	Una molécula con alta probabilidad de formar *activity cliffs* y en donde todas las estructuras fueron estudiadas bajo el mismo ensayo biológico	[18]
Matched molecular pair	Par de compuestos que sólo difieren en una subestructura ubicada en un sitio especifico	[19]

2.2 Epigenética

El término epigenética fue utilizado por primera vez en 1942 por Conrad Waddington, quien intento describir las interacciones de los genes con su ambiente [20]. Sin embargo, la definición se ha ido transformando hasta referirse al estudio de las características fenotípicas heredables que resultan de las modificaciones en el cromosoma, que no alteren la secuencia del código genético. Una función primaria de la epigenética es el empaquetamiento del DNA en las células, en este sentido, las histonas tienen un papel importante. Al complejo que forman las histonas con el DNA para organizar el material genético se conoce como cromatina y a la unidad fundamental de este como nucleosoma. El nucleosoma está formado por un octámero de histonas (proteínas tipo histona: H1, H2A, H3 y H4). Por tal motivo la epigenética puede referirse al estudio de todos los elementos que modifican la regulación nucleosoma-cromatina y por la tanto que afecten la expresión genética [21].

2.2.1 Tipos de modificaciones epigenéticas

Las modificaciones son llevadas a cabo por tres tipos de proteínas: 1) *epigenetic writers*, 2) *epigenetic readers* y 3) *epigenetic erasers*. Los tres tipos de modificadores son enzimas que en general colocan, modulan y eliminan, respectivamente, las modificaciones químicas en las histonas [22]. En la Tabla 2 son resumidos los tipos de modificador, el mecanismo que tienen y un ejemplo de enzima que lleva a cabo la acción:

Tabla 2. Modificaciones epigenéticas más comunes. Modificado de Pande, V. J. Med. Chem. 2016, 59:1299-1307.

Writers	**Acción**	***Readers***	**Módulo al que se unen**	***Erasers***	**Acción**
Ac(et)il transferasas	Ac(et)ilación de lisinas, treoninas y serinas	Bromo	Lisinas acetiladas	Deac(et)ilasas de histona	Deac(et)ilación de lisinas
Metil transferasas	Metilación de lisinas, argininas y citosinas	PHD y tudor	Lisinas y argininas metiladas	Desmetilasas de histona	Desmetilación de lisinas y argininas
			Metil citosinas en el DNA	TET desmetilasas	Desmetilación del DNA

2.2.2 Bromodominios (BRDs)

La acetilación de las histonas está generalmente asociada con el aumento en la accesibilidad del DNA y también de los factores de transcripción. La acetilación debilita las interacciones de la histona-DNA mediante la neutralización de la carga positiva de los residuos de lisina (K) produciendo cambios sutiles en la estructura de la histona. La acetilación de la histona produce el reclutamiento de factores de transcripción y de remodelación de la cromatina que generan un aumento de la actividad transcripcional. El reclutamiento es llevado a cabo por los BRDs, un lector epigenético (*epigenetic reader*) que reconoce específicamente a los residuos de la lisina ε-N-acetiladas (Kac) [23].

El nombre bromodominio proviene de la proteína Brahma de *Drosophila sp.* en donde fue descubierto por primera vez [24]. Se caracteriza por presentar un plegamiento globular y está formado por cuatro α-hélices (nombradas Z, A, B y C) que son regiones conservadas en toda la familia (Figura 4). La cavidad (*pocket*) en donde se une la Kac es hidrofóbica y está formada por dos bucles (*loops*), uno largo αZ-αA (ZA *loop*) y otro corto αB-αC (BC *loop*).

Existen dos residuos responsables del reconocimiento de la Kac: en la mayoría de los casos uno es de tirosina (Y) en el bucle ZA, y otro de asparagina (N) en el bucle BC [23].

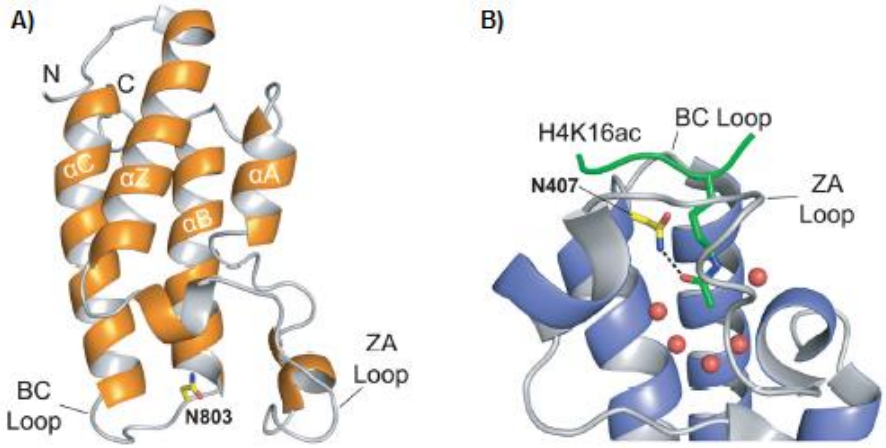


Figura 4. Estructura general de los BRDs y modo de unión con la Kac. (A) Estructura del BRD de PCAF obtenida por RMN (PDB ID: 1N72). Se muestran las cuatro hélices, así como los extremos terminales N y C, y los *loops* ZA y BC. (B) Estructura cristalográfica del BRD de GCN5 (azul) en unión con el péptido acetilado H4K16ac (verde, PDB ID: 1E6I), cinco moléculas de agua conservadas en el sitio de unión se representan con esferas rojas. Se observa el puente de hidrógeno entre la asparagina, N407, y la lisina acetilada. Tomada de Smith, S.G. and M.-M. Zhou. ACS Chem. Biol. 2016, 11: 598-608.

El proteoma humano codifica 46 proteínas que contienen BRDs. En tales proteínas se han observado 61 BRDs diferentes, los cuales están clasificados en ocho grandes familias (I-VIII) de acuerdo a su secuencia y a su homología estructural [23]. El presente trabajo está enfocado en el estudio de la familia II, conocida como familia BET (*Bromodomain and Extra-Terminal*), la cual contiene BRDs formados por dos tándem de bromodominio (BD1 y BD2) y un *C-terminal Extra-Terminal* (ET); BRD2, BRD3, BRD4 y BRDT forman parte de esta subfamilia [25].

2.2.3 Inhibidores de bromodominios

En los últimos años los inhibidores de BRDs (iBRDs) han adquirido mayor relevancia ya que potencialmente pueden ser usados como tratamiento para varias enfermedades [23].

En cáncer se ha encontrado que estructuras como JQ1 (iBRDs) inhiben al complejo BRD4-NUT, una oncoproteína generada de la fusión entre BRD4 y NUT (*Nuclear protein in testis*) que resulta de una translocación en el NMC (*NUT Midline Carcinoma*), un tipo de cáncer muy agresivo. También ha sido probado para el tratamiento de leucemia mieloide aguda, mieloma múltiple y linfoma de células B (tipos de cáncer de índole hematológico); el mecanismo general es la

inhibición de BRD4. Como se mencionó anteriormente los BRDs, en este caso BRD4, al unirse a la Kac reclutan factores de elongación hacia la cromatina; uno de ellos P-TEFb (*Positive Transcriptional Elongation Factor b*) el cual es una cinasa que fosforila varias regiones importantes para el control transcripcional. Dicho factor de elongación propicia la liberación de RNA polimerasa II (Pol II) y por lo tanto la transcripción de genes. Al inhibir a BRD4 la transcripción de algunos genes (proto-oncogenes) es suprimida o disminuida [23].

En enfermedades inflamatorias (artritis reumatoide y osteoartritis) ha sido demostrado que compuestos como I-BET762 (iBRDs) pueden suprimir o bien disminuir la expresión de citocinas pro-inflamatorias (TNFα, IL1B, IL6 e IL12A) así como de FTs (factores de transcripción) involucrados en la modulación de la respuesta inflamatoria. Además, se sabe que la estimulación de los macrófagos por lipopolisacárido (LPS) activa al receptor TLR4 (*Toll-like receptor*) permitiendo la expresión de FTs pro-inflamatorios específicos, los iBRDs como I-BET762 han mostrado alta selectividad en la supresión de un conjunto de genes inducidos por el LPS y por lo tanto la cadena de señales para liberar citocinas se ve afectada [23].

Se ha observado que en enfermedades como fallo cardíaco la transcripción es inducida por el estrés generado y es coactivada por BRD4. BRD4 activa FTs como NF-κB que favorecen el progreso de la enfermedad. Algo distinto sucede con otras cardiopatías como la ateroesclerosis, en donde su progresión es reducida por el transporte del colesterol de la arteria al hígado para su excreción por procesos como el transporte reverso de colesterol. Proteínas como ApoA1 y lipoproteínas de alta densidad (HDL, *High-Density Lipoproteins*) actúan como aceptores y transportadores del colesterol. RVX-208, fue el primer iBRD probado en ensayos clínicos, este inhibidor estimula la transcripción de ApoA1 [23]. En la Tabla 3 se resumen iBRDs y la fase de desarrollo clínico en donde se encuentran.

Tabla 3. iBRDs de la familia BET en ensayos clínicos. Modificado de Ferri, E., C. Petosa, and C.E. McKenna. Biochem. Pharm. 2016 ,106:1-18.

Inhibidor	Patrocinador	Fase	Padecimiento	Inicio	Estado	Ensayo clínico
RVX-208	Resverlogix	I, II	AT, DP y EAC	Oct 08	Completado	NCT00768274
		II	AT y EAC	Dic 09	"	NCT01058018
		IIb	EAC	Sep 11	"	NCT01067820
		IIb	DP y EAC	Ago 11	"	NCT01423188
		II	DB	Nov 12	"	NCT01728467
		II	DP y EAC	May 13	Terminado	NCT1863225
		III	DMT2 y EAC	Oct 15	Reclutamiento	NCT02586155
I-BET762	GSK	I	NMC	Mar 12	Reclutamiento	NCT01587703
		I	Hematológico	May 14	"	NCT01943851
OTX-015	Oncoethix (Merck)	I	LMA y LDCB	Dic 12	Reclutamiento	NCT01713582
		Ib	NMC y CPRC	Oct 14	"	NCT02259114
		IIa	Glioblastoma	Oct 14	Activo	NCT02296476
CPI-0610	Constellation Pharmaceuticals	I	Linfoma	Sep 13	Reclutamiento	NCT01949883
		I	Mieloma múltiple	Jul 14	"	NCT02157636
		I	LMA y SMD	Jun 14	"	NCT02158858
TEN-010	Tensha Therapeutics	I	NMC	Oct 13	Reclutamiento	NCT01987362
			LMA y SMD	Oct 14	"	NCT02308761
BAY 1238097	Bayer	I	Cáncer avanzado	Mar 15	Reclutamiento	NCT02369029
ABBV-075	AbbVie	I	Cáncer de mama	Abr 15	Reclutamiento	NCT02391480
INCB 054329	Incyte	I/II	Cáncer avanzado	May 15	Reclutamiento	NCT02431260
BMS-986158	Bristol-Myers Squibb	I/IIA	Cáncer de ovário y CPRC	Jun 15	Reclutamiento	NCT02419417
FT-1101	Forma Therapeutics	I	LMA y SMD	Sep 15	Reclutamiento	NCT02543879

AT=Aterosclerosis, CPRC= Cáncer de Próstata Resistente a la Castración, DB= Diabetes, DMT2= Diabetes Mellitus Tipo 2, DP=Dislipidemia, EAC= Enfermedad Arterial Coronaria, LDCB=Linfoma Difuso de Células B, LMA= Leucemia Mieloide Aguda, NMC= *NUT Midline Carcinoma* y SMD= Síndrome Mielodisplásico.

3. OBJETIVOS

3.1 Objetivo general

Determinar relaciones estructura-actividad de una colección pública de inhibidores de bromodominios (familia BET) mediante el desarrollo de un análisis quimioinformático automatizado de panoramas de actividad.

3.2 Objetivos particulares

- Desarrollar un programa (*script*) que permita la automatización del análisis de panoramas de actividad incluyendo la detección de acantilados de actividad.

- Describir y caracterizar cuantitativamente el panorama de actividad de 88 iBRDs de la familia BET reportados en bases de datos públicas.

- Identificar subestructuras químicas que generen cambios muy pronunciados en la actividad de los iBRDs.

4. METODOLOGÍA

La metodología se divide en cuatro partes principales: a) Preparación o curado de bases de datos. 2) Estrategia general del modelado de panoramas de actividad. 3) Relaciones de estructura-actividad de iBRDs. 4) El desarrollo de un *script* para automatizar parte del análisis de las relaciones estructura actividad utilizando el concepto de panoramas de actividad.

4.1 Curado de la base de datos

Las estructuras químicas y los valores de IC_{50} de los diferentes iBRDs (hacia las isoformas BRD2, BRD3 y BRD4) fueron obtenidos de un trabajo previamente publicado por Prieto-Martínez et al., 2016. Los autores obtuvieron la información y las estructuras de ChEMBL versión 20 (*European Molecular Biology Laboratory*): una gran base de datos estructurales de libre acceso. Las moléculas fueron descargadas como SMILES (*Simplified Molecular Input Line Entries*) y después convertidas a estructuras en 3D utilizando MOE (*Molecular Operating Enviroment),* versión 2013. Para curar la base de datos se utilizó el módulo *Wash* implementado en MOE. Con dicho módulo se estandarizaron las estructuras de la siguiente forma: se removieron sales metálicas y componentes minoritarios (moléculas de solventes y contraiones) y se recalcularon los estados de protonación (desprotonando ácidos fuertes y/o protonando bases fuertes). Se utilizó la estereoquímica codificada en la notación SMILES [26]. De las 207 estructuras encontradas sólo 88 fueron utilizadas debido a que las demás presentaban actividad para otras familias de BRDs. Para clasificar las actividades como: inactivas, de actividad media y activas se utilizaron como criterios de IC_{50}: inactivo ≥ 10.0 μM, actividad media entre 1.0-9.99 μM y activo < 1.0 Mm [26]. Los valores de IC_{50} para los tres receptores se resumen en el Apéndice A.2, al igual que el identificador interno (ID) utilizado y el asignado por ChEMBL.

4.2 Modelado de los panoramas de actividad (*SAS maps*)

Para representar la estructura molecular se utilizó como huella digital ECFP de diámetro cuatro. Para este fin se usó el *script* desarrollado por MayaChem *Tools*, *ExtendedConnectivityFingerprints*.pl [11]. Dicho *script* generó un vector binario para cada compuesto. Una vez obtenidos todos los vectores, se obtuvo la matriz

de similitud empleando el Ct. Tanto la matriz de similitud como el coeficiente fueron obtenidos con el *script* de MayaChem *Tools SimilarityMatricesFingerprints*.pl. Es importante notar que tanto el cálculo de similitud estructural como el del Ct se obtiene por comparaciones pareadas. El cálculo del Ct fue obtenido de la siguiente manera:

$$\frac{SUM\ (Xai*Xbi)}{(SUM\ (Xai^2)+SUM\ (Xbi^2)-SUM\ (Xai*Xbi))} \qquad (4)$$

En donde, SUM (Xai * Xbi) es la suma de los bits que comparten ambos compuestos. Los términos Xai^2 y Xbi^2 en el denominador tiene el objetivo de normalizar los valores de 0 a 1 [11].

Para el tratamiento de las actividades, se usó el valor absoluto de las diferencias entre los cologaritmos de los IC_{50} (Ecuación 3), en vez de la similitud de actividades. Es importante notar que todos los valores de IC_{50} fueron tratados en molar (M).

Para poder definir los límites que generan las diferentes regiones del SAS *map* fue importante tomar en cuenta la información de la base de datos y la manera en la que ésta se generó. No existe un criterio único ni absoluto para definirlos, sin embargo, los más utilizados están establecidos por una o dos unidades logarítmicas de $|\Delta pIC50(T)a,b|$, en el presente trabajo se utiliza una unidad logarítmica. En el caso de la similitud estructural se utilizó la mediana [10].

4.3 Análisis de las relaciones estructura-actividad de iBRDs

Para poder analizar las relaciones estructura-actividad de los diferentes iBRDs se identificaron los núcleos base de los compuestos, se cuantificó el valor del SAR en general y se analizaron los AC de mayor importancia y la información que de ellos se puede obtener (el valor de SALI, actividades, selectividad hacia algún receptor, subestructuras de mayor importancia, etc.).

4.3.1 *Scaffolds* moleculares

Los núcleos base o *scaffolds* moleculares fueron generados con DataWarrior, version 4.4.3, un programa quimioinformático para la visualización y análisis de información química [27]. Se utilizaron los módulos de *Analyse Scaffolds* y *Murcko Scaffols*. El análisis por Murcko localizó todos los sistemas cíclicos de

las moléculas, manteniendo los átomos y enlaces que se encontraban directamente conectados a los diferentes anillos.

4.3.2 Análisis del SAR global

El índice SARI (*Structure-Activity Relationship Index*) está diseñado para cuantificar las características del SAR de un conjunto de compuestos activos con actividad hacia una diana específica [28]. Está constituido por dos tipos de función de puntuación que evalúan la continuidad ($score_{cont}$) y la discontinuidad ($score_{disc}$). El valor total de la continuidad (raw_{cont}) mide la actividad ponderada mientras que el de la discontinua (raw_{disc}) determina el promedio de las diferencias de actividad de pares de compuestos similares.

$$raw_{cont} = 1 - \frac{\sum_{compuesto\ i \neq j} w_{i,j}\ sim\ (i.j)}{\sum_{compuesto\ i \neq j}\ w_{i,j}} \qquad (5)$$

El valor ponderado para cada par de compuestos (i, j) está definido como:

$$w_{i,j} = \frac{pIC_{50}(i)\ x\ pIC_{50}(j)}{1+|pIC_{50}(i)\ -\ pIC_{50}(j)|} \qquad (6)$$

El cálculo de la discontinuidad se realizó utilizando la siguiente ecuación:

$$raw_{disc} = \frac{\sum_{i,j\ |sim\ (i,j)>mediana\ i \neq j}\ |pIC_{50}(i)\ -\ pIC_{50}(j)|\ x\ \ sim\ (i.j)}{\#\ de\ compuestos\ que\ cumplen\ con{:}\ i,j\ |sim\ (i,j)>mediana\ i \neq j} \qquad (7)$$

Los valores de raw_{cont} y raw_{disc} fueron normalizados en una escala de 0 a 1 con una base de referencia (Apéndice A.2) para poder obtener el $score_{cont}$ y el $score_{disc}$. Estos datos fueron sustituidos en la siguiente expresión, para conocer el índice del SAR [28]:

$$SARI = \frac{1}{2}\ (\ score_{cont} + (1 -\ score_{disc})) \qquad (8)$$

4.3.3 Análisis de *activity cliffs*

Con los AC de mayor relevancia se realizó el estudio de los cambios subestructurales de mayor importancia. También, se realizó la comparación de aquellos AC seleccionados, hacía dos y tres receptores simultáneamente.

4.3.3.1 Análisis individuales, hacia un solo BRD: BRD2, BRD3 y BRD4

Los compuestos clasificados como AC fueron aquellos que presentaron un valor de $|\Delta pIC50(T)a,b|$ mayor a una unidad logarítmica y en donde la similitud fue mayor a la mediana de todos los valores del Ct. Una métrica utilizada para la identificación de AC fue el Índice de los Panoramas de Estructura-Actividad (en inglés SALI: *Structure-Activity Landscape Index*), el cual permitió identificar y cuantificar los AC. Para el cálculo de SALI se empleó la siguiente ecuación [29]:

$$SALI_{i,j} = \frac{|A_i - A_j|}{1 - sim(i,j)} \qquad (9)$$

En donde A_i y A_j son las actividades de la i-ésima y del j-ésima molécula, y sim(i, j) es el Ct.

Con los pares catalogados como AC se realizó la identificación de los *activity cliff generators* (ACG). Es importante notar que a pesar de que no todas las actividades fueron obtenidas bajo el mismo ensayo biológico, es posible utilizar el término. Para la identificación de los ACG se consideraron todas aquellas moléculas que presentaron una frecuencia superior al valor promedio más dos desviaciones estándar del número de pares de la región III. Esto permitió asegurar de forma estadística que las moléculas identificadas presentaban una frecuencia mayor al valor esperado [18], [30].

Una vez conocidos todos los AC se procedió a clasificarlos en dos categorías de acuerdo a los criterios de la Tabla 4, el valor de IC_{50} es de al menos un compuesto [15].

Tabla 4. Criterios de clasificación de los *activity cliffs.*

Categoría	Criterios		
***Deep* AC**	$\|\Delta pIC50(T)a,b\| \geq 2.0$	Ct > Mediana	IC_{50} < 1.0 µM
***Shallow* AC**	1.0 < $\|\Delta pIC50(T)a,b\|$ < 2.0	Ct > Mediana	IC_{50} < 1.0 µM

Para los *deep* AC identificados se realizó el análisis SAR, con el objetivo identificar las subestructuras a las cuales se les atribuyeron las grandes diferencias en pIC_{50}.

4.3.3.2 Análisis pareado: BRD2-BRD3, BRD2-BRD4 y BRD3-BRD4

Se realizaron mapas de tipo DAD (*Dual-activity difference maps*) con el objetivo de identificar cambios estructurales que favorecen la actividad hacia un receptor, para ello se seleccionaron compuestos que presentaron actividad hacia ambas dianas. Con las actividades se calculó la diferencia de $pIC_{50}(T)_{i,j}$, de cada par de compuestos, el criterio de división utilizado fue un intervalo de una unidad logarítmica [31]. Respecto a la similitud estructural, los pares de compuesto se filtraron seleccionando sólo aquellos AC con una similitud molecular mayor a la mediana de la base de datos. De las diferentes regiones del DAD *map* se estudiaron aquellos pares localizados en Z2, esto con el objetivo de identificar cambios estructurales que invirtieran la selectividad de dos compuestos similares (*selectivity switch*). En la región Z1 los cambios estructurales favorecen las actividades en un solo sentido y las regiones Z3 y Z4 corresponden a pares de compuestos con actividades similares hacia un receptor; la región Z5 contiene pares con actividades similares hacia ambos receptores [10].

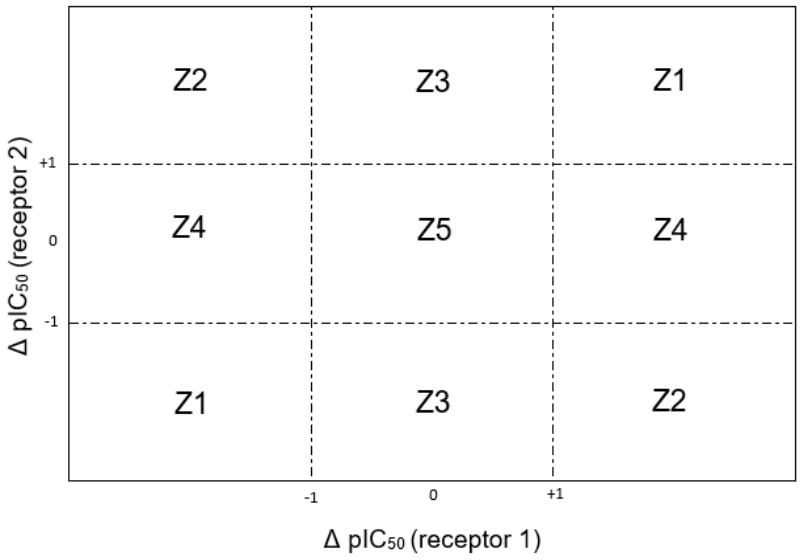


Figura 5. DAD *map* y las zonas que lo constituyen. En las zonas Z1, cambios grandes o pequeños en las estructuras generan un impacto similar hacia ambos receptores. En Z2, los cambios estructurales afectan de manera opuesta a las actividades. Z3 y Z4 representan zonas en donde las actividades son similares para un blanco, pero distintas para el otro y en Z5 se encuentran los pares que presentan actividades similares para ambos receptores. Modificado de Medina-Franco, JL. J. Chem. Inf. Model., 2012, 52:2485-2493

4.3.3.3 Análisis simultáneo hacía los tres receptores: BRD2-BRD3-BRD4

Se llevó a cabo el análisis de los gráficos TAD (*Triple activity-difference*) (Figura 6) que son una extensión de los DAD *maps*. Se interpretaron los TAD *maps* agrupando los DAD *maps* generados de la combinación de dos receptores: BRD2-BRD3, BRD2-BRD4 y BRD3-BRD4 [31].

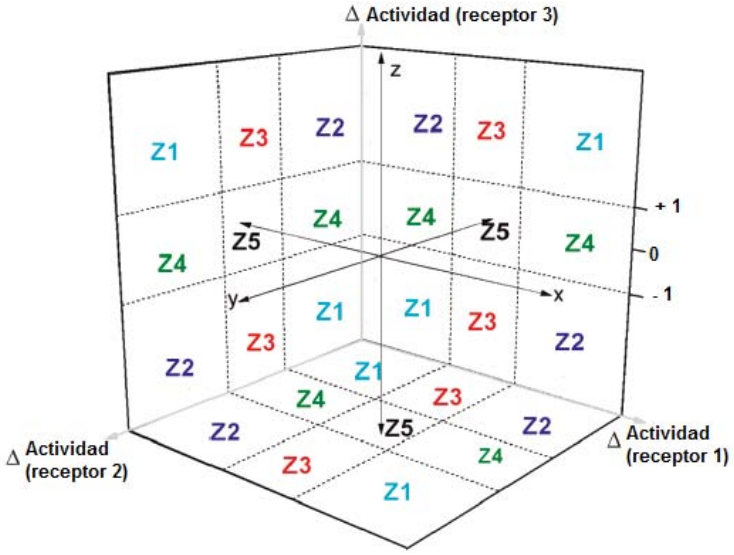


Figura 6. TAD *map* y las zonas que lo constituyen. Tomado de Medina-Franco, J.L., et al. J. Chem. Inf. Model. 2011, 51:2427-243

4.4 Elaboración del *script* para la automatización del análisis SAR

Con el objetivo de automatizar una parte del análisis de los panoramas de actividad se elaboró un *script* en el lenguaje Python 3.5.2. El cual genera a partir de archivos con extensión .sdf (estructuras) y .csv (actividades) las matrices de similitud molecular mediante el cálculo del Ct y el del valor $|\Delta pIC50(T)a,b|$. Además, permite la selección del *fingerprint* (ECFP de radio 2, 3 o 4 y MACCS *keys*). Cuantifica la cantidad de pares en cada región del SAS *map* y genera la representación gráfica del mismo. Por último, exporta todos los datos en un archivo .csv, el cual contiene la siguiente información: ID de cada compuesto en el par, su valor de Ct, el valor de $|\Delta pIC50(T)a,b|$ y la región a la que pertenece el par de compuestos. En la Figura 7 se muestra el flujo general de tareas que se llevan a cabo. El *script* completo se encuentra en el Apéndice A.3.

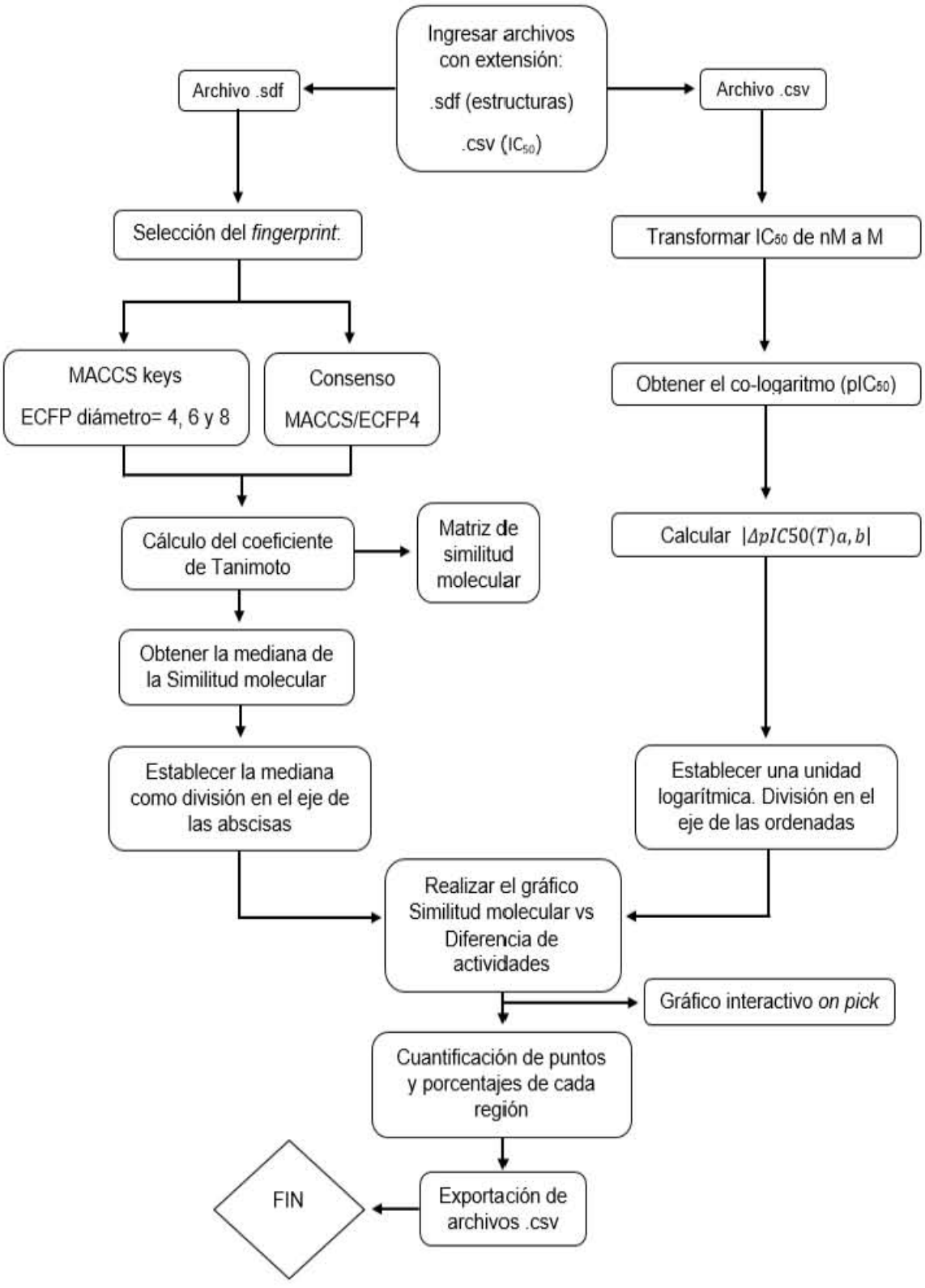


Figura 7. Diagrama de flujo de la automatización parcial del análisis SAR. Se colocan las principales tareas que lleva a cabo el *script*.

5. RESULTADOS Y DISCUSIÓN

5.1 Base de datos

En la mayoría de los casos las actividades de los inhibidores hacia los diferentes BRDs fue distinta. Aquellos compuestos cuyo IC_{50} no fue reportado se eliminaron del estudio, según el receptor del que se tratara. La Tabla 5 resume el número de compuestos estudiados y la distribución de la actividad biológica.

Tabla 5. Resumen general de las actividades en las bases de datos.

iBRD	Compuestos	% inactivo	% actividad intermedia	% activo	Valor mínimo*	Valor máximo*	Media*
iBRD2	84	23.81	63.10	13.10	0.0299	100	7.41
iBRD3	79	12.66	59.49	27.85	0.0284	100	4.96
iBRD4	86	18.60	65.12	16.28	0.0008	51.2	6.07

* Los valores representan a la IC_{50} en μM.

Es notorio que uno de los compuestos (**Cmpd_87**) de la base de iBRD4 tuvo el valor más pequeño de actividad reportada entre todas las bases de datos (IC_{50} = 0.0008 μM). Lo anterior influyó en el valor de $|\Delta pIC50(T)a,b|$ y por lo tanto en la región del SAR en donde se localizó el par de compuestos que contenía a **Cmpd_87**. Los compuestos más abundantes fueron los de actividad media con un porcentaje de aproximadamente el 60%. Además, los iBRD2 e iBRD4 presentaron un porcentaje de compuestos activos menor a los inactivos, algo que no sucedió con los iBRD3. Lo anterior posiblemente se atribuye a la afinidad de los compuestos para ese receptor y al tipo de compuestos depositados a la fecha en ChEMBL.

5.2 Modelado de los panoramas de actividad (SAS *maps*)

Del modelado de los panoramas de actividad se obtuvieron tres SAS *maps* (Figura 8), uno para cada isoforma de BRD estudiada. Dado que la cantidad y tipo de compuestos analizados en cada SAS *map* fue distinto (84, 79 y 86, para BRD2, BRD3 y BRD4, respectivamente) el valor de la mediana de similitud estructural cambió (0.23, 0.21 y 0.23, respectivamente). Gráficamente es posible observar que las regiones en donde se encuentra el mayor número de pares de compuestos son las regiones I y II (*Scaffold hops* y *Smooth SAR*, respectivamente*)* con porcentajes entre 33.79 y 42.08%, indicando que entre

ambas zonas del SAR se encuentra del 73.6% al 80.18% del total de las combinaciones. Este resultado se obtuvo con los tres mapas en la Figura 8.

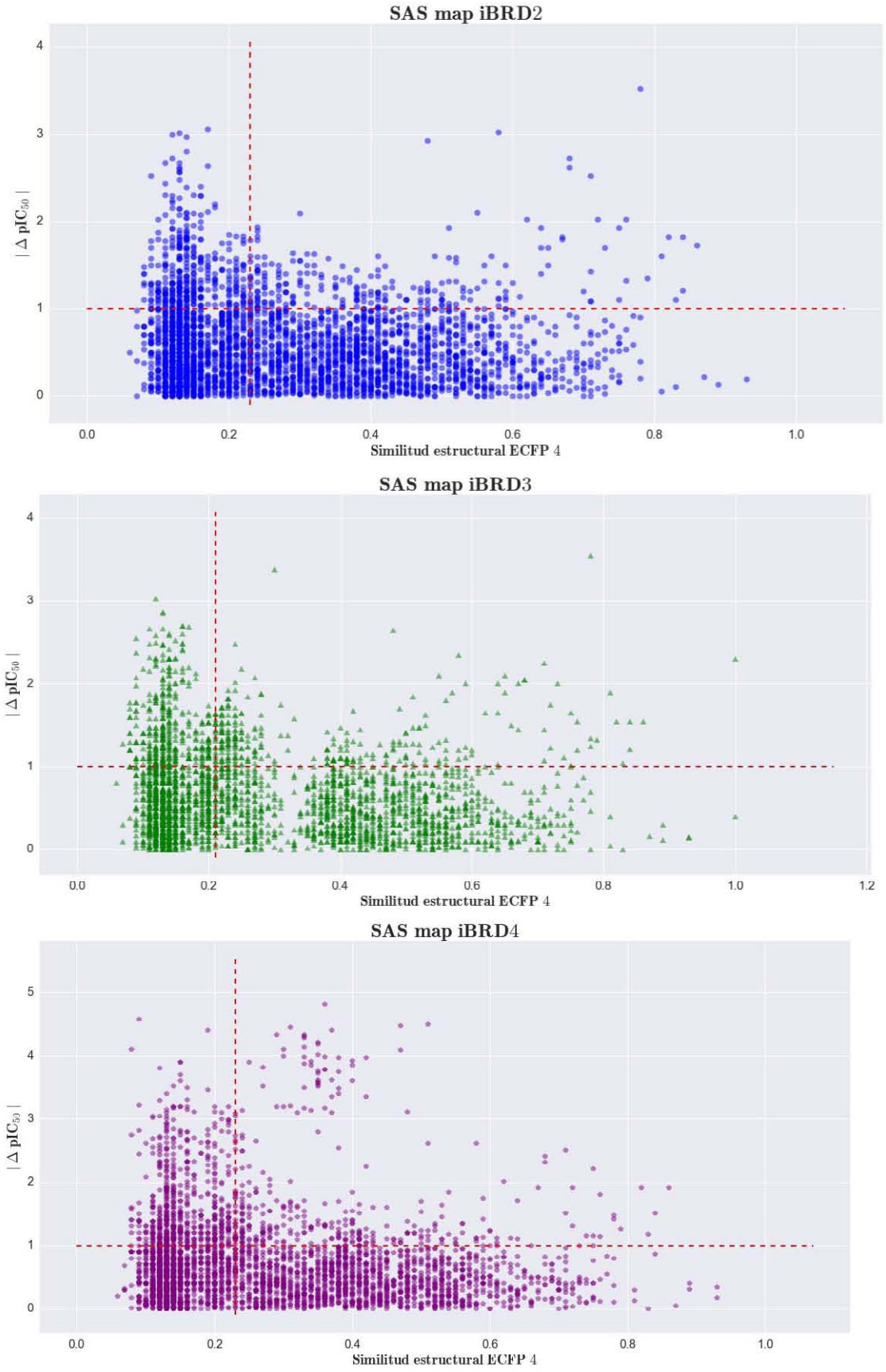


Figura 8. SAS *maps* de los iBRDs (BRD2, BRD3 y BRD4).

La región III (*activity cliffs*) es la que tuvo el menor número de compuestos con un porcentaje promedio de 8.02 ± 1.34%. En la Tabla 6 se resumen los valores obtenidos para cada región, así como el número de combinaciones generadas.

Tabla 6. Porcentajes de las combinaciones en las diferentes regiones del SAS *map.* Para los iBRDs (BRD2, BRD3 y BRD4).

BRD2			BRD3			BRD4			
			N= 88						
n utilizada:	84		n utilizada:	79		n utilizada:	86		
n despreciada:	4		n despreciada:	9		n despreciada:	2		
Valor mediana:	0.23		Valor mediana:	0.21		Valor mediana:	0.23		
Total de combinaciones:	3486	100.0 %	Total de combinaciones:	3081	100.0 %	Total de combinaciones:	3655	100.0 %	
Regiones del SAS *map*			**Regiones del SAS *map***			**Regiones del SAS *map***			Media
Activity cliffs:	226	6.48 %	Activity cliffs:	274	8.89 %	Activity cliffs:	318	8.70 %	8.02 ± 1.34 %
Scaffold hops:	1328	38.10 %	Scaffold hops:	1115	36.19 %	Scaffold hops:	1235	33.79 %	30.03 ± 2.16 %
SAR continuo:	1467	42.08 %	SAR continuo:	1210	39.27 %	SAR continuo:	1455	39.81 %	40.39 ± 1.49 %
No descriptiva:	386	11.07 %	No descriptiva:	393	12.76 %	No descriptiva:	560	15.32 %	13.05 ± 2.14 %
Indeterminados:	79	2.27 %	Indeterminados:	89	2.89 %	Indeterminados:	87	2.38 %	2.51 ± 0.33 %

N= Total de compuestos en la base de datos, n= compuestos utilizados.

Una región antes no descrita, la de los indeterminados, fue introducida en la clasificación ya que en ella se encuentran pares que se localizan sobre alguno de los ejes. Si bien, depende de la selección de los limites, la cantidad y tipo de combinaciones “indeterminadas” debe ser tomada en cuenta ya que *per se* no forma parte de ninguna región y por lo tanto está fuera del análisis. La región IV o no descriptiva no fue considerada en el análisis debido a que presenta compuestos con grandes diferencias de pIC_{50} y poca similitud estructural, estas combinaciones no proporcionan información de relevancia en el análisis estructura-actividad.

En general, se observa que en la representación gráfica para BRD4 existen pares de compuestos cuyo valor de $|\Delta pIC50(T)a,b|$ sobrepasa las cuatro unidades logarítmicas, esto debido a la presencia de un compuesto con la actividad más baja en toda la base de datos (**Cmpd_87**). En el panorama de actividad de BRD3 dos pares de compuestos presentaron una similitud cercana a 1.0, uno de los cuales está en región III (**Cmpd_2** y **Cmpd_3**). Además, debido a que es el SAS *map* en donde se eliminaron más compuestos con el mismo núcleo base (*scaffold*) se observa la separación en dos grupos. Respecto al modelado para iBRD2, la región III presentó la menor cantidad de combinaciones de todas las representaciones gráficas.

5.3 Análisis de las relaciones estructura-actividad de iBRDs

En esta sección se describen los análisis generales realizados (tales como: *scaffolds* moleculares y SAR global), así como el relacionado con los AC encontrados.

5.3.1 *Scaffolds* moleculares

El análisis de los *scaffolds* moleculares permitió conocer la diversidad de la base de datos respecto a los sistemas cíclicos de los compuestos. En este análisis se encontró que el núcleo base más frecuente fue el 4-fenil-1,2-oxazol con una proporción de 13/88 (14.77 %). En contra parte se observó que la proporción más grande (20.45 %, fracción en color morado) estaba constituida por compuestos con *scaffold* único (18 núcleos base distintos). En la Figura 9 se observa una fracción color azul turquesa que indica el porcentaje (15.92%) de estructuras que comparten el *scaffold* con otra, es decir, dos compuestos con el mismo *scaffold*. Una situación similar ocurre con la fracción color rosa (17.05%), que indica que cinco compuestos comparten un mismo *scaffold* (de los cuales se encontraron tres núcleos diferentes, 15 compuestos). Las fracciones restantes (11.36%-azul claro y 13.64%-amarilla) representan a 12 y 10 compuestos respectivamente, que comparten el mismo *scaffold* cada una. De los 33 *scaffolds* localizados en toda la base de datos 30 (60.24%) están representados por más de un compuesto.

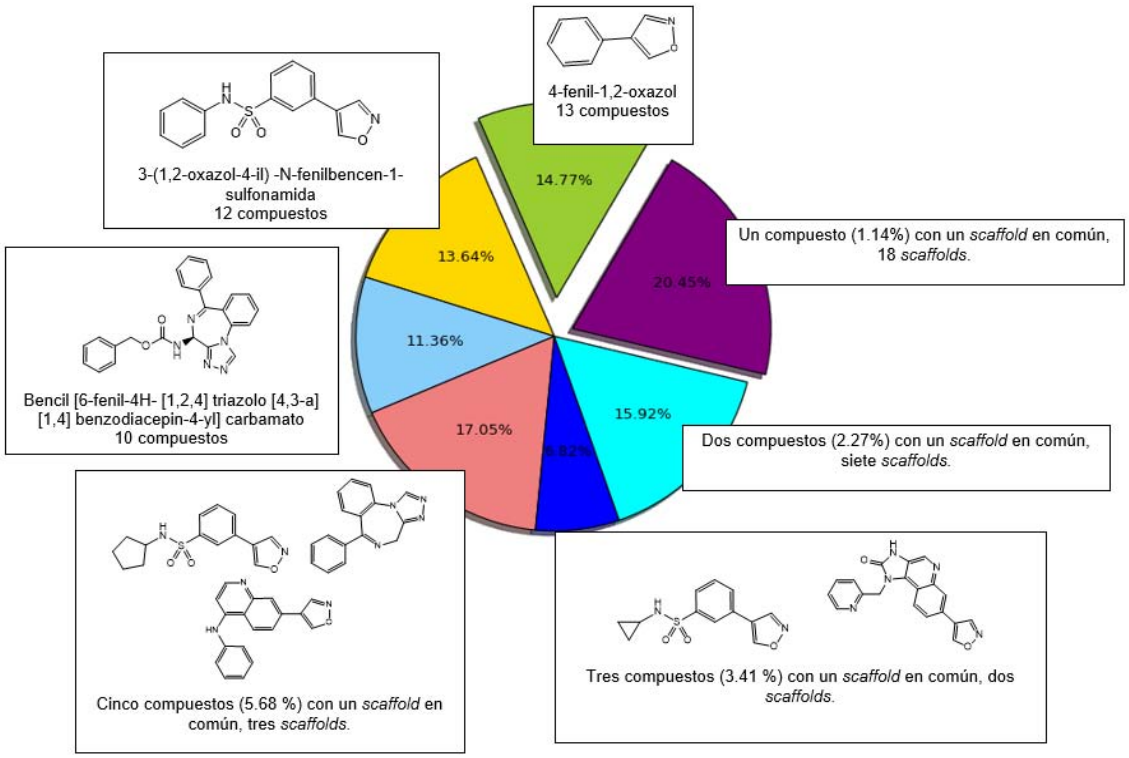


Figura 9. Frecuencia de *scaffolds* moleculares de los 88 iBRDs.

5.3.2 Análisis del SAR global

Para poder cuantificar el SAR de las bases de iBRDS se utilizó el índice SARI. Este indice toma en cuenta la distribución de la diversidad estructural y las diferencias en las actividades [28]. En la Tabla 7 se indican los *scores* y el valor de SARI obtenido para cada isoforma.

Tabla 7. Valores de SARI obtenidos para cada conjunto de compuestos iBRDs

iBRD	$Score_{con}$	$Score_{disc}$	SARI
iBRD2	0.52	0.37	0.57
iBRD3	0.48	0.31	0.58
iBRD4	0.48	0.54	0.47

De acuerdo con Peltason L. y Bajorath J. (2007) entre mayor es el valor de SARI mayor es el carácter de continuidad; dicho de otra manera, las diferencias estructurales entre los compuestos no generan cambios abruptos en la actividad de los mismos. Respecto a los resultados obtenidos (Tabla 7), los inhibidores de BRD2 y BRD3 fueron catalogados como SAR continuo. Indicando que el cambio en las actividades entre pares de compuestos es gradual al cambio de sus estructuras, lo cual es consistente con el principio de “compuestos semejantes tiene propiedades semejantes” [28]. Sin embargo, el cálculo del *score* de continuidad pondera el impacto de las actividades sobre el de la similitud, por lo que pares de compuestos con diferencias estructurales significativas son tomados en cuenta siempre y cuando exista similitud entre las actividades. Esto facilita la identificación de subestructuras, también conocidas como bioisósteros, que permiten mantener la actividad con una disminución de toxicidad o incremento de la afinidad.

En cuanto a los iBRD4 el valor de SARI fue de 0.47 lo que es representativo de un SAR heterogéneo. En otras palabras, las relaciones estructura-actividad entre estos compuestos presentan una influencia de ambos tipos de SAR (continuo y discontinuo). A diferencia de los dos casos anteriores, los iBRD4 tienen el valor de $score_{disc}$ más alto, lo cual hace referencia a que las diferencias de actividades son considerablemente mayores a los de las otras bases. Esto se observa en el SAS *map* de este receptor debido a que es el único que presenta valores de $|\Delta pIC50(T)a,b|$ mayores a tres unidades logarítmicas.

5.3.3 Análisis de *activity cliffs*

En esta sección se analizan los AC para cada receptor, así como las representaciones gráficas de DAD y TAD.

5.3.3.1 Análisis individuales, hacia un solo BRD: BRD2, BRD3 y BRD4

La cantidad de AC encontrados para BRD2, BRD3 y BRD4 fue de 226, 274 y 318, respectivamente. La Tabla 8 resume el número de compuestos y la clasificación a la que pertenecen según el tipo de AC, al igual que el valor mínimo y máximo de SALI.

Tabla 8. Cantidad de *activity cliffs* y categoría a la que pertenecen.

BRDs	Ct > Mediana y $\lvert\Delta pIC50(T)a,b\rvert$ (AbsAct)			Actividad	Cantidad de compuestos		SALI	
							Min.	Máx.
BRD2	226	AbsAct ≥ 2.0	11	Inact-Act	8	*Deep* AC	2.99	16.00
				Inter-Act	3		5.32	8.42
		1.0 < AbsAct < 2.0	215	Inact-Act	75	*Shallow* AC	1.49	5.51
				Inter-Act	22		1.34	12.29
				Act-Act	2		1.96	4.9
				Inact-Inter	116		1.32	8.42
BRD3	274	AbsAct ≥ 2.0	18	Inact-Act	8	*Deep* AC	2.66	2.3 x10^{13}
				Inter-Act	6		2.85	7.76
				Inact-Inter	4		4.88	7.41
		1.0 < AbsAct < 2.0	256	Inact-Act	81	*Shallow* AC	1.59	6.09
				Inter-Act	111		1.29	11.07
				Act-Act	6		1.54	5.68
				Inact-Inter	48		1.45	10.00
BRD4	318	AbsAct ≥ 2.0	75	Inact-Act	17	*Deep* AC	2.91	9.16
				Inter-Act	55		2.83	8.84
				Act-Act	3		3.29	4.31
		1.0 < AbsAct < 2.0	243	Inact-Act	32	*Shallow* AC	1.84	6.73
				Inter-Act	69		1.35	13.64
				Inact-Inter	142		1.36	9.44

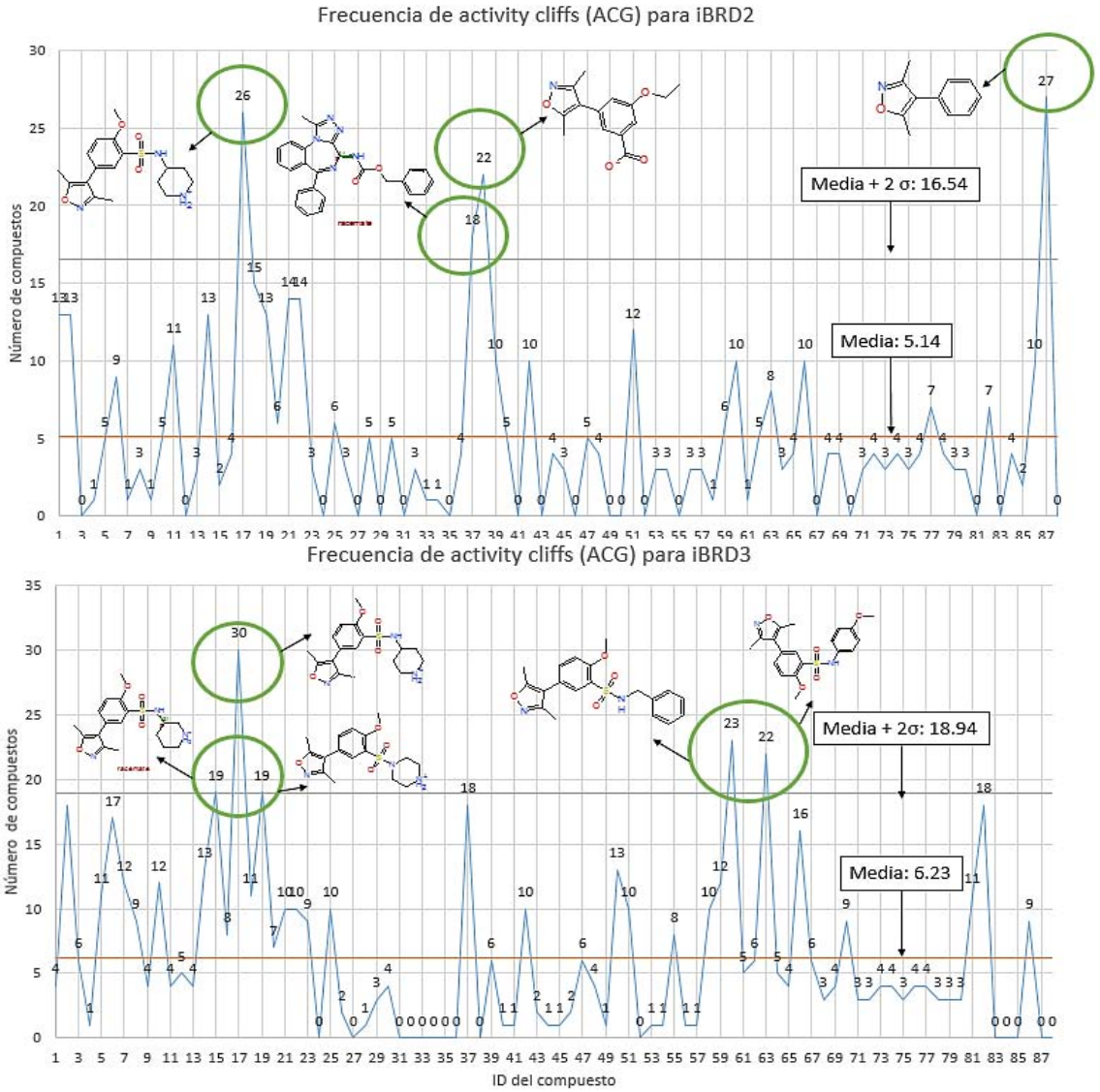
Inact= Inactivo, Inter= Actividad Intermedia y Act=Activo.

En la Tabla 8 se observan los AC que fueron subclasificados (de acuerdo al nivel de actividad de ambos compuestos en cada par) como: Inact-Act, Inter-Act, Act-Act e Inact-inter. Lo anterior se realizó con el objetivo de discriminar entre aquellos pares que no presentaron por lo menos un compuesto activo. Aquellos pares que cumplieron con los criterios para *deep* AC o AC profundos están indicados con color gris.

Respecto a los valores de SALI, en BRD3 se encuentra el par con el valor más alto de todas las bases (2.3 x10^{13}), el cual será discutido más adelante. Si bien un valor alto de SALI es indicativo de la presencia de un AC, no siempre es el de

mayor relevancia. En los inhibidores de BRD2 tenemos que para la subclasificación de los *deep* AC en donde los pares tienen un compuesto de actividad intermedia y el otro con actividad (Inter-Act), tres compuestos donde el valor más alto de SALI es de 8.42. Para el mismo receptor y subclasificación (Inter-Act), pero respecto a los *shallow* AC, hay 22 compuestos en donde el valor más alto de SALI es de 12.29. Con lo anterior se ejemplifica que un valor de SALI alto no necesariamente hace referencia a un AC de mayor relevancia. Además, es posible tener valores de SALI similares en las diferentes regiones del SAS *map*, por lo cual SALI sólo es utilizado como un orientador en la identificación de AC más no como un criterio único.

Con los pares de compuestos catalogados como AC (sin distinguir entre los *deep* y *shallow*) se identificaron aquellas estructuras ACG. Estas moléculas y su frecuencia se resumen en la Figura 10.



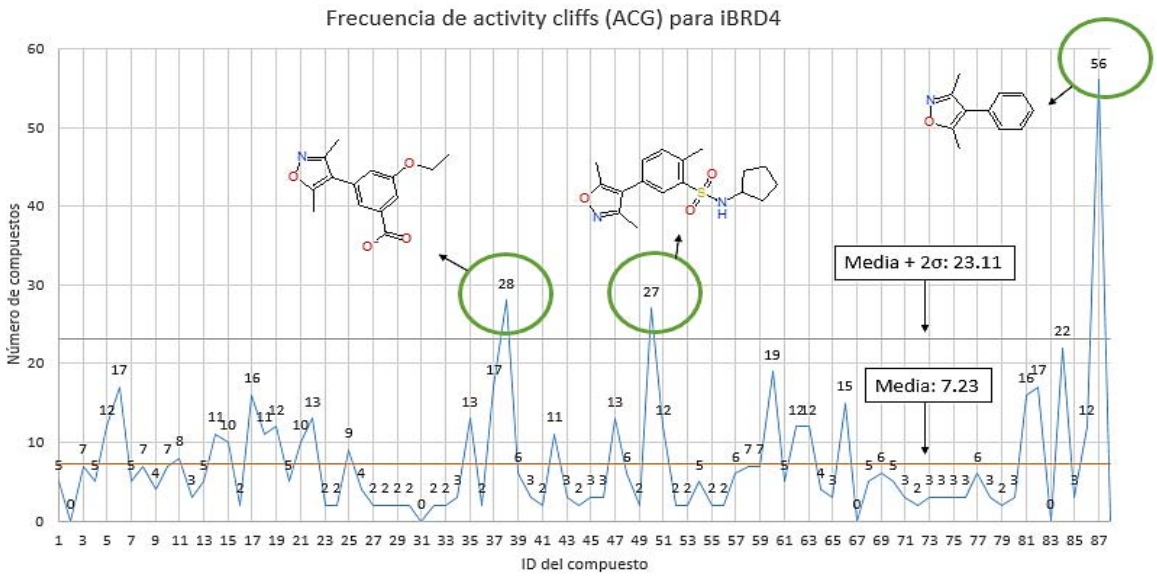


Figura 10. Frecuencia de los *activity cliffs generators* (ACG) (continuación)

Para los iBRD2 se identificaron como ACG a los compuestos **Cmpd_17**, **Cmpd_37**, **Cmpd_38** y **Cmpd_87**, con frecuencias de 26, 18, 22 y 27, respectivamente. De los cuatro ACG, **Cmpd_38** y **Cmpd_87** presentan la misma subestructura (3,5-dimetil-4-fenil-1,2-oxazol); **Cmpd_37** es el único de los ACG con un núcleo triazolobenzodiacepínico y **Cmpd_17** comparte la subestructura 5-(3,5-dimetil-1,2-oxazol-4-il)-2-metoxi-N-**R-**benceno sulfonamida con otros ACG, en donde **R** es un sustituyente distinto.

Para los iBRD3 los ACG fueron los compuestos **Cmpd_15**, **Cmpd_17**, **Cmpd_19**, **Cmpd_60** y **Cmpd_63**, con frecuencias de 19, 30, 19, 23 y 22, respectivamente. Los cinco ACG presentaron la misma estructura que **Cmpd_27**. Para los iBRD4 los ACG son los compuestos **Cmpd_38**, **Cmpd_50** y **Cmpd_87** con frecuencias de 28, 27 y 56, respectivamente. Los tres compuestos tienen la misma subestructura (3,5- dimetil-4-fenil-1,2-oxazol).

La generación de los ACG es consecuencia de compuestos con actividades muy distintas en comparación con el otro compuesto con el que forman el par. Es probable que esta diferencia interfiera con la ecuación matemática que correlaciona las propiedades de los compuestos en modelos QSAR [14]. Por tal motivo el estudio SAR previo a la realización de estos modelos es una herramienta útil que permite eliminar aquellos compuestos que impidan una adecuada correlación de la información.

De acuerdo con los criterios de clasificación se consideraron a los *deep* AC como aquellos AC de mayor importancia dada la diferencia entre las actividades de los pares de compuestos (dos o más unidades logarítmicas de pIC_{50}). Para BRD2, BRD3 y BRD4 se obtuvieron 11, 14 y 75 *deep* AC, respectivamente. En BRD3 se descartaron cuatro pares con un valor de $|\Delta pIC50(T)a,b|$ superior a dos unidades logarítmicas ya que no cumplían con el criterio de que por lo menos uno de los compuestos fuera activo. En BRD4 se observó la subclasificación Act-Act para tres pares siendo el único caso que se presenta, esto debido a que el conjunto de compuestos contiene la actividad más pequeña de todas las bases de datos, permitiendo que existan grandes diferencias incluso entre compuestos catalogados como activos.

La Figura 11 muestra las estructuras de los 11 *deep* AC para BRD2. En general se aprecia que cualquier cambio en la estructura de **Cmpd_37** genera un aumento en la IC_{50} lo que provoca que disminuya su actividad al requerir una mayor concentración del inhibidor. En cada comparación se observaron cambios tanto en las subestructuras como en la estereoquímica. La modificación más representativa es generada al cambiar el bencilo de **Cmpd_37** (considerado ACG) por un metilo en **Cmpd_2**, lo que produce que el segundo compuesto alcance una concentración inhibitoria de 100 μM (generando el valor más alto de SALI para los AC de este receptor, 16.0). Los compuestos **Cmpd_13**, **Cmpd_71**, **Cmpd_73** y **Cmpd_79** se caracterizaron por tener sustituciones en los bencenos del núcleo triazolobenzodiacepínico, lo cual afectó el valor de IC_{50} llevando la clasificación de las actividades de los compuestos hasta intermedia e inactiva. Las comparaciones de **Cmpd_37** con **Cmpd_1**, **Cmpd_80** y **Cmpd_86** muestran que posiblemente el carbamato y las modificaciones en la estructura de carbamato tienen un rol importante en la actividad de estos inhibidores. Como lo muestra la Figura 11, la modificación en el carbamato de **Cmpd_37** modificó la concentración requerida para inhibir al receptor, alcanzando concentraciones inhibitorias por encima de 25 μM. De todos los *deep* AC sólo aquellos formados con **Cmpd_37** generaron *Matched Molecular Pairs* (MMP), que son pares de compuestos que sólo difieren en una subestructura ubicada en un sitio especifico [18]. Estos compuestos son de mucho interés debido a que pueden dirigir y facilitar la optimización de los compuestos en estudio.

Respecto a los pares **Cmpd_77** y **Cmpd_82** que hacen AC con **Cmpd_2** es notorio que el cambio de carbamato por etilacetamida mejora la actividad al igual

que el bencilo en el carbamato (como se había observado anteriormente).

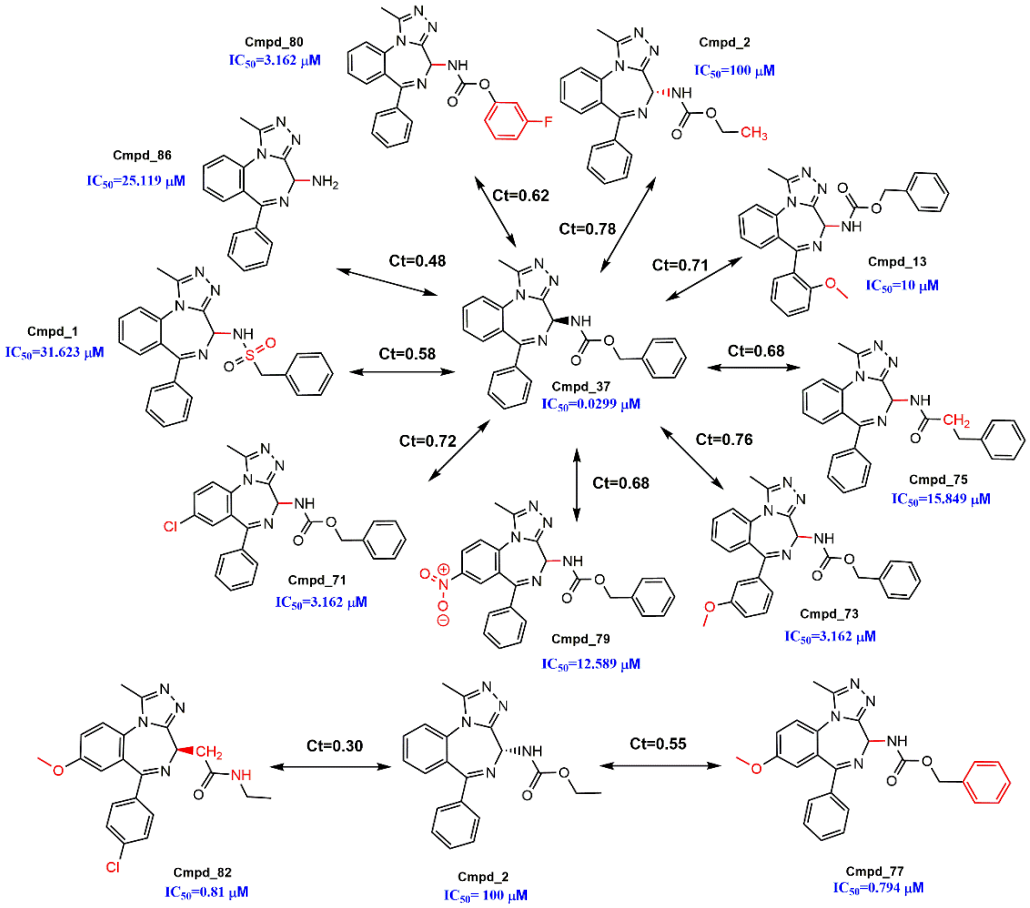


Figura 11. *Deep* AC de los iBRD2.

La Figura 12 muestra los 18 *deep* AC encontrados para BRD3. De manera similar a BRD2, **Cmpd_37** formó parte de la mayoría de los pares. Sin embargo, este compuesto no es considerado como ACG ya que su frecuencia se encuentra por debajo de la media + 2σ. Los inhibidores con los que formó los *deep* AC fueron los mismos que con BRD2, manteniendo un comportamiento similar de las actividades. El par **Cmpd_2** con **Cmpd_3** generó el valor más alto de SALI calculado entre todas las bases de datos (2.3 x10^{13}). Dicho valor fue producido por la alta similitud estructural, siendo la estereoquímica entre los compuestos la única diferencia (**Cmpd_2** con la forma enantiomérica (*S*) y **Cmpd_3** (*R*)). Esta disconformidad generó que se requiera una menor concentración (0.501 μM) del enantiómero (*R*) en relación con la del estereoisómero (*S*) que fue de 100 μM. Otros compuestos como **Cmpd_65**, **Cmpd_77** y **Cmpd_82** mejoraron su actividad en relación con **Cmpd_2** al cambiar la orientación del enlace entre el carbamato y el núcleo triazolobenzodiacepínico, cambiando de actividades

consideradas inactivas a activas. De las comparaciones realizadas con **Cmpd_82** es notorio que resultó favorable la sustitución del carbamato, presente en prácticamente todos los compuestos, por etilacetamida. Lo anterior se observa en el hecho de que en semejanza con **Cmpd_13**, **Cmpd_1** y **Cmpd_86**, mostró una actividad en el rango de los activos. El último de los *deep* AC encontrados para este receptor fue el constituido por **Cmpd_17** (un ACG) y **Cmpd_66**, los cuales conservaron la estructura 3,5-dimetil-4-fenil-1,2-oxazol. La única diferencia entre este par de inhibidores se debió al cambio de la imidazoquinolinona (**Cmpd_66**) por una piperidinil bencenosulfonamida (**Cmpd_17**), aumentando la actividad de 0.20 a 21.0 µM.

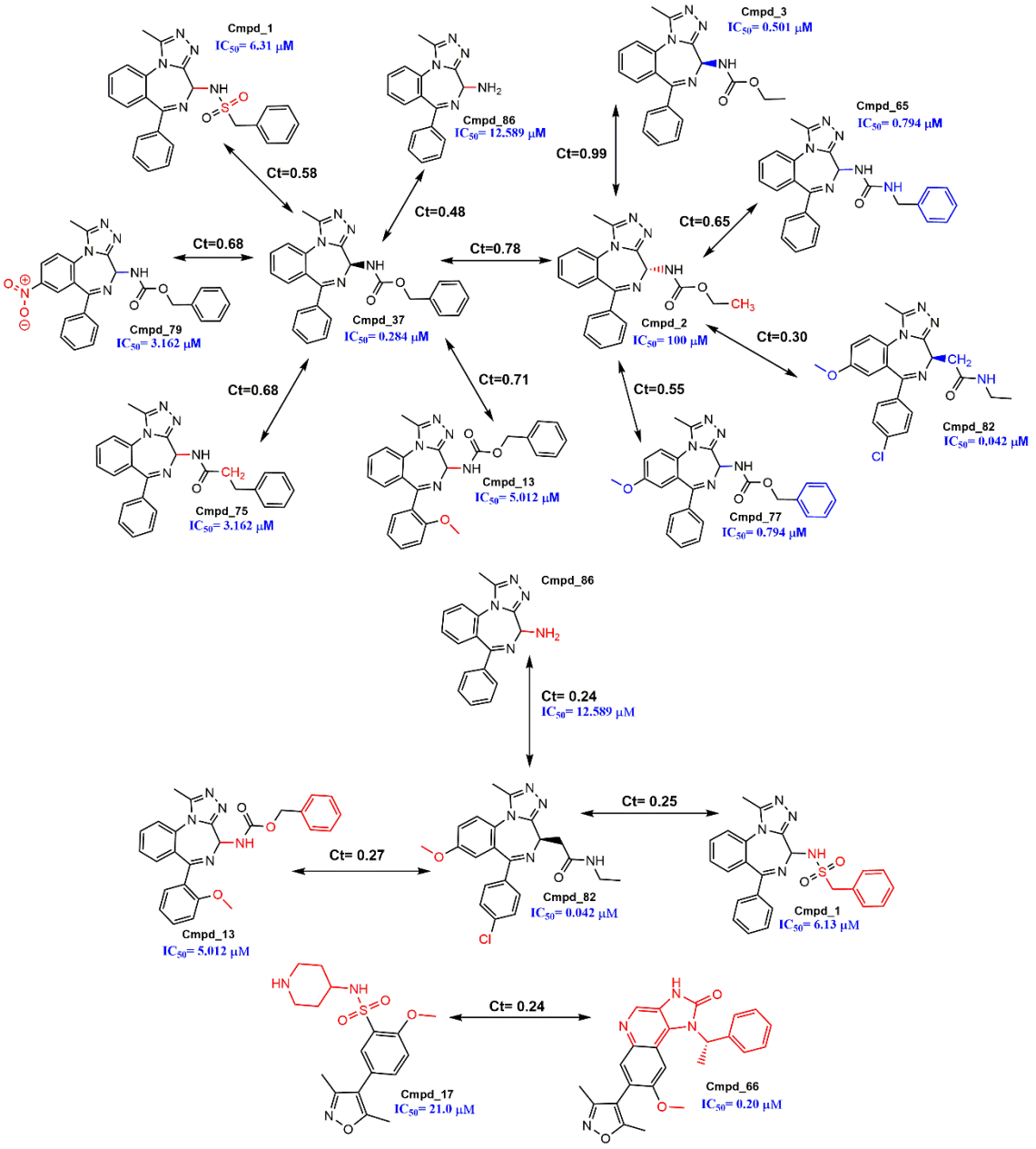


Figura 12. *Deep* AC de los iBRD3

En la Figura 13 se muestran ejemplos de *deep* AC generados con **Cmpd_87** (el compuesto que más AC generó para esta base de datos). En 56 de los 75 pares, **Cmpd_87** formó parte, incluyendo al par con el mayor valor de SALI (9.16, **Cmpd_87** y **Cmpd_84**). En las 56 comparaciones se conservó la estructura 3,5-dimetil-4-fenil-1,2-oxazol y cualquier modificación en ésta generó un aumento de IC_{50}. Además, debido a el valor tan pequeño de concentración inhibitoria 50 (0.008 µM) fue posible encontrar valores de $|\Delta pIC50(T)a,b|$ mayores a las cuatro unidades logarítmicas. Una característica particular para este receptor fue que se encontraron pares de compuestos cuyas actividades se mantuvieron en la clasificación de activas. Esto se llevó a cabo entre **Cmpd_87** y **Cmpd_6**, **Cmpd_50** y **Cmpd_66** (Figura 13); los valores de actividad de estos compuestos cambiaron de 0.008 hasta 0.398 µM permitiendo tener valores de $|\Delta pIC50(T)a,b|$ mayores a las dos unidades logarítmicas.

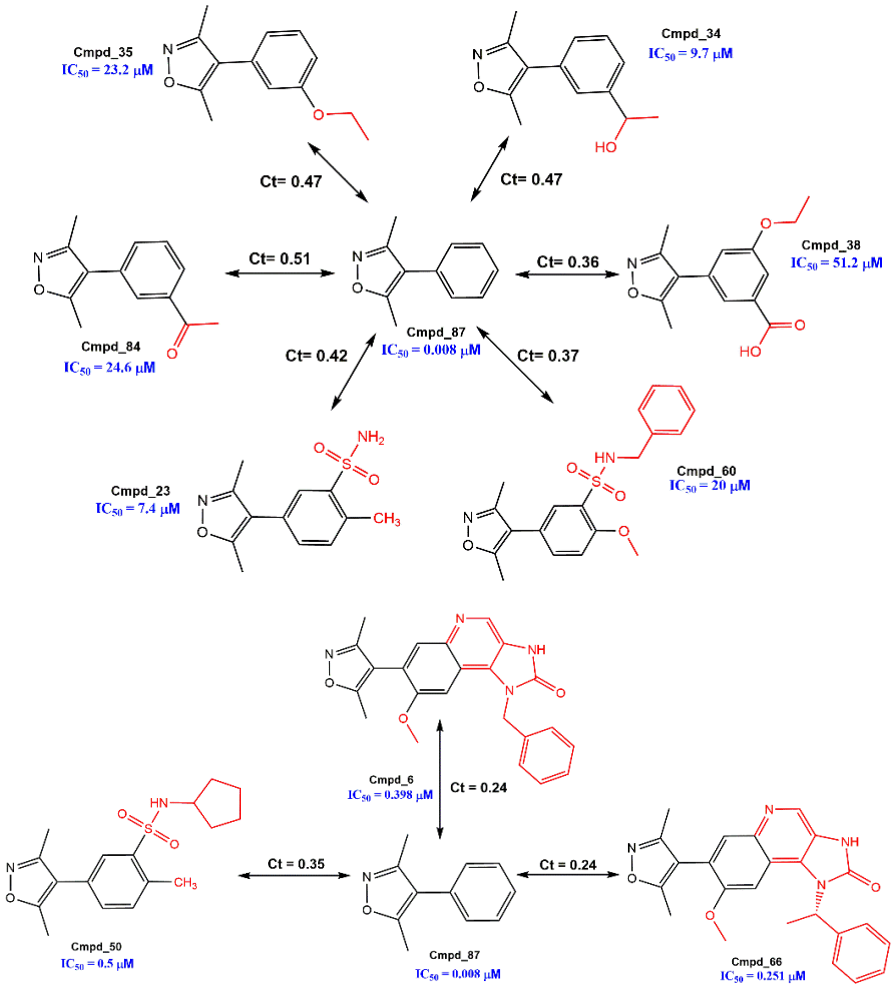


Figura 13. *Deep* AC de los iBRD4 con **Cmpd_87**.

Se encontró que **Cmpd_37** estaba presente en ocho de los 75 pares (Figura 14). De estos pares cabe destacar que adiciones sobre los bencenos del núcleo triazolobenzodiacepínico y cambios del carbamato por otros grupos aumentan el IC_{50} de los inhibidores. En los pares restantes, **Cmpd_81**, **Cmpd_82** y **Cmpd_38** generaron *deep* AC con frecuencias respectivas de cinco, cinco y uno. Cabe destacar que todos los pares comparados con tales compuestos tuvieron un valor de $|\Delta pIC50(T)a,b|$ menor a las tres unidades logarítmicas y cambios subestructurales en más de un sitio.

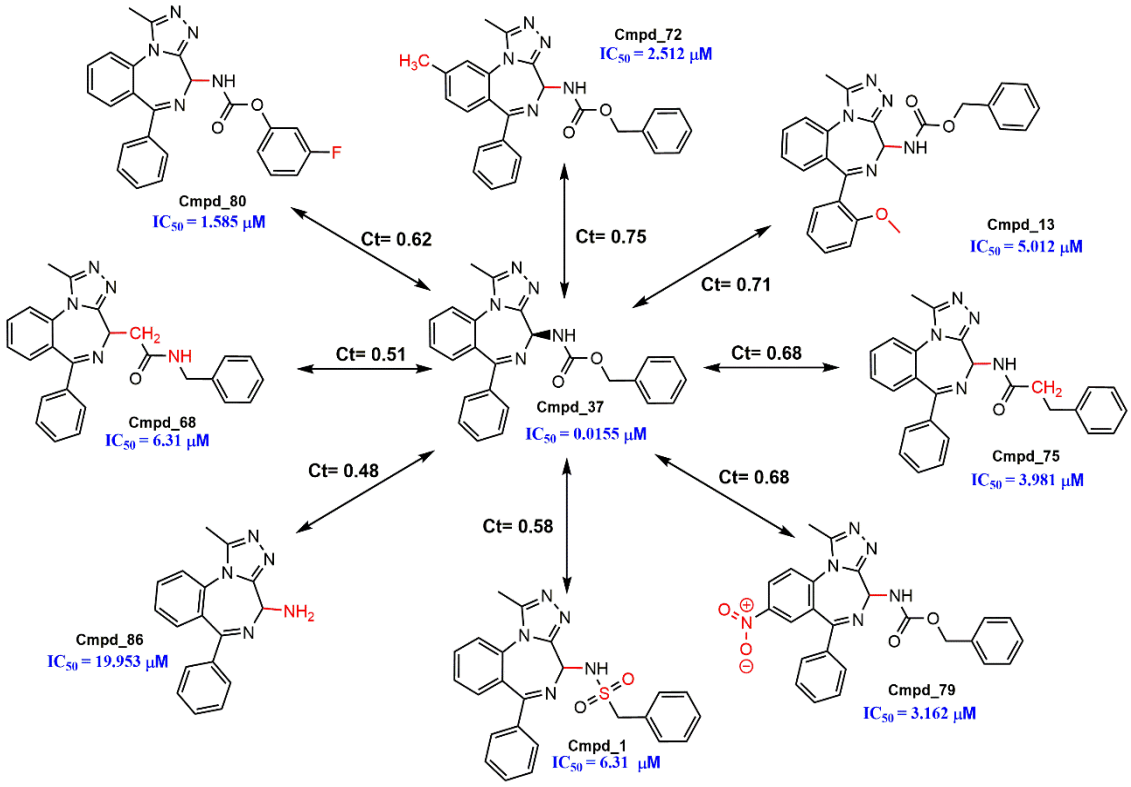


Figura 14. *Deep* AC de los iBRD4 con **Cmpd_37**.

5.3.3.2 Análisis pareado: BRD2-BRD3, BRD2-BRD4 y BRD3-BRD4

Como se observa en los tres DAD *maps* de la Figura 15 la mayoría de los pares (cerca del 80 %) se localizaron en la región Z5 (región central de los mapas). Esto indica que son compuestos con actividades similares hacia ambas dianas moleculares y por lo tanto tiene un impacto nulo o muy pequeño en la actividad hacia un receptor en específico. Las regiones Z4 y Z3 correspondieron a pares de inhibidores con actividades similares entre si hacia un receptor, pero diferentes hacia el otro. Debido a las características de similitud estructural y

diferencias de actividades seleccionadas, los pares localizados en esta región son considerados AC específicos hacia una diana molecular. En la región Z1 los cambios estructurales entre los compuestos afectan en el mismo sentido la actividad, ya se aumentándola o disminuyéndola, por lo que se afirma que los pares comparten el mismo SAR [10]. Respecto a la región Z2 únicamente el DAD *map* generado por BRD2-BRD4 contuvo pares de compuestos. De los 27 pares, todos incluyeron a **Cmpd_87** (*selectivity switch*). Esto indica que los cambios estructurales que son realizados en **Cmpd_87** aumentan la selectividad hacia BRD2, pero la disminuyen para BRD4.

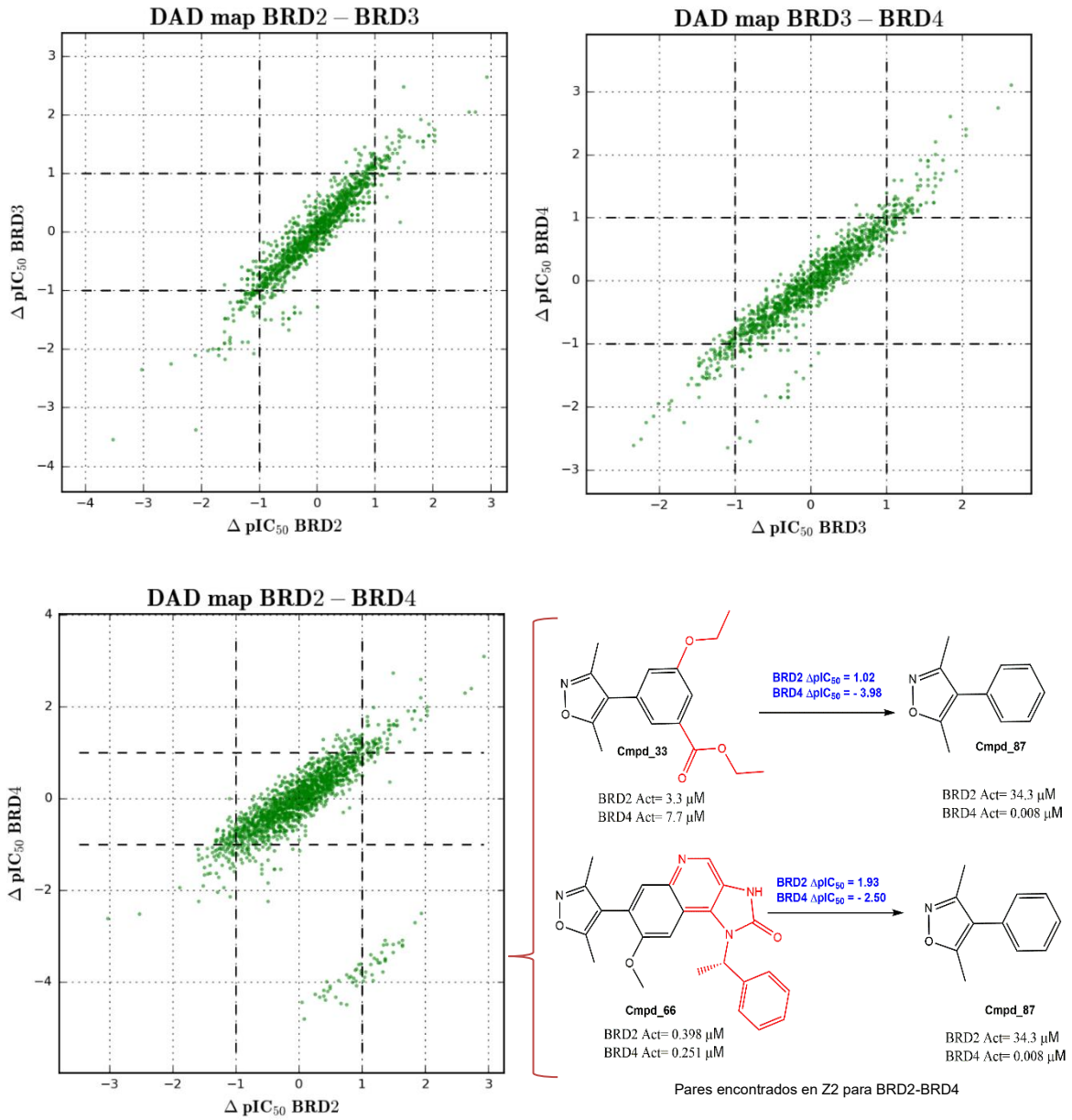


Figura 15. *Dual-Activity Difference maps* (DAD *maps*).

En la Figura 15 también se representan dos ejemplos de estos pares: **Cmpd_33-Cmpd_87** y **Cmpd_66-Cmpd_87**. En ellos se observa que los cambios estructurales en **Cmpd_87** favorecen la inhibición de BRD2 y la perjudican para BRD4, debido a que las concentraciones de IC_{50} disminuyen y aumentan, respectivamente.

5.3.3.3 Análisis simultáneo hacía los tres receptores: BRD2-BRD3-BRD4

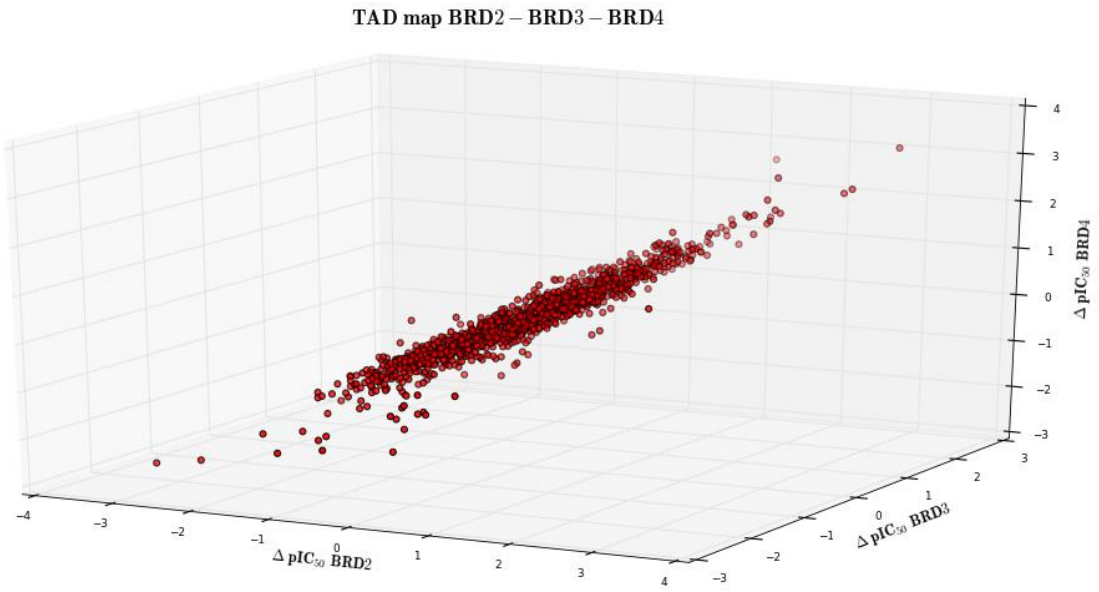


Figura 16. *Triple-Activity Difference map* (TAD *map*).

La Figura 16 es un TAD *map* en donde se analizan simultáneamente las tres representaciones de los DAD *maps* mostradas en la sección anterior. La mayoría de las comparaciones se encuentra dentro del SAR continuo, debido a que están en la zona Z5 de por lo menos uno de los ejes. La información de mayor relevancia obtenida con los TAD *maps* es la identificación de AC específicos hacia los tres o bien un receptor [29]. De todos lo AC encontrados para cada receptor (BRD2, BRD3 y BRD4), 91 fueron los compartidos entre las tres dianas moleculares. Por otra parte, los AC específicos fueron 40, 86 y 107 para BRD2, BRD3 y BRD4, respectivamente. El gran número de AC para BRD4 es atribuido a **Cmpd_87**, por las razones discutidas anteriormente. A pesar de que los TAD *maps* generan información interesante, el uso de los DAD *maps* es preferido ya que los primeros son más difíciles de visualizar.

6. CONCLUSIONES

- Este trabajo contribuye al desarrollo del campo de investigación emergente de la epi-informática [32]. En especial, el proyecto facilita la elucidación de las relaciones estructura-actividad epigenéticas que a la fecha han sido estudiadas en forma limitada, pero que tienen cada vez mayor interés y relevancia [33].
- Mediante la programación de un *script* en Python, se facilitó la automatización del análisis e identificación de los pares de compuestos; así como los ID de los mismos, los valores de Ct y $|\Delta pIC50(T)a,b|$ y la región donde se localizan.
- El empleo de los SAS *maps* como modelado de los panoramas de actividad permitió la caracterización cuantitativa de los iBRDs reportados en una base de datos pública de la familia BET.
- Se identificó que, en el grupo de compuestos estudiados, cualquier modificación a la estructura de **Cmpd_37** esta asociada a cambios importantes en la actividad del inhibidor, pasando de ser considerado como un compuesto activo a inactivo o de actividad intermedia. Esto se cumple en los tres BRDs.
- La estereoquímica de los compuestos juega un papel importante en la inhibición del receptor, siendo el enantiómero (*S*) de **Cmpd_2** (inactivo) el caso más representativo, esto en los iBRD3.
- Se determinó que los cambios estructurales en **Cmpd_87** para los iBRD4 afectan notablemente la actividad del inhibidor, clasificado como activo.
- Se encontró mediante el análisis dual de relaciones-estructura actividad con dos dianas biológicas que modificaciones realizadas en **Cmpd_87** favorecen la actividad hacia BRD2, pero la desfavorecen hacia BRD3.

7. PERSPECTIVAS

Con la información generada en este trabajo se realizarán estudios QSAR con aquellos compuestos que se localicen en la región II del SAS también conocida como región *Smooth* SAR o del SAR continuo, caracterizada por tener diferencias del valor de $|\Delta pIC50(T)a,b|$ no mayores a una unidad logarítmica y similitudes estructurales mayores a la mediana de todos los datos. Además, se pretende eliminar aquellos compuestos catalogados como generadores de acantilados de actividad y ver el impacto que esto tiene en el estudio QSAR.

Otra perspectiva de este trabajo es plantear hipótesis a nivel estructural de los acantilados de actividad. Para corroborar estas hipótesis se realizarán estudios de acoplamiento molecular automatizado (*docking*) con estructuras de rayos X de los BRDs utilizados (BRD2, BRD3 y BRD4) y reportados en la literatura. Con la información obtenida se pretenden buscar las posibles interacciones estructurales de aquellos compuestos que sean selectivos hacia alguno de los BRDs.

El método de automatización del análisis de panoramas de actividad mediante los SAS y DAD *maps* se usará para analizar otros grupos de datos epigenéticos u otros conjuntos de moléculas con actividad reportada hacia otros blancos de interés terapéutico.

Se pretende optimizar y generar una interface visual y ejecutable (EXE) del *script* desarrollado, con el objetivo de facilitar su uso y aplicación.

8. REFERENCIAS

1. Saldívar-González, F., F.D. Prieto-Martínez, and J.L. Medina-Franco, *Descubrimiento y desarrollo de fármacos: un enfoque computacional.* Educación Química. 2017, 28(1):51-58

2. Filippakopoulos, P. and S. Knapp, *Targeting bromodomains: epigenetic readers of lysine acetylation.* Nature Reviews Drug Discovery, 2014. 13(5): p. 337-356.

3. Smith, S.G. and M.-M. Zhou, *The Bromodomain: A New Target in Emerging Epigenetic Medicine.* ACS Chemical Biology, 2016. 11(3): p. 598-608.

4. Brown, F.K., *Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery*, in *Annual Reports in Medicinal Chemistry*, A.B. James, Editor. 1998, Academic Press. p. 375-384.

5. Gasteiger, J., *Introduction*, in *Chemoinformatics*. 2004, Wiley-VCH Verlag GmbH & Co. KGaA. p. 1-13.

6. Aguayo-Ortiz, R. and E. Fernández-de Gortari, *Chapter 2 - Overview of Computer-Aided Drug Design for Epigenetic Targets A2 - Medina-Franco, José L*, in *Epi-Informatics*. 2016, Academic Press: Boston. p. 21-52.

7. Wassermann, A.M., M. Wawer, and J. Bajorath, *Activity Landscape Representations for Structure−Activity Relationship Analysis.* Journal of Medicinal Chemistry, 2010. 53(23): p. 8209-8223.

8. Bajorath, J., *Modeling of activity landscapes for drug discovery.* Expert Opinion Drug Discovery, 2012. 7(6):463-73.

9. Shanmugasundaram, V. and G.M. Maggiora, *Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. .* Proceedings of 222nd American Chemical Society National Meeting, Division of Chemical Information, Chicago, IL, August 26-30, 2001; American Chemical Society: Washington, DC, 2001; abstract no.77.

10. Medina-Franco, J.L., *Scanning Structure–Activity Relationships with Structure–Activity Similarity and Related Maps: From Consensus Activity Cliffs to*

Selectivity Switches. Journal of Chemical Information and Modeling, 2012. 52(10): p. 2485-2493.

11. Sud, M., *MayaChemTools: An Open Source Package for Computational Drug Discovery.* Journal of Chemical Information and Modeling, 2016.

12. Rogers, D. and M. Hahn, *Extended-Connectivity Fingerprints.* Journal of Chemical Information and Modeling, 2010. 50(5): p. 742-754.

13. Medina-Franco, J.L. and G.M. Maggiora, *MOLECULAR SIMILARITY ANALYSIS*, in *Chemoinformatics for Drug Discovery*. 2013, John Wiley & Sons, Inc. p. 343-399.

14. Maggiora, G.M., *On Outliers and Activity CliffsWhy QSAR Often Disappoints.* Journal of Chemical Information and Modeling, 2006. 46(4): p. 1535-1535.

15. Cruz-Monteagudo, M., et al., *Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?* Drug Discovery Today, 2014. 19(8): p. 1069-1080.

16. Pérez-Villanueva, J., et al., *Towards a systematic characterization of the antiprotozoal activity landscape of benzimidazole derivatives.* Bioorganic & Medicinal Chemistry, 2010. 18(21):7380-7391.

17. Stumpfe, D. and J. Bajorath, *Exploring Activity Cliffs in Medicinal Chemistry.* Journal of Medicinal Chemistry, 2012. 55(7):2932-2942.

18. Pérez-Villanueva, J., et al., *Activity cliffs and activity cliff generators based on chemotype-related activity landscapes.* Molecular Diversity, 2015. 19(4):1021-1035.

19. Hu, Y., D. Stumpfe, and J. Bajorath, *Advancing the activity cliff concept.* F1000Research, 2013. 2:199.

20. Waddington, C.H., *The Epigenotype.* International Journal of Epidemiology, 2012. 41(1):10-13.

21. Dueñas-González, A., J. Jesús Naveja, and J.L. Medina-Franco, *Chapter 1 - Introduction of Epigenetic Targets in Drug Discovery and Current Status of*

Epi-Drugs and Epi-Probes, in *Epi-Informatics*. 2016, Academic Press: Boston. p. 1-20.

22. Pande, V., *Understanding the Complexity of Epigenetic Target Space.* Journal of Medicinal Chemistry, 2016. 59(4):1299-1307.

23. Ferri, E., C. Petosa, and C.E. McKenna, *Bromodomains: Structure, function and pharmacology of inhibition.* Biochemical Pharmacology, 2016. 106:1-18.

24. Tamkun, J.W., et al., *brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2.* Cell, 1992. 68(3):561-72.

25. Padmanabhan, B., et al., *Bromodomain and extra-terminal (BET) family proteins: New therapeutic targets in major diseases.* Journal of Biosciences, 2016. 41(2):295-311.

26. Prieto-Martinez, F.D., et al., *A chemical space odyssey of inhibitors of histone deacetylases and bromodomains.* RSC Advances, 2016. 6(61):56225-56239.

27. Sander, T., et al., *DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis.* Journal of Chemical Information and Modeling, 2015. 55(2):460-473.

28. Peltason, L. and J. Bajorath, *SAR Index: Quantifying the Nature of Structure−Activity Relationships.* Journal of Medicinal Chemistry, 2007. 50(23):5571-5578.

29. Guha, R. and J.H. Van Drie, *Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs.* Journal of Chemical Information and Modeling, 2008. 48(3):646-658.

30. Aguayo Ortiz, R., sustentante, *Análisis quimioinformático y modelado molecular de derivados del bencimidazol para la selección eficiente de compuestos giardicidas y reposicionamiento en otras enfermedades.* Tesis que para obtener el grado de Maestría en Ciencias Químicas, 2015. Tutor principal de tesis Rafael Castillo Bocanegra.

31. Medina-Franco, J.L., et al., *Multitarget Structure–Activity Relationships Characterized by Activity-Difference Maps and Consensus Similarity Measure.* Journal of Chemical Information and Modeling, 2011. 51(9): p. 2427-2439.

32. Medina-Franco, J.L. and J. Yoo, *Chapter 15 - The Road Ahead of the Epi-Informatics Field*, in *Epi-Informatics*. 2016, Academic Press: Boston. p. 399-418.

33. García-Sánchez, M., Cruz-Monteagudo M, Medina-Franco JL., *Quantitative Structure-Epigenetic Activity Relationships.* "Challenges and Advances in Computational Chemistry and Physics", 2017. Advances in QSAR modeling with Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences".

APÉNDICE

A.1 Carta de aceptación y primeras páginas del capítulo de libro: *Quantitative Structure-Epigenetic Activity Relationships*

Prof. Kunal Roy MPharm, PhD

Professor
Division of Medicinal and Pharmaceutical Chemistry
Department of Pharmaceutical Technology
Jadavpur University
Raja S C Mullick Road, Jadavpur,
Kolkata 700 032 (INDIA)

Former Commonwealth Academic Staff Fellow,
University of Manchester, UK
Former Marie Curie International Incoming Fellow,
University of Manchester, UK

Editor-in-Chief, *International Journal of Quantitative Structure-Property Relationships* (IGI Global)
Associate Editor, *Molecular Diversity* (Springer)
Member of the Editorial Advisory Board, *European Journal of Medicinal Chemistry* (Elsevier); *Chemical Biology and Drug Design* (Wiley), *Expert Opinion on Drug Discovery* (Informa)

Email: kunalroy_in@yahoo.com
URL: http://sites.google.com/site/kunalroyindia/
Phone (O): +91 33 2414 6666 Ext 2053
Mobile: +91 98315 94140

September 29, 2016

José L. Medina-Franco, Ph.D.
Department of Pharmacy,
School of Chemistry
Universidad Nacional Autónoma de México (UNAM)
Mexico

Sub: Book Chapter entitled "**Quantitative Structure-Epigenetic Activity Relationships"**
(Authors: Mario Omar García-Sánchez, Maykel Cruz-Monteagudo, José L. Medina-Franco)

Dear Prof. Medina-Franco:

I am pleased to inform you that your above mentioned chapter has been accepted to be included in the upcoming Springer Book "*Advances in QSAR modeling with Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*" Edited by Kunal Roy (Under the Series: Challenges and Advances in Computational Chemistry and Physics, Series Editor: Prof. Jerzy Leszczynski; http://www.springer.com/series/6918). Your article will be forwarded to the Publisher for further processing in due course.

Best regards,

Kunal Roy
Editor, Advances in QSAR modeling with Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences, Springer (Forthcoming)

K. Roy, Editor. Volume "Advances in QSAR modeling with Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences" under the book series "Challenges and Advances in Computational Chemistry and Physics"

Chapter X

Quantitative Structure-Epigenetic Activity Relationships

Mario Omar García-Sánchez,[1] Mykel Cruz-Monteagudo,[2,3,4] José L. Medina-Franco[1,*]

[1]Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, México

[2]Instituto de Investigaciones Biomédicas (IIB), Universidad de Las Américas, 170513 Quito, Ecuador

[3]CIQUP/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal

[4]REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal

[5]Instituto de Química, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, México

[*]Corresponding author. E-mail: medinajl@unam.mx; jose.medina.franco@gmail.com

Tel. +5255-5622-3899, ext. 44458

Abstract

The relevance of epigenetic drug discovery has increased during the past few years as revealed by the augmenting number of related publications and the amount of structure epigenetic activity data in compound databases. This chapter discusses the current status of epigenetic target-based therapies. It is also analyzed the progress of QSAR models developed for compounds databases screened with epigenetic targets. A special emphasis is made on compounds directed to inhibitors of DNA methyltransferases, one of the first epigenetic target families associated with therapeutic potential. Novel approaches applied to develop models for inhibitors of bromodomains, other epigenetic target families with high relevance in modern drug discovery programs, are also discussed. The chapter analyses epigenetic activity landscape modeling, activity cliffs, and activity cliff generators and their relevance to develop QSAR models. Computational methods applied to elucidate Quantitative Structure-Epigenetic Activity Relationships are in line with the increasing and developing research area of Epi-informatics.

Keywords: Activity cliffs, Activity landscape, Bromodomains, DNA methyltransferase, Epigenetics, Epi-informatics, HDAC, SEARS

List of abbreviations: ACG: activity cliffs generators, ALM: activity landscape modeling, BRD: bromodomain, DNMT: DNA methyltransferase, ERCS: Epigenetic Relevant Chemical Space, FDA: Food and Drug Administration, HDAC: histone lysine deacetylase, ISMs: instances that should be misclassified, MMP: matched molecular pairs, MODI: modelability index, NSG: Network-like similarity graphs, PLIF: protein-ligand interaction fingerprint, PLM: property landscape modeling, QSAR: quantitative structure-activity relationship, SALI: structure activity landscape index, SAR: structure-activity relationship, SAS: structure-activity similarity, SEARS: structure-epigenetic activity relationships, SARI: structure-activity relationship index, SmAR: structure-multiple activity relationship, SVMs: support vector machines.

A.2 Bases de datos (IC_{50} para BRD2, BRD3 y BRD4) y referencia para SARI

Las concentraciones son reportadas en nanoMolar, N/A se aplica en aquellos casos en donde no hay IC_{50} reportada:

Compuesto ID	CHEMBL ID	BRD2 (IC_{50})	BRD3 (IC_{50})	BRD4 (IC_{50})
1	CHEMBL2430887	31623	6310	6310
2	CHEMBL2430872	100000	100000	N/A
3	CHEMBL2430873	N/A	501.19	398
4	CHEMBL2181819	3162.28	1584.9	6309.6
5	CHEMBL2181818	1584.89	631	1258.9
6	CHEMBL2017285	501.19	251.2	398.1
7	CHEMBL2017277	1995.26	501.2	1258.9
8	CHEMBL2017283	501.19	398.1	501.2
9	CHEMBL2017281	1000	501.2	794.3
10	CHEMBL2017273	501.19	316.2	631
11	CHEMBL2017264	12589.25	3162.3	7943.3
12	CHEMBL2017282	2511.8899	631	1259
13	CHEMBL2430893	10000	5012	5012
14	CHEMBL2017269	794.33002	501.2	1259
15	CHEMBL2181714	9000	13000	17000
16	CHEMBL2181713	11000	8400	10000
17	CHEMBL2181712	31000	21000	22000
18	CHEMBL2181711	19000	11000	15000
19	CHEMBL2181708	16000	11000	17000
20	CHEMBL2181707	14000	8200	12000
21	CHEMBL2181704	800	700	1300
22	CHEMBL2181703	800	700	1000
23	CHEMBL2181701	8100	8600	7400
24	CHEMBL2181741	5600	4200	6500
25	CHEMBL2181740	12000	8600	13000
26	CHEMBL2181720	1900	1300	3000
27	CHEMBL2181717	3600	2200	5200
28	CHEMBL2181716	1500	1400	2500
29	CHEMBL2181715	3600	3000	3200
30	CHEMBL2181718	1500	1100	2600
31	CHEMBL1950956	2000	2500	2000
32	CHEMBL1828985	2300	N/A	7500
33	CHEMBL1828983	3300	N/A	7700
34	CHEMBL1828982	4200	N/A	9700
35	CHEMBL1828981	7400	N/A	23200

Las concentraciones son reportadas en nanoMolar, N/A se aplica en aquellos casos en donde no hay IC_{50} reportada:

(continuación de A.2)

Compuesto ID	CHEMBL ID	BRD2 (IC_{50})	BRD3 (IC_{50})	BRD4 (IC_{50})
36	CHEMBL1828986	1600	N/A	4800
37	CHEMBL1738926	29.9	28.4	15.5
38	CHEMBL1828984	28200	N/A	51200
39	CHEMBL2181724	1100	1000	1800
40	CHEMBL2181730	1500	2100	2800
41	CHEMBL2181709	4200	3200	4200
42	CHEMBL2181706	1100	800	1200
43	CHEMBL2181736	4700	3800	6600
44	CHEMBL2181726	1800	2000	3000
45	CHEMBL2181702	2200	1500	2400
46	CHEMBL2181738	4400	3300	4400
47	CHEMBL2181719	1500	1000	1000
48	CHEMBL2181705	1700	1100	1800
49	CHEMBL2181739	3900	3000	3300
50	CHEMBL2181721	N/A	500	500
51	CHEMBL2181723	900	700	1100
52	CHEMBL2181729	3600	4000	4500
53	CHEMBL2181728	2500	1900	3300
54	CHEMBL2181742	2500	1800	2000
55	CHEMBL2181734	6000	5900	5600
56	CHEMBL2181732	2000	1900	2600
57	CHEMBL2181725	2400	2200	5600
58	CHEMBL2181731	4900	5400	8200
59	CHEMBL2181710	11000	6500	11000
60	CHEMBL2181737	13000	14000	20000
61	CHEMBL2181727	4300	4800	7600
62	CHEMBL2181722	1400	1000	1200
63	CHEMBL2181735	14000	13000	16000
64	CHEMBL2181733	5200	4700	6000
65	CHEMBL2430886	1995.26	794	1260
66	CHEMBL2017288	398.10999	200	251.2
67	CHEMBL2181820	1584.89	631	N/A
68	CHEMBL2430888	2511.8899	1995	6310
69	CHEMBL2430869	1995.26	1000	501
70	CHEMBL2017286	N/A	501.19	1000

Las concentraciones son reportadas en nanoMolar, N/A se aplica en aquellos casos en donde no hay IC_{50} reportada:

(continuación A.2)

Compuesto ID	CHEMBL ID	BRD2 (IC_{50})	BRD3 (IC_{50})	BRD4 (IC_{50})
71	CHEMBL2430890	3162.28	1585	1260
72	CHEMBL2430891	2511.8899	1259	2511.9
73	CHEMBL2430894	3162.28	1000	1000
74	CHEMBL2430896	1585	1000	1259
75	CHEMBL2430885	15849	3162	3981
76	CHEMBL2430895	1995.26	1000	1259
77	CHEMBL2430892	794	794	501
78	CHEMBL2430874	2511.8899	1259	794
79	CHEMBL2430889	12589	3162	3162
80	CHEMBL2430870	3162.28	1260	1585
81	CHEMBL2017291	N/A	251.19	18
82	CHEMBL1232461	810	42	36
83	CHEMBL2205766	10000	10000	10000
84	CHEMBL1828980	6000	N/A	24600
85	CHEMBL1828978	3000	N/A	4800
86	CHEMBL2430884	25119	12589	19953
87	CHEMBL1828979	34300	N/A	0.8
88	CHEMBL1950957	30000	30000	30000

Base de referencia para el cálculo de SARI:

		ECFP4			Actividad (pIC50)			*Score*	
	# Cmpds	Min.	Máx.	Media	Min.	Máx.	Media	Cont	Disct
ECA	365	0.02	1.00	0.22	4.00	10.53	6.88	0.77	0.61
AR	652	0.02	1.00	0.15	4.00	9.88	6.45	0.84	0.39
HMG-CoA	199	0.02	1.00	0.22	4.12	10.85	7.29	0.73	0.65
PGD2	98	0.08	1.00	0.38	4.00	9.56	6.53	0.59	1.00
ASK-1	122	0.03	0.90	0.18	4.08	8.22	6.17	0.79	0.33
OATL4	262	0.04	0.82	0.29	4.00	9.0	6.79	0.69	0.73

En orden de mención: ECA= Enzima Convertidora de Angiotensina, AR=Aldosa Reductasa, HMG-CoA= 3-hidroxi-3-metilglutaril-coenzima A, PGD2= Receptor Prostaglandina D2, ASK-1= Cinasa 1 reguladora de la señal de apoptosis y OATL4= Proteina transportadora 4 de aniones orgánicos.

A.3 Script elaborado en Python 3.5.2

```
import os
import csv
import itertools
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math
import seaborn as sns

print("\n")
print("#############################################################################################")
print("####Antes de usar obtener MayaChem Tools e incluirlo en variables de entorno                 ####")
print("####Script generado para trabajar con actividad en nM                                       ####")
print("####Se requiere el formato de las actividades sin ceros, espacios en blanco y con ext.csv   ####")
print("#############################################################################################")

# Cambiar wd a cwd:
cd=os.getcwd()
os.chdir('%s'%(cd))

# Calcular FP y SM con script de MayaChem Tools:
print("\n")
archivoSDF=input("Dame el nombre del archivo .sdf: ")
actividades=input("Dame el nombre del archivo .csv con la lista de actividades: ")

print("\n")
print(" Elige un FP para calcular el SAS map:\n")
print(" 1: MACCS key 166 bits")
print(" 2: ECFP d=4")
print(" 3: ECFP d=6")
print(" 4: ECFP d=8")
print(" 5: Consenso: MACCS key 166 & ECFP4")
FP=int(input(">"))

while FP <1 or FP>5:
        print (" ¡¡¡¡No es una opción correcta!!!!!")
        print("\n")
        print(" Elige un FP para calcular el SAS map:\n")
        print(" 1: MACCS key 166 bits")
        print(" 2: ECFP d=4")
        print(" 3: ECFP d=6")
        print(" 4: ECFP d=8")
        print(" 5: Consenso: MACCS Key 166 & ECFP4 r=7")
        FP=int(input(">"))

if FP == 1 :
        os.system ("MACCSKeysFingerprints.pl -r MACCSFP_%s -o %s.sdf" %(archivoSDF,archivoSDF))
        data=os.system ("SimilarityMatricesFingerprints.pl -o --InputDataMode LoadInMemory --
        OutMatrixFormat RowsAndColumns --OutMatrixType LowerTriangularMatrix MACCSFP_%s.csv"
        %(archivoSDF))
        SMnombre="MACCSFP_%sTanimotoSimilarity"%(archivoSDF)
        graf="MACCSFP_%sTanimotoSimilarity"%(archivoSDF)
```

```
elif FP== 2:
        os.system ("ExtendedConnectivityFingerprints.pl -r ECFPR4_%s -n 2 -o %s.sdf" %(archivoSDF,archivoSDF))
        data=os.system ("SimilarityMatricesFingerprints.pl -o --InputDataMode LoadInMemory --
        OutMatrixFormat RowsAndColumns --OutMatrixType LowerTriangularMatrix ECFPR4_%s.csv"
        %(archivoSDF))
        SMnombre="ECFPR4_%sTanimotoSimilarityAlgebraicForm"%(archivoSDF)
        graf="ECFPR4_%sTanimotoSimilarity"%(archivoSDF)

elif FP== 3:
        os.system ("ExtendedConnectivityFingerprints.pl -r ECFPR5_%s -n 3 -o %s.sdf" %(archivoSDF,archivoSDF))
        data=os.system ("SimilarityMatricesFingerprints.pl -o --InputDataMode LoadInMemory --
        OutMatrixFormat RowsAndColumns --OutMatrixType LowerTriangularMatrix ECFPR5_%s.csv"
        %(archivoSDF))
        SMnombre="ECFPR5_%sTanimotoSimilarityAlgebraicForm"%(archivoSDF)
        graf="ECFPR5_%sTanimotoSimilarity"%(archivoSDF)

elif FP== 4:
        os.system ("ExtendedConnectivityFingerprints.pl -r ECFPR6_%s -n 4 -o %s.sdf" %(archivoSDF,archivoSDF))
        data=os.system ("SimilarityMatricesFingerprints.pl -o --InputDataMode LoadInMemory --
        OutMatrixFormat RowsAndColumns --OutMatrixType LowerTriangularMatrix ECFPR6_%s.csv"
        %(archivoSDF))
        SMnombre="ECFPR6_%sTanimotoSimilarityAlgebraicForm"%(archivoSDF)
        graf="ECFPR6_%sTanimotoSimilarity"%(archivoSDF)

elif FP== 5:
        os.system ("MACCSKeysFingerprints.pl -r MACCSFP_%s -o %s.sdf" %(archivoSDF,archivoSDF))
        data=os.system ("SimilarityMatricesFingerprints.pl -o --InputDataMode LoadInMemory --
        OutMatrixFormat RowsAndColumns --OutMatrixType LowerTriangularMatrix MACCSFP_%s.csv"
        %(archivoSDF))
        SMnombre1="MACCSFP_%sTanimotoSimilarity"%(archivoSDF)
        graf1="MACCSFP_%sTanimotoSimilarity"%(archivoSDF)
        os.system ("ExtendedConnectivityFingerprints.pl -r ECFPR4_%s -n 2 -o %s.sdf" %(archivoSDF,archivoSDF))
        data=os.system ("SimilarityMatricesFingerprints.pl -o --InputDataMode LoadInMemory --
        OutMatrixFormat RowsAndColumns --OutMatrixType LowerTriangularMatrix ECFPR4_%s.csv"
        %(archivoSDF))
        SMnombre2="ECFPR4_%sTanimotoSimilarityAlgebraicForm"%(archivoSDF)
        graf2="ECFPR4_%sTanimotoSimilarity"%(archivoSDF)

# Acomodar datos de la matriz de similitud en una lista [SM]:
if FP==5:
        data1 = pd.read_csv('%s.csv'%(SMnombre1),header=None)
        data1.drop([0],axis=0,inplace=True)
        data1.drop([0],axis=1,inplace=True)
        data1.fillna(0,inplace=True)
        np.fill_diagonal(data1.values, 0)
        r1=data1.values.T.tolist()
        merged1 = list(itertools.chain.from_iterable(r1))
        s1 = [x for x in merged1 if x != 0]
        #print(s1)
        SM1=[]
        for i in s1:
                i = float(i)
                SM1.append(i)
        data2 = pd.read_csv('%s.csv'%(SMnombre2),header=None)
        data2.drop([0],axis=0,inplace=True)
        data2.drop([0],axis=1,inplace=True)
        data2.fillna(0,inplace=True)
        np.fill_diagonal(data2.values, 0)
        r2=data2.values.T.tolist()
        merged2 = list(itertools.chain.from_iterable(r2))
        s2 = [x for x in merged2 if x != 0]
```

```
            #print(s2)
            SM2=[]
            for h in s2:
                        h = float(h)
                        SM2.append(h)
            SMCons= [(sum(i)/2) for i in zip (SM1,SM2)]


else:
            data = pd.read_csv('%s.csv'%(SMnombre),header=None)
            data.drop([0],axis=0,inplace=True)
            data.drop([0],axis=1,inplace=True)
            data.fillna(0,inplace=True)
            np.fill_diagonal(data.values, 0)
            r=data.values.T.tolist()
            merged = list(itertools.chain.from_iterable(r))
            s = [x for x in merged if x != 0]
            #print(s)
            SM=[]
            for i in s:
                        i = float(i)
                        SM.append(i)

  ##############################################################################################
  # Calcular los pIC50 y sus diferencias en valor absoluto, guárdalas en una lista [AbsAct]    ##
  ##############################################################################################

# Extraer los datos de actividad y generar una sola lista:
A = pd.read_csv('%s.csv'%(actividades),header=None)
A1=A.values.T.tolist()
A1 = list(itertools.chain.from_iterable(A1))
# Convertir la actividad nM en M:
A2=[]
for f in A1:
            f= (f/1000000000)
            A2.append(f)
# Convertir la actividad M en -log():
A3=[]
for z in A2:
            z= math.log10(z)
            z *=-1
            A3.append(z)
# Generar las combinaciones y calcular la diferencia de pIC50 de cada combinación:
A4=[]
for c in itertools.combinations(A3,2):
            A4.append(c)
AbsAct=[]
for i in itertools.starmap(lambda x,y: (abs(y-x)),A4):
            AbsAct.append(i)

if FP==5:
            # Calcular la Mediana de los valores de SM1:
            F1= sorted(SM1)
            if len(F1) % 2 == 0:
               n1 = len(F1)
               medianaSM1 = (F1[n1//2-1]+ F1[n1//2] )/2
            else:
               medianaSM1 = (F1[len(F1)//2])
            # Calcular la Mediana de los valores de SM2:
            F2= sorted(SM2)
            if len(F2) % 2 == 0:
               n2 = len(F2)
```

```
            medianaSM2 = (F2[n2//2-1]+ F2[n2//2] )/2
        else:
            medianaSM2 = (F2[len(F2)//2])
        FCons= sorted(SMCons)
        #Calcular la Mediana de los valores de SMCons
        if len(FCons) % 2 == 0:
            n3 = len(FCons)
            medianaSMCons = (FCons[n3//2-1]+ FCons[n3//2] )/2
        else:
            medianaSMCons = (FCons[len(FCons)//2])
else:
  # Calcular la Mediana de los valores de SM:
        F= sorted(SM)
        if len(F) % 2 == 0:
            n = len(F)
            medianaSM = (F[n//2-1]+ F[n//2] )/2
        else:
            medianaSM = (F[len(F)//2])
#Plot maximo valor del eje y:
if max(AbsAct)<=1:
        maxvalueAbsAct= 1.15
else:
        maxvalueAbsAct= max(AbsAct)*1.15

                                    ############################
                                    # Generar del gráfico:  #####
                                    ############################

if FP==5:
        # Convertir las listas en arrays:
        x1= np.asarray(SM1)
        y1= np.asarray(AbsAct)
        # Definir el plt como fig:
        fig, ax= plt.subplots()
        # Generar el plot:
        line,= plt.plot(x1,y1, "o", color="green", alpha=0.5, picker=5)
        maxvalueSM1= max(SM1)
        # Definir cuadrantes, adición de eje x y eje y:
        plt.plot([medianaSM1, medianaSM1], [-0.1,maxvalueAbsAct], '--', color="red")
        plt.plot([0,maxvalueSM1*1.15], [1,1], "--",color="red")
        # Definicón de nombre de los ejes, margenes y cuadriculado:
        plt.xlabel("%s"%(graf1))
        plt.ylabel("|Delta pIC50|")
        plt.title("SAS map %s\n" %(archivoSDF))
        plt.margins(y=0.05, x=0.05)
        plt.grid(True)
        # Generar la lista de etiquetas:
        p1=len(A3)
        q1= range(1,p1+1)
        k1=[]
        for c in itertools.combinations(q1,2):
                k1.append(c)


                        ####################################################
                        #En caso de querer las etiquetas sobre cada punto  #
                        #habilitar los siguientes comandos:                #
                        ####################################################

        #for h, txt in enumerate(k1):
        #       plt.annotate(txt, (x[h],y[h]))
```

```
# Para al dar click mostrar el número de los compuestos:
r1=np.asarray(k1)
def onpick1(event):
   thisline = event.artist
   ndata = thisline.get_xdata()
   zdata = thisline.get_ydata()
   ind = event.ind
   points = tuple(zip(ndata[ind], zdata[ind]))
   # Para hacer aparecer las coordenas de cada punto:
   print ('onpick scatter MACCSKeys:', r1[ind], points)
   # print ('onpick scatter MACCSKeys:', r1[ind])
fig.canvas.mpl_connect('pick_event', onpick1)
# Para mostrar el gráfico:
# plt.show()
# Convertir las listas en arrays:
x2= np.asarray(SM2)
y2= np.asarray(AbsAct)
# Definir el plt como fig:
fig, ax= plt.subplots()
# Generar el plot:
line,= plt.plot(x2,y2, "o", color="blue", alpha=0.5, picker=5)
maxvalueSM2= max(SM2)
# Definir cuadrantes, adición de eje x y eje y:
plt.plot([medianaSM2, medianaSM2], [-0.1,maxvalueAbsAct], '--', color="red")
plt.plot([0,maxvalueSM2*1.15], [1,1], "--",color="red")
# Definicón de nombre de los ejes, margenes y cuadriculado:
plt.xlabel("%s"%(graf2))
plt.ylabel("|Delta pIC50|")
plt.title("SAS map %s\n" %(archivoSDF))
plt.margins(y=0.05, x=0.05)
plt.grid(True)
# Generar la lista de etiquetas:
p2=len(A3)
q2= range(1,p2+1)
k2=[]
for c in itertools.combinations(q2,2):
        k2.append(c)
            ##################################################
            #En caso de querer las etiquetas sobre cada punto #
            #habilitar los siguientes comandos:               #
            ##################################################

#for h, txt in enumerate(k2):
#        plt.annotate(txt, (x[h],y[h]))
# Para al dar click mostrar el número de los compuestos:
r2=np.asarray(k2)
def onpick2(event):
   thisline = event.artist
   ndata = thisline.get_xdata()
   zdata = thisline.get_ydata()
   ind = event.ind
   points = tuple(zip(ndata[ind], zdata[ind]))
   # Para hacer aparecer las coordenas de cada punto:
   print ('onpick scatter ECPF4:', r2[ind], points)
   # print ('onpick scatter ECPF4:', r2[ind])
fig.canvas.mpl_connect('pick_event', onpick2)
# Para mostrar el gráfico:
# plt.show()
# Para generar el consensous map:
# Convertir las listas en arrays:
x3= np.asarray(SMCons)
```

```
y3= np.asarray(AbsAct)
# Definir el plt como fig:
fig, ax= plt.subplots()
# Generar el plot:
line,= plt.plot(x3,y3, "o", color="grey", alpha=0.5, picker=5)
maxvalueSMCons= max(SMCons)
# Definir cuadrantes, adición de eje x y eje y:
plt.plot([medianaSMCons, medianaSMCons], [-0.1,maxvalueAbsAct], '--', color="red")
plt.plot([0,maxvalueSMCons*1.15], [1,1], "--",color="red")
# Definicón de nombre de los ejes, margenes y cuadriculado:
plt.xlabel("MACCS & ECFP4 TanimotoSimilarity")
plt.ylabel("|Delta pIC50|")
plt.title("SAS map %s\n" %(archivoSDF))
plt.margins(y=0.05, x=0.05)
plt.grid(True)
# Generar la lista de etiquetas:
p3=len(A3)
q3= range(1,p3+1)
k3=[]
for c in itertools.combinations(q3,2):
        k3.append(c)

            ##################################################
            #En caso de querer las etiquetas sobre cada punto #
            #habilitar los siguientes comandos:               #
            ##################################################

#for h, txt in enumerate(k3):
# plt.annotate(txt, (x[h],y[h]))
# Para al dar click mostrar el número de los compuestos:
r3=np.asarray(k3)
def onpick3(event):
   thisline = event.artist
   ndata = thisline.get_xdata()
   zdata = thisline.get_ydata()
   ind = event.ind
   points = tuple(zip(ndata[ind], zdata[ind]))
   # Para hacer aparecer las coordenas de cada punto:
   print ('onpick scatter MACCS\ECFP4:', r3[ind], points)
   # print ('onpick scatter MACCS\ECFP4:', r3[ind])
fig.canvas.mpl_connect('pick_event', onpick3)
# Para cauntificar el total de datos del Consensus SAS map:
longk=(len(k3))
lCuadranteI= []
lCuadranteII= []
lCuadranteIII= []
lCuadranteIV= []
lSobrelog1menormediana= []
lSobrelog1mayormediana= []
lSobremedianamenorlog1= []
lSobremedianamayorlog1= []
lPuntosenelcentro= []
for x,y in zip(SMCons,AbsAct):
        CuadranteI= x<medianaSMCons and y<1
        CuadranteII= x>medianaSMCons and y<1
        CuadranteIII= x>medianaSMCons and y>1
        CuadranteIV= x<medianaSMCons and y>1
        Sobrelog1menormediana= x<medianaSMCons and y==1
        Sobrelog1mayormediana= x>medianaSMCons and y==1
        Sobremedianamenorlog1= (x==medianaSMCons and y<1)
        Sobremedianamayorlog1= (x==medianaSMCons and y>1)
```

```
                    Puntosenelcentro= x==medianaSMCons and y==1
                    lCuadranteI.append(CuadranteI)
                    lCuadranteII.append(CuadranteII)
                    lCuadranteIII.append(CuadranteIII)
                    lCuadranteIV.append(CuadranteIV)
                    lPuntosenelcentro.append(Puntosenelcentro)
                    lSobrelog1menormediana.append(Sobrelog1menormediana)
                    lSobrelog1mayormediana.append(Sobrelog1mayormediana)
                    lSobremedianamenorlog1.append(Sobremedianamenorlog1)
                    lSobremedianamayorlog1.append(Sobremedianamayorlog1)
            print ("VALORES DEL CONSENSUS SAS MAP")
            print ("El total de puntos en el cuadrante I (scaffold or R-hopping):", lCuadranteI.count(True), "y su
            porcentaje es:",(lCuadranteI.count(True)/longk)*100)
            print ("El total de puntos en el cuadrante II (Smooth SAR):", lCuadranteII.count(True), "y su porcentaje
            es:",(lCuadranteII.count(True)/longk)*100)
            print ("El total de puntos en el cuadrante III (Activity cliffs):", lCuadranteIII.count(True), "y su porcentaje
            es:",(lCuadranteIII.count(True)/longk)*100)
            print ("El total de puntos en el cuadrante IV (Nondescript):", lCuadranteIV.count(True), "y su porcentaje
            es:",(lCuadranteIV.count(True)/longk)*100)
            print ("El total de puntos sobre log(1) y menor a la mediana:", lSobrelog1menormediana.count(True), "y
            su porcentaje es:",(lSobrelog1menormediana.count(True)/longk)*100)
            print ("El total de puntos sobre log(1) y mayor a la mediana:", lSobrelog1mayormediana.count(True), "y su
            porcentaje es:",(lSobrelog1mayormediana.count(True)/longk)*100)
            print ("El total de puntos sobre la mediana y menor a log(1) :", lSobremedianamenorlog1.count(True), "y
            su porcentaje es:",(lSobremedianamenorlog1.count(True)/longk)*100)
            print ("El total de puntos sobre la mediana y mayor a log(1) :", lSobremedianamayorlog1.count(True), "y
            su porcentaje es:",(lSobremedianamayorlog1.count(True)/longk)*100)
            print ("El total de puntos en el centro son:", lPuntosenelcentro.count(True), "y su porcentaje
            es:",(lPuntosenelcentro.count(True)/longk)*100)
            print ("Total de puntos a graficar:", longk)
            # Para generar la lista en .csv:
            k3.insert(0, ("Compuesto1", "Compuesto2"))
            SM1.insert(0,"TantMACCKeys")
            SM2.insert(0,"TantECFP4")
            AbsAct.insert(0,"DifAbsAct")
            SMCons.insert(0,"PromMACCS_ECFP4")
            blanco= [" "]* (len(SMCons)+1)
            listacompleta=[]
            for i in zip(k3, SM1, SM2, AbsAct, SMCons, blanco):
                    i=i
                    listacompleta.append(i)
            listacompleta1=[listacompleta]
            with open('Tabla_comparativa_de_datos_%s.csv' %(archivoSDF), "w") as mycsvfile:
              thedatawriter = csv.writer(mycsvfile, delimiter="\n")
              for i in listacompleta1:
                thedatawriter.writerow(i)
            # Para mostrar el gráfico:
            plt.show()
    else:
            # Convertir las listas en arrays:
            x= np.asarray(SM)
            y= np.asarray(AbsAct)
            # Definir el plt como fig:
            fig, ax= plt.subplots()
            # Generar el plot:
            line,= plt.plot(x,y, "o", color="green", alpha=0.5, picker=5)
            maxvalueSM= max(SM)
            # Definir cuadrantes, adición de eje x y eje y:
            plt.plot([medianaSM, medianaSM], [-0.1,maxvalueAbsAct], '--', color="red")
            plt.plot([0,maxvalueSM*1.15], [1,1], "--",color="red")
            # Definicón de nombre de los ejes, margenes y cuadriculado:
```

```
plt.xlabel("%s"%(graf))
plt.ylabel("|Delta pIC50|")
plt.title("SAS map %s\n" %(archivoSDF))
plt.margins(y=0.05, x=0.05)
plt.grid(True)
# Generar la lista de etiquetas:
p=len(A3)
q= range(1,p+1)
k=[]
for c in itertools.combinations(q,2):
        k.append(c)

            ####################################################
            #En caso de querer las etiquetas sobre cada punto  #
            #habilitar los siguientes comandos:                #
            ####################################################

#for h, txt in enumerate(k):
# plt.annotate(txt, (x[h],y[h]))
# Para al dar click mostrar el número de los compuestos:
r=np.asarray(k)
def onpick(event):
   thisline = event.artist
   # ndata = thisline.get_xdata()
   # zdata = thisline.get_ydata()
   ind = event.ind
   # points = tuple(zip(ndata[ind], zdata[ind]))
   # Para hacer aparecer las coordenas de cada punto:
   # print ('onpick3 scatter:', r[ind], points)
   print ('onpick scatter:', r[ind])
fig.canvas.mpl_connect('pick_event', onpick)
# Para cuantificar el total de datos del Consensus SAS map:
longk=(len(k))
lCuadranteI= []
lCuadranteII= []
lCuadranteIII= []
lCuadranteIV= []
lSobrelog1menormediana= []
lSobrelog1mayormediana= []
lSobremedianamenorlog1= []
lSobremedianamayorlog1= []
lPuntosenelcentro= []
for x,y in zip(SM,AbsAct):
        CuadranteI= x<medianaSM and y<1
        CuadranteII= x>medianaSM and y<1
        CuadranteIII= x>medianaSM and y>1
        CuadranteIV= x<medianaSM and y>1
        Sobrelog1menormediana= x<medianaSM and y==1
        Sobrelog1mayormediana= x>medianaSM and y==1
        Sobremedianamenorlog1= (x==medianaSM and y<1)
        Sobremedianamayorlog1= (x==medianaSM and y>1)
        Puntosenelcentro= x==medianaSM and y==1
        lCuadranteI.append(CuadranteI)
        lCuadranteII.append(CuadranteII)
        lCuadranteIII.append(CuadranteIII)
        lCuadranteIV.append(CuadranteIV)
        lPuntosenelcentro.append(Puntosenelcentro)
        lSobrelog1menormediana.append(Sobrelog1menormediana)
        lSobrelog1mayormediana.append(Sobrelog1mayormediana)
        lSobremedianamenorlog1.append(Sobremedianamenorlog1)
        lSobremedianamayorlog1.append(Sobremedianamayorlog1)
print ("VALORES DEL SAS MAP")
```

```
print ("El total de puntos en el cuadrante I (scaffold or R-hopping):", lCuadranteI.count(True), "y su
porcentaje es:",(lCuadranteI.count(True)/longk)*100)
print ("El total de puntos en el cuadrante II (Smooth SAR):", lCuadranteII.count(True), "y su porcentaje
es:",(lCuadranteII.count(True)/longk)*100)
print ("El total de puntos en el cuadrante III (Activity cliffs):", lCuadranteIII.count(True), "y su porcentaje
es:",(lCuadranteIII.count(True)/longk)*100)
print ("El total de puntos en el cuadrante IV (Nondescript):", lCuadranteIV.count(True), "y su porcentaje
es:",(lCuadranteIV.count(True)/longk)*100)
print ("El total de puntos sobre log(1) y menor a la mediana:", lSobrelog1menormediana.count(True), "y
su porcentaje es:",(lSobrelog1menormediana.count(True)/longk)*100)
print ("El total de puntos sobre log(1) y mayor a la mediana:", lSobrelog1mayormediana.count(True), "y su
porcentaje es:",(lSobrelog1mayormediana.count(True)/longk)*100)
print ("El total de puntos sobre la mediana y menor a log(1) :", lSobremedianamenorlog1.count(True), "y
su porcentaje es:",(lSobremedianamenorlog1.count(True)/longk)*100)
print ("El total de puntos sobre la mediana y mayor a log(1) :", lSobremedianamayorlog1.count(True), "y
su porcentaje es:",(lSobremedianamayorlog1.count(True)/longk)*100)
print ("El total de puntos en el centro son:", lPuntosenelcentro.count(True), "y su porcentaje
es:",(lPuntosenelcentro.count(True)/longk)*100)
print ("Total de puntos a graficar:", longk)
print ("El valor de la mediana de SM es:", medianaSM)
# Para generar la lista en .csv:
k.insert(0, ("Compuesto1", "Compuesto2"))
SM.insert(0,"%s" %(graf))
AbsAct.insert(0,"DifAbsAct")
blanco= [" "]* (len(SM)+1)
listacompleta=[]
for i in zip(k, SM, AbsAct, blanco):
        listacompleta.append(i)
listacompleta1=[listacompleta]
with open('Tabla_comparativa_de_datos_%s.csv' %(archivoSDF), "w") as mycsvfile:
  thedatawriter = csv.writer(mycsvfile, delimiter="\n")
  for i in listacompleta1:
    thedatawriter.writerow(i)
plt.show()
```