UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

DISEÑO COMPUTACIONAL DE MODULADORES DE DNMT-3B

PROYECTO DE INVESTIGACIÓN

PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA:

Químico Oscar Palomino Hernández

Dr. José Luis Medina Franco
Facultad de Química, UNAM

Ciudad de México, Enero 2018

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

DISEÑO COMPUTACIONAL DE MODULADORES DE DNMT-3B

PROYECTO DE INVESTIGACIÓN

PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA:

Químico Oscar Palomino Hernández

Dr. José Luis Medina Franco
Facultad de Química, UNAM

Ciudad de México, Enero 2018

Este trabajo se realizó en el cubículo 108 del Edificio F de la H. Facultad de Química de la UNAM

Durante la realización de este proyecto fueron publicados los siguientes tres artículos, los cuales se anexan al final de este trabajo escrito:

- Fernanda I. Saldívar-González, J. Jesús Naveja, **Oscar Palomino-Hernández**, José L. Medina-Franco (2017). Getting SMARt in Drug Discovery: Chemoinformatics Approaches for Mining Structure-Multiple Activity Relationships. RSC Advances, 7: 632-641.
- **Oscar Palomino-Hernández**, A. Christiaan Jardínez-Vera and José L. Medina- Franco (2017). Progress on the Computational Development of Epigenetic Modulators of DNA Methyltransferases 3A and 3B. Journal of the Mexican Chemical Society, 61(3): 266-272
- **Oscar Palomino-Hernández** and José L. Medina- Franco (2017). Comparative Cheminformatic Analysis of Inhibitors of DNA Methyltransferases. Chemical Informatics, 3(2): 4

Resultados parciales de este proyecto fueron presentados en las siguientes conferencias:

- "Comparative Cheminformatic Analysis of DNA Methyltransferase Inhibitors". **Oscar Palomino-Hernández** and José Luis Medina-Franco. 254th American Chemical Society (ACS) 2017 National Meeting and Exposition. Chemical Information Division. 20 al 24 de agosto de 2017 en Washington, D.C., Estados Unidos.
- "Diversity metrics in focused libraries: Self-Organizing Maps for drug discovery". **Oscar Palomino-Hernández** and José Luis Medina-Franco. 52° Congreso Mexicano de Química y 36° Congreso Nacional de Educación Química. División de Química Teórica y Computacional. 26 al 29 de septiembre de 2017 en Puerto Vallarta, Jalisco, México.

Agradecimientos

Antes que nada quisiera agradecer a mi familia, en especial a mis padres y a mi hermana, por apoyarme durante mis estudios de maestría. Gracias a sus consejos, su paciencia y sus voces de ánimo he llegado hasta este punto, manteniéndome siempre atento a las nuevas oportunidades que la vida puede brindar. Gracias por todo.

Al Dr. José Luis Medina Franco por abrirme las puertas de su grupo de investigación y apoyarme con sus consejos y observaciones. Me siento muy satisfecho de haber sido parte de este grupo en este tiempo, y me llevo no solo conocimiento de quimioinformática, sino de muchas cosas más. Muchísimas gracias.

A todos mis compañeros del grupo DIFACQUIM por las charlas y los seminarios. Muchas gracias a todos por las pláticas amenas y los buenos momentos compartidos, tanto en la primera generación (Fernando, Fernanda, Mario Omar, Mariana), como en la segunda (Bárbara, Norberto, Eduardo). Me siento muy feliz de haberlos conocido, y sé que estamos en un grupo lleno de gente talentosa. En especial me gustaría agradecer al Dr. Oscar Méndez Lucio y al M.C. Jesús Naveja por sus aportaciones a mi crecimiento académico.

A mis amigos de la maestría y de la licenciatura: muchas gracias por los momentos agradables que hemos pasado juntos. En especial me gustaría mencionar a Karina Zarco, por llegar a mi vida y brindarme su amistad, y a Bárbara Díaz, por seguir en ella. Me siento muy feliz por haberles conocido.

A mis sinodales, por su apoyo en la rápida revisión de este trabajo y sus acertados consejos.

Al proyecto "Programa de Apoyo a Proyectos de Investigación e Innovacioón Tecnológica"(PAPIIT) IA204016, por el apoyo para la adquisición de la licencia de *Molecular Operating Environment* (MOE).

Al Posgrado en Ciencias Químicas, por el apoyo para al asistencia al 254th ACS Meeting, en Washington, D.C.

Y finalmente a Conacyt por la beca 606003, la que da lugar a todo este trabajo.

Índice general

Índice de figuras

Resumen

La metilación en el ADN es un mecanismo epigenético mediado por una familia de enzimas llamadas ADN metiltransferasas. Existen tres ADN metiltransferasas con actividad catalítica: DNMT1, DNMT3A y DNMT3B, con blancos moleculares, localizaciones celulares y niveles de expresión bien diferenciados, lo que las convierte a cada una en un blanco particular. Se conoce que la inhibición de estos blancos moleculares está correlacionada con una reducción en la tasa de tumorigénesis y con una expresión aumentada de genes supresores de tumores. De esta manera, estas proteínas emergen como dianas biológicas modulables para el tratamiento del cáncer y de otas enfermedades.

Con línea en lo anterior, se han realizado varios esfuerzos con respecto al desarrollo de moduladores selectivos para las diferentes ADN metiltransferasas; sin embargo, existen pocos estudios comparativos entre las diferentes enzimas de esta familia. Por lo tanto, como un primer paso hacia el diseño de moléculas selectivas hacia DNMT3B, en este informe se reporta una caracterización quimioinformática completa de diferentes bibliotecas moleculares con inhibidores de DNMT1, DNMT3A y DNMT3B. Estos conjuntos de moléculas fueron analizadas en términos de propiedades fisicoquímicas, quimiotipos moleculares, relaciones estructura-actividad y espacio químico. La disponibilidad de información permitió analizar el enriquecimiento en actividad de diversos quimiotipos y en consecuencia sugerir motivos estructurales privilegiados. Los resultados de este trabajo indican diferencias significativas en los compuestos con actividad hacia las diferentes dianas moleculares. Además, los resultados dan origen a estudios posteriores para el diseño racional de inhibidores selectivos a las diferentes ADN metiltransferasas, y en particular hacia la ADN metiltransferasa 3B.

Capítulo 1

Antecedentes

Metilación en el ADN

La metilación del ADN ha sido identificada como una modificación epigenética clave de diversos procesos biológicos, los cuales abarcan desde la diferenciación y el desarrollo celular, hasta la inestabilidad del ADN y el desarrollo del cáncer[1,2]. Se ha observado que patrones de metilación anormales están involucrados en la transformación del tumor y su consecuente progresión, así indicando que estas disrupciones epigenéticas están asociadas con la tumorigénesis[3]. Estos patrones de metilación no son de naturaleza estocástica, puesto que se ha observado que tienden a silenciar a los genes supresores de tumores. Así, la inhibición de estos niveles de metilación anormales ha sido activamente empleada como quimioterapia en un intento para reactivar estos genes supresores de tumores[4].

ADN metiltransferasas

La metilación en el ADN se lleva a cabo por una serie de ADN metiltransferasas (DNMTs, por sus siglas en inglés), las cuales donan un grupo metilo de la *S*-adenosilmetionina a la quinta posición de la citosina (Figura 1.1)[5,6].

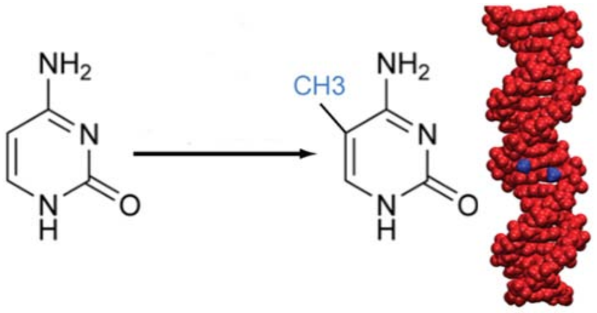


Figura 1.1: Reacción de metilación en el ADN

Tres ADN metiltransferasas, DNMT11, DNMT3A y DNMT3B, poseen esta habilidad catalítica en mamíferos[7]. En particular, DNMT1 es responsable de la metilación de hebras de ADN hemimetilado, por lo que esta enzima es responsable del mantenimiento de la metilación, mientras que la DNMT3A y DNMT3B participan tanto en el mantenimiento de la metilación, como en la metilación *de novo*[8,9]. Con respecto al contexto celular, existen diferentes roles para la DNMT3A y la DNMT3B, dado que la primera tiene una preferencia por la metilación de regiones pericentroméricas del ADN, mientras que la segunda promueve la metilación de regiones centroméricas del ADN[10].

Moduladores de ADN metiltransferasas

El tratamiento terapéutico más directo para tratar el cáncer que presenta patrones de hipermetilación es la reducción de la tasa de metilación del ADN, i.e., la inhibición de la actividad de las ADN metiltransferasas[11]. A la fecha, (noviembre 2017), la FDA ha aprobado dos fármacos que tienen como blancos moleculares a las DNMTs: azacitidina and decitabina, ambas para síndromes mielodisplásticos[12,13]. Sin embargo, estos fármacos actúan como inhibidores covalentes que se acompañan con múltiples efectos secundarios. Por ende, el diseño y desarrollo de inhibidores no covalentes de DNA metiltransferasa sigue en aumento[14,15].

Capítulo 2

Justificación y objetivo

Existen estudios previos con respecto al análisis de bases de datos de compuestos epigenéticos [16,17]; sin embargo, estos estudios no examinan a fondo las diferencias moleculares existentes entre los inhibidores para las tres ADN metiltransferasas, tales como propiedades fisicoquímicas y quimiotipos privilegiados. Más aún, una gran cantidad de compuestos han sido publicados en los últimos años como inhibidores de DNMTs, sin existir un estudio comparativo que considere la actividad y la selectividad hacia las diferentes enzimas.

Por tanto, el objetivo de este proyecto de investigación es la caracterización quimioinformática de un conjunto representativo de inhibidores de DNMT1, DNMT3A y DNMT3B, con el fin de describir las características químicas presentes para todos los inhibidores, organizarlos por actividad biológica, y comparar elementos estructurales para un mejor diseño racional.

Capítulo 3

Metodología

Creación y curación de la base de datos

La base de datos de compuestos con actividad hacia las tres DNMTs fue construída recabando información tanto de las bases de datos públicas más empleadas como ChEMBL [18], BindingDB [19] y HEMD [20], como de otros buscadores tales como "Web of Science" [21] y "SciFinder" [22]. La búsqueda se enfocó en artículos publicados desde 2010 hasta noviembre 2017.

El curado de este conjunto de moléculas se llevó a cabo siguiendo un protocolo previamente reportado [23]: una estructura canónica en notación lineal (InChI y SMILES) se obtuvo para cada molécula. Posteriormente, las moléculas fueron preparadas empleando el *software* Molecular Operating Environment (MOE) [24], el cual conserva el fragmento molecular más largo, remueve metales, neutraliza estados de protonación, y analiza si en la base de datos existen duplicados. Para compuestos idénticos con valores de actividad cercanos se empleó el promedio de los valores de actividad inhibitoria. Tras este procedimiento, se encontraron 351 moléculas únicas para DNMT1, 192 para DNMT3A y 86 para DNMT3B.

Al realizar el curado de las bases de datos para DNMT3A y DNMT3B, se encontró que una gran cantidad de compuestos no poseían valores de IC_{50}[1] pero tenían valores de porcentajes de inhibición a diferentes concentraciones de ligando. Con el fin de comparar entre diferentes medidas de actividad se clasificaron los valores continuos de actividad basada en un criterio heurístico: índice 4 si el pIC_{50} era mayor a 5.5, o el valor de inhibición era mayor al 75 %; índice 3 si el pIC_{50} era mayor a 5, o el valor de inhibición era mayor al 50 %; índice 2 si el pIC_{50} era mayor a 4, o el valor de inhibición era mayor al 25 %; e índice 1 si el pIC_{50} era menor a 4, o el valor de inhibición era menor al 25 %.

[1] Se define IC_{50} como la concentración necesaria de un compuesto para generar la inhibición de al 50 % de una o varias funciones biológicas. Con el fin de remover unidades y facilitar la comparación entre entidades químicas, se define pIC_{50} = -log IC_{50}.

Propiedades fisicoquímicas

Distribución de propiedades químicas

Como un primer paso en el análisis quimioinformático, fueron calculados descriptores químicos relevantes con ayuda de MOE y de utilidades de R Core Team[25] implementadas en RStudio[26]. Los descriptores químicos seleccionados fueron coeficiente de partición octanol/agua (logP), número de enlaces rotables (RB), número de donadores de puente de hidrógeno (HBD), número de aceptores de puente de hidrógeno (HBA), área total polar superficial (TPSA), y peso molecular (MW). Estos seis descriptores son comúnmente empleados para describir las propiedades farmacocinéticas de moléculas tipo fármaco (*drug-like*), tal como se describe por Lipinski[27] y Veber[28], y son igualmente empleados por diversos grupos, como el nuestro, para comparar entre diversas bases de datos moleculares. En adición a lo anterior, con el fin de analizar diferencias de índole estructural para estos grupos de moléculas, se calcularon seis descriptores topológicos: plano de mejor encaje (PBF, *plane of best fit*)[29], globularidad, fracción de carbonos sp^3, densidad de masa, radio de giro e índice de Wiener. Para estos descriptores se empleó una conformación de baja energía.

Análisis de correlación

La correlación entre dos descriptores X_1 y X_2 fue obtenida con el coeficiente de correlación de Pearson (r) con la siguiente ecuación

$$r_{X_1,X_2} = \frac{cov(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sum_{i=1}^{n}(x_{1i}-\overline{x_1})(x_{2i}-\overline{x_2})}{\sqrt{\sum_{i=1}^{n}(x_{1i}-\overline{x_1})^2\sum_{i=1}^{n}(x_{2i}-\overline{x_2})^2}} \tag{3.1}$$

donde $\overline{x_1}$ es el valor promedio del descriptor X_1, x_{1i} se refiere al dato i-ésimo presente para el descriptor X_1, y σ_{X_1} se refiere a la desviación estándar del descriptor X_1. Esta notación puede ser extrapolada para el descriptor X_2.

Para este análisis, se obtuvieron dos subconjuntos de compuestos activos (índice de actividad igual o mayor a 3) e inactivos (índice de actividad menor que 3), y para cada subconjunto se obtuvo una matriz de correlación de los predictores. Con lo anterior, un producto de Hadamard entre cada par de matrices se obtuvo, generando diversas matrices con valores r^2. Esta matriz puede tener valores en el rango de [-1,1], y la información que puede obtenerse se basa tanto en el signo (un signo positivo indica que la correlación del par de variables se mantiene, mientras que un signo negativo indica que la correlación del par de variables cambia), y en el valor (a mayor valor, la correlación es más fuerte y viceversa).

Análisis de diversidad estructural

Similitud intra e intergrupal

Para analizar la similitud intragrupal (esto es, de todos los inhibidores que actúan sobre una de las proteínas), la similitud para cada par de compuestos de la misma base de datos se calculó, expresándose como la función de distribución acumulada (*cumulative distribution function*, CDF), para diferentes huellas digitales (*fingerprints*) tales como Molecular Access System (MACCS) Keys, Extended Connectivity Fingerprints (ECFP, radio 4), y PubChem FP. De esta manera, los valores estadísticos obtenidos para el CDF fueron empleados para la comparación de las bases de datos.

El criterio de similitud seleccionado para la comparación fue el coeficiente de Tanimoto/Jaccard Index[30] donde en el numerador se evalúan los elementos presentes tanto en los compuestos A y B, mientras que en el denominador se evalúa la diferencia de los elementos presentes en A y B menos los elementos presentes en ambos elementos. Lo anterior se define como:

$$S_{AB} = \frac{|A \cap B|}{|A \cup B| - |A \cap B|} \tag{3.2}$$

Para analizar la similitud intergrupal, la similitud de un compuesto en uno de los conjuntos se calculó contra todos los compuestos de un segundo conjunto de moléculas. Dos técnicas de fusión de datos se emplearon para cada compuesto (*promedio* y *máximo*), lo que permitió la construcción de mapas de similitud multifusión[31].

Contenido de quimiotipos y evaluación de diversidad

Bajo la aproximación de Bemis y Murcko[32], las cadenas laterales de cada molécula fueron removidas, y el quimiotipo se obtuvo para cada compuesto. Empleando RStudio, se asignó un identificador único para cada quimiotipo. De esta forma, estos sistemas cíclicos representan clases químicas equivalentes, y cada molécula puede asignarse a sólo un quimiotipo. Para este análisis se calculó la similitud para cada par de quimiotipos, y se seleccionaron los quimiotipos más frecuentes.

Análisis de enriquecimiento de quimiotipos

Tomando en cuenta los valores de índice de actividad, para un cierto quimiotipo λ presente en uno de los conjuntos moleculares C, la actividad intrínseca del quimiotipo se obtuvo como

$$Act[C_\lambda] = \frac{1}{n_\lambda}\sum_{i=1}^{n_\lambda}[Indice\ de\ Actividad]_i \tag{3.3}$$

donde n_λ es el número de compuestos incluídos en el quimiotipo λ. Este valor muestra la actividad ponderada para cada quimiotipo, lo que permitió separar los quimiotipos en activos, intermedios e inactivos.

La actividad basal o promedio de un conjunto molecular se calculó como:

$$Act[C] = \frac{1}{n}\sum_{i=1}^{n}[Indice\ de\ Actividad]_i \tag{3.4}$$

donde n es el número total de compuestos para el conjunto molecular C.

El factor de enriquecimiento (E.F.) para el quimiotipo λ se obtuvo como:

$$EF[C_\lambda] = \frac{Act[C_\lambda]}{Act[C]} \tag{3.5}$$

el cual indica cuántas veces un quimiotipo es más (o menos) activo al compararlo con la actividad promedio de un conjunto de moléculas. Valores altos de E.F. para algún quimiotipo indican motivos estructurales muy atractivos en diseño de fármacos, dado que son estructuras con actividad mayor al promedio.

Panoramas de actividad

Se obtuvieron sendos panoramas de actividad[33,34] para las dianas epigenéticas estudiadas, describiendo en particular las áreas pertenecientes a las zonas de SAR continuo y acantilados de actividad. Este análisis se basa en realizar comparaciones tanto de similitud como de actividad entre pares de compuestos pertenecientes al mismo grupo molecular, permitiendo observar su distribución en un mapa de panorama de actividad: pares de compuestos similares en estructura pero no en actividad pertenecen a la zona de *acantilados de actividad*, mientras que pares de compuestos similares tanto en estructura como en actividad pertenecen a la zona de *SAR continuo*, y compuestos similares en actividad pero no en estructura pertenecen a la zona de *acantilados de similitud*. Los pares de estructuras que no caen en las zonas anteriores, se clasifican en la zona *no descriptiva*.

Espacio químico

Una reducción de la dimensionalidad del espacio de propiedades se realizó con análisis de componentes principales (PCA) y con mapas de autoorganización (SOM). El preprocesamiento se realizó empleando la paquetería de *caret*[35] en RStudio, mientras que la visualización se hizo con la paquetería *ggplot2*[36]. Los descriptores empleados para esta sección fueron los calculados en la sección de propiedades químicas.

Capítulo 4

Resultados

Creación y curado de la base de datos

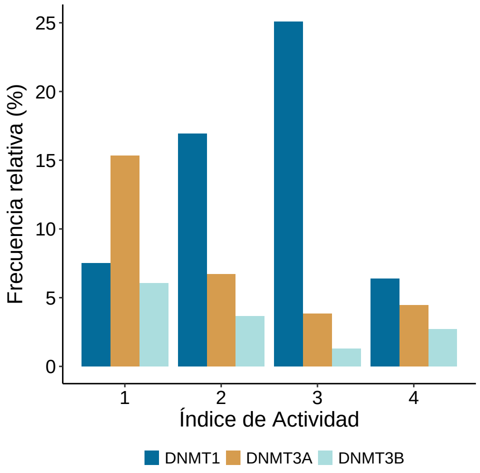


Figura 4.1: Distribución relativa de actividad de los compuestos seleccionados para las tres proteínas estudiadas.

Tabla 4.1: Resumen de la base de datos con la que se realizó el presente trabajo.

DNMT	Tamaño(n)	n(IC50)	n(IC50)	n(INDEX=4)	n(INDEX=3)	n(INDEX=2)	n(INDEX=1)
DNMT1	350	350	0	40 (11.5 %)	157 (45.0 %)	106 (30.0 %)	47 (13.5 %)
DNMT3A	190	35	155	28 (15.0 %)	24 (12.5 %)	42 (22.0 %)	96 (50.5 %)
DNMT3B	86	61	25	17 (20.0 %)	8 (9.0 %)	23 (27.0 %)	38 (44.0 %)

Se indican los inhibidores que existen para las tres proteínas en los cuatro niveles de actividad definidos. Igualmente, se muestran los porcentajes relativos a la base de datos de cada enzima.

En la Figura 4.1 se muestra la distribución de los compuestos en las diferentes categorías de la actividad, mientras que en la Tabla 4.1 se muestran la numeralia correspondiente. Se puede observar que se han publicado inhibidores con mejor actividad para DNMT1, lo cual puede deberse a que hay un mayor número de compuestos que se han desarrollado para dicha diana.

Distribución de las propiedades fisicoquímicas

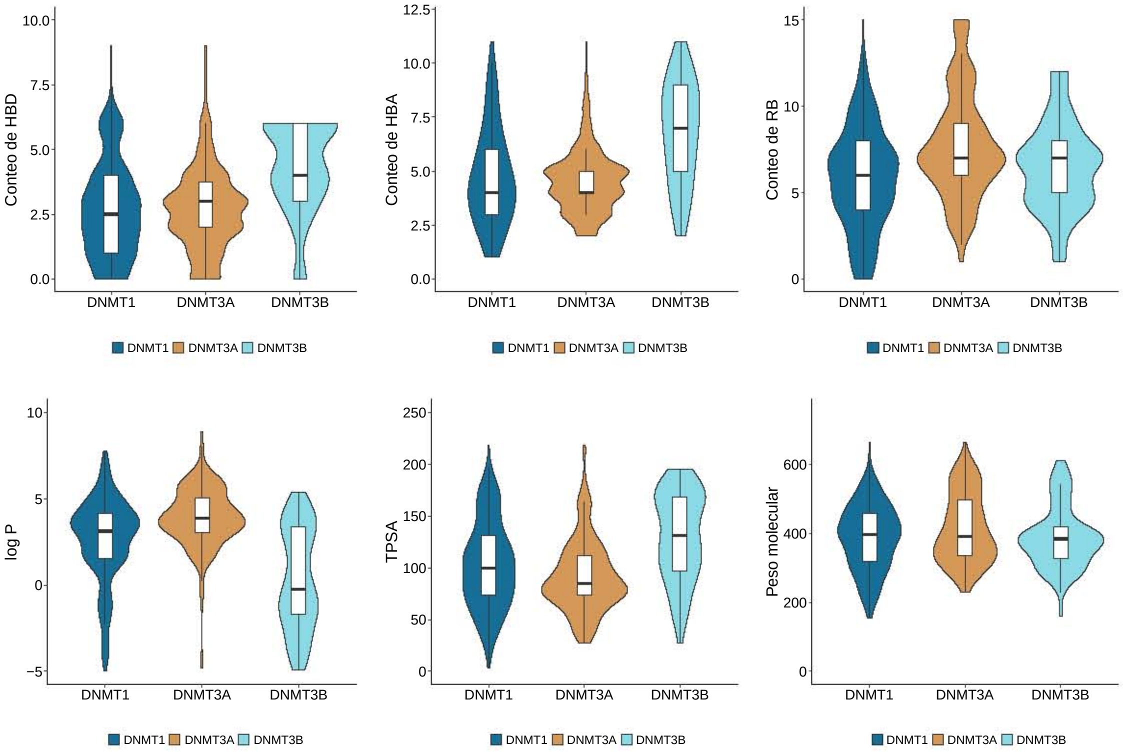


Figura 4.2: Distribución de las propiedades fisicoquímicas calculadas para las tres bases de datos.

Con la combinación de diagramas de caja y diagramas de violín se realizó la visualización de la distribución de propiedades fisicoquímicas para los tres conjuntos estudiados (Figura 4.2).

Para conteo de HBD (número de átomos donadores de puente de hidrógeno), DNMT1 y DNMT3A son similares en sus rangos y distribuciones, mientras que DNMT3B tiene valores superiores, lo que indica que este grupo molecular tiene más átomos electronegativos con hidrógenos polares al compararse con los otros grupos moleculares. Para conteo de HBA (ńumero de átomos aceptores de puente de hidrógeno), la tendencia se repite, mostrando que DNMT3B posee más átomos con pares de electrones libres. El conteo de RB (enlaces rotables) nos indica que los tres conjuntos poseen valores similares, lo que indicaría que, en general, las moléculas suelen ser igual de flexibles para las tres isoformas.

Los valores de logP muestras una clara separación para los inhibidores de DNMT3B, mostrándolos mucho más hidrofílicos que sus contrapartes para DNMT1 y DNMT3A. Con respecto a TPSA, DNMT1 y DNMT3B cubren un amplio rango, mientras que DNMT3A está mas centrado en un valor cercano a 80 Å^2. Finalmente, para el peso molecular, se observa que para DNMT1 y DNMT3A los valores están repartidos sobre un amplio rango, mientras que para DNMT3B tienden a estar centrados alrededor de 400 Da.

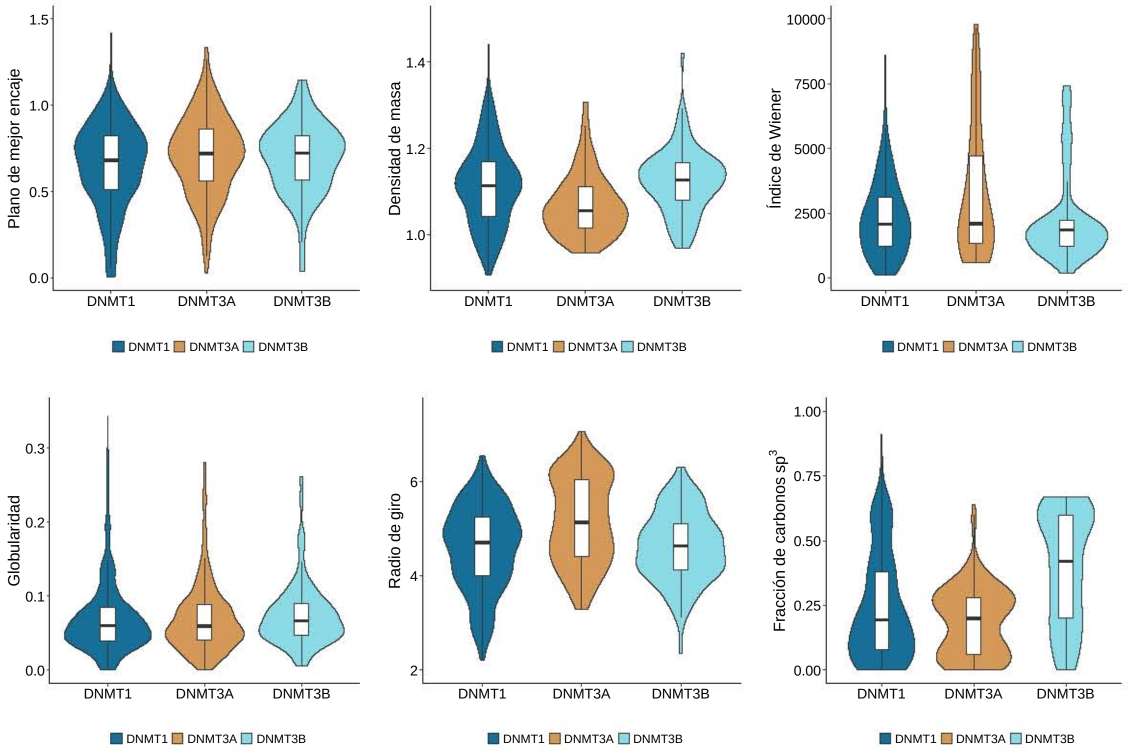


Figura 4.3: Distribución de las propiedades topológicas calculadas para las tres bases de datos.

Para los descriptores topológicos se observa que hay poca diferencia entre las diferentes distribuciones calculadas, tal como se muestra en la Figura 4.3. En particular se observa que DNMT3A tiene valores más altos de radio de giro, índice de Wiener, y menores valores de densidad de masa y fracción de carbonos sp^3; mientras que DNMT3B tiene mayores valores de fracción de carbonos sp^3. Estas diferencias pueden ser significativas para diseñar posteriormente moléculas selectivas a alguna de las ADN metiltransferasas.

Análisis de correlación

El analisis de correlación (obtención de la matriz de descriptores r^2) se realizó para los subconjuntos de DNMT1$_{\text{(activos vs. inactivos)}}$, DNMT3A$_{\text{(activos vs. inactivos)}}$ y DNMT3B$_{\text{(activos vs. inactivos)}}$, con el fin de observar si la tendencia entre los descriptores para un cierto grupo molecular se conservaba entre las moléculas activas e inactivas. De igual manera, tomando solamente los subconjuntos activos, se evaluaron las matrices r^2 entre DNMT1·DNMT3A, DNMT1· DNMT3B y DNMT3A·DNMT3B. Las matrices se muestran en la Figura 4.4. Para los inhibidores de DNMT1 se observan correlaciones positivas para la mayoría de los descriptores, lo que indica que el comportamiento de los descriptores no varía al comparar subconjuntos activos o inactivos. Se observa que tomando en cuenta todas las correlaciones, algunos descriptores siguen un comportamiento muy similar, tales como HBA y TPSA, o MW y Wiener, lo que puede implicar que para este grupo molecular, esos descriptores considerarse redundantes. Los inhibidores de DNMT3A muestran valores negativos importantes, como en el caso de HBD-RB o Wiener-TPSA, los cuales indican que la correlación cambia entre los subconjuntos activos e inactivos, haciendo estos pares de descriptores útiles en la separación del espacio activo correspondiente a DNMT3A. Se observa que HBD y TPSA tienen comportamiento global similar para este conjunto molecular. Para los compuestos con actividad hacia DNMT3B se encontraron valores negativos de correlación para R.Gyr-HBD y MW-Dens y tendencias globales similares para HBA y TPSA.

El análisis de subconjuntos activos mostró que para diferenciar entre DNMT1 y DNMT3A existen múltiples pares de descriptores, como HBD-Wiener, entre otros; para diferenciar entre DNMT1 y DNMT3B existen menos pares, tales como Dens-*f*sp^3; y para diferenciar entre DNMT3A y DNMT3B existen varios descriptores como RB-HBD entre otros. De esta forma, para separar el espacio químico de compuestos activos e inactivos entre tres dianas epigenéticas se pueden emplear descriptores como Dens, *f*sp^3, HBD y RB, los cuales maximizan la separación entre compuestos activos e inactivos.

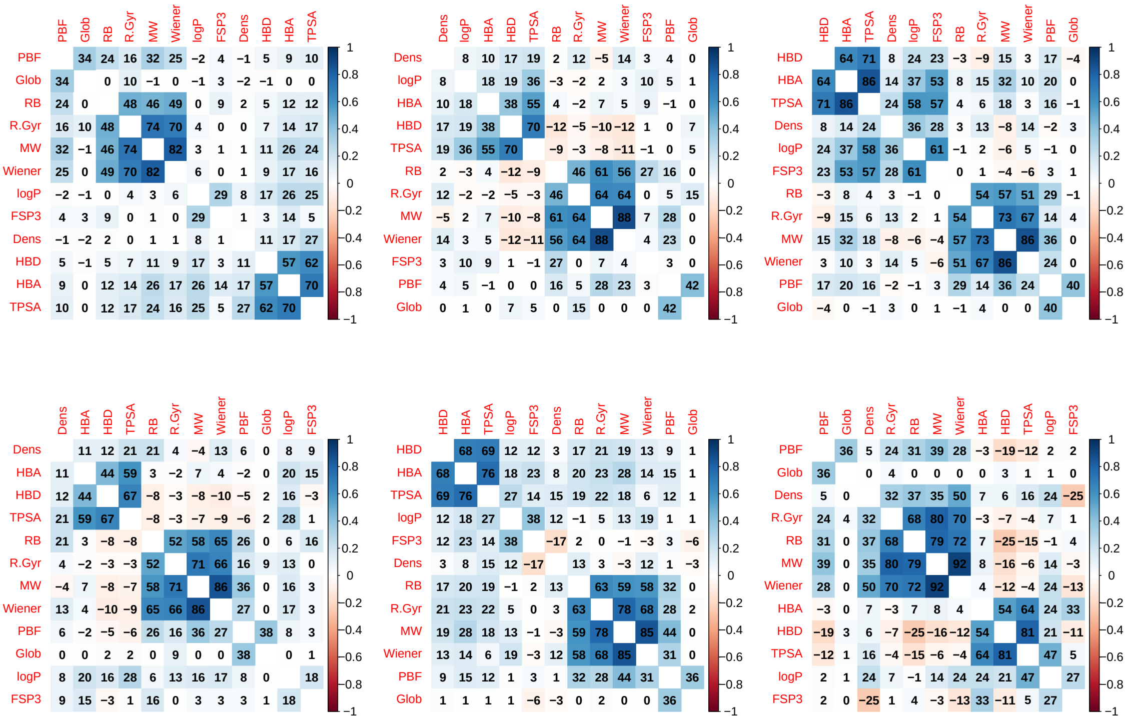


Figura 4.4: Análisis de correlación entre subconjuntos. **Arriba:** Correlaciones entre descriptores obtenidas al comparar subconjuntos activos e inactivos de DNMT1 (izquierda), DNMT3A (centro) y DNMT3B (derecha). **Abajo:** Correlaciones entre descriptores obtenidas al comparar subconjuntos activos entre DNMT1-DNMT3A (izquierda), DNMT1-DNMT3B (centro) y DNMT3A-DNMT3B (derecha).

Análisis de diversidad estructural

Similitud intragrupal

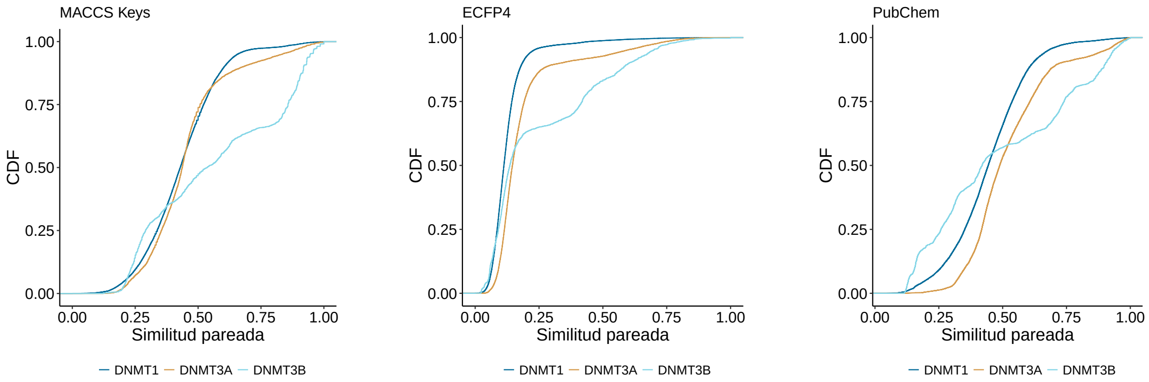


Figura 4.5: Funciones de distribución acumulada (CDF) sobre la similitud pareada para los tres conjuntos de inhibidores estudiados, representadas con MACCS keys (izquierda), ECFP4 (centro) and PubChem FP (derecha).

En la Figura 4.5 se observan las tres bases de datos con las diferentes representaciones de similitud, mientras que en la Tabla 4.2 se tienen valores estadísticos relevantes. Se observa que los valores de desviación estándar (SD.) son muy similares para DNMT1 y DNMT3A, pero mayores para DNMT3B; indicando que esta ultima distribución no está tan localizada en un punto. De igual manera, se observa que para algunos cuantiles DNMT1 es la más diversa, mientras que en otros cuantiles es DNMT3B. De esta manera, para realizar apropiadamente las comparaciones entre grupos moleculares, se decidió cuantificar el área bajo la curva (ABC.) para las distintas distribuciones. Se observó que, en general, los valores de ABC son mayores para DNMT1, posteriormente para DNMT3A y finalmente para DNMT3B, en la mayoría de los *fingerprints* empleados, lo que sugiere que el grupo de moléculas más diverso - estructuralmente hablando - es DNMT1, posteriormente DNMT3A y finalmente DNMT3B. De esta análisis, se observó que el *fingerprint* que mostraba un mejor comportamiento era ECFP4, por lo que se seleccionó para estudios posteriores.

Tabla 4.2: Valores estadísticos de las CDF para la similitud pareada.

DNMTs	DNMT1			DNMT3A			DNMT3B		
Fingerprint	MACCS	PubChem	ECFP4	MACCS	PubChem	ECFP4	MACCS	PubChem	ECFP4
Min.	0.00	0.04	0.00	0.12	0.12	0.03	0.10	0.10	0.00
1er Qu.	0.34	0.35	0.09	0.35	0.42	0.11	0.29	0.26	0.09
Mediana	0.43	0.44	0.11	0.44	0.49	0.15	0.54	0.42	0.14
Media	0.43	0.44	0.13	0.46	0.52	0.19	0.57	0.50	0.25
3er Qu.	0.52	0.54	0.14	0.51	0.61	0.19	0.86	0.74	0.42
Max.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SD.	0.14	0.15	0.08	0.16	0.16	0.15	0.27	0.28	0.22
ABC.	0.57	0.55	0.87	0.54	0.47	0.81	0.43	0.50	0.75

Similitud intergrupal

Se realizaron mapas de multifusión de similitud, los cuales se muestran en la Figura 4.6. Para el primer caso, al comparar tanto DNMT3A y como DNMT3B con DNMT1, se observa que DNMT3B tiende a formar agregados en la parte inferior izquierda del diagrama, mientras que DNMT3A cubre una amplia parte del diagrama. Esto implica que estructuralmente, hay un traslape estructural relativamente continuo de DNMT1 con DNMT3A, mientras que para DNMT3B tiende a ser menos similar. Para el segundo caso, al comparar DNMT1 y DNMT3B con respecto a DNMT3A, se observa que están distribuídas en forma similar, lo que indica que este grupo molecular está cercano - estructuralmente hablando - tanto de DNMT1 como de DNMT3B, con valores de similitud promedio cercanos a 0.12. Finalmente, al comparar DNMT1 y DNMT3A con respecto a DNMT3B, se observan dos aglomerados, uno con valores de similitud máxima mayores perteneciente a DNMT3A, y otro con valores de similitud máxima menores, perteneciente a DNMT31, lo que indica que DNMT3A es estructuralmente más similar a DNMTB que DNMT1.

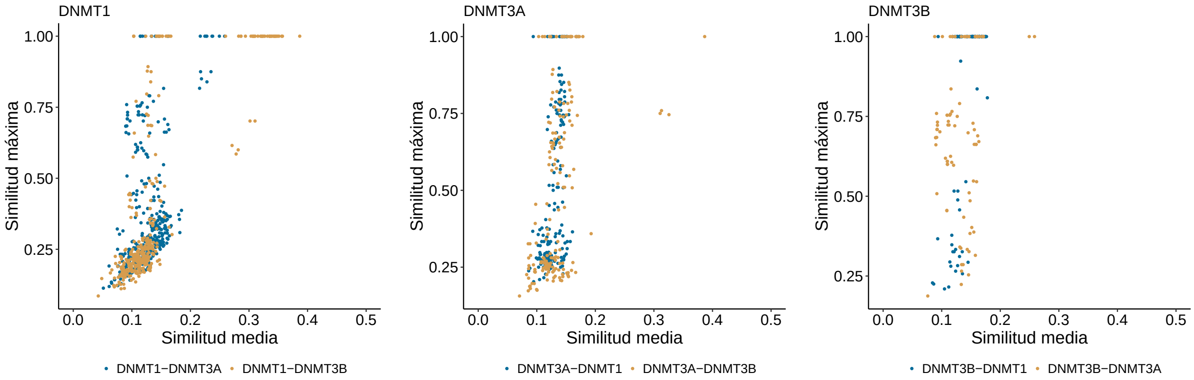
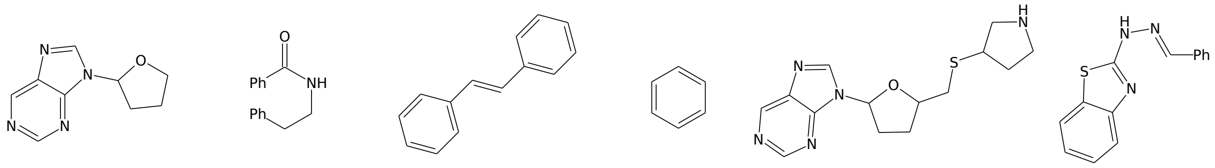


Figura 4.6: Mapas de multifusión de similitud basados en DNMT1 (izquierda), DNMT3A (centro) y DNMT3B (derecha).

Contenido de quimiotipos y evaluación de diversidad

Para los tres grupos de moléculas se obtuvieron los quimiotipos correspondientes, y se calculó la similitud intragrupal, indicando que la mediana de las funciones de distribución acumulada eran 0.125, 0.158 y 0.125 para DNMT1, DNMT3A y DNMT3B, respectivamente. Al comparar estos valores con los obtenidos anteriormente en la sección de **Similitud intragrupal**, se observa que el valor de la mediana de la similitud aumentó para DNMT1 y DNMT3A, mientras que disminuyó para DNMT3B. Esto indica que, en general, los conjuntos de los quimiotipos de DNMT1 y DNMT3A son ligeramente más diversos, mientras que el conjunto de quimiotipos de DNMT3B es ligeramente más similar entre sí.

Para los tres grupos moleculares se obtuvieron los quimiotipos más frecuentes (Figura 4.7). Estructuralmente se puede observar que los quimiotipos más frecuentes en DNMT1 están relacionados con el cofactor *S*-adenosilmetionina (SCAFF78 y SCAFF75), mientras que los quimiotipos con mayor frecuencia en DNMT3A tienden a ser dos fragmentos aromáticos unidos por un segmento carbonado de longitud variable. Los quimiotipos más frecuentes para DNMT3B igualmente están relacionados estructuralmente con el cofactor.



SCAFF78
DNMT1: 11 (3.14%)
DNMT3A: 4 (2.01%)
DNMT3B: 12 (12.90%)

SCAFF101
DNMT1: 1 (0.29%)
DNMT3A: 24 (12.06%)
DNMT3B: 1 (1.08%)

SCAFF315
DNMT1: 1 (0.29%)
DNMT3A: 12 (6.03%)
DNMT3B: 12 (12.90%)

SCAFF7
DNMT1: 8 (2.28%)
DNMT3A: 3 (1.50%)
DNMT3B: 3 (3.23%)

SCAFF75
DNMT1: 11 (3.13%)
DNMT3B: 11 (11.83%)

SCAFF320
DNMT3A: 14 (7.04%)

Figura 4.7: Quimiotipos más frecuentes para los grupos analizadas. Se muestra el número de compuestos con determinado quimiotipo para cada grupo, así como su frecuencia relativa en porcentaje

Diversidad de quimiotipos según su actividad

Como se especificó en el capítulo de **Metodología**, se obtuvo la actividad intrínseca de cada quimiotipo. Este cálculo permitió agruparlos en quimiotipos inactivos (con actividad intrínseca menor o igual a 2), intermedios (con actividad intrínseca mayor a dos pero menor o igual a 3), y activos (con actividad intrínseca mayor a 3). Tras el procedimiento anterior, se obutvieron las curvas de recobro de quimiotipos, las cuales se muestran en la Figura 4.8, mientras que en la Tabla 4.3 se muestran valores para caracterizar la diversidad. Para una discusión acerca de las métricas de diversidad, se refiere al lector a otro artículo[37].

Tabla 4.3: Métricas para la diversidad de scaffold para cada DNMT.

DNMTs	Actividad	M	N	Nsing	N/M	Nsing/M	Nsing/N	f_{50}	ABC	Mediana (ECFP4)
DNMT1	Inactiva	127	107	93	0.84	0.73	0.87	0.411	0.570	0.1014
	Intermedia	179	123	99	0.69	0.55	0.80	0.276	0.640	0.1053
	Activa	44	28	22	0.64	0.50	0.79	0.214	0.656	0.0978
DNMT3A	Inactiva	138	68	51	0.49	0.37	0.75	0.117	0.728	0.1465
	Intermedia	22	15	10	0.68	0.45	0.67	0.333	0.618	0.2115
	Activa	30	18	16	0.60	0.53	0.89	0.167	0.681	0.2187
DNMT3B	Inactiva	39	21	16	0.54	0.41	0.76	0.190	0.703	0.0750
	Intermedia	37	10	6	0.27	0.16	0.60	0.200	0.701	0.1379
	Activa	10	7	5	0.70	0.50	0.71	0.286	0.614	0.3793

N: número de quimiotipos; M: número de moléculas; Nsing: número de singuletes; f_{50}: fracción de quimiotipos que contienen el 50 % del grupo de moléculas; ABC: área bajo la curva.

Para DNMT1, el valor de la mediana de la similitud entre pares de quimiotipos para los tres subconjuntos es muy similar al obtenido en la sección anterior (0.125), lo que indica que los quimiotipos en cada subconjunto son estructuralmente diversos. Las medidas de diversidad global f_{50} y ABC indican que el subconjunto de quimiotipos inactivos es el más diverso. Un análisis más detallado indica que el subconjunto de quimiotipos inactivos tiene los valores más altos de N/M (quimiotipos por molécula), Nsing/M (quimiotipos singuletes por molécula) y Nsing/N (quimiotipos singuletes por quimiotipo). Para los otros dos subconjuntos, los valores de ABC y f_{50} indican que el subconjunto intermedio es más diverso que el subconjunto activo, siendo que esta diversidad proviene de los altos valores de N/M, Nsing/M y Nsing/N que tiene el subconjunto de quimiotipos con actividad intermedia. Para DNMT3A, el valor de la mediana de la similitud es menor para el subconjunto inactivo, pero aumenta para los subconjuntos intermedio y activo, lo . Esto sugiere que la diversidad estructural de DNMT3A está muy influenciada por el subconjunto de quimiotipos inactivos, y que los quimiotipos intermedios y activos son mucho menos diversos. Los valores de ABC y f_{50} indican que el subconjunto intermedio es el más diverso (por su alto valor de N/M), seguido del subconjunto activo (por su valor de Nsing/M y Nsing/N), y finalmente el subconjunto inactivo. Para DNMT3B, el valor de la mediana para el subconjunto inactivo es menor que el obtenido en la sección anterior (0.125), indicando que este subconjunto es el más diverso estructuralmente. En cambio, el subjunto activo tiene un valor mucho mayor, lo que lo convierte en el subconjunto con menor diversidad estructural. Las medidas de diversidad globales indican que el subconjunto activo es el más diverso en conteo de scaffolds (por su alto valor de N/M), seguido del subconjunto intermedio (por su valor de Nsing/M y Nsing/N), y finalmente el subconjunto inactivo.

En resumen, se observa que los inhibidores de DNMT1 tienden a cubrir amplias áreas del espacio químico, mientras que los inhibidores de DNMT3A y DNMT3B cubren áreas menores, y tienden a estar enfocados (DNMT3A y DNMT3B tienen quimiotipos activos poco diversos estructuralmente).

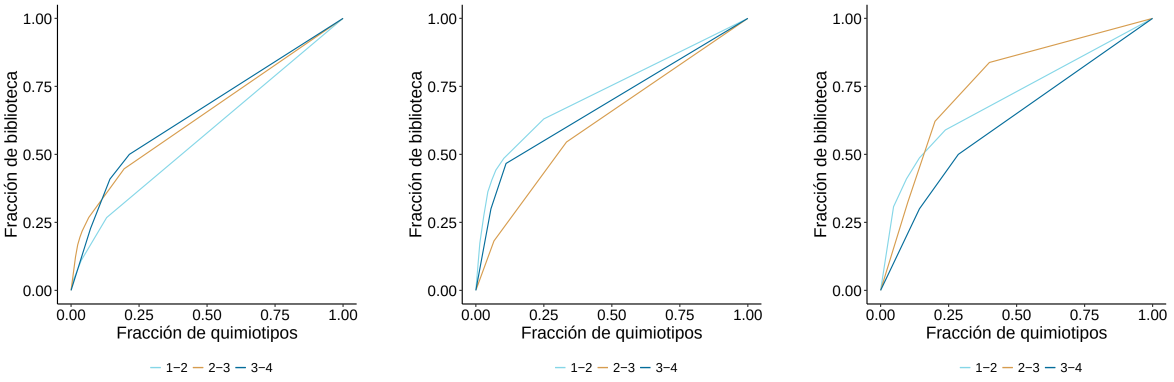


Figura 4.8: Curvas de recobro de quimiotipos para DNMT1 (izquierda), DNMT3A (centro) y DNMT3B (derecha).

Análisis de enriquecimiento de quimiotipos

Las actividades intrínsecas de cada quimiotipo fueron divididas entre la actividad de fondo de cada grupo, lo que permitió obtener el factor de enriquecimiento (E.F. por sus siglas en inglés) de cada quimiotipo. De esta manera, para cada grupo de moléculas se obtuvieron sus gráficos de enriquecimiento de quimiotipos[38], mostrados en la figura 4.9.

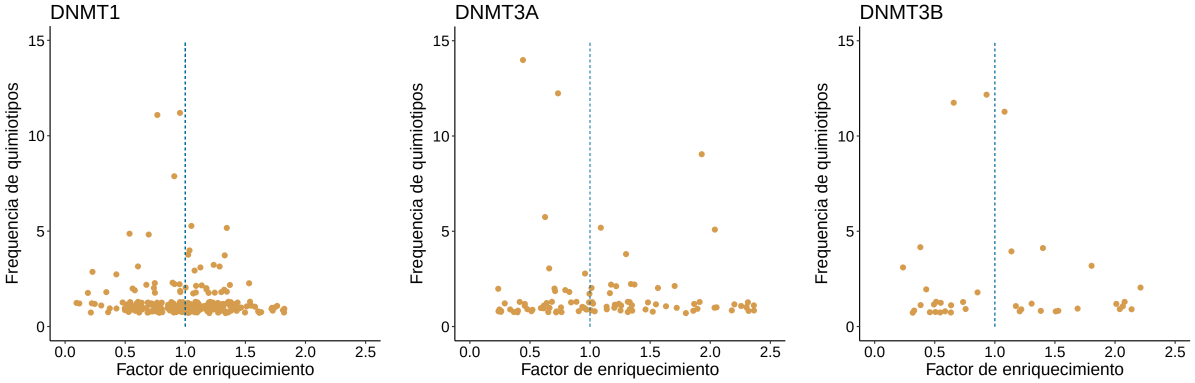


Figura 4.9: Gráficos de enriquecimiento de quimiotipos para DNMT1 (izquierda), DNMT3A (centro) y DNMT3B (derecha).

Para DNMT1 se observa que los quimiotipos están distribuídos de manera simétrica alrededor del valor de factor de enriquecimiento de 1 (*i.e.*, los quimiotipos en este valor tienen la misma actividad que el promedio del grupo molecular), con cerca del 55 % de los quimiotipos con valores de EF mayores que la unidad. Los quimiotipos de mayor frecuencia corresponden a SCAFF75, SCAFF78 y SCAFF7 (*vide supra*).

Para DNMT3A, la distribución de los quimiotipos se observa menos simétrica, con cerca de 61 % de los quimiotipos con valores de EF mayores a 1. Los quimiotipos más frecuentes corresponden a SCAFF101, SCAFF320 y SCAFF315. Se observa que para esta distribución existen varios quimiotipos con EF menor a 1, pero una frecuencia alta, lo que desplaza el valor del punto de corte hacia la izquierda.

Para DNMT3B, solamente 44 % de los quimiotipos tienen valores de EF mayores que 1. Los quimiotipos más frecuentes corresponden a SCAFF78, SCAFF315 y SCAFF75. De igual manera, se observa que hay varios quimiotipos con EF bajos y frecuencias altas, lo que puede desplazar el valor del punto de corte de actividad hacia la izquierda.

En general, estos resultados indican que DNMT1 ha sido explorada de manera uniforme con respecto a la actividad basal del grupo de moléculas, mientras que DNMT3A posee varios quimiotipos ampliamente explorados con EF diversos. Este resultado permite el estudio de regiones estructura-actividad continua con índole predictiva.

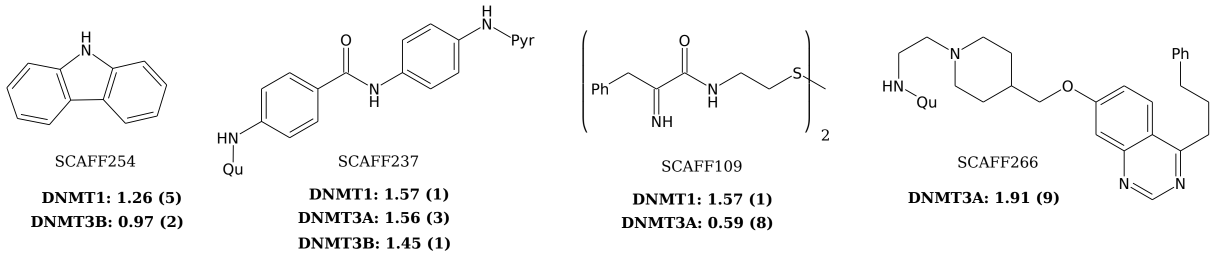


Figura 4.10: Otros quimiotipos encontrados en la base de datos. Se muestra el factor de enriquecimiento, y la frecuencia de cada quimiotipo en paréntesis para cada grupo (Ph = Fenilo, Qu = Quinolina, Pyr = Pirimidina).

Como se concluyó en la sección anterior, los quimiotipos más frecuentes mostrados en la Figura 4.7 muestran valores bajos de factor de enriquecimiento: SCAFF101 tiene un valor de EF de 0.52 (24 compuestos) para DNMT3A; SCAFF320 tiene un valor de EF de 0.70 (14 compuestos) para DNMT3A; y SCAFF315 tiene un EF de 0.78 (1 compuesto) para DNMT1 y de 0.60 (12 compuestos en cada grupo molecular) para DNMT3A y DNMT3B.

Otros quimiotipos interesantes encontrados se muestran en la Figura 4.10. En particular, SCAFF254 y SCAFF109 muestran selectividad para DNMT1; SCAFF266 muestra un alto valor de EF para DNMT3A, y SCAFF237 tiene valores altos de EF para las tres DNMTs.

Enriquecimiento sobre el quimiotipo del cofactor

Al analizar los quimiotipos obtenidos, se encontró que muchos de ellos tenían una estructura máxima común, la cual pertenecía al quimiotipo SCAFF78. Por tanto, tomando este quimiotipo como linea base, se analizó cómo los cambios sobre esta estructura influenciaban el factor de enriquecimiento. En la Figura 4.11 se muestra el quimiotipo base, y las modificaciones estructurales que se encontraron: en el lado izquierdo se muestran las modificaciones en **R1**, y en el lado derecho se muestran las modificaciones en **R2**. Para este análisis, se emplearon quimiotipos con un valor de frecuencia mayor o igual a 3.

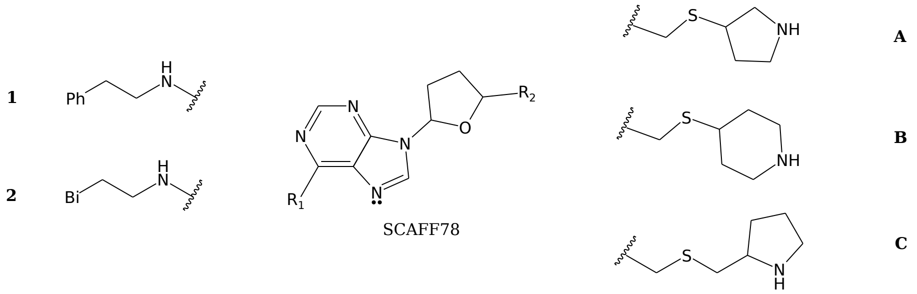


Figura 4.11: Quimiotipos encontrados en la base de datos con la máxima estructura de SCAFF78 (Ph = Fenilo, Bi = Bifenilo).

Tomando el EF del quimiotipo SCAFF78 como referencia (0.89 para DNMT1, 1.56 para DNMT3A y 1.2 para DNMT3B), y dejando fija la posición de **R2** como hidrógeno, se observó que cuando **R1** = **1**, el EF incrementaba sustancialmente para DNMT3B (1.18 para DNMT1 y 1.6 for DNMT3B), y cuando **R1** = **2**, el EF disminuía para DNMT1, pero aumentaba para DNMT3B (0.78 para DNMT1 y 1.9 para DNMT3B). Esto sugiere que elongar la cadena del quimiotipo sobre el punto **R1** puede mejorar la selectividad para análogos del cofactor para DNMT3B con respecto a DNMT1.

Al contrario, cuando **R1** es hidrógeno, se observa que la sustitución de **R1** = **A** no mejora el EF para las dianas (0.96 para DNMT1 y 1.05 para DNMT3B), mientras que la sustitución de **R2** = B disminuye el EF para DNMT1 mientras se conserva para DNMT3B (0.65 para DNMT1 y 1.09 for DNMT3B), y la sustitución con **R2** = **C** disminuye el EF para las dianas (0.78 para DNMT1 y 0.73 para DNMT3B). Esto implica que un sustituyente con una cadena más larga en **R2** tiende a disminuir la actividad, y que conservar un ciclo hexamembrado favorece la selectividad para DNMT3B respecto a DNMT1.

Los resultados anteriores se ven combinados en el quimiotipo SCAFF77, el cual tiene **R1** = **2** y **R2** = **A**, y tiene un EF para DNMT1 de 1.18 y por DNMT3B de 1.45. Dado que el valor de EF no aumentó tan drásticamente, se sugiere que los efectos entre **R1** y **R2** no interactúan de manera sinérgica positiva. Finalmente, se observa que sustituír el átomo de nitrógeno marcado con el par electrónico (Figura 4.11) por un átomo de carbono incrementa el EF para DNMT1 y DNMT3B a 1.57 y 1.94, respectivamente.

Panoramas de actividad

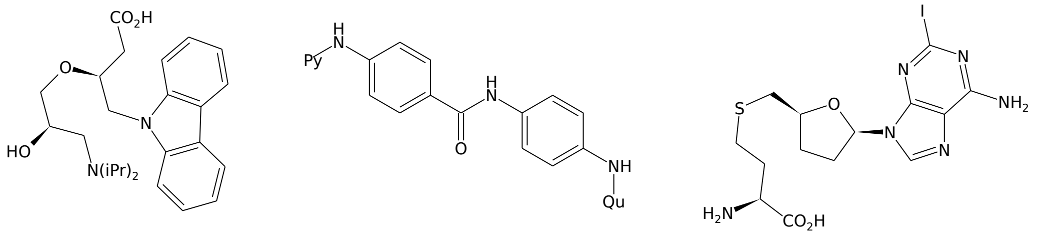
Para este análisis, el límite para determinar si los compuestos eran estructuralmente similares se escogió como el tercer cuartil de la distribución de similitud. El límite para determinar si los valores de actividad eran similares se escogió como 3 unidades de diferencia, con el fin de seleccionar los pares de compuestos que eran acantilados de actividad más prometedores. Los resultados acerca de estos límites se muestran en la tabla 4.4.

Tabla 4.4: Porcentaje de comparaciones entre pares para las tres DNMTs.

DNMTs	DNMT1			DNMT3A			DNMT3B		
Fingerprint	MACCS	PubChem	ECFP4	MACCS	PubChem	ECFP4	MACCS	PubChem	ECFP4
Acant. de act.	0.79 %	0.76 %	0.72 %	2.77 %	2.71 %	2.41 %	3.28 %	2.38 %	2.71 %
SAR continuo	24.33 %	24.24 %	24.17 %	22.30 %	22.27 %	22.42 %	22.46 %	22.68 %	22.35 %
Acant. de sim.	72.59 %	72.68 %	72.75 %	62.73 %	62.76 %	62.61 %	59.86 %	59.64 %	59.97 %
No descriptivos	2.29 %	2.32 %	2.36 %	12.20 %	12.26 %	12.56 %	14.39 %	15.29 %	14.97 %

Este análisis mostró que, independientemente del *fingerprint* empleado, los porcentajes de pares de compuestos (puntos en la gráfica) en las diferentes áreas tienden a ser relativamente similares para la misma proteína, en los tres casos. En particular, DNMT1 muestra un bajo porcentaje de pares de compuestos en la región de acantilados de actividad al compararse con DNMT3A y DNMT3B, lo que sugiere que el principio de similitud (compuestos con estructura similar presentarán actividad similar) se cumple en menor medida para los grupos de moléculas de DNMT3A y DNMT3B que para DNMT1. Se observa que para la zona perteneciente al SAR continuo, DNMT1 tiene una mayor cantidad de pares de compuestos que DNMT3A y DNMT3B, lo que indica que DNMT1 es más adecuada para estudios predictivos de estructura-activdad que los otros dos grupos moleculares. Para la zona de acantilados de similitud se observa un porcentaje de puntos mayores para DNMT1 que para DNMT3A y DNMT3B, lo que indica que existen más compuestos estructuralmente diversos y con la misma actividad para DNMT1.

El análisis de acantilados de actividad permitió la identificación de compuestos generadores de acantilados de actividad - compuestos que consistentemente aparecen en la región de acantilados de actividad[39]. Entre los compuestos encontrados (Figura 4.12), se tiene que la mayoría de generadores de acantilados de actividad para DNMT1 se relacionaban con compuestos altamente activos semejantes al cofactor o al mostrado en la figura. Con respecto a DNMT3A los compuestos generadores eran moléculas con actividad alta, estructuralmente similares a la molécula mostrada. Sin embargo, para DNMT3B, se observó que los compuestos generadores, a pesar de estar relacionados al cofactor, mostraban bajos niveles de actividad.



Índice de actividad = 4
DNMT1

Índice de actividad = 4
DNMT3A

Índice de actividad = 1
DNMT3B

Figura 4.12: Estructuras de compuestos generadores de acantilados de actividad (iPr = Isopropilo, Qu = Quinolina, Pyr = Pirimidina).

Espacio químico

El espacio químico de la base de datos fue estudiado por métodos de reducción de dimensionalidad tales como PCA y SOM, con el fin de generar una visualización de la distribución de los datos para los compuestos de las tres proteínas estudiadas.

Con respecto al análisis basado en PCA, el cual se muestra en la Figura 4.13, se realizaron diferentes visualizaciones basadas tanto en los descriptores del espacio químico oral, como en un subconjunto de descriptores obtenidos por el análisis de correlación. El primer componente principal muestra influencias de TPSA, HBA y HBD, mientras que el segundo muestra influencias de RB, MW y log P.

La visualización muestra que los compuestos presentes en DNMT1 cubren una mayor área que con respecto a DNMT3A y DNMT3B, lo que convierte al primero en el conjunto más diverso en términos de propiedades. Al observar sólo los compuestos más activos - compuestos con índice de actividad igual o mayor a 3 - se aprecia que aparecen agregados separados para DNMT3A y DNMT3B. Este resultado indica que es posible diferenciar entre compuestos activos para los tres grupos en el espacio químico oral, y que modificaciones en sus propiedades pueden favorecer la selectividad hacia las diferentes dianas.

Del análisis de correlación se seleccionaron dos pares de descriptores que permitieron la separación de compuestos activos de los diferentes grupos moleculares, siendo éstos HBD, $f$sp^3, RB y Dens. Con lo anterior se realizó un análisis de componentes principales, el cual se muestra en la Figura 4.13. La visualización mostró que los compuestos de la base de datos se encontraban menor dispersos, con DNMT1 siendo el grupo molecular que poseía la mayor diversidad en términos de propiedades. Al visualizar sólo los compuestos activos, se muestra que los aglomerados para DNMT3A y DNMT3B se vuelven mucho más definidos, lo que sugiere que este segundo grupo de descriptores es mejor para la clasificación de las zonas activas para las DNMT3A y DNMT3B. De esta forma, el uso de estos descriptores se recomienda para el análisis de nuevos compuestos con potencial uso para estas dianas biológicas.

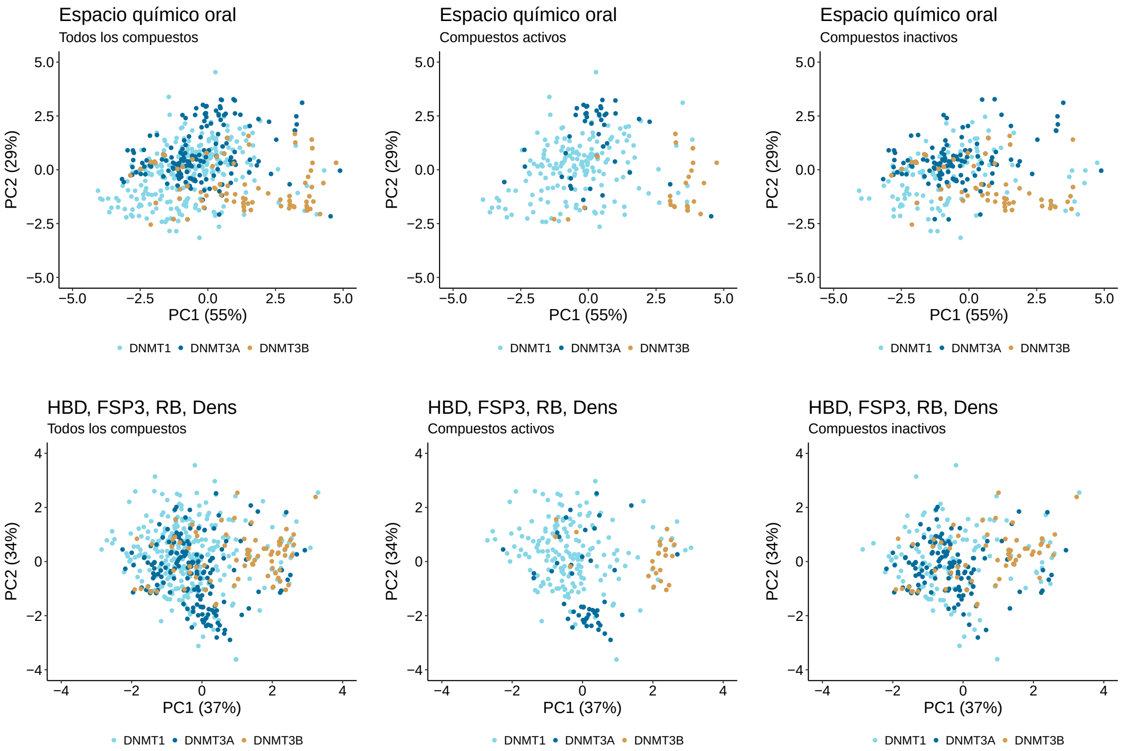


Figura 4.13: Visualización del espacio químico por componentes principales para todos los compuestos (izquierda), compuestos activos (centro) y compuestos inactivos (derecha). **Arriba:** Espacio basado en los descriptores de biodisponibilidad oral (MW, logP, RB, TPSA, HBD and HBA). **Abajo:** Espacio basado en los descriptores obtenidos en el análisis de correlación (TPSA, HBD, RB and Dens).

Con respecto al mapa de autoorganización (Figura 4.14), empleando los descriptores del espacio químico oral, se observa que los compuestos inactivos de DNMT1 y DNMT3A tienden a cubrir una amplia parte del espacio químico basado en celdas. DNMT3B, en cambio, tiende a estar en localizado en una sección de este espacio químico. En particular, al observar los compuestos activos, se puede observar que las moléculas pertenecientes a DNMT1 cubren igualmente una amplia región del espacio químico, mientras que DNMT3A y DNMT3B tienden a agruparse en zonas opuestas del espacio químico, en consenso con la representación anterior. Se observa igualmente que muchos de los compuestos activos de DNMT3A y DNMT3B suelen tener un perfil de propiedades altamente similar, dado que en el espacio de celdas tienden a estar muy cercanos.

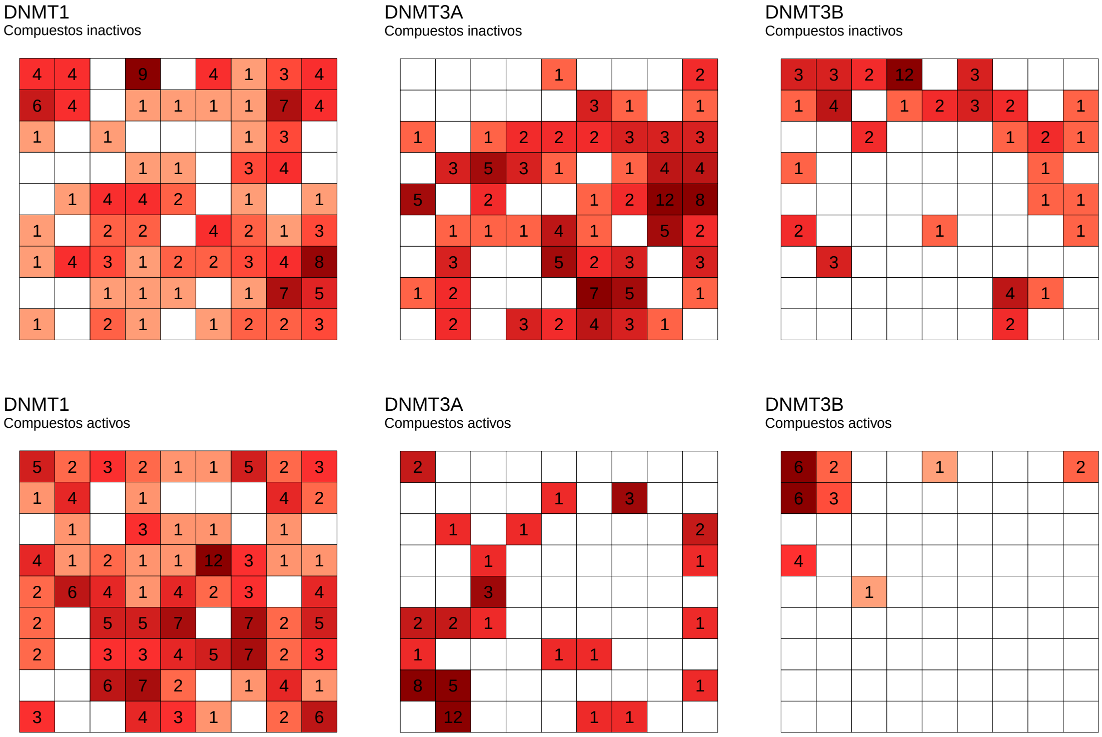
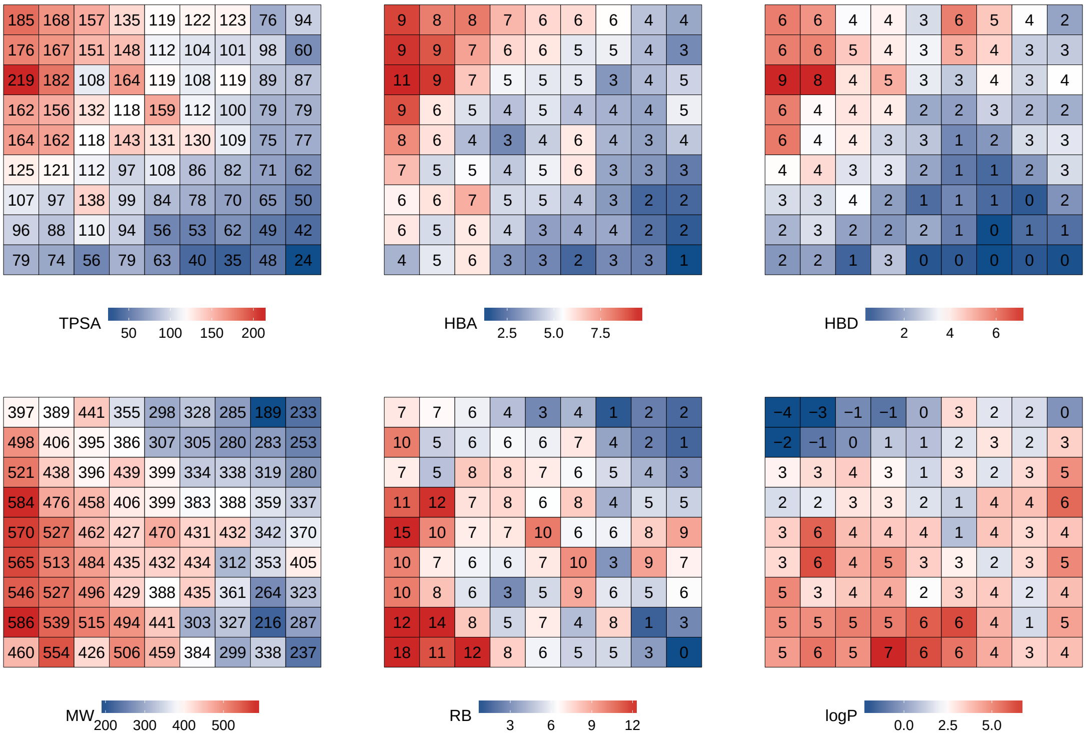


Figura 4.14: Visualización del espacio químico por mapas de autoorganización para las moléculas con actividad hacia DNMT1 (izquierda), DNMT3A (centro) y DNMT3B (derecha).**Arriba:** Compuestos inactivos. **Abajo:** Compuestos activos.

Para analizar con más profundidad los resultados del mapa de autoorganización, en la Figura 4.15 se ilustra la organización de la distribución de propiedades fisicoquímicas sobre la base de datos. En general, se observa que hay una correlación entre TPSA, HBA y HBD, ya que se observa que hay una distribución similar entre valores altos y bajos para cada uno de los descriptores. El mismo fenómeno se observa para MW y RB, sugiriendo una correlación entre estos descriptores.

Al realizar la proyección de los datos, no se aprecia una tendencia para los compuestos inactivos de DNMT1 y DNMT3A, mientras que se observa que la mayoría de los compuestos inactivos de DNMT3B tienden a tener valores bajos de MW y de RB, y valores intermedios de TPSA, HBA y HBD. Con respecto a los compuestos activos de DNMT1, no se aprecia igualmente una tendencia. Los compuestos activos hacia DNMT3A tienden a tener altos valores de MW, RB y logP, mientras que para DNMT3B los compuestos activos suelen tener valores altos de TPSA, HBA y HBD.

De esta manera, el análisis del espacio químico nos permitió detectar patrones y puntos clave para el desarrollo de inhibidores activos y selectivos hacia las diferentes DNMT.



TPSA

185	168	157	135	119	122	123	76	94
176	167	151	148	112	104	101	98	60
219	182	108	164	119	108	119	89	87
162	156	132	118	159	112	100	79	79
164	162	118	143	131	130	109	75	77
125	121	112	97	108	86	82	71	62
107	97	138	99	84	78	70	65	50
96	88	110	94	56	53	62	49	42
79	74	56	79	63	40	35	48	24

HBA

9	8	8	7	6	6	6	4	4
9	9	7	6	6	5	5	4	3
11	9	7	5	5	5	3	4	5
9	6	5	4	5	4	4	4	5
8	6	4	3	4	6	4	3	4
7	5	5	4	5	6	3	3	3
6	6	7	5	5	4	3	2	2
6	5	6	4	3	4	4	2	2
4	5	6	3	3	2	3	3	1

HBD

6	6	4	4	3	6	5	4	2
6	6	5	4	3	5	4	3	3
9	8	4	5	3	3	4	3	4
6	4	4	4	2	2	3	2	2
6	4	4	3	3	1	2	3	3
4	4	3	3	2	1	1	2	3
3	3	4	2	1	1	1	0	2
2	3	2	2	2	1	0	1	1
2	2	1	3	0	0	0	0	0

MW

397	389	441	355	298	328	285	189	233
498	406	395	386	307	305	280	283	253
521	438	396	439	399	334	338	319	280
584	476	458	406	399	383	388	359	337
570	527	462	427	470	431	432	342	370
565	513	484	435	432	434	312	353	405
546	527	496	429	388	435	361	264	323
586	539	515	494	441	303	327	216	287
460	554	426	506	459	384	299	338	237

RB

7	7	6	4	3	4	1	2	2
10	5	6	6	6	7	4	2	1
7	5	8	8	7	6	5	4	3
11	12	7	8	6	8	4	5	5
15	10	7	7	10	6	6	8	9
10	7	6	6	7	10	3	9	7
10	8	6	3	5	9	6	5	6
12	14	8	5	7	4	8	1	3
18	11	12	8	6	5	5	3	0

logP

−4	−3	−1	−1	0	3	2	2	0
−2	−1	0	1	1	2	3	2	3
3	3	4	3	1	3	2	3	5
2	2	3	3	2	1	4	4	6
3	6	4	4	4	1	4	3	4
3	6	4	5	3	3	2	3	5
5	3	4	4	2	3	4	2	4
5	5	5	5	6	6	4	1	5
5	6	5	7	6	6	4	3	4

Figura 4.15: Visualización de la distribución de las propiedades fisicoquímicas en el espacio químico basado en celdas del mapa de autoorganización.

Capítulo 5

Conclusiones

Se realizó una comparación quimioinformática global de tres grupos de inhibidores reportados en bases de datos públicas de DNMT1, DNMT3A y DNMT3B. El objetivo fue encontrar compuestos ya sean de índole fisicoquímico o estructurales para un diseño racional de moléculas con una mejor selectividad.

Se concluyó que los descriptores de propiedades fisicoquímicas permitieron una diferenciación entre moléculas activas e inactivas para DNMT3A y DNMT3B. El análisis de correlación permitió encontrar otros pares de descriptores empleables en la separación entre moléculas activas de DNMT3A y DNMT3B.

Con respecto al análisis de panoramas de actividad, se concluyó que los inhibidores de DNMT1 permiten estudios de índole predictivo con mayor certeza que los conjuntos de moléculas para DNMT3A y DNMT4B.

Con respecto a la estructura, el análisis de diversidad mostró que DNMT1 es el conjunto más diverso estructuralmente, mientras DNMT3B es el menos diverso estructuralmente. El análisis de quimiotipos permitió el análisis de motivos estructurales que permitieran la separación entre compuestos activos de DNMT1 y DNMT3B.

Los resultados de este proyecto de investigación respaldan la posibilidad de obtener moduladores selectivos para las diferentes proteínas de esta familia, y proveen varios puntos clave con el fin de generar un mejor diseño racional de nuevas entidades químicas capaces de inhibir a las diferentes ADN metiltransferasas.

Capítulo 6

Perspectivas

Los resultados obtenidos en este proyecto permiten generar propuestas a largo plazo, las cuales incluyen, pero no se limitan a las siguientes:

- Generar filtros con las diferentes propiedades para realizar cribado virtual. Esto con el objetivo de encontrar nuevas entidades químicas con actividad hacia las diferentes DNMT.
- Analizar las posibilidades de reemplazos isostéricos sobre las diferentes moléculas con actividad, con el fin de generar nuevas estructuras con posible actividad moduladora.
- Realizar modelos de farmacóforo para las diferentes ADN metiltransferasa. El objetivo será entender puntos clave en el reconocimiento molecular, y poder realizar un diseño selectivo hacia cada diana molecular.

Bibliografía

[1] Christoph Plass, Stefan M. Pfister, Anders M. Lindroth, Olga Bogatyrova, Rainer Claus, and Peter Lichter. **Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer**. *Nature Reviews Genetics*, **14**:765–780, 2013. 2

[2] L. Morera, M. Lübbert, and M. Jung. **Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy**. *Clinical epigenetics*, **8**:57, 2016. 2

[3] Wu Zhang and Jie Xu. **DNA methyltransferases and their roles in tumorigenesis**. *Biomarker Research*, **5**(1):1, Jan 2017. 2

[4] Ke Liu, Yanli Liu, Johnathan L. Lau, and Jinrong Min. **Epigenetic targets and drug discovery Part 2: Histone demethylation and DNA methylation**. *Pharmacology & Therapeutics*, **151**(Supplement C):121–140, 2015. 2

[5] Timothy H. Bestor. **The DNA methyltransferases of mammals**. *Human Molecular Genetics*, **9**(16):2395–2402, 2000. 2

[6] Mary Grace Goll and Timothy H. Bestor. **EUKARYOTIC CYTOSINE METHYLTRANSFERASES**. *Annual Review of Biochemistry*, **74**(1):481–514, 2005. 2

[7] Sergio Valente, Yiwei Liu, Michael Schnekenburger, Clemens Zwergel, Sandro Cosconati, Christina Gros, Maria Tardugno, Donatella Labella, Cristina Florean, Steven Minden, Hideharu Hashimoto, Yanqi Chang, Xing Zhang, Gilbert Kirsch, Ettore Novellino, Paola B. Arimondo, Evelina Miele, Elisabetta Ferretti, Alberto Gulino, Marc Diederich, Xiaodong Cheng, and Antonello Mai. **Selective Non-nucleoside Inhibitors of Human DNA Methyltransferases Active in Cancer Including in Cancer Stem Cells**. *Journal of Medicinal Chemistry*, **57**(3):701–713, 2014. PMID: 24387159. 3

[8] Peter A. Jones and Gangning Liang. **Rethinking how DNA methylation patterns are maintained**. *Nature Reviews Genetics*, **10**:805–811, 2009. 3

[9] Lisa D. Moore, Thuc Le, and Guoping Fan. **DNA Methylation and Its Basic Function**. *Neuropsychopharmacology*, **38**:23–38, 2012. 3

[10] Renata Zofia Jurkowska, Tomasz Piotr Jurkowski, and Albert Jeltsch. **Structure and Function of Mammalian DNA Methyltransferases**. *ChemBioChem*, **12**(2):206–222, 2011. 3

[11] Pan Xu, Guang Hu, Cheng Luo, and Zhongjie Liang. **DNA methyltransferase inhibitors: an updated patent review (2012-2015)**. *Expert Opinion on Therapeutic Patents*, **26**(9):1017–1030, 2016. PMID: 27376512. 3

[12] Paul W. Hollenbach, Aaron N. Nguyen, Helen Brady, Michelle Williams, Yuhong Ning, Normand Richard, Leslie Krushel, Sharon L. Aukerman, Carla Heise, and Kyle J. MacBeth. **A Comparison of Azacitidine and Decitabine Activities in Acute Myeloid Leukemia Cell Lines**. *PLOS ONE*, **5**(2):1–10, 02 2010. 3

[13] Ehab Atallah, Hagop Kantarjian, and Guillermo Garcia-Manero. **The role of decitabine in the treatment of myelodysplastic syndromes**. *Expert Opinion on Pharmacotherapy*, **8**(1):65–73, 2007. 3

[14] Omar Castillo-Aguilera, Patrick Depreux, Ludovic Halby, Paola B. Arimondo, and Laurence Goossens. **DNA Methylation Targeting: The DNMT/HMT Crosstalk Challenge**. *Biomolecules*, **7**(1):1–21, 2017. 3

[15] Oscar Palomino-Hernandez, A. Christiaan Jardinez-Vera, and Jose L. Medina-Franco. **Progress on the Computational Development of Epigenetic Modulators of DNA Methyltransferases 3A and 3B**. *Journal of the Mexican Chemical Society*, **61**(3):266–272, 2017. 3

[16] Eli Fernandez-de Gortari and Jose L. Medina-Franco. **Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases**. *RSC Adv.*, **5**:87465–87476, 2015. 4

[17] Fernando D. Prieto-Martinez, Eli Fernandez-de Gortari, Oscar Mendez-Lucio, and Jose L. Medina-Franco. **A chemical space odyssey of inhibitors of histone deacetylases and bromodomains**. *RSC Adv.*, **6**:56225–56239, 2016. 4

[18] A. Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, Felix A. Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, Michal Nowotka, George Papadatos, Rita Santos, and John P. Overington. **The ChEMBL bioactivity database: an update**. *Nucleic Acids Research*, **42**(D1):D1083, 2014. 5

[19] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. **BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities**. *Nucleic Acids Research*, **35**(suppl_1):D198, 2007. 5

[20] Zhimin Huang, Haiming Jiang, Xinyi Liu, Yingyi Chen, Jiemin Wong, Qi Wang, Wenkang Huang, Ting Shi, and Jian Zhang. **HEMD: An Integrated Tool of Human Epigenetic Enzymes and Chemical Modulators for Therapeutics**. *PLOS ONE*, **7**(6):1–6, 06 2012. 5

[21] HTTP://WWW.WEBOFKNOWLEDGE.COM/. *Web of Science*. 5

[22] HTTP://SCIFINDER.CAS.ORG/. *Scifinder*. 5

[23] Denis Fourches, Eugene Muratov, and Alexander Tropsha. **Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research**. *Journal of Chemical Information and Modeling*, **50**(7):1189–1204, 2010. PMID: 20572635. 5

[24] Chemical Computing Group. *Molecular Operating Environment (MOE) 2014.09*, 2014. [link]. 5

[25] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. ISBN 3-900051-07-0. 6

[26] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016. 6

[27] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1**. *Advanced Drug Delivery Reviews*, **46**(1):3 – 26, 2001. Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998. 6

[28] Daniel F. Veber, Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple. **Molecular Properties That Influence the Oral Bioavailability of Drug Candidates**. *Journal of Medicinal Chemistry*, **45**(12):2615–2623, 2002. PMID: 12036371. 6

[29] Nicholas C. Firth, Nathan Brown, and Julian Blagg. **Plane of Best Fit: A Novel Method to Characterize the Three-Dimensionality of Molecules**. *Journal of Chemical Information and Modeling*, **52**(10):2516–2525, 2012. PMID: 23009689. 6

[30] José L. Medina-Franco and Gerald M. Maggiora. *MOLECULAR SIMILARITY ANALYSIS*, pages 343–399. John Wiley & Sons, Inc, 2013. 7

[31] José L. Medina-Franco, Gerald M. Maggiora, Marc A. Giulianotti, Clemencia Pinilla, and Richard A. Houghten. **A Similarity-based Data-fusion Approach to the Visual Characterization and Comparison of Compound Databases**. *Chemical Biology & Drug Design*, **70**(5):393–412, 2007. 7

[32] Guy W. Bemis and Mark A. Murcko. **The Properties of Known Drugs. 1. Molecular Frameworks**. *Journal of Medicinal Chemistry*, **39**(15):2887–2893, 1996. PMID: 8709122. 7

[33] José L. Medina-Franco. **Scanning Structure–Activity Relationships with Structure–Activity Similarity and Related Maps: From Consensus Activity Cliffs to Selectivity Switches**. *Journal of Chemical Information and Modeling*, **52**(10):2485–2493, 2012. 8

[34] Fernanda I. Saldivar-Gonzalez, J. Jesus Naveja, Oscar Palomino-Hernandez, and Jose L. Medina-Franco. **Getting SMARt in drug discovery: chemoinformatics approaches for mining structure-multiple activity relationships**. *RSC Adv.*, **7**:632–641, 2017. 8

[35] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2017. R package version 6.0-76. 8

[36] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. 8

[37] Jose L. Medina-Franco, Karina Martinez-Mayorga, Andreas Bender, and Thomas Scior. **Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure**. *QSAR & Combinatorial Science*, **28**(11-12):1551–1560, 2009. 15

[38] Jose Luis Medina-Franco, Joachim Petit, and Gerald M. Maggiora. **Hierarchical Strategy for Identifying Active Chemotype Classes in Compound Databases**. *Chemical Biology & Drug Design*, **67**(6):395–408, 2006. 17

[39] Oscar Mendez-Lucio, Jaime Perez-Villanueva, Rafael Castillo, and Jose L. Medina-Franco. **Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps**. *Molecular Informatics*, **31**(11-12):837–846, 2012. 20

RSC Advances

REVIEW

Cite this: *RSC Adv.*, 2017, 7, 632

Getting SMARt in drug discovery: chemoinformatics approaches for mining structure–multiple activity relationships†

Fernanda I. Saldívar-González,[a] J. Jesús Naveja,[abc] Oscar Palomino-Hernández[a] and José L. Medina-Franco*[a]

Received 3rd November 2016
Accepted 6th December 2016

DOI: 10.1039/c6ra26230a

www.rsc.org/advances

In light of the high relevance of polypharmacology, multi-target screening is a major trend in drug discovery. As such, the increasing amount of available structure–activity data requires the application of chemoinformatic approaches to mine structure–multiple activity relationships. To this end, activity landscape methods, initially developed to explore the structure–activity relationships for compounds screened against one target, have been adapted to mine Structure–Multiple Activity Relationships (SMARt). Herein, we survey advances in the chemoinformatic approaches to retrieve SMARt from screening data sets. Case studies relevant to modern drug discovery are discussed. The methods covered in this survey are general and can be implemented to explore the SMARt of other data sets screened across multiple biologically endpoints.

Introduction

Analysis of structure–activity relationships (SAR) is a common practice in many areas of chemistry. Most medicinal and computational chemists working on drug discovery obtain SAR of compound data sets on a routine basis. This is true not only in academic settings but also in the pharmaceutical industry and research institutes. In several current drug discovery projects, compound data sets are screened across more than one biological endpoint. Depending on the project, it is desirable to identify selective compounds or identify molecules with activity across multiple endpoints. Moreover, in light of the increasing awareness of polypharmacology[1] and multi-target drug discovery,[2] screening small compound data sets or large chemical libraries across more than one biological endpoint is a fundamental task. Therefore, getting Structure–Multiple Activity Relationships (SMARt) is a common need in drug discovery.

Methods to get SMARt can be broadly classified into qualitative and quantitative. Qualitative approaches can be applied without the need of computational tools and depend on the experience of the chemist analyzing the data. Thus, qualitative methods are suitable to handle small-to-medium size data sets. In contrast, large screening data sets, in particular those tested across several endpoints, usually require the application of computational procedures in addition to the experience of the chemist.[3] In these cases, *in silico* methods can be performed for either predictive or descriptive purposes. As discussed previously, understanding the SAR of compound data sets should be performed before developing predictive models[4] such as QSAR and QSPR in order to predict novel, potent, and selective compounds.[5,6] In this regard, new computational models that combine multi-target QSAR with machine learning such as artificial neural network algorithms have been developed with the aim of predict the interactions of multiple molecules to targets involved in many diseases and processes of neuroprotection.[5,7]

Activity landscape modeling (ALM) is a chemoinformatic strategy to mine the SAR of compound data sets and it is actively used in academia, industry and other research settings. ALM can be regarded as part of computer-aided drug design and it is an important component in medicinal chemistry.[8] For more than ten years several groups have worked on the development of ALM. These approaches relay on the quantitative comparison of structure similarity with activity similarity (or potency difference) for all pairs of compounds in a screening data set. Over the years a large number of quantitative and visual methods have been developed. Most of these methods started with the main goal of identifying 'activity cliffs': pairs of compounds with very similar structure but unexpected high activity difference.[9] Activity cliffs have a 'dual face' with a large impact in medicinal and computational chemistry.[10] It has been

[a] *Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico. E-mail: medinajl@unam.mx; jose.medina.franco@gmail.com; Tel: +52-55-5622-3899 ext. 44458*

[b] *Facultad de Medicina, PECEM, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City, 04510, Mexico*

[c] *Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, 85764, Neuherberg, Germany*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6ra26230a

largely advised that after identifying activity cliffs, molecular modeling studies should be conducted that help to explain, at the molecular level, the reason associated with the large change in activity due to a small modification the chemical structure. Such studies are highly valuable because add three-dimensional information to the system. To this end, mechanistic studies towards the structural interpretation of activity cliffs in three dimensions have been published.[11,12] Overall, the specific reasons that are associated with the formation of the activity cliffs depend on the system. An alternative approach to add three-dimensional information to the system and consider additional effects of functional groups, conformations and configurations, molecular descriptors that take into account the coordinates space of the compounds or even using several different conformations of the molecules in the data set have been reported.[13,14]

ALM seeks not only to identify activity cliffs but other significant areas of the activity landscape such as 'similarity cliffs' (which are related to scaffold hops)[15] and other continuous regions of the activity landscape. The broad applicability of ALM in medicinal chemistry has been reviewed.[16] Initially developed to describe SARs, ALM has been tested for predictive purposes.[17,18] Similarly, ALM was originally applied to describe the SAR of compound data sets screened for one biological endpoint, for instance, for a single target. However, several methods used in ALM have been adapted to mine SMARts.

The goal of this work is to survey the progress of ALM to get SMARt in drug discovery. We put special emphasis on the development and application of Structure–Activity Similarity (SAS) maps which were one of the first approaches used in ALM.[19] Four years ago the authors reviewed the development of SAS maps to explore SARs.[20] In contrast, this review covers the most recent developments and applications aimed to explore SMARts. As part of the recent developments the concept of 'pro-activity cliffs' is introduced. The manuscript is organized in five main sections: after this introduction a brief overview of the SAS maps is presented with special emphasis on the development of density SAS maps and activity landscape sweeping strategies. The section after that describes the adaptation of ALM from single to multi-target activity analysis. This section is followed by a discussion of future trends in SAR and SMARt analysis using ALM. Concluding remarks are presented at the end.

Structure–activity similarity (SAS) maps

SAS maps were proposed in 2001.[19] The basic idea of a SAS map is to plot in two-dimensions (2D) the pairwise structure similarity (usually plotted on the *X*-axis) and activity difference (plotted on the *Y*-axis) for all pairs of compounds in a data set. A general form of a SAS map is shown in Fig. 1A. To aid in the interpretation, a SAS map can be roughly divided in four major quadrants each one distinguishing pairs of compounds with high/low activity difference and high/low structure similarity. Activity cliffs are located in the quadrant that identifies pairs of molecules with high structure similarity and high activity difference (region IV). Compound pairs with a smooth SAR have high structure similarity and low activity difference (region II). Scaffold hops (or similarity cliffs) are located in the opposite quadrant of the activity cliffs (region I). Noteworthy, even in the absence of the thresholds with formally defined quadrants, SAS maps are helpful to differentiate major regions in the landscape.

One of the known limitations of the SAS maps is the quantitative criteria to define the thresholds along the *X*- and *Y*-axis. A number of approaches to address this issue are discussed elsewhere.[20] Briefly, the thresholds that define high/low activity difference depend on the goal of the project. Usual cutoffs are one, two or more potency units. The thresholds to define high/low structure similarity can be set up based on the distribution of the similarity values of the data set. In some instances, heuristic values of similarity are considered based on author's experience.

Another limitation of the SAS maps is the large amount of data points that could be generated. Therefore, for large data sets it is challenging the visual interpretation of the SAS maps. To address this issue several strategies have been proposed which are discussed below.

Density SAS maps

To aid in the visualization of the SAS and related maps three major strategies have been developed: (1) categorical SAS maps;[13] (2) filtered SAS maps showing only the most relevant data points (for instance, the 'active pairs' of compounds defined as pair of molecules containing at least one active compound in the pair) and, more recently (3) density SAS maps that display the amount of data points using a continuous color scale.[21] Fig. 1 shows examples of 'simplified' SAS maps: categorical, filtered and density SAS map for a data set of 140 pyrimidine hydroxyl amide compounds tested with histone deacetylase 1 (HDAC1). These compounds were synthesized and tested as part of a program of optimization to find potent and selective inhibitors of HDAC6, enzyme required for the formation of the aggresome and survival of cancer cells.[22] HDAC is a major epigenetic target and the computational analysis of the SAR can be regarded as part of the emerging research field of Epi-Informatics.[23] SAR analysis of HDAC inhibitors is particularly useful for the treatment of proliferative diseases and disorders by protein deposition, likewise, it is useful for probing biological pathways. A full discussion of the SAR of HDAC inhibitors is out of the scope of this Short Review that is focused on ALM. Fig. S1 in the ESI† shows additional examples of simplified SAS maps for a data set of 91 compounds tested against the parasite *Giardia intestinalis*. Note that density SAS maps provide better information regarding the general distribution of the data points, though sacrificing the chance of including information regarding the individual activity of any of the compounds in the pair.

Several analyses have shown that the similarity cliff region is one of the most populated for several data sets.[13] Results of Maggiora *et al.* further confirmed these observations analyzing many data sets.[15] This is also the case in the activity landscape depicted in Fig. 1 and S1.† Density SAS maps have been employed to analyze the ALM of 5α-alpha reductase inhibitors[24]

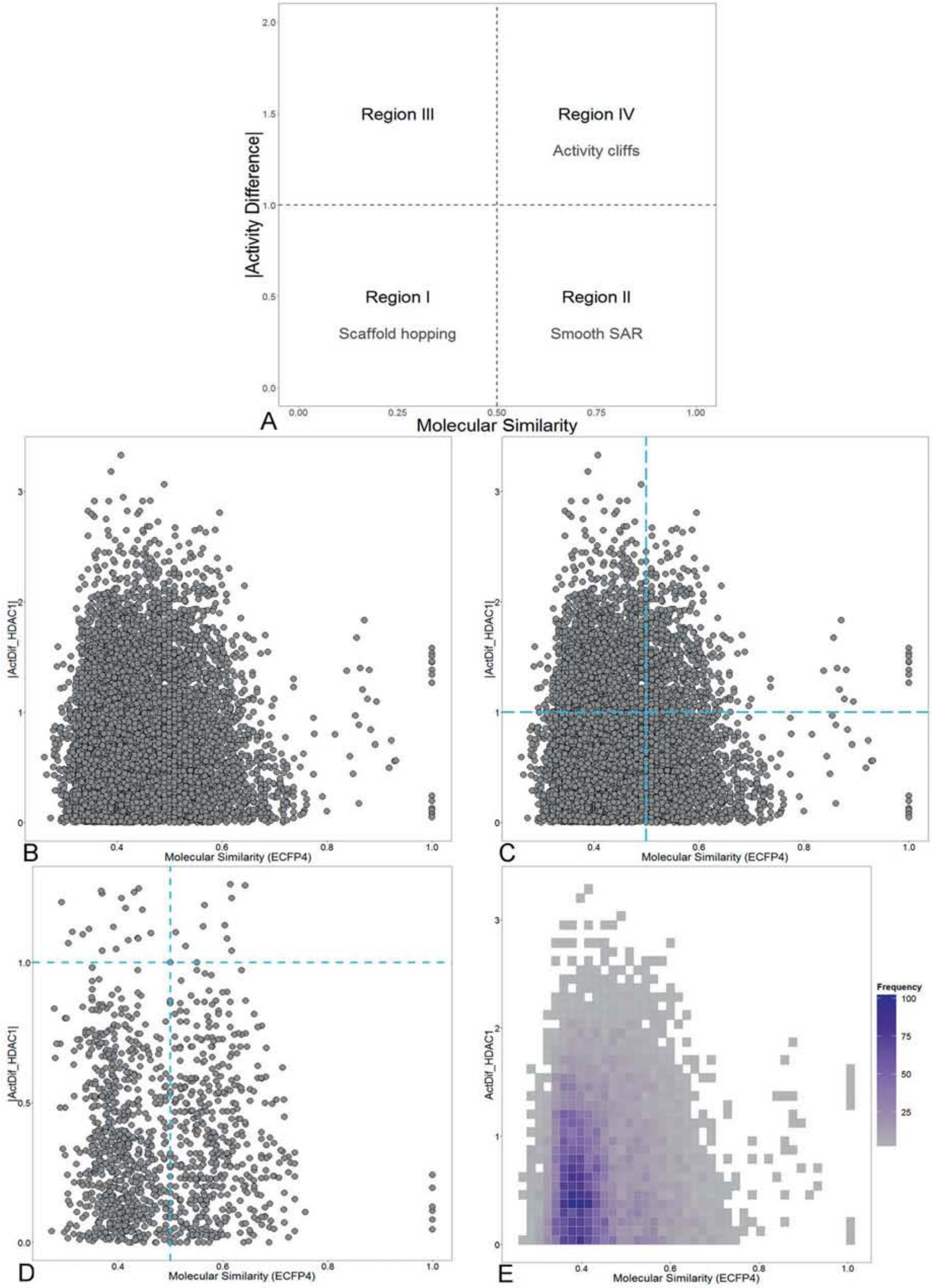


Fig. 1 (A) General form of the structure–activity similarity (SAS) maps showing four major regions. Regions I and II are associated with scaffold hopping and smooth SAR, respectively. Region III does not provide relevant information and region IV indicates discontinuous SAR and activity cliffs. Actual (B) and simplified SAS maps for a data set of 140 compounds tested with HDAC1. (C) Categorical map showing the distribution of the data point in each of the four quadrants of the SAS map; (D) filtered map displaying the 'active regions' of the landscape *i.e.*, pairs of compounds that contain at least one active molecular in the pair; and (E) density map that shows the amount of data points in each region using a continuous color scale from purple color (more data points) to grey color (less data points). The simplified SAS maps are designed to aid in the visual representation and interpretation of the SAS maps.

and inhibitors of DNA metiltransferases (DNMTs), other major epigenetic target.[21]

Activity landscape sweeping

Activity landscape sweeping is a strategy recently developed to 'clean' the SAR/SMARt of a data set by filtering first the compounds that are considered to analyze the landscape. An approach is to classify the compounds by the types of molecular scaffold[25] or the relative position in chemical space, to name two criteria. Then, the ALM would be centered on the local SAR of the filtered molecules. In a broad sense, activity landscape sweeping is an approach to analyze local models of SAR/SMARts. Despite the fact that such models are not general, activity landscape sweeping gives rise to focused analysis of the most interpretable areas of the activity landscape.

In order to illustrate the filtering of compounds before ALM, *i.e.*, activity landscape sweeping, Fig. 2 shows a visual representation of the chemical space of a series of 140 pyrimidine hydroxyl amide compounds synthesized and evaluated as HDAC6 inhibitors. Two main clusters (A: circles, B: triangles) are readily distinguished: compounds of cluster A correspond to formulas IV–VIII described by Van Duzer *et al.* while compounds of cluster B correspond to formulas I–III described in the same work.[22] The main difference between these two groups is the carbon attached to the nitrogen of 2-amino-*N*-hydroxypyrimidine-5-carboxamide. In group A, this carbon is tertiary, while in group B is primary or secondary. Representative chemical structures are shown in Fig. S2 of the ESI.†

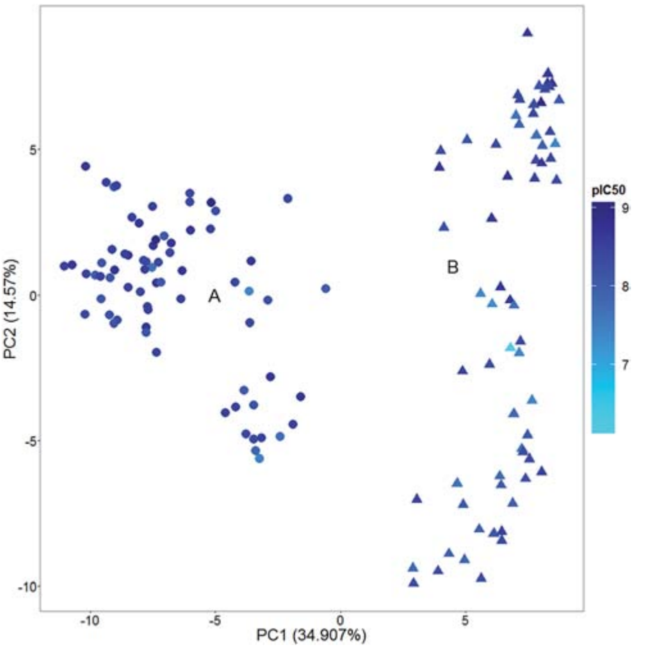


Fig. 2 Example of an activity landscape sweeping. Visual representation of the chemical space of the 140 inhibitors of histone deacetylase 6 (HDAC6). The visualization was obtained by principal component (PC) analysis of the similarity matrix computed with extended connectivity fingerprint 4 (ECFP4). The percentage of variance explained by each PC is indicated in the corresponding axis. Data points are colored by the pIC_{50} values of HDAC6 in a continuous scale.

Activity landscape sweeping has been recently applied to DNMT inhibitors[21] and 5α-reductase inhibitors.[24] In both instances activity landscape sweeping was used in conjunction with SAS maps. This approach helped to 'clean' the landscape and facilitated the visual analysis of the SAS maps. Activity landscape sweeping has been used in conjunction with SAS maps but could be implemented in combination with any other ALM strategy such as Structure–Activity Landscape Index (SALI)[26] or other methods.

SAS maps and PLIFS

Protein–ligand interaction fingerprints (PLIFS) are convenient representations to capture protein–ligands contacts in a systematic manner. PLIFS are at the interface of chemoinformatics and molecular modeling[27] and have been designed to 'capture a 1D representation of the interactions between ligand and protein either in complexes of known structure or in docked poses'.[28] Recently SAS maps have been adapted to analyze structure–protein ligand interactions giving rise to the protein–ligand interaction cliffs.[27] These are defined as pairs of compounds with high structure similarity, high protein–ligand contact similarity but very different activity profile. That study was conducted for a series of kinase inhibitors. In that work, Méndez-Lucio *et al.* integrated PLIFS to a multi-target kinase activity landscape analysis. Three data sets, containing the crystallographic structure of the ligand bound to a kinase were used. The authors employed three data sets, containing the crystallographic structure of the ligand bound to a kinase. Pairwise interaction similarity was assessed using PLIFs and the Tanimoto coefficient, whereas twelve 2D and 3D molecular descriptors were used to compute pairwise molecular similarity. Pairwise structure-similarity analysis revealed no correlation with interaction similarity in none of the data sets despite the fact that the kinase ATP binding site is highly conserved. On average, only 33% of the molecular pairs categorized as highly similar showed similar interactions. This approach not only provided structural information of activity cliffs but it also was useful to identify hot spots in the target protein associated with selectivity.[27,29]

Tuning ALM to get SMARt

In addition to SAS maps several other methods have been developed for ALM analysis.[16,20,26,30,31] For instance SALI, the first index developed to rapidly identify activity cliffs, is calculated with the expression:[26]

$$\mathrm{SALI}_{i,j} = \frac{|A_i - A_j|}{1 - \mathrm{sim}(i,j)}$$

where A_i and A_j are the activities of the ith and jth molecules, and sim(i,j) is the similarity coefficient between the two molecules. Also, the research group of Bajorath has developed a large number of approaches for ALM.[16]

Several of ALM methods have been adapted to handle SMARt. For instance, a straightforward extension of SALI to measure SMARt is replacing the numerator of the SALI with the

Table 1 Examples of case studies of SMARt studies conducted with SAS-like maps

Study	Major outcome (method)	Major outcome (interpretation)	Ref.
SMART of >50 benzimidazoles tested with *T. vaginalis* and *G. intestinalis*	Dual activity difference maps with fingerprint and sub-structure representation	'Activity switches' are introduced: pairs of compounds where one small change in the structure is associated with a different and opposite change in the activity of two biological endpoints	34 and 35
SMART of a series of purine analogs screened against the cysteine protease cathepsins	Triple activity difference maps	The concept of structure–property–activity (SPA) similarity in SAR studies are introduced. SPA maps are analyzed to determine the extent to which property similarities could be applied to characterize SARs	14
SMART of compounds in PubChem	Structure multiple Activity Similarity (SmAS) maps	Bioassay activity landscape is introduced to study the relationship between the structure and bioactivity profiles	37
ADMET analysis of 166 compounds screened for kappa-opioid receptor activity	ADMET property–activity pairwise similarity maps with ADMET descriptors and dimensional 'violation bit vector' representing	Study of the range of ADMET property violations that arise from structural changes, subtle and significant	41
SMART of 15 252 compounds screened across 100 diverse proteins reported by Clemons *et al.*[38]	SPID measure (Structure–Promiscuity Index Difference)	Structure promiscuity index is introduced to identify the pairs of compounds with high structure–similarity but large activity difference	39

biological profile similarity of the compound pair computed with the Tanimoto coefficient (giving rise to a Structure–Multiple Activity Landscape Index).[32] Representative case studies of the adaptation of ALM to get SMARt are summarized in Table 1 and discussed in the next sections.

SMARt with few biological endpoints

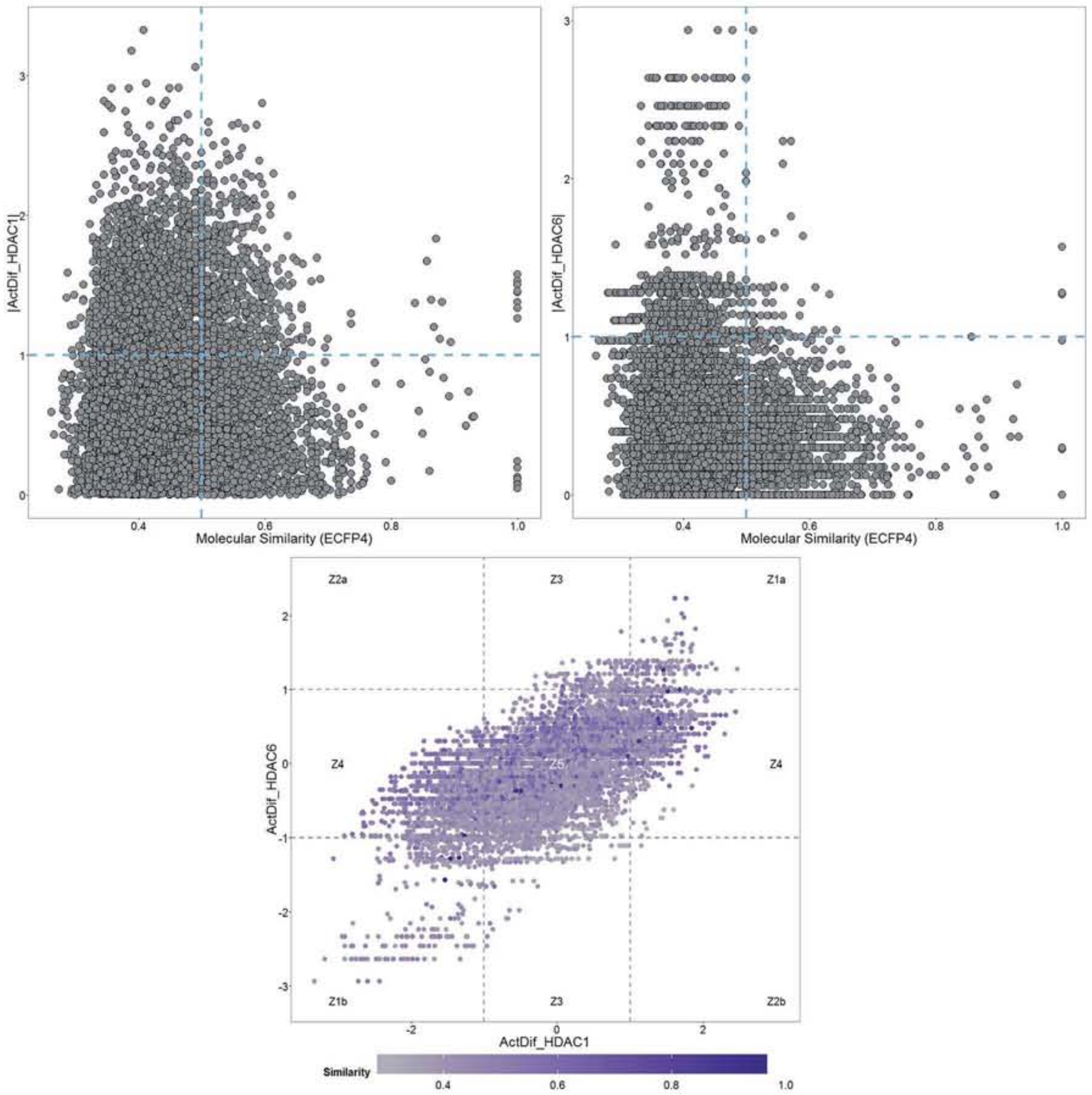
One of the first applications of SAS maps applied to analyze data sets across more than one biological endpoint was the SMARt exploration of more than 50 benzimidazole analogues tested for their ability to inhibit the growth of the protozoa *Trichomonas vaginalis* and *Giardia intestinalis.*[33] A tool to analyze simultaneously the difference in activity data for both parasites was the Dual Activity Difference (DAD) maps. DAD maps represent in 2D changes in potency difference for two targets.[34] One of the major outcomes of the DAD maps are 'activity switches' defined as pairs of compounds where one small change in the structure is associated with a very different but opposite change in the activity for both biological endpoints. Activity switches have been reviewed in detail.[20] Triple-Activity Difference (TAD) maps where developed later as a natural extension of the DAD maps to analyze SMARts.[14]

More recently, DAD maps were used to analyze systematically the activity landscape of a series of 91 benzimidazoles tested with the parasites *T. vaginalis* and *G. intestinalis.*[35] In that work the chemical structure of the 91 benzimidazoles was encoded using a fragment-based approach that indicated the presence or absence of six substituents around a common benzimidazole nucleus. Using DAD maps, single and dual substitutions around the benzimidazole scaffold were identified that were associated with large changes in potency for each of the two parasites. Furthermore, single and dual substitutions associated with large and opposite changes in activity for the two parasites were found.[35]

To illustrate a DAD map, Fig. 3 shows a plot of a data set of 140 molecules tested as HDAC1 and HDAC6 inhibitors.[22] As reference, Fig. 3 also shows the corresponding SAS maps for HDAC1 and HDAC6. In general, the DAD map in Fig. 3 shows that the larger amount of pairs of compounds are located in the region Z5 of the plot (close to 68%), indicating that most of the compounds show activity values very similar for both enzymes. The pairs identified in the Z3 and Z4 regions (simple activity cliffs) suggest that changes in the scaffold are more susceptible to present changes in activity against HDAC1 compared with HDAC6. The increased presence of pairs of compounds in the Z1 region compared to Z2 region indicates that there is a greater likelihood that the modifications affect the activity of both enzymes in the same magnitude and direction.

SMARt with many biological endpoints

ALM have also been applied to analyze the SMARt of screening collections tested across a large number of biological endpoints. Different ALM methods have been used including SAS maps. For instance, SAS maps were employed to analyze the SMARt obtained from Pubchem.[32] In a proof-of-concept study, Medina-Franco and Wadell analyzed the bioassay activity landscape of 618 molecules tested across 244 confirmatory bioassays. One of the particular challenges in that work was that each bioassay in PubChem has its own specific definition of active, inactive, or inconclusive. A second major challenge was



Region	Interpretation	9730 pairs total
Z1	Substitution (s) result in a significant decrease or increase of activity in both targets	474 (4.87%)
Z2	Substitution (s) increase activity for one target, while decreasing activity for the other target significantly	1 (0.01%)
Z3	Substitution (s) result in significant changes in activity on HDAC6 but not an appreciable change HDAC1	379 (3.89%)
Z4	Substitution (s) result in significant changes in activity on HDAC6 but not an appreciable change HDAC1	2262 (23.25%)
Z5	Substitution does not change significantly the activity for HDAC1 and HDAC6	6614 (67.98%)

Fig. 3 Example of SAS and DAD maps of a data set of 140 compounds tested across two biological endpoints (HDAC1 and HDAC6). Each data point represents a pairwise comparison. The table shows the interpretation and number and percentage of data points in each region of the map for compound pairs.

that not all 618 compounds were tested in all 244 bioassays. A distinctive feature of the SAS-like maps proposed to address those two challenges was the calculation of a pairwise bioassay activity profile similarity (bAPS): for each of the 618 compounds tested in any of the 244 confirmatory assays the bioassay activity profile was represented as a multiset fingerprint encoding of the activity data as follows: 'active' was set to '2'; 'inactive' as '1'; inconclusive or not tested as '0'; the pairwise bAPS was calculated using the Tanimoto coefficient:[36]

$$\mathrm{bAPS}(i,j) = \frac{\sum_{k=1}^{n} \min[m_k(i), m_k(j)]}{\sum_{k=1}^{n} \max[m_k(i), m_k(j)]}$$

where bAPS(i,j) is the bioassay activity profile similarity of the ith and jth molecules, $m_k(i)$ and $m_k(j)$ are the activity encodings of the ith and jth molecules, respectively, and n is the total number of assays that the molecules were screened across. This encoding of the activity data enabled the systematic structure- and bioprofile activity similarity and identified bioassay activity profile cliffs *i.e.*, pairs of compounds with high structure similarity but very different bioassay activity profiles.[37]

In a separate work Yongye *et al.* analyzed the ALM of a chemogenomics data set released by Clemons *et al.* The data set contained more than 15 000 compounds from different sources (commercial compounds, natural products and synthetic molecules) that where screened across 100 sequence-unrelated proteins.[38] SMARt analysis using SAS maps led to the identification of structural changes that differentiated highly specific from promiscuous compounds. It was also concluded that, in general, similar synthetic structures from academic groups showed greater promiscuity differences than do commercial compounds and natural products.[39] A characteristic metric employed in that work was the Structure-Promiscuity Index Difference (SPID); for each pair of compounds, the relationship between structure similarity and the different number of proteins to which each compound in the pair binds was computed using the expression:

$$\mathrm{SPID}(X_\mathrm{a}, X_\mathrm{b}) = \frac{|P_{X_\mathrm{a}} - P_{X_\mathrm{b}}|}{1 - T_n(X_\mathrm{a}, X_\mathrm{b})}$$

where P_{X_a} and P_{X_b} are the number of proteins to which compounds X_a and X_b are bound and $T_n(X_\mathrm{a},X_\mathrm{b})$ is the pairwise Tanimoto structure similarities of both compounds. The SPID metric is reminiscent of SALI (see above). Noteworthy, SPID focuses on the change in the number of proteins bound associated with a change in the molecular structure but does not account for the specific proteins involved, such as the metric 'binding profile similarity'.[40] In order to address the identity of the proteins Yongye *et al.* also computed the pairwise binding profile similarities employing the binary profile of each compound as a 100-dimensional vector *e.g.*, a pairwise Tanimoto similarity. As such it was also analyzed the multiple–assay profile SAR of the data set using the modified version of SALI: Structure–Multiple Activity Landscape Index (*vide supra*).

Similar to activity landscape analysis with one biological endpoint, the structural interpretation of SMARt with many biological endpoints would require further molecular modeling studies with the three dimensional structures of the targets, if available. An alternative is to incorporate three dimensional molecular descriptors to describe the chemical structures. It is particularly interesting to provide a further rationale of the source of selectivity or promiscuity of the compounds.

Future directions

In principle, methods employed in ALM can be implemented to explore the SAR or SMART of any screening data evaluated across multiple biological endpoints. Moreover, several methods can be extended to mine biological fingerprints. SMART studies can be further extended to analyze properties such as toxicity. In this regard, Austin *et al.* introduced ADMET property–activity pairwise similarity maps to analyze the relationships between activity, structure and ADMET violations/compliance with particular emphasis on determining structural changes that have a large impact on the ADMET compliance.[41]

In drug discovery, big data is typically obtained from high-throughput screening (HTS). HTS usually is conducted in two general steps: assays at a single dose followed by confirmatory assays at multiple doses. Despite the fact that biological assays at single-dose concentrations have not been considered for activity landscape analysis,[42] such assays do provide valuable information that could be considered in preliminary activity landscape studies. We propose that this is a relevant future direction not only in ALM but in SMARt studies in general.

Pro-activity cliffs

Relevant regions in the activity landscape are analyzed using high quality biological activity data that are obtained after multiple-dose inhibition assays. Due to its rigorous determination, it has been proposed that only those values can be used within the realm of ALM methods in order to minimize errors.[42,43] While the latter remains as the ideal case, there are several cases where only single-dose biological activity data for a given target is available. Herein is proposed that this data can be used for a preliminary activity landscape analysis in order to identify potential areas of interest and guide the next steps towards the acquisition of high quality information. Thus, identification of potential activity cliffs *i.e.*, *pro-activity cliffs* is valuable, as they can be prioritized for additional experimental evaluation.

To temporarily address both the lack of multiple-dose/high quality data and the error involved in the biological activity measurement, the percentage of inhibition frequently obtained at single dose evaluations can be distributed in different categories; for instance, potentially very active, active, inactive and potentially very inactive. Integer indices can be assigned to the different classes: *e.g.*, an integer index of 1 for the least active compounds and 4 for the most active ones. The limits of the inhibitory activity can be fitted to the distribution of the data set; *e.g.*, those compounds with less than 25% of inhibition can be regarded as potentially inactive (*e.g.*, activity index of 1),

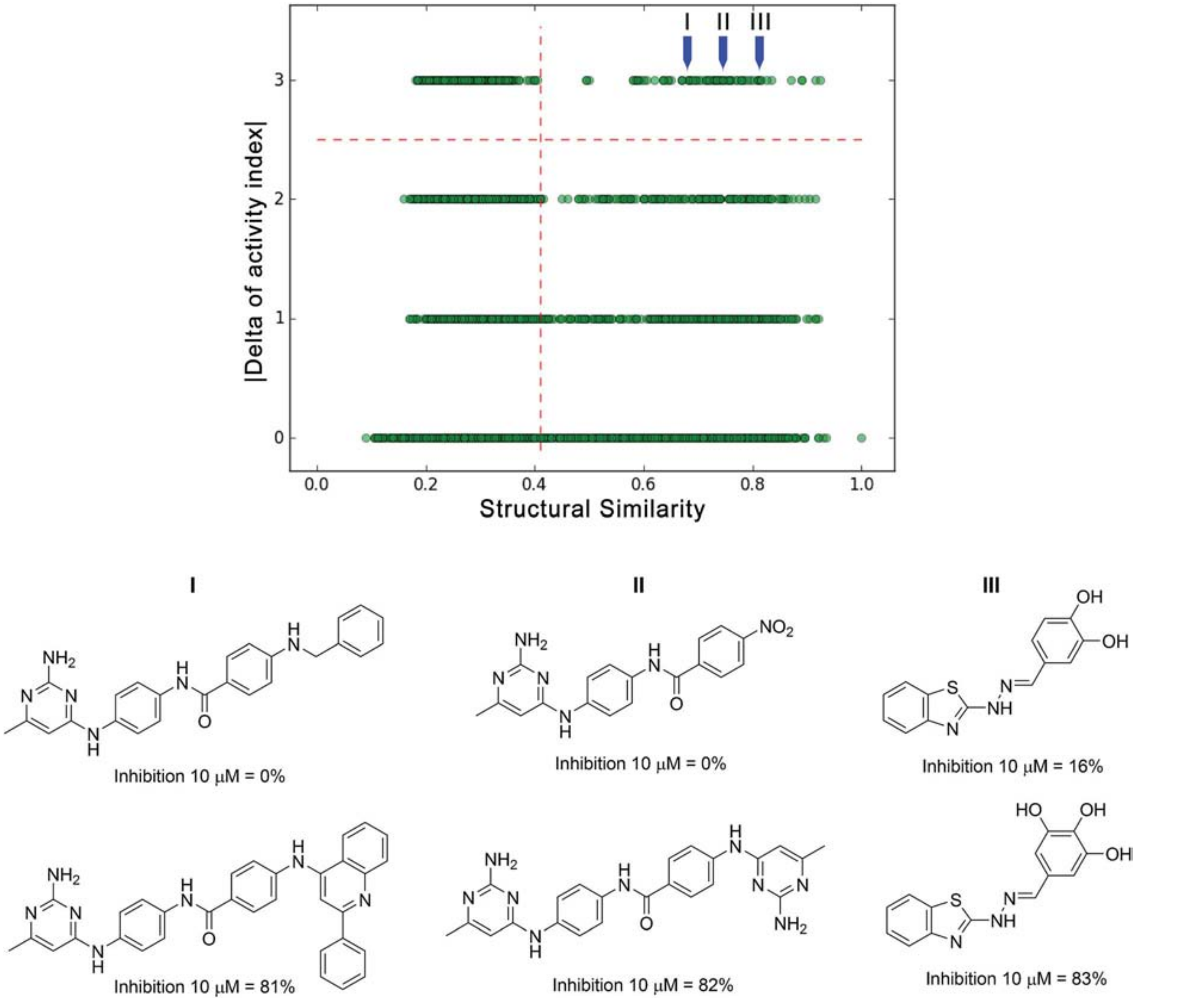


Fig. 4 Examples of pro-activity cliffs for a set of 106 compounds tested as inhibitors of DNMT3A.

while those with more than 75% inhibition can be considered as potentially active compounds (*e.g.*, activity index of 4). After classifying activity data and generating a categorical structure–activity similarity map, four horizontal zones can be defined as the result of comparing the activity index: 0 (as the result of comparing members of the same group), 1 (by comparing members of one unit of difference in the activity index), and so on. Thus, pro-activity cliffs can be defined as pairs of compounds with high structure similarity where one is highly probable to be active and the other is highly probable to be inactive. To illustrate this point, an actual set of single-dose activity data is exemplified for a group of inhibitors of DNA methyltransferase 3A (DNMT3A); for a large number of compounds, only percentages of inhibition obtained at single dose are available (10 μM). Fig. 4 shows a categorical SAS map for 106 compounds tested as potential modulators of DNMT3A. The SAS map in this figure has 5565 data points; the *x*-axis represents the pair-wise structure similarity computed as the mean of the Tanimoto similarity values computed with Extended Connectivity Fingerprints (radius 2) and MACCS keys (166 bits). The *y*-axis represents the four regions defined by the difference of the activity indices. The vertical dashed line is marked in the 3rd quartile of the pair-wise mean similarity values of the data set (mean similarity of 0.41). In Fig. 4 upper right quadrant identifies the pro-activity cliffs. The same figure illustrates three specific examples of pro-activity cliffs. As shown in Fig. 4, the three pairs of compounds show a remarkable resemblance, and a high difference in their inhibition activities. For instance, the only structural difference in pro-activity cliff "III" is a hydroxyl group. Further multiple-dose testing would confirm or not the status of the potential activity cliffs.

Concluding remarks

ALM is a quantitative approach to analyze systematically SAR of compound data sets. In many drug discovery programs compound data sets are screened against two, three or many more biological endpoints. To rapidly mine the usually large data generated, ALM have been adapted to analyze the associated SMARt. Among ALM approaches, SAS maps have evolved rapidly to address the increasing need of analyzing SMARt. To date, several successful applications have been reported including the analysis of SAR, SMARt and protein–ligand interaction cliffs. As part of the development of the SAS maps, a number of metrics and visualization approaches have been developed. Since SMARt analysis usually involves analysis of large amount of data, getting smart in drug discovery may require using information available in large screening campaigns that include incomplete chemogenomics data sets or activity data obtained at single concentrations. Bioactivity-profile similarity, activity landscape sweeping and pro-activity cliffs are examples of recently proposed concepts to advance the SMARt analysis in drug discovery. One of the major perspectives in the field is to incorporate the principles of quantum mechanics to refine the SMARt models and further improve their applicability in drug discovery projects.

Conflict of interest

The authors declare that they do not have any conflict of interest related to this manuscript.

Acknowledgements

This work was supported by the Universidad Nacional Autónoma de México (UNAM) grant 'Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza' (PAPIME) PE200116 and grant 'Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica' (PAPIIT) IA204016. We also thank funding from the 'Programa de Apoyo a la Investigación y el Posgrado' (PAIP) 50009163, Facultad de Química, UNAM. OP-H is grateful to CONACyT for the fellowships granted No. 765897/606003. JJ-N is grateful with PECEM MD PhD program for the organization of the fellowship. Rich discussions with Oscar Méndez-Lucio, Eli Fernández-de-Gortari and Mario Omar García-Sánchez are highly acknowledged.

References

1 O. Mendez-Lucio, J. J. Naveja, H. Vite-Caritino, F. D. Prieto-Martinez and J. L. Medina-Franco, *J. Mex. Chem. Soc.*, 2016, **60**, 168–181.

2 J. L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker and R. A. Houghten, *Drug Discovery Today*, 2013, **18**, 495–501.

3 X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 1138–1145.

4 J. L. Medina-Franco, G. Navarrete-Vázquez and O. Méndez-Lucio, *Future Med. Chem.*, 2015, **7**, 1197–1211.

5 H. González-Díaz, D. M. Herrera-Ibatá, A. Duardo-Sánchez, C. R. Munteanu, R. A. Orbegozo-Medina and A. Pazos, *J. Chem. Inf. Model.*, 2014, **54**, 744–755.

6 G. M. Casañola-Martin, H. Le-Thi-Thu, F. Pérez-Giménez, Y. Marrero-Ponce, M. Merino-Sanjuán, C. Abad and H. González-Díaz, *Mol. Diversity*, 2015, **19**, 347–356.

7 F. Durán, N. Alonso, O. Caamaño, X. García-Mera, M. Yañez, F. Prado-Prado and H. González-Díaz, *Int. J. Mol. Sci.*, 2014, **15**, 17035.

8 F. Saldívar-González, F. D. Prieto-Martínez and J. L. Medina-Franco, *Educ. Quim.*, 2017, DOI: 10.1016/j.eq.2016.06.002.

9 G. M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535.

10 M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro and F. Borges, *Drug Discovery Today*, 2014, **19**, 1069–1080.

11 O. Méndez-Lucio, J. Pérez-Villanueva, R. Castillo and J. L. Medina-Franco, *Mol. Inf.*, 2012, **31**, 837–846.

12 J. J. Naveja and J. L. Medina-Franco, *RSC Adv.*, 2015, **5**, 63882–63895.

13 J. L. Medina-Franco, K. Martinez-Mayorga, A. Bender, R. M. Marin, M. A. Giulianotti, C. Pinilla and R. A. Houghten, *J. Chem. Inf. Model.*, 2009, **49**, 477–491.

14 A. Yongye, K. Byler, R. Santos, K. Martínez-Mayorga, G. M. Maggiora and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2011, **51**, 1259–1270.

15 P. Iyer, D. Stumpfe, M. Vogt, J. Bajorath and G. M. Maggiora, *Mol. Inf.*, 2013, **32**, 421–430.

16 D. Dimova and J. Bajorath, *Mol. Inf.*, 2016, **35**, 181–191.

17 R. Guha, *J. Chem. Inf. Model.*, 2012, **52**, 2181–2191.

18 J. Husby, G. Bottegoni, I. Kufareva, R. Abagyan and A. Cavalli, *J. Chem. Inf. Model.*, 2015, **55**, 1062–1076.

19 V. Shanmugasundaram and G. M. Maggiora, presented in part at the *222nd ACS National Meeting*, Chicago, IL, United States, August 26–30, 2001.

20 J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2012, **52**, 2485–2493.

21 J. J. Naveja and J. L. Medina-Franco, *Expert Opin. Drug Discovery*, 2015, **10**, 1059–1070.

22 J. H. Van duzer, R. Mazitschek, Y. Ding, N. Yu, Y. Cao and Y. Liu, *Pyrimidine Hydroxy Amide Compounds as Protein Deacetylase Inhibitors and Methods of Use Thereof*, Acetylon Pharmaceuticals, Inc., 2014, EP2640709.

23 A. Dueñas-González, J. Jesús Naveja and J. L. Medina-Franco, in *Epi-Informatics*, Academic Press, Boston, 2016, pp. 1–20.

24 J. J. Naveja, F. Cortés-Benítez, E. Bratoeff and J. L. Medina-Franco, *Mol. Diversity*, 2016, **20**, 771–780.

25 J. Pérez-Villanueva, O. Méndez-Lucio, O. Soria-Arteche and J. Medina-Franco, *Mol. Diversity*, 2015, **19**, 1021–1035.

26 R. Guha and J. H. VanDrie, *J. Chem. Inf. Model.*, 2008, **48**, 646–658.

27 O. Méndez-Lucio, A. J. Kooistra, C. d. Graaf, A. Bender and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2015, **55**, 251–262.

28 S. C. Brewerton, *Curr. Opin. Drug Discovery Dev.*, 2008, **11**, 356–364.

29 J. L. Medina-Franco, O. Méndez-Lucio and K. Martinez-Mayorga, *Adv. Protein Chem. Struct. Biol.*, 2014, **96**, 1–37.

30 R. Guha, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 829–841.

31 V. F. Kuyoc-Carrillo and J. L. Medina-Franco, *Drug Dev. Res.*, 2014, **75**, 313–323.

32 J. Waddell and J. L. Medina-Franco, *Bioorg. Med. Chem.*, 2012, **20**, 5443–5452.

33 J. Perez-Villanueva, R. Santos, A. Hernandez-Campos, M. A. Giulianotti, R. Castillo and J. L. Medina-Franco, *Bioorg. Med. Chem.*, 2010, **18**, 7380–7391.

34 J. Pérez-Villanueva, R. Santos, A. Hernández-Campos, M. A. Giulianotti, R. Castillo and J. L. Medina-Franco, *MedChemComm*, 2011, **2**, 44–49.

35 R. Aguayo-Ortiz, J. Perez-Villanueva, A. Hernandez-Campos, R. Castillo, N. Meurice and J. L. Medina-Franco, *Future Med. Chem.*, 2014, **6**, 281–294.

36 G. M. Maggiora and V. Shanmugasundaram, in *Chemoinformatics and Computational Chemical Biology, Methods in Molecular Biology*, ed. J. Bajorath, Springer, New York, 2011, vol. 672, pp. 39–100.

37 J. L. Medina-Franco and J. Waddell, *J. Mex. Chem. Soc.*, 2012, **56**, 163–168.

38 P. A. Clemons, N. E. Bodycombe, H. A. Carrinski, J. A. Wilson, A. F. Shamji, B. K. Wagner, A. N. Koehler and S. L. Schreiber, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 18787–18792.

39 A. B. Yongye and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2012, **52**, 2454–2461.

40 A. Steffen, T. Kogej, C. Tyrchan and O. Engkvist, *J. Chem. Inf. Model.*, 2009, **49**, 338–347.

41 A. B. Yongye and J. L. Medina-Franco, *Drug Discovery Today*, 2013, **18**, 732–739.

42 D. Stumpfe and J. Bajorath, *J. Med. Chem.*, 2012, **55**, 2932–2942.

43 J. L. Medina-Franco, G. M. Maggiora, M. A. Giulianotti, C. Pinilla and R. A. Houghten, *Chem. Biol. Drug Des.*, 2007, **70**, 393–412.

J. Mex. Chem. Soc. **2017**, *61(3)*, 266-272
© 2017, Sociedad Química de México
ISSN 1870-249X

Progress on the Computational Development of Epigenetic Modulators of DNA Methyltransferases 3A and 3B

Oscar Palomino-Hernández, A. Christiaan Jardínez-Vera, and José L. Medina-Franco*

Departamento de Farmacia, Facultad de Química, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico
* Corresponding author: *Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*
E-mails: medinajl@unam.mx; jose.medina.franco@gmail.com.
Tel. +5255-5622-3899, Ext. 44458

Received January 6th, 2017; Accepted April 21st, 2017.

Abstract. Inhibitors of DNA methyltransferases 3A and 3B (DNMT3A/3B) are promising candidates for the treatment of cancer and other diseases. Selective inhibitors of DNMT3A/3B are also attractive as small-molecule probes. During the past few years has increased significantly the research towards the development of DNMT1 inhibitors. However, there are no reviews of the recent progress in the development of small-molecule inhibitors of DNMT3A/B. Herein we review the status of inhibitors of DNMT3A/3B with emphasis on computational guided approaches. We discuss in a critical manner compound databases containing structure-activity information, crystallographic structures of DNMT3s, structure-guided studies, and virtual (*in silico*) screening coupled with experimental validation that have led to the identification or development of selective inhibitors. Perspectives in the field are also discussed.

Keywords: Drug discovery; Epi-informatics; epigenetics; DNA methyltransferase; structure-activity relationships.

Resumen. Inhibidores de metiltransferasas de ADN 3A y 3B (DNMT3A/3B) son candidatos prometedores para el tratamiento de cáncer y otras enfermedades. Inhibidores selectivos de DNMT3A/3B también son atractivos como moléculas sonda. En los últimos años se ha incrementado el desarrollo de inhibidores de DNMT1. Sin embargo, no hay trabajos de revisión que abarquen el desarrollo de inhibidores de DNMT3A/3B. En este trabajo se revisa el estado actual de la investigación de inhibidores de DNMT3A/3B con énfasis en las estrategias guiadas por métodos computacionales. Se discuten en forma crítica bases de datos moleculares que contienen información de relaciones estructura-actividad, estructuras cristalográficas de DNMT3s, estudios guiados por la estructura y estudios de evaluación virtual unidos a caracterización experimental para la identificación de inhibidores selectivos. También se discuten perspectivas en este campo.

Palabras clave: Descubrimiento de fármacos; Epi-informática; epigenética; ADN metiltransferasa; relaciones estructura-actividad.

1. Introduction

DNA methylation is one of the many epigenetic modifications associated with genetic expression in higher eukaryotes, regulating the processes of DNA transcription and peptide synthesis due to chromatin remodelling. The covalent modification is attributed to the catalytic action of DNA methyltransferases (DNMTs) [1], enzymes which can either promote *de novo* methylation (DNMT3A and DNMT3B) or maintenance methylation (DNMT1), following DNA replication. To date, only these three types of DNMTs with catalytic activity have been identified in the human genome [2], whereas the fourth type, namely DNMT3L, is a highly homologous nuclear protein that lacks the amino acid residues required for catalytic activity.

Regular function of DNMTs enzymes are crucial for DNA methylation in both cellular development and mitosis. However, overexpression of DNMTs can lead to aberrant methylation patterns [3], which have been associated with many cancer types including lung, colorectal, prostate, breast, cervical and pancreatic cancer, among others [4-6]. Experimentally, it has been observed that DNMT inhibition reactivates silenced (hypermethylated) genes, particularly tumor-suppressor genes [7]. Hence, DNMTs are promising biological targets for the design of anti-cancer agents. Also, DNMTs are becoming relevant targets in the pathology of other diseases such as diabetes, obesity [8], Alzheimer’s disease [9], and other central nervous system disorders [10].

De novo methylation is the main activity of the DNMT3A and DNMT3B enzymes [11], enabling key epigenetic modifications essential for processes such as cellular differentiation and embryonic development, transcriptional regulation, heterochromatin formation, X-chromosome inactivation, imprinting, and genome stability. Albeit very similar in structure, these two enzymes appear to have different roles in mammals [12].

In 2015 Medina-Franco *et al*., reviewed inhibitors of DNMTs with emphasis on inhibitors of DNMT1. In that work, the authors discussed different *in-silico* studies used in the development of DNMTs. Examples of techniques discussed were molecular docking, virtual screening, pharmacophore model, molecular dynamics, and similarity searching [13]. In 2016 Prieto-Martínez, *et al*. discussed advances in computational approaches applied mostly to DNMT1 [14]. However, there are no reviews focused on the computational development on DNMT3A and 3B inhibitors. This is because the development of inhibitors of DNMT3A and DNMT3B has been slower. In this work, we summarize recent developments of small molecule inhibitors of DNMT3A and 3B as potential therapeutic agents or molecular probes. The review focuses on the role of computational methods that have contributed to such developments.

2. DNMTs: structure and mechanism

Mammalian DNMTs have a high level of homology, possessing a large N-terminal region of variable size and a C-terminal catalytic portion. Whereas the N-terminal region encodes regulatory functions, the C-terminal region is involved in both cofactor binding and substrate catalysis. These enzymes rely on *S*-adenosyl-L-methionine (AdoMet or SAM) as a methyl group donor and contain several highly conserved structural features such as the SAM binding pocket, the DNA-cytosine binding pocket, and a vicinal proline-cysteine pair important for the reaction mechanism. However, there are structural differences between DNMTs: while DNMT1 relies on its N-terminal domain [15] to recognize hemimethylated strands (hence its function in maintaining the methylation status), DNMT3A and DNMT3B do not show any preference about the methylation status of the DNA, rendering them as *de novo* transferases. The 3D coordinates of DNMTs with different domains have already been deposited in the Protein Data Bank (PDB) [16]. Based on the structural information, mostly obtained from X-ray crystallography, it is plausible to consider at least three types of inhibitors, namely; (a) molecules binding in the co-factor binding pocket (allosteric inhibition), (b) molecules binding in the DNA binding pocket (competitive inhibition), and (c) molecules interacting directly with the catalytic cysteine (irreversible inhibition). Fig. 1 shows the catalytic domains of DNMT3A and DNMT3B. For illustration purposes, this figure shows the DNA-binding domain of human and bacteria DNMT3A and 3B, respectively.

3. Modulators of DNMT3A and DNMT3B

In order to explore the chemical space of the current inhibitors a molecular database of inhibitors of DNMT3A and DNMT3B was assembled following a previously published protocol [17]. Six public databases and two sources of scientific literature were explored. The public databases explored were (1) ChEMBL [18], (2) Therapeutic Target Database (TTD) [19], and (3) ChEpiMod [20] using the query text 'DNMT3A' or 'DNMT3B'; (4) HEMD [21] with the information located in the enzyme browser, option DNA and submenu DNA (cytosine-5)-methyltransferase 3A or 3B; (5) Binding Database [22] searching in the IC_{50} menu, submenu DNA methyltransferase A or B; and (6) ChromoHub [23] clicking on DNMT, option 'browse inhibitors'. To retrieve additional compounds not reported in those public databases, we also reviewed the scientific literature using Web of Science (webofscience.com) and Chemical Abstracts (scifinder.cas.org). After data curation, 269 unique structures with reported activity were identified. It was found that ChEMBL, HEMD and Chemical Abstracts contained the information provided by the other chemical databases, therefore only these three sources are shown in Table 1. The analysis of the database allowed the identification of two classifications of inhibitors: (a) nucleosidic or (b) non-nucleosidic. Fig. 2 shows representative chemical structures of molecules that have been proposed as nucleoside and non-nucleoside inhibitors.

Chemoinformatic tools enable the management and mining of the chemical information in compound databases [24]. Computer-assisted molecular design methods have been used to calculate the distribution of molecular properties of inhibitors of DNMT1 [17,25]. Therefore, it is warranted the chemoinformatic analysis of modulators of DNMT3A/3B. These analyses contribute to the ongoing effort of charting the epigenetic-relevant chemical space (ERCS) [25].

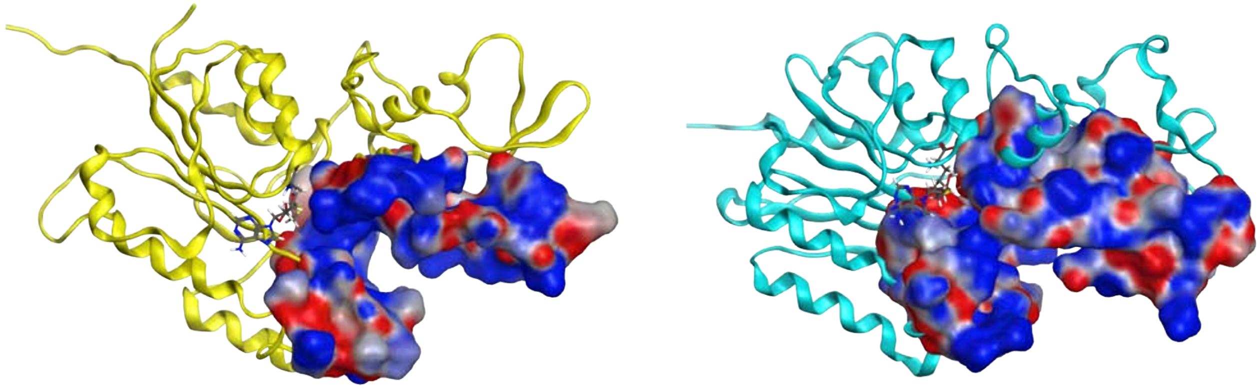


Fig. 1. Catalytic domains of DNMT3A (yellow, PDB ID: 2QRV) and DNMT3B (blue, PDB ID: 5CIY). The electrostatic molecular surface shows the DNA-binding domain. The reduced cofactor SAH is shown in both structures.

Table 1. Sources to build the database of inhibitors of DNMT3.

Source type	Database	Compounds
Public database	ChEMBL	83
Public database	HEMD	25
Literature search	Chemical Abstracts	161
Total number of unique compounds		269

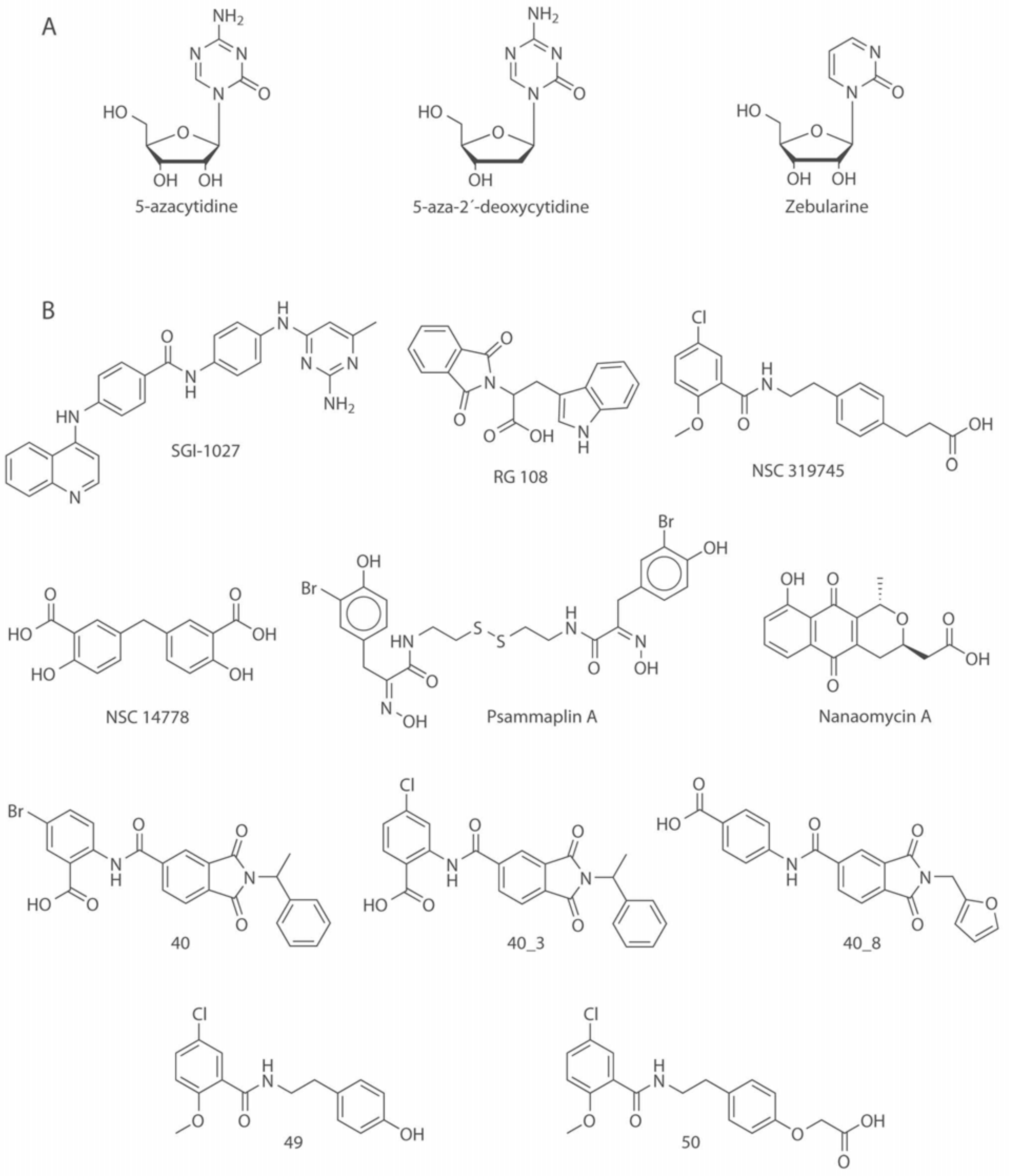


Fig. 2. Chemical structures of selected DNMT3 inhibitors (DNMTi) and other compounds associated with hypomethylating properties. Representative compounds in this figure are roughly classified as A) nucleosidic and B) non-nucleosidic inhibitors.

4. Role of computational methods towards the development of DNMTs

Computational methods have become a corner stone in the development of bioactive compounds [26]. The *in silico* development of inhibitors of DNMT is not exception as demonstrated by the emerging 'Epi-informatics' field [27]. Computer-aided approaches can be divided in two major groups depending on the experimental information used. Structure-based approaches rely on the availability of three-dimensional information of the molecular target, whereas the ligand-based approaches depend mostly on the structure-activity information of small molecule modulators. In this section, we discuss the progress on computational studies towards the development of DNMT3A/3B inhibitors. Several methods can be classified as structure-based methods but representative ligand-based approaches have been developed as well.

4.1 Structure-based analysis of modulators of DNMT3A/3B

As of December 2016, the PDB contained 15 structures associated with DNMTs enzymes of *Homo sapiens*: 12 for DNMT3A and three for DNMT3B (Table 2).

A number of studies have employed the structure of DNMT3A (PDB ID: 2QRV, 4U7P, 4U7T) as a starting point for the construction of a complex between DNMT3A and SAH [13,28-31]. Furthermore, several authors have reported homologous structures of DNMT3B obtained from genetic modifications of the structure of DNMT3A. This is because there is no report of a crystallographic structure that includes the catalytic domain of human DNMT3B [31-35]. These studies highlight the relevance of a set of amino acids such as Ser, Gly, Glu, Cys, Pro, Lys and Arg, showing the affinity of inhibitors against DNMT3B [30,33,36].

Erez-Rechavi *et al*. recently reported a clinical study coupled with a bioinformatics analysis in which expressed a mutation in the protein structure of DNMT3B. It was found that the structural constraints of Ala585 prevent the structure from undergoing mutations or changes in amino acids largely because this alanine residue has a small size. The small number of mutations in this region retains its high degree of structural similarity, which optimizes the binding site of the cofactor and the catalytic efficiency [37].

In 2014 Sheng-Chao *et al.*, reported that antroquinonol D, a ubiquinone derivative, may act as a selective DNMT1 inhibitor. In the study, a docking model of antroquinonol D with DNMT1 was superimposed with a crystallographic structure of DNMT3A and with a model structure of DNMT3B. It was observed that antroquinol D could not fit in the cavity of binding site of DNMT3B. These results provided a rationale of the observed low inhibition of antroquinol D in DNMT3B [35].

In a different study, Kuck *et al.* also overlapped the structures of DNMT1 and DNMT3B and found that the Gln89 residue in DNMT1 is located towards the interior of the binding site, while the corresponding residue in DNMT3B (Asn652) is located about 4 Å outside the binding pocket. Authors of that work concluded that Gln89 acts as a hydrogen bond donor available to bind to inhibitors. In contrast, in the DNMT3B active site this hydrogen bonding interaction is absent [32]. Combinations of structural domains of DNMTs have been reported

Table 2. Crystallographic structures of human DNMT3A and 3B available in Protein Data Bank.

Structure	Domain	PDB ID	Co-crystal ligand	Resolution Å
DNMT3A-DNMT3L	C-terminal	2QRV	SAH.	2.89
DNMT3A	ADD	3A1A	EDO, Zn.	2.30
DNMT3A	ADD with histone H3	3A1B	EDO, Zn.	2.29
DNMT3A	PWWP	3LLR	BTB, SO4.	2.30
DNMT3AK47me2	MPP8	3SVM	MLY.	2.31
DNMT3AK44me2	GLP-SET	3SWC	MLY, SAH, Zn.	2.33
DNMT3AK44me0	GLP-SET	3SW9	SFG, Zn.	3.05
DNMT3A-DNMT3L	ADD	4U7P	SAH, Zn.	3.82
DNMT3A-DNMT3L	ADD with histone H3	4U7T	SAH, Zn.	2.90
DNMT3A	ADD bound to H3	4QBQ	Zn.	2.41
DNMT3A	ADD G550D bound H3	4QBR	Zn.	1.90
DNMT3A	ADD E54SR bound H3	4QBS	SO4, TPO, Zn.	1.80
DNMT3B	PWWP	3FLG	---	1.80
DNMT3B	PWWP	3QKL	BTB, SO4.	2.04
H3K36me3–DNMT3B	PWWP	5CIU	GOL, M3L.	2.24

SAH (S-Adenosyl-L-Homocysteine), EDO (Ethylene glycol), Zn (Zinc ion), BTB (Bis-Tris Buffer), SO4 (Sulfate ion), MLY (N-Dimethyllysine), SFG (Sinefungin), TPO (Phosphothreonine), UNX (Unknown Atom or ion), GOL (Glycerol), M3L (N-Trimethyllysine).

that emphasize the importance of arginine residues in the interaction of DNA with DNMT3A [38].

4.2 Molecular docking and dynamics

The study of active DNMT3A and DNMT3B inhibitors using molecular docking has helped to explore, at the molecular level, the protein-ligand interactions associated with compound affinity and, in some cases, selectivity.

Induced-fit docking (IFD) has been conducted with DNMT3A focusing the analysis on the binding site of the cofactor SAH. The results have shown that inhibitors could occupy the binding site of the cofactor SAH [13,29,31,39]. In particular, hydrogen bonds are formed in the catalytic site between ligands and Val754, Phe636, Thr641, Glu660, Val661, Val683, Arg684, Ser704, Cys706, Asn711, Leu726, Glu752, Arg883, Leu884, and Arg887 [29,39]. Some of these interactions are like those observed with SAH and Arg684, Thr641 y Glu660.

Fahad-Aldawsari *et al.* performed a docking study of resveratrol analogues with DNMT3B. Authors concluded that the synthetic analogues of the natural product exhibited π-π type of interactions with Trp889 and Trp834 in DNMT3A and DNMT3B, respectively [30] and hydrogen bonds interactions with the amino acids of the pocket: Ser111, Gly112, Arg157, Arg193, Pro650, Gly697, Arg731, Arg733, Lys828, Gly831, Arg832 and Cys651. In particular, the hydrogen bond with the catalytic Cys651 could prevent nucleophilic attack of cysteine to the target cytosine. The authors discussed that these hydrogen bonds play a key role in the stabilization of the protein-ligand complex, which could suppress the function of DNMT3B [30,32,33]. Fahad-Aldawsari *et al.* showed that, in the presence of the cofactor, resveratrol analogues make interactions with the catalytic cysteine Cys706, Glu752, and Arg788 in DNMT3A; and with Cys651, Glu697 and Arg733 in DNMT3B. The binding energy calculated with docking for resveratrol analogues were favored for DNMT3A and DNMT3B but not for DNMT1. This result was in agreement with the observed experimental selectivity [30].

The presence of water molecules in the binding site of the cofactor seems to play an important role. To further investigate the role of water molecules, molecular dynamics have been reported allowing optimizing the selectivity and binding affinity of the hit to lead molecules against DNMTs. Evans *et al.*, pointed out that the conformational entropy of proteins increases the binding affinity of DNMT with its cofactor which is assisted by SAM. When the structure of the protein becomes rigid, the entropic contribution of the cofactor decreases [34].

Caulfield *et al*. reported a molecular dynamics study of the inhibitor nanaomycin A (Fig. 2) with DNMT3B. The study of nanaomycin A was performed in the presentence and absence of the cofactor SAM. The energy profiles showed less stability for the complexes without SAM (not interactions with water molecules) in contrast to those simulations performed in the presence of SAM. Authors also concluded that in the SAM-DNMT3B complex the presence of water molecules favor binding of nanaomycin A with the thiol group of Cys651. Nanaomycin A also exhibited interactions with the amino acids Arg731, Arg733, Arg832, and Cys651 [36].

4.3 Virtual screening

Virtual screening of chemical libraries, followed by experimental validation of hit compounds is a technique commonly used in drug discovery. Virtual screening has been used to identify hit molecules that directly or indirectly have led to the identification of inhibitors of DNMT3A/3B. For instance, Kuck, *et al*., reported a virtual screening of a collection of more than 65,000 compounds from the National Cancer Institute [32]. For that study, authors used lead-like filters and molecular docking. One of the hit compounds was later used as a starting point of a hit optimization program reported few years later by Kabro *et al.*, These authors developed two new compounds **49** and **50** (using the labeling in the original papers), analogous of the virtual screening hit NSC319745 (Fig. 2) as inhibitors of DNMT3A [40].

Maldonado-Rojas *et al*. performed a virtual screening of natural products in three main steps: (1) QSAR based on Linear Discriminant Analysis (LDA), (2) molecular docking, and (3) cluster analysis. Six natural products with new scaffolds were selected as virtual DNMTi hits: 9, 10-dihydro-12-hydroxygambogic acid, phloridzin, 2, 4-dihydroxychalcone 4-glucoside, daunorubicin, pyrromycin, and centaurein (Fig. 3). Experimental testing of the hit compounds is warranted [39].

The increasing amount of structural data available for the catalytic domain of human DNMT3A/3B encourage the continued identification of novel inhibitors using structure-based virtual screening of compound databases followed by experimental testing.

4.3.1 Ligand-based virtual screening

Ligand-based virtual screening can be used to select series of molecules or compounds that have shown activity against a receptor target. In order to identify new inhibitors of DNMT3A, Shao *et al*., performed an analogous search using ligand-based virtual screening of the SPECS database. Compound **40** (using the labeling in the original papers, Fig. 2) was used as reference. Pharmacophore-based mapping and molecular docking were combined to conduct the search. Two compounds **40_3** and **40_8** (Fig. 2) were identified as effective inhibitors [31].

In an independent study, Méndez-Lucio *et al.*, conducted three-dimensional similarity searching using NSC14778 (Fig. 2) as reference. Authors identified olsalazine (Fig. 3) as an experimentally validated hypomethylating agent with potential to be an inhibitor of DNMT. The later because of its high structural similarity compared to NSC14778 [33]. It remains to conduct the enzymatic inhibition assays to test if this approved anti-inflammatory drug is selective towards DNMT3A and/or 3B [33].

4.4 Structure-activity relationships (SAR)

Quantitative approaches have been used to identify descriptors associated with the activity or specificity of compounds. For

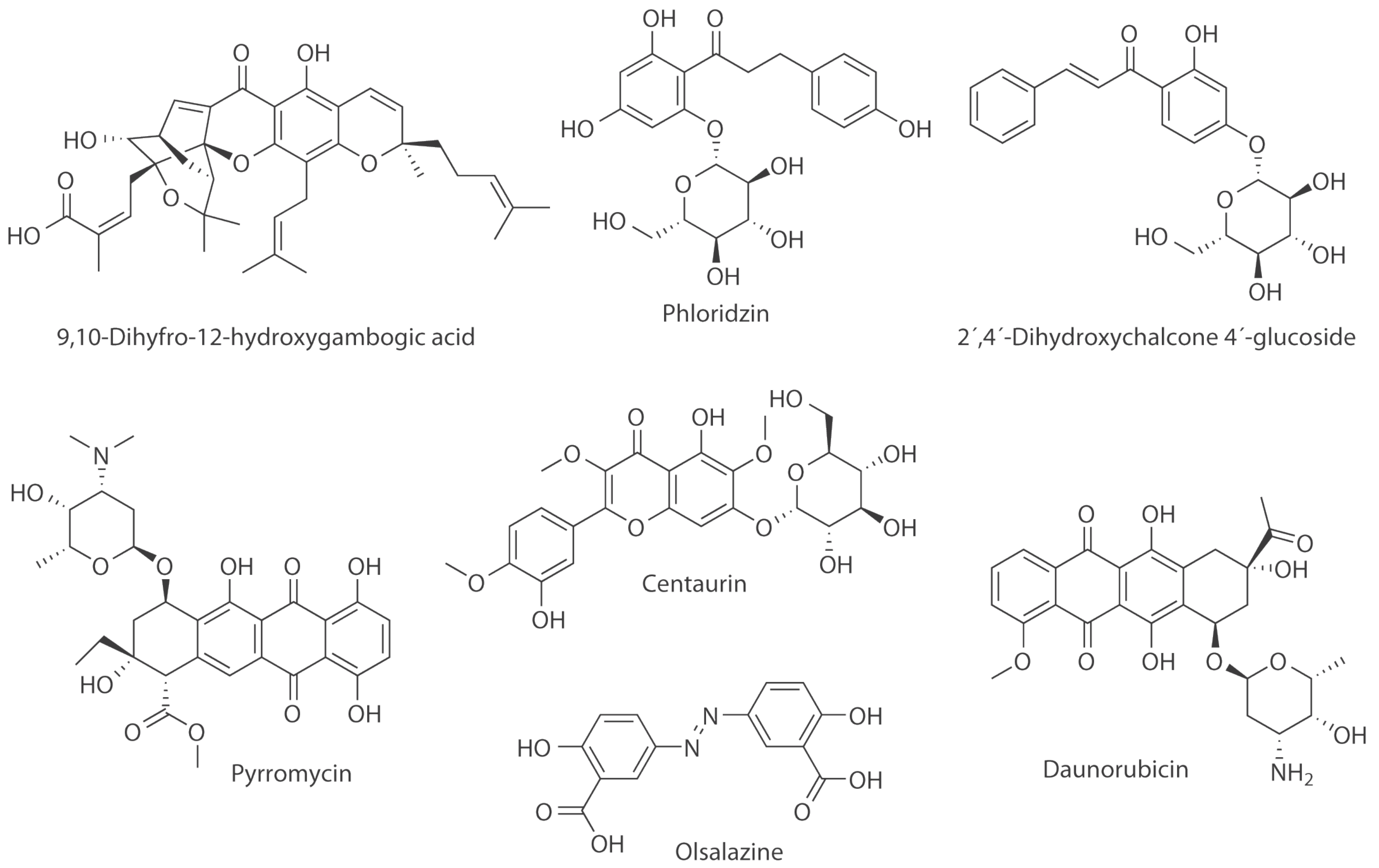

9,10-Dihyfro-12-hydroxygambogic acid
Phloridzin
2´,4´-Dihydroxychalcone 4´-glucoside
Pyrromycin
Centaurin
Olsalazine
Daunorubicin

Fig. 3. Computational hits with potential inhibitory activity of DNMTs.

instance, as commented above, Maldonado-Rojas *et al.* performed a QSAR study based on LDA to select natural products with putative activity against DNMTs. For that study, authors selected a total of 47 compounds dividing them into a training and test sets. Six different types of molecular descriptors, calculated with the software Dragon 5.5, were included in the model: information index (SIC2), 3DmoRSE descriptor (Mor13m), 2D autocorrelation (GATS5m), Randi´c molecular profile (SHP2), topological descriptor (ZM2V), and atomic centered-fragment (H-047). The QSAR model had a R= 0.94, Q= 9.9, F(6,25)= 16.262 and $p < 0.00001$ [39]. The model was later used to predict the activity of new compounds with potential active molecules (*vide supra*). Overall, it is anticipated that the increasing amount of structure-activity data published for DNMT3A/3B (Table 1) can be used to continue developing quantitative models to explore SAR.

5. Conclusions and future directions

In the past few years, there has been a significant upsurge in information correlating the structure and function(s) of DNMT3 enzymes, with the structure-activity relationships defining the potency and selectivity for DNMT3A and DNMT3B inhibitors. Some of this information is publicly available in several databases, including the Protein Data Bank and small-molecule databases such as ChEMBL, Binding Database, and other epigenetic-specific databases such as ChromoHub and Human Epigenetic Enzyme and Modulator Database. Consequently, the complexity associated with the analysis of large sets of data on the wide variety of DNMT3A/3B inhibitors requires the chemoinformatic characterization of the chemical space, including the quantification of its structural diversity. This effort will continue to expand the characterization of the ERCS that has been already started [25]. Structural data of DNMT3 is a key component in the computational-assisted design of inhibitors of DNMT3A/B. The combination of computational screening, with medicinal chemistry-guided optimization and experimental validation, has led to the identification of novel and specific inhibitors of DNMT3. These inhibitors may be promising candidates for the therapeutic treatment of a number of diseases, including cancer, or as molecular probes. It is anticipated that the increasing amount of SAR data will speed up the development of new and more specific inhibitors. It is also expected that the increasing amount of SAR will favor the development of qualitative and quantitative models such as activity landscape models and QSAR predictive models.

Acknowledgments

This work was supported by the Universidad Nacional Autónoma de México (UNAM) grant ‘Programa de Apoyo a Proyectos de Investigación e Inovación Tecnológica’ (PAPIIT) IA204016. We also thank funding from the ‘Programa de Apoyo a la Investigación y el Posgrado’ (PAIP) 50009163, Facultad de Química, UNAM. ACJ-V acknowledges CONACyT for the postdoctoral funding agreement No. 291053. OP-H is grateful to CONACyT for the fellowships granted No. 765897/606003.

References

1. Bestor, T. H. *Human Molecular Genetics* **2000**, *9*, 2395.
2. Yokochi, T.; Robertson, K. D. *J. Biol. Chem.* **2002**, *277*, 11735.
3. Denis, H.; Ndlovu, M. N.; Fuks, F. *EMBO reports* **2011**, *12*, 647.
4. Gao, J.; Wang, L.; Xu, J.; Zheng, J.; Man, X.; Wu, H.; Jin, J.; Wang, K.; Xiao, H.; Li, S.; Li, Z. *J. Exp. Clin. Cancer Res.* **2013**, *32*, 86.
5. Gnyszka, A.; Jastrzebski, Z.; Flis, S. *Anticancer Res.* **2013**, *33*, 2989.
6. Subramaniam, D.; Thombre, R.; Dhar, A.; Anant, S. *Front. Oncol.* **2014**, *4*.
7. Karpf., A. R.; Jones., D. A. *Oncogene* **2002**, *21*, 5496.
8. Arguelles, A. O.; Meruvu, S.; Bowman, J. D.; Choudhury, M. *Drug Discovery Today* **2016**, *21*, 499.
9. Cacabelos, R.; Torrellas, C. *Exp. Opin. Drug Discov.* **2014**, *9*, 1059.
10. Szyf, M. *Eur. Neuropsychopharmacol.* **2015**, *25*, 682.
11. Jia, Y.; Li, P.; Fang, L.; Zhu, H.; Xu, L.; Cheng, H.; Zhang, J.; Li, F.; Feng, Y.; Li, Y.; Li, J.; Wang, R.; Du, J. X.; Li, J.; Chen, T.; Ji, H.; Han, J.; Yu, W.; Wu, Q.; Wong, J. *Cell Discovery* **2016**, *2*, 16007.
12. Takeshima, H.; Suetake, I.; Shimahara, H.; Ura, K.; Tate, S.-i.; Tajima, S. *J. Biochem.* **2006**, *139*, 503.
13. Medina-Franco, J. L.; Méndez-Lucio, O.; Yoo, J.; Dueñas, A. *Drug Discovery Today* **2015**, *20*, 569.
14. Prieto-Martínez, F.; Peña-Castillo, A.; Méndez-Lucio, O.; Fernández-de Gortari, E.; Medina-Franco, J. L. *Adv. Protein Chem. Struct. Biol.* **2016**, *105*, 1.
15. Araujo, F. D.; Croteau, S.; Slack, A. D.; Milutinovic, S.; Bigey, P.; Price, G. B.; Zannis-Hajopoulos, M.; Szyf, M. *J. Biol. Chem.* **2001**, *276*, 6930.
16. Erdmann, A.; Arimondo, P. B.; Guianvarc’h, D. In *Epi-Informatics*; Medina-Franco, J. L., Ed.; Academic Press: London, UK, 2016, p 53.
17. Fernandez-de Gortari, E.; Medina-Franco, J. L. *RSC Adv.* **2015**, *5*, 87465.
18. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. *Nucleic Acids Res.* **2014**, *42*, D1083.
19. Yang, H.; Qin, C.; Li, Y. H.; Tao, L.; Zhou, J.; Yu, C. Y.; Xu, F.; Chen, Z.; Zhu, F.; Chen, Y. Z. *Nucleic Acids Res.* **2016**, *44*, D1069.
20. Meslamani, J.; Smith, S. G.; Sanchez, R.; Zhou, M.-M. *Bioinformatics* **2014**, *30*, 1481.
21. Huang, Z.; Jiang, H.; Liu, X.; Chen, Y.; Wong, J.; Wang, Q.; Huang, W.; Shi, T.; Zhang, J. *PLoS One* **2012**, *7*, e39917.
22. Liu, T. Q.; Lin, Y. M.; Wen, X.; Jorissen, R. N.; Gilson, M. K. *Nucl. Acids Res.* **2007**, *35*, D198.
23. Liu, L.; Zhen, X. T.; Denton, E.; Marsden, B. D.; Schapira, M. *Bioinformatics* **2012**, *28*, 2205.
24. Medina-Franco, J. L. *Drug Dev. Res.* **2012**, *73*, 430.
25. Prieto-Martinez, F. D.; Gortari, E. F.-d.; Mendez-Lucio, O.; Medina-Franco, J. L. *RSC Adv.* **2016**, *6*, 56225.
26. Saldívar-González, F.; Prieto-Martínez, F. D.; Medina-Franco, J. L. *Educ. Quim.* **2017**, *28*, 51.
27. Medina-Franco, J. L.; Yoo, J. In *Epi-Informatics*; Academic Press: Boston, 2016, p 399.
28. Erdmann, A.; Halby, L.; Fahy, J.; Arimondo, P. B. *J. Med. Chem.* **2015**, *58*, 2569.
29. Yoo, J.; Choi, S.; Medina-Franco, J. L. *PLoS One* **2013**, *8*, e62152.
30. Aldawsari, F. S.; Aguayo-Ortiz, R.; Kapilashrami, K.; Yoo, J.; Luo, M.; Medina-Franco, J. L.; Velázquez-Martínez, C. A. *J. Enzyme Inhib. Med. Chem.* **2016**, *31*, 695.
31. Shao, Z.; Xu, P.; Xu, W.; Li, L.; Liu, S.; Zhang, R.; Liu, Y.-C.; Zhang, C.; Chen, S.; Luo, C. *Bioorg. Med. Chem. Lett.* **2017**, *27*, 342.
32. Kuck, D.; Singh, N.; Lyko, F.; Medina-Franco, J. L. *Bioorg. Med. Chem.* **2010**, *18*, 822.
33. Méndez-Lucio, O.; Tran, J.; Medina-Franco, J. L.; Meurice, N.; Muller, M. *ChemMedChem* **2014**, *9*, 560.
34. Evans, D. A.; Bronowska, A. K. *Theor. Chem. Acc.* **2010**, *125*, 407.
35. Wang, S.-C.; Lee, T.-H.; Hsu, C.-H.; Chang, Y.-J.; Chang, M.-S.; Wang, Y.-C.; Ho, Y.-S.; Wen, W.-C.; Lin, R.-K. *J. Agric. Food Chem.* **2014**, *62*, 5625.
36. Caulfield, T.; Medina-Franco, J. L. *J. Struct. Biol.* **2011**, *176*, 185.
37. Rechavi, E.; Lev, A.; Eyal, E.; Barel, O.; Kol, N.; Barhom, S. F.; Pode-Shakked, B.; Anikster, Y.; Somech, R.; Simon, A. J. *J. Clin. Immunol.* **2016**, *36*, 801.
38. Gowher, H.; Loutchanwoot, P.; Vorobjeva, O.; Handa, V.; Jurkowska, R. Z.; Jurkowski, T. P.; Jeltsch, A. *J. Mol. Biol.* **2006**, *357*, 928.
39. Maldonado-Rojas, W.; Olivero-Verbel, J.; Marrero-Ponce, Y. *J. Mol. Graphics Modell.* **2015**, *60*, 43.
40. Kabro, A.; Lachance, H.; Marcoux-Archambault, I.; Perrier, V.; Dore, V.; Gros, C.; Masson, V.; Gregoire, J. M.; Ausseil, F.; Cheishvili, D.; Laulan, N. B.; St-Pierre, Y.; Szyf, M.; Arimondo, P. B.; Gagnon, A. *MedChemComm* **2013**, *4*, 1562.

Research Article

iMedPub Journals
www.imedpub.com

Chemical informatics
ISSN 2470-6973

2017
Vol. 3 No. 2:4

DOI: 10.21767/2470-6973.100025

Comparative Cheminformatic Analysis of Inhibitors of DNA Methyltransferases

Oscar Palomino-Hernández and José L. Medina-Franco.

School of Chemistry, Department of Pharmacy, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico

*Corresponding author:
Jose L Medina Franco

jose.medina.franco@gmail.com (or)
medinajl@unam.mx

Tel: +441454325530

School of Chemistry, Department of Pharmacy,
Universidad Nacional Autónoma de México,
Avenida Universidad 3000, Mexico City 04510,
Mexico.

Citation: Franco JLM, Hernández OP (2017) Comparative Cheminformatic Analysis of Inhibitors of DNA Methyltransferases. Chem Inform Vol. 3 No. 2:4.

Abstract

DNA methylation is an epigenetic mechanism mediated by a family of the enzymes DNA methyltransferases (DNMTs): DNMT1, DNMT3A and DNMT3B. These enzymes are emerging targets for the treatment of cancer and other diseases. Over the past few years several inhibitors of the three enzymes have been reported. Herein, we present a comprehensive chemoinformatic characterization of data sets of inhibitors of DNMT1, DNMT3A and DNMT3B assembled in this work. The compound data sets were analyzed in terms of physicochemical properties, structural fingerprints, and molecular scaffolds. As part of the characterization, a scaffold enrichment analysis was performed as well as visual representation of the chemical space. It was found that inhibitors of DNMT1 are the most diverse covering a broad area of the chemical space. Scaffold diversity analysis showed that inhibitors of DNMT1 and DNMT3A have a larger number of molecular scaffolds as compared to DNMT3B. It was also concluded that for all inhibitors there are molecular scaffolds enriched with active molecules and thus represent promising starting points for additional drug development.

Keywords: Chemical space; Epi-informatics; Epigenetics; Molecular scaffolds; Structure-activity relationships

Received: November 29, 2017; **Accepted:** December 04, 2017; **Published:** December 09, 2017

Introduction

DNA methylation has been identified as a key epigenetic modification responsible for several biological processes including cell differentiation and development, DNA instability, and cancer development [1]. Aberrant methylation patterns are involved in tumor transformation and progression, thus indicating that these epigenetic disruptions are associated with tumorigenesis [2]. This methylation patterns are not stochastic, as they tend to silence tumor suppressor genes. Thus, inhibition of these abnormal methylation levels has been sought as a strategy to reactivate tumor suppressor genes [3,4].

DNA methylation is carried out by DNA-methyltransferases (DNMTs), which donate a methyl group from S-adenosylmethionine (SAM) to the fifth position of cytosine [5]. The enzymes DNMT1, DNMT3A and DNMT3B possess this catalytic ability in mammals [6]. In particular, DNMT1 is responsible for methylating partially methylated DNA strands and thus this it is responsible for DNA-methylation maintenance, whereas DNMT3A and DNMT3B participate in both maintenance and de novo DNA methylation [7].

As of now, the most attractive approach for treating hypermethylation-based cancer is the inhibition of DNA methyl transferases [4,8]. To date, the Food and Drug Administration of the United States has approved two drugs that target DNMTs: azacitidine and decitabine, both for myelodysplastic syndromes [9]. However, these drugs act as covalent inhibitors and are associated with several unwanted effects. Therefore, the design and development of non-covalent DNMT inhibitors is still on the rise [4,10].

Previous studies of the chemical space of epigenetic compounds have been performed [11,12]. However, these studies do not delve further into the molecular differences between the three DNMTs. Moreover, several inhibitors of DNMTs have been recently published and there are not comparative studies regarding their chemical structures and properties. Herein, we report a comprehensive cheminformatic characterization of compound data sets with inhibitors of DNMT1, DNMT3A, and DNMT3B. The characterization was based on physicochemical properties of pharmaceutical relevance, molecular fingerprints, and molecular scaffolds.

Methods

Data sets

A compound database of inhibitors for all three DNMTs was assembled by collecting information from ChEMBL [13], BindingDB [14] and HEMD [15]. Additional searching was done in Web of Science and SciFinder focusing on papers published from 2010 to the time of writing (November 2017). The curation of the datasets was performed in Molecular Operating Environment (MOE) using a published protocol [16,17]. Briefly, a linear notation canonical structure (InChI and SMILES) was obtained for each molecule. Then, molecules were prepared by keeping the largest molecular fragment, removing metals, neutralizing protonation states, and removing duplicates. For identical compounds with close but different activity values, the mean activity value was kept. After data curation, the data sets had 351 unique molecules for DNMT1, 192 for DNMT3A and 86 for DNMT3B.

Several compounds reported for DNMT3A and DNMT3B only had annotated percentages of inhibition. To be able to compare different activity measures, a manual binning of continuous data was performed based on a heuristic criterion: compounds were distributed into four classes (numbered 1-4) as follows: potentially very active, active, inactive, and potentially very inactive. For this analysis, the boundaries were: class 4 if the pIC_{50} was larger than 5.5, or inhibition value was larger than 75%; class 3 if the pIC_{50} was larger than 5, or inhibition value was larger than 50%; class 2 if the pIC_{50} was larger than 4, or inhibition value was larger than 25%; and class 1 if the pIC_{50} was lower than 4, or inhibition value was lower than 25%.

Distribution of relevant chemical properties

Relevant chemical descriptors were computed using MOE and R Core Team utilities [18] in RStudio [19]. Six molecular properties of pharmaceutical interest were computed [20,21]: partition coefficient octanol/water (logP), rotatable bonds (RB), hydrogen-bond donors (HBD), hydrogen-bond acceptors (HBA), topological polar surface area (TPSA), and molecular weight (MW). Six additional topological descriptors were calculated: Plane of Best Fit, globularity, fraction of sp^3 carbons, mass density, radius of Gyration and Wiener Index. For most of these descriptors, a low-energy conformation was used. Data visualization was done using RStudio.

Statistical analysis: The statistical comparison of the descriptors was carried in RStudio with R Core Team and the lawstat, PMCMR, and dunn.test packages. The statistical analysis were a Shapiro test to determine normality of distributions, a Levene test for the evaluation of heteroskedacity of the descriptors, a Kruskal-Wallis test as a non-parametric ANOVA, and Dunn test for post-hoc testing. To assess the impact of heteroskedacity of the distribution of chemical properties, the variance of the distributions for the three libraries was obtained.

Correlation analysis: In order to analyze if the tendency among descriptors is constant within the library, a correlation analysis for detecting subtle differences was used. The correlation between two descriptors X_1 and X_2 was computed using the Pearson product-moment correlation coefficient. For this analysis, the three compound data sets were divided into active and inactive subsets. A correlation analysis was performed generating a correlation matrix for each subset. A Hadamard product was performed for the two matrices, obtaining a matrix with r^2 value for each correlation.

Fingerprint-based diversity

The similarity for all pair of compound in a database was computed using three distinct molecular fingerprints: Molecular Access System (MACCS) keys, Extended Connectivity Fingerprints (ECFP, radius 4), and PubChem fingerprints. The similarity coefficient for fingerprint comparison was the Tanimoto/Jaccard index [22]. The distribution of the similarity values was analyzed with cumulative distribution functions (CDF).

To analyze inter-set similarity, the similarity of a compound in a given set was computed against all the compounds in the other set. The mean and maximum similarity values were recorded and multi-fusion similarity maps [23] were generated.

Scaffold content and diversity

Using the Bemis and Murcko's approach [24] the side chains from the molecules were removed and the molecular scaffold for each molecule was obtained. A unique identifier for each scaffold was assigned with RStudio.

Scaffold enrichment

The molecular scaffolds present in each of the three data sets were classified in terms of their intrinsic activity. Considering a given data set C with n elements and with λ different scaffolds (chemotypes), the intrinsic activity for the λ-th specific chemotype C_λ was calculated as [25]:

$$Act\left[C_\lambda\right] = \frac{1}{n_\lambda}\sum_{i=1}^{n_\lambda}\left[Activity\ Index\right]_i$$

where n_λ is the number of molecules included in the chemotype λ.

The background activity of the data set C was calculated as:

$$Act\left[C\right] = \frac{1}{n}\sum_{i=1}^{n_\lambda}\left[Activity\ Index\right]_i$$

where n is the total number of compounds in the set.

The enrichment factor (EF) for the λ-th specific chemotype was then calculated as:

$$EF\left[C_\lambda\right] = \frac{Act\left[C_\lambda\right]}{Act\left[C\right]}$$

EF indicates how many times a scaffold λ is more active than the mean activity of the compound data set. Thus, scaffolds high EF values are attractive for drug discovery.

Visual representation of chemical space

Visual representations of the chemical space were performed using principal components analysis (PCA) and self-organizing maps (SOMs). Preprocessing of the data was performed using the caret package in RStudio. The visualization of the first PCs and the respective loadings was performed in RStudio with the ggplot2 package. The features used for these methods were the computed chemical descriptors and molecular fingerprints.

Results

Data set creation and curation

Table 1 shows the distribution of the activity values of the three data sets e.g., inhibitors of DNMT1, DNMT3A, and DNMT3B. Results in **Table 1** indicate that, in general, more active compounds have been identified for DNMT1 as compared to DNMT3A and 3B (e.g., larger number of compound in activity class 4). This result may be related to the larger number of compounds developed for DNMT1.

Distribution of relevant chemical properties

Figure 1 shows the distribution of the six properties of pharmaceutical relevance (log P, RB, HBD, HBA, TPSA, and MW). The distributions are shown as a combination of boxplots and violin plots. The figure suggests that the sets of inhibitors of DNMT1 and DNMT3A have similar distributions of HBD and HBA, while DNMT3B has slightly higher values. All sets have comparable distributions of RB. Compounds in the DNMT3B set are slightly less lipophilic (lower logP values) than the other two sets. Regarding TPSA, inhibitors of DNMT1 cover a large range of values, while inhibitors of DNMT3A are centered near the mean of the distribution. The median TPSA values for DNMT3B inhibitors is higher than for the other two sets. **Figure 1** also indicates that all three data sets have comparable distribution of MW.

Figure 2 shows the distribution of selected topological descriptors. Some of the topological descriptors showed small differences between the data sets, as illustrated by the distributions of Plane of Best Fit Index and Globularity. It appears that inhibitors of DNMT3A tend to have a larger volume, as evidenced by the higher values of radius of gyration, Wiener index, and lower mass density. Inhibitors of DNMT3B tend to have higher values of fraction of sp^3 atoms than the other sets.

The distribution of the molecular properties was also analyzed considering the four activity classes of each set. For DNMT1, active compounds tend to have higher values of HBD. For DNMT3A, the inactive compounds tend to have lower values of HBD while the actives have larger values of HBA, RB, log P, and MW. For DNMT3B, the most active compounds tend to have higher values of HBA, HBD, RB, MW and TPSA. Also, the most active compounds are less lipophilic with lower values of log P.

For most topological descriptors there was no relevant difference. Overall, inactive compounds tend to have higher values of globularity and fraction of sp^3 atoms that the other data sets. Some topological descriptors also show that active compounds for DNMT3B have a high Wiener Index, high mass density, a high fraction of sp^3 carbons, and a high radius of gyration.

Statistical analysis: Only the distribution of MW for DNMT1, RG for DNMT3B, and PBF for DNMT3A and DNMT3B had p-values larger than 0.05, indicating that most of the distributions of chemical descriptors for the three enzymes deviate from normality. The Levene test indicated that only RB, MW, PBF and Glob could be considered as having similar variances, rendering the other distributions of descriptors as heteroskedastic, but without high heteroskedastic effects (see Methods). The Kruskal-Wallis analysis indicated that only MW and Glob had similar ranks for the three proteins. The post-hoc Dunn test indicated that between DNMT1 and DNMT3A only HBA and HBD were comparable, while for DNMT1 and DNMT3B Wiener Index and RG had larger p- values than 0.05. Comparing DNMT3A and DNMT3B, only PBF had similar ranks. These results suggested that, in general, the distributions of chemical properties of the three data sets show significant differences.

Correlation analysis: For the three data sets of inhibitors of DNMT1, 3A and 3B compounds were considered active if they had an activity index of 3 or 4, and inactive otherwise. The results of the correlation analysis indicate that the DNMTs show different tendencies between active and inactive subsets in several chemical descriptor. In particular, HBD and radius of gyration showed negative correlation between active and inactive subsets of DNMT3A and DNMT3B, which indicates that this descriptor pair is able to discriminate between active and inactive molecules. For the cross-correlation, HBD and Wiener index were able to distinguish active subsets of DNMT1 and DNMT3A.

Fingerprint-based diversity

Intra-set comparisons: Figure 3 shows the CDF of the pairwise similarity for all the compounds in the DNMT1, 3A, and 3B sets computed with the Tanimoto coefficient and three different fingerprints (see the Methods section). **Table 2** summarizes representative statistics of the distributions.

According to MACCS keys, both DNMT1 and DNMT3A have similar diversity. DNMT3B shows, in general, higher quantile values and higher standard deviation, indicating that compounds in the DNMT3B set are less diverse. According to PubChem and ECFP4 fingerprints, DNMT1 is the most diverse set and DNMT3A is the least diverse. The larger diversity of DNMT1 can be associated with the larger amount of compounds in this set. Interestingly, Pubchem and ECFP4 fingerprints were able to differentiate the data sets. This is associated with the better resolution of these fingerprints as compared to MACCS keys.

Inter-set comparisons: Multi-fusion similarity maps (**Figure 4**) were used to compare the data sets to each other based on fingerprints. When comparing the similarity values of DNMT3A and DNMT3B with DNMT1 as the reference set, DNMT3B tends to cluster in the left bottom area of the plot, with the largest values of mean fusión similarity and the lowest values of max fusion similarity. In contrast, DNMT3A covers a broader area regarding the maximum fusion value. This result indicates that there is a smooth structural overlap between compounds of DNMT3A with DNMT1, while DNMT3B is overall less similar to DNMT1. Taking DNMT3A as reference (**Figure 4**, middle), DNMT1 and DNMT3B have comparable distribution in the multi-fusion similarity map, with some compounds in the DNMT3B set with higher values of mean fusion similarity. The map indicated that most of the molecules in DNMT1 and DNMT3B have, on average, a value

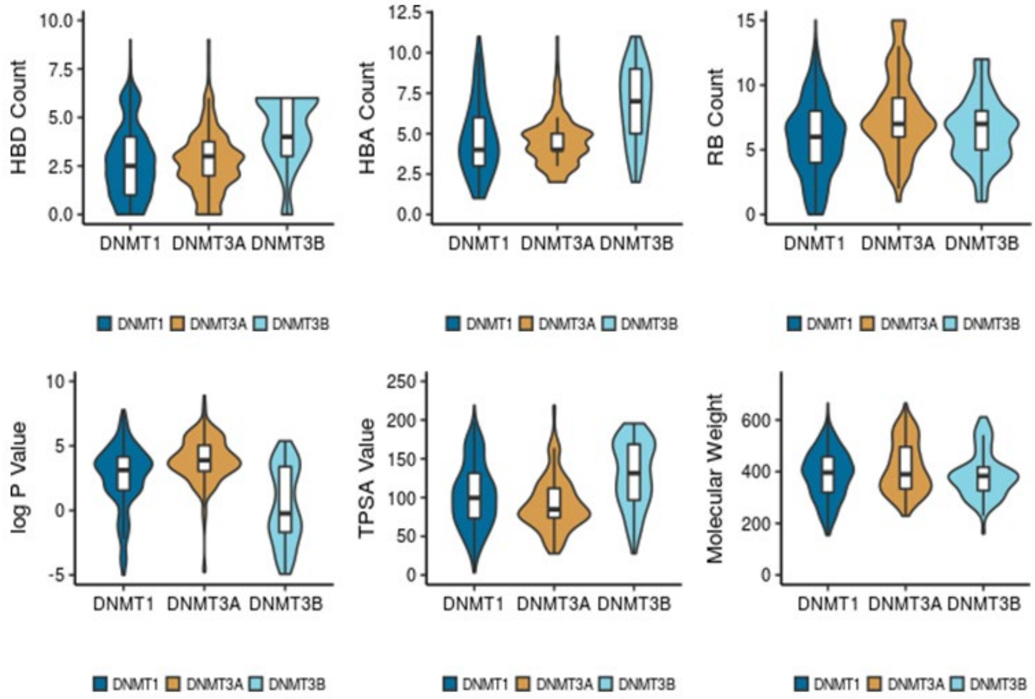


Figure 1 Distribution of pharmaceutical properties of pharmaceutical relevance.

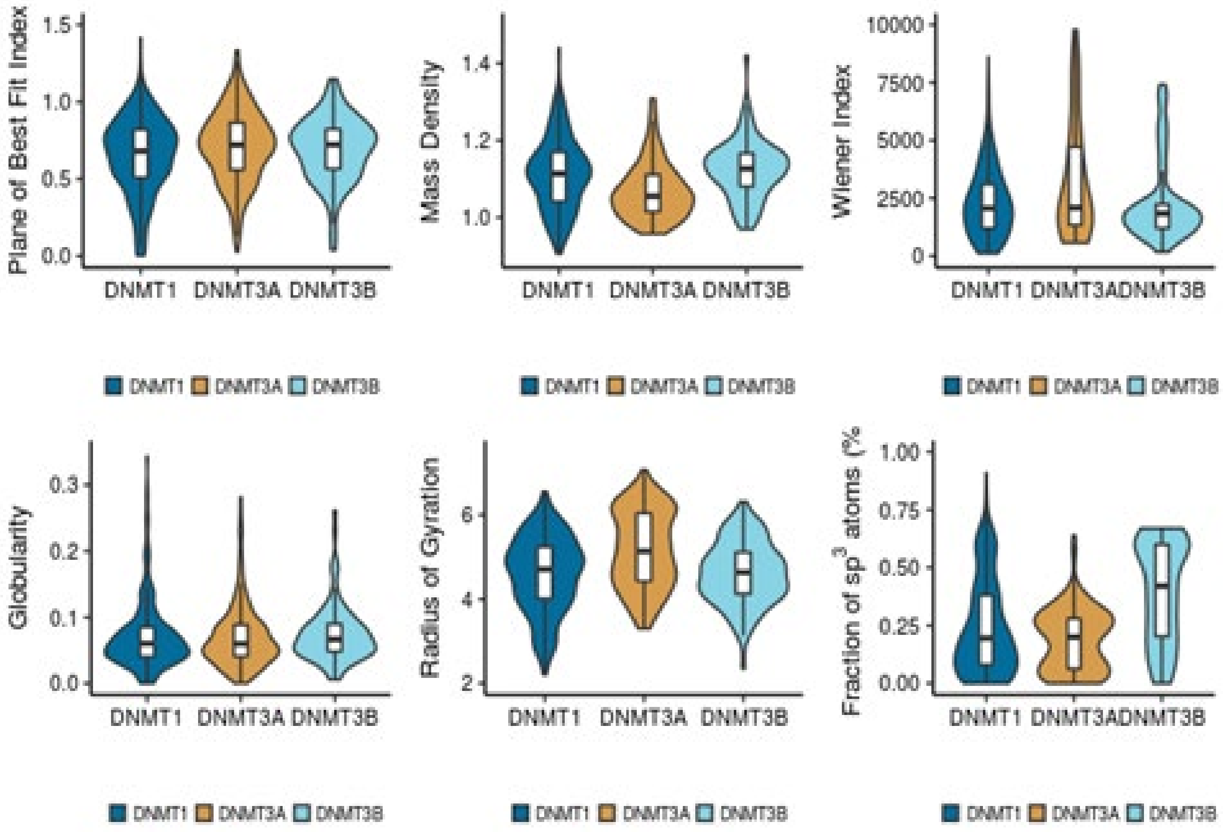


Figure 2 Boxplots and violin plots of topological descriptors.

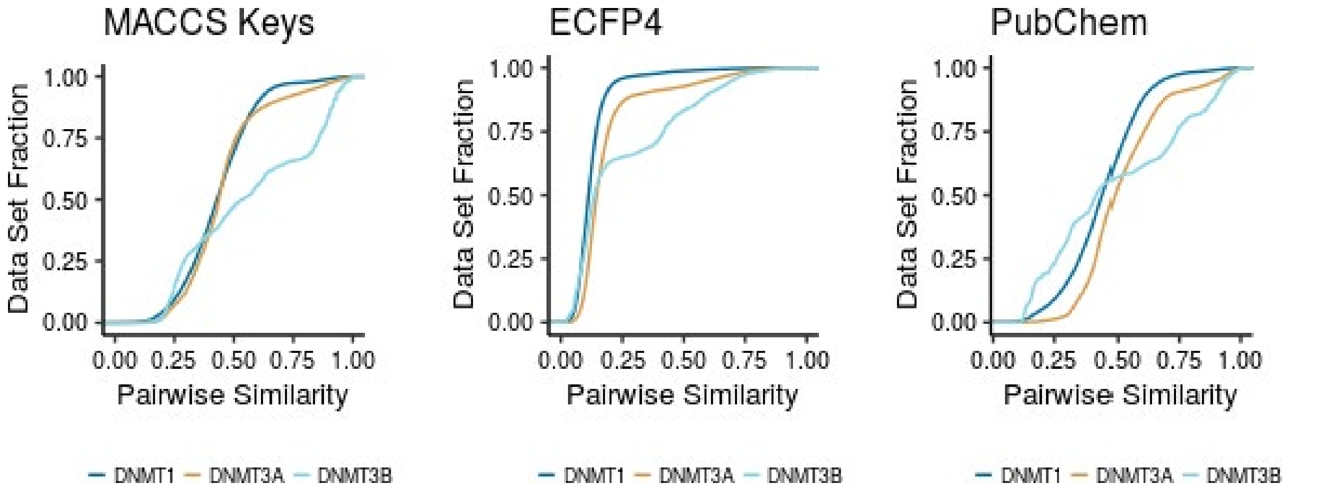


Figure 3 Empirical cumulative distribution functions for the pairwise similarity of compounds in the three data sets calculated with the Tanimoto coefficient and MACCS keys, ECFP4, and PubChem FP.

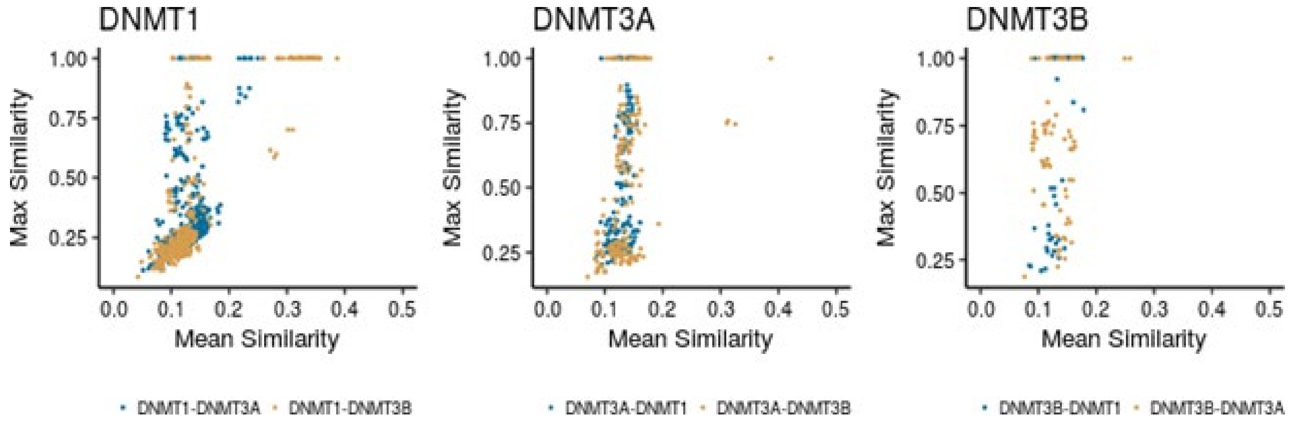


Figure 4 Multi-fusion similarity maps with DNMT1, DNMT3A, and DNMT3B as reference data sets.

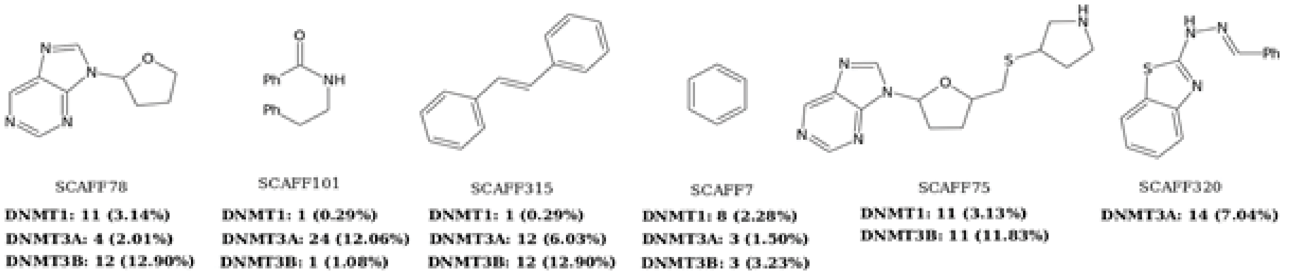


Figure 5 Most frequent scaffolds found in the three sets. The frequency and percentage relative to all scaffolds in the data set are indicated. ’Ph’ = phenyl ring.

of c.a. 0.13 of similarity when compared to DNMT3A, but some compounds in the DNMT3B set are more similar. Considering DNMT3B as reference (**Figure 4),** the DNMT1 set has lower values of maximum fusion similarity. This result indicates proximity between the compoubds of DNMT3A with respect to DNMT3B, and the higher distance between compounds of DNMT1 and DNMT3B.

Scaffold content, diversity and enrichment

Scaffold content: Figure 5 shows the three most frequent scaffolds retrieved for each data set. In agreement with previous scaffold content analysis [16], most of the scaffolds identified in this work were previously found such as SCAFF78, SCAFF75 and SCAFF7. However, additional interesting scaffolds were identified (vide infra).

Scaffold diversity: For each set of inhibitors of DNMT1, 3A, and 3B, the scaffold diversity analysis was done for three sub-sets. The sub-sets were organized based on the reported activity as inactive (intrinsic activity lower than 2), intermediate (intrinsic activity equal or higher than 2, but lower than 3), and active (intrinsic activity equal or higher than 3). Scaffold recovery curves were obtained for each of the nine sub-groups (**Figure 6**). **Table 3** summarizes the results of scaffold diversity for each sub group as determined by different metrics [26].

The scaffold analysis revealed that inhibitors of DNMT1 have a high scaffold diversity, in particular the inactive subset (**Table 3**). In contrast, the active sub-set of DNMT1 is the least diverse. For DNMT3A, the active set and compounds with intermediate activity showed, in general, larger scaffold diversity than the inactive compounds. For DNMT3B, the active set had the largest scaffold diversity. When comparing the active-scaffold subsets from the three enzymes, diversity measures indicated that DNMT3B is the most diverse, followed by DNMT1 and DNMT3A.

Scaffold enrichment: Chemotype-enrichment plots [25] were generated for each set by plotting the scaffold frequency vs. the EF (see the Methods section). The chemotype-enrichment plots are shown in **Figure 7**.

For DNMT1, nearly 55% of the chemotypes have EF values larger than one. The three most frequent scaffolds are SCAFF75, SCAFF78 and SCAFF7 **(Figure 5**). For DNMT3A, 61% of the chemotypes have EF values higher than one. In contrast, for DNMT3B, only 44% of the chemotypes have values larger than one. These results indicated that DNMT3A has been explored more in terms of scaffolds given that it has chemotypes with high frequency and high EF. This figure also shows the existence of some chemotypes with high values of EF and low values of frequency, which could indicate areas of opportunity regarding the development of new SAR studies for the three DNMTs.

Figure 8 shows additional attractive scaffolds: SCAFF254 and SCAFF109 has selectivity for DNMT1; SCAFF266 has high EF for DNMT3A; SCAFF237 has high EF for all three DNMTs.

SAR analysis based on selected scaffolds: Analysis of cofactor-related scaffolds revealed that a substructure of SCAFF78 was present in several chemotypes. Thus, considering SCAFF78 as

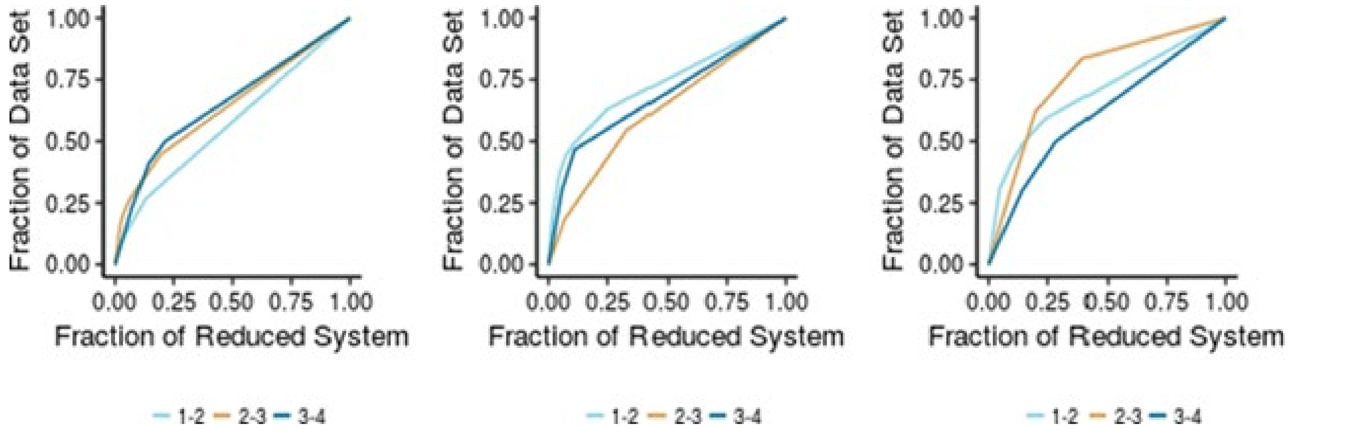


Figure 6 Scaffold recovery curves for DNMT1 (left), DNMT3A (center), and DNMT3B (right), analyzed in terms of highly active scaffolds (3-4), moderately active scaffolds (2-3), and inactive scaffolds (1-2).

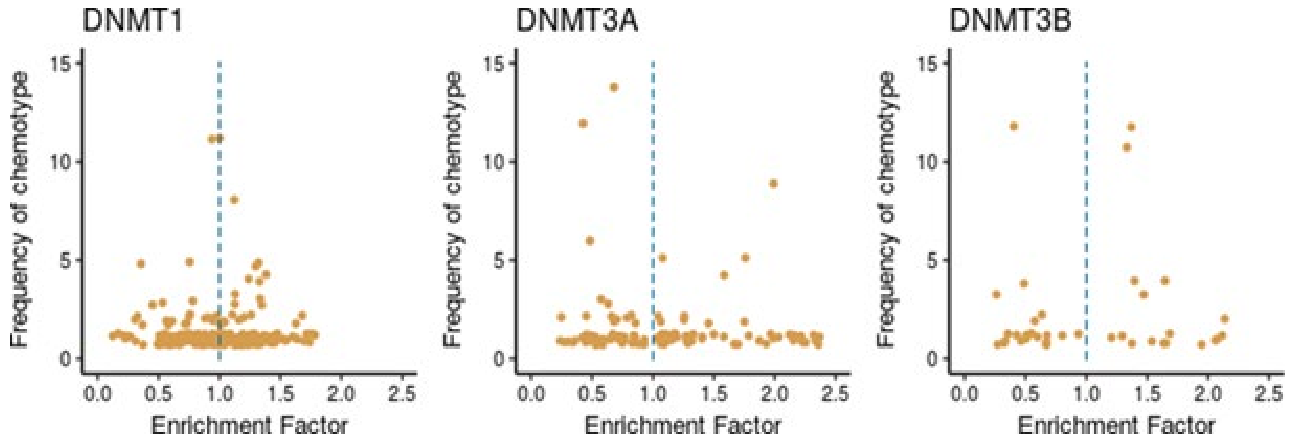


Figure 7 Chemotype-enrichment plots for DNMT1, DNMT3A, and DNMT3B.

reference, the EF values were used as a guide to explore selectivity among the three DNMTs. **Figure 9** shows the core nucleosidic scaffold with two side chains, R_1 and R_2. For this analysis, we used scaffolds with a chemotype frequency equal or larger than three.

Taking the EF of SCAFF78 as a baseline (0.89 for DNMT1, 1.56 for DNMT3A and 1.2 for DNMT3B), and leaving R_2 fixed as a hydrogen, it was found that when R_1=2, the EF improved substantially for DNMT3B (1.18 for DNMT1 and 1.6 for DNMT3B). When R_1=3, the EF decreased for DNMT1 while improving for DNMT3B (0.78 for DNMT1 and 1.9 for DNMT3B). This suggests that elongating the side chain of the scaffold can improve selectivity for DNMT3B against DNMT1. Keeping fixed R_1=1, it was found that the substitution R_2=**A** (**Figure 9**) did not improve the EF (0.96 for DNMT1 and 1.05 for DNMT3B). The substitution R_2=B decreased the EF for DNMT1 while being similar for DNMT3B (0.65 for DNMT1 and 1.09 for DNMT3B). The substitution R_2=**C** diminished overall the EF (0.78 for DNMT1 and 0.73 for DNMT3B). These results suggest that a longer linker in R_2 tends to decrease the overall activity, and that keeping a constrained cycle of six can also favor selectivity for DNMT3B against DNMT1. These results can be found combined in SCAFF77, which has R_1=3 and R_2=A, and has an EF for DNMT1 of 1.18 and for DNMT3B of 1.45, implying that the previous effects cannot interact in synergy. Finally, it was also noted that removing the nitrogen atom marked with the electron pair can increase both EF of DNMT1 and DNMT3B to 1.57 and 1.94, respectively.

Visual representation of the chemical space

Figure 10 shows a visual representation of the chemical space based on PCA of six properties of pharmaceutical relevance i.e., HBA, HBD, TPSA, RB, logP and MW. The first principal component is largely associated with TPSA, HBA and HBD, while the second principal component is associated with RB, MW, and LOGP. **Figure 10** shows that the three data sets share a common space, with

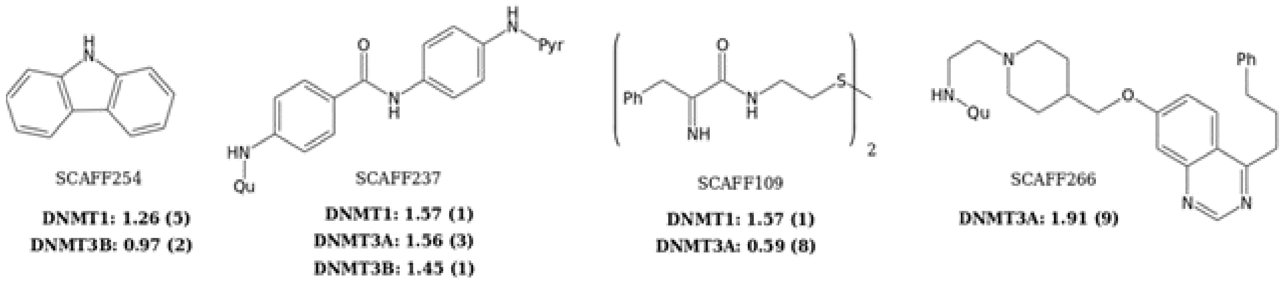


Figure 8 Other representative scaffolds in the datasets (Ph=Phenyl, Qu=Quinoline, Pyr=Pyrimidine). For each scaffold is shown the enrichment factor and scaffold frequency (in parenthesis).

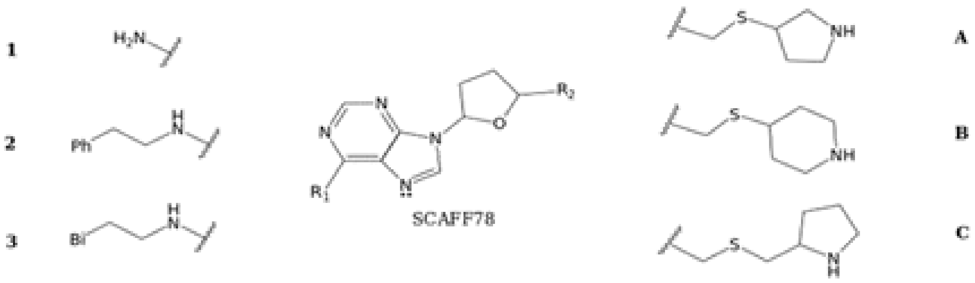


Figure 9 Scaffolds found in the dataset, with the maximum common substructure of SCAFF78 (Ph=Phenyl, Bi=Biphenyl).

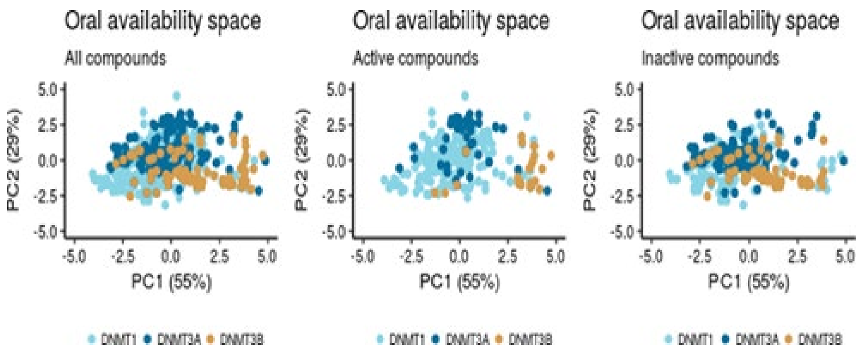


Figure 10 Visualization of the chemical space based on oral availability descriptors (MW, logP, RB, TPSA, HBD and HBA) and a principal component analysis. Left: All compounds. Center: Only active compounds (activity index equal or greater than 3). Right: Only inactive compounds (activity index lower than 3).

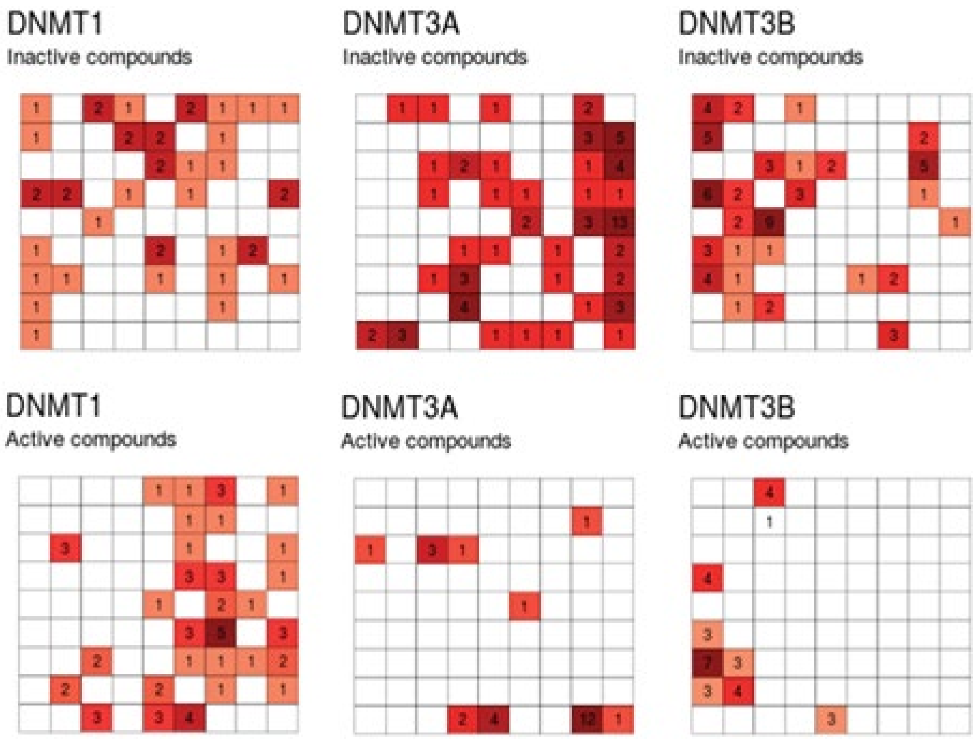


Figure 11 Visualization of the chemical space based on oral availability descriptors (MW, logP, RB, TPSA, HBD and HBA) projected on a self-organizing map. Top: only inactive compounds. Bottom: only active compounds.

Table 1 Distribution of the activity values of the inhibitors of DNMT1, 3A and 3B considered in this work. The percentage is relative to the total elements in each data set.

Library	Size (n)[a]	n (IC_{50})[a]	n (%)[a]	n (class 4)[b]	n (class 3)	n (class 2)	n (class 1)
DNMT1	350	350	-	40 (11.5%)	157 (45%)	106 (30%)	47 (13.5%)
DNMT3A	190	35	155	28 (15%)	24 (12.5%)	42 (22%)	96 (50.5%)
DNMT3B	86	61	25	17 (20%)	8 (9%)	23 (27%)	38 (44%)

[a] Size, total number of compounds; n(IC_{50}), number of compounds with IC_{50} values; n (%), number of compounds with activity data as percentage.

[b] Activity classes: Class 4 if pIC_{50} was larger than 5.5, or inhibition value was larger than 75%; Class 3 if pIC_{50} was larger than 5, or inhibition value was larger than 50%; Class 2 if pIC_{50} was larger than 4, or inhibition value was larger than 25%, and Class 1 if pIC_{50} was lower than 4, or inhibition value was lower than 25%.

Table 2 Statistics of pairwise similarity distributions computed with three fingerprints and the Tanimoto coefficient.[a]

	DNMT1			DNMT3A			DNMT3B		
	MACCS	PubChem	ECFP4	MACCS	PubChem	ECFP4	MACCS	PubChem	ECFP4
Min.	0.00	0.04	0.00	0.12	0.12	0.03	0.10	0.10	0.00
1st Qu.	0.34	0.35	0.09	0.35	0.42	0.11	0.29	0.26	0.09
Median	0.43	0.44	0.11	0.44	0.49	0.15	0.54	0.42	0.14
Mean	0.43	0.44	0.13	0.46	0.52	0.19	0.57	0.50	0.25
3rd Qu.	0.52	0.54	0.14	0.51	0.61	0.19	0.86	0.74	0.42
Max.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SD.	0.14	0.15	0.08	0.16	0.16	0.15	0.27	0.28	0.22

[a] Qu, quartile; SD, standard deviation.

Table 3 Summary table for metrics of scaffold diversity of each DNMT.

Library	Set	M	N	Nsing	N/M	Nsing/M	Nsing/N	*f*50	AUC	Median(ECFP4)
DNMT1	Inactive	127	107	93	0.84	0.73	0.87	0.411	0.570	0.1014
	Intermediate	179	123	99	0.69	0.55	0.80	0.276	0.640	0.1053
	Active	44	28	22	0.64	0.50	0.79	0.214	0.656	0.0978
DNMT3A	Inactive	138	68	51	0.49	0.37	0.75	0.117	0.728	0.1465
	Intermediate	22	15	10	0.68	0.45	0.67	0.333	0.618	0.2115
	Active	30	18	16	0.60	0.53	0.89	0.167	0.681	0.2187
DNMT3B	Inactive	39	21	16	0.54	0.41	0.76	0.190	0.703	0.0750
	Intermediate	37	10	6	0.27	0.16	0.60	0.200	0.701	0.1379
	Active	10	7	5	0.70	0.50	0.71	0.286	0.614	0.3793

N: number of cyclic systems; M: number of molecules; Nsing: number of singletons; *f*50: fraction of cyclic systems that contains 50% of the data set; AUC: area under the curve

DNMT1 inhibitors being the most diverse. When analyzing only the most active compounds (**Figure 10**) - compounds with activity index equal or greater than 3 – the three active subsets appear to cluster in different regions of the chemical space.

Figure 11 shows a visualization of the chemical space based on SOM. In this plot, inactive compounds in the three sets tend to span over the map. However, when showing only the active compounds (active index equal or greater than 3), it shows that active compounds of DNMT3A and DNMT3B are not covering the same chemical space.

Conclusions

A global cheminformatic comparison of three data sets of inhibitors of DNMT1, DNMT3A and DNMT3B is reported in this work. Analysis of physicochemical properties and molecular diversity based on fingerprints showed that inhibitors of DNMT1 cover broader areas of the chemical space. In contrast, DNMT3A and DNMT3B cover smaller areas. Analysis with topological descriptors revealed that inhibitors of DNMT3A have larger volume than inhibitors of DNMT1 and 3B. Inhibitors of DNMT3B also had higher values of fraction of sp^3 atoms than the other sets. Visual representation of the chemical space revealed that all three sets of inhibitors of DNMT1, 3A and 3B share a common space. Scaffold diversity analysis indicated that inhibitors of DNMT1 and DNMT3A have a larger number of molecular scaffolds as compared to DNMT3B. For all three data sets, there are molecular scaffolds enriched with active molecules representing promising starting points for drug development.

Acknowledgments

This work was supported by the Universidad Nacional Autónoma de México (UNAM) grant 'Programa de Apoyo a Proyectos de Investigación e Inovación Tecnológica' (PAPIIT) IA204016. We also thank funding from the 'Programa de Apoyo a la Investigación y el Posgrado' (PAIP) 50009163, Facultad de Química, UNAM. OP-H is grateful to CONACyT for the fellowships granted No. 765897/606003.

References

1 Robertson KD (2001) DNA methylation, methyltransferases, and cancer. Oncogene 20: 3139-3155.

2 Zhang W, Xu J (2017) DNA methyltransferases and their roles in tumorigenesis. Biomarker Research 5: 1.

3 Liu K, Liu Y, Lau JL, Min J (2015) Epigenetic targets and drug discovery Part 2: Histone demethylation and DNA methylation. Pharmacol Ther 151: 121-140.

4 Dueñas-González A, Jesús Naveja J, Medina-Franco JL (2016) Introduction of epigenetic targets in drug discovery and current status of epi-drugs and epi-probes: Epi-Informatics. Boston: Academic Press, pp: 1-20.

5 Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74: 481-514.

6 Valente S, Liu YW, Schnekenburger M, Zwergel C, Cosconati S, et al. (2014) Selective non-nucleoside inhibitors of human DNA methyltransferases active in cancer including in cancer stem cells. J Med Chem 57: 701-713.

7 Jurkowska RZ, Jurkowski TP, Jeltsch A (2011) Structure and function of mammalian DNA methyltransferases. ChemBioChem 12: 206-222.

8 Xu P, Hu G, Luo C, Liang Z (2016) DNA methyltransferase inhibitors: an updated patent review. Expert Opin Ther Pat 26: 1017-1030.

9 Hollenbach PW, Nguyen AN, Brady H, Williams M, Ning Y, et al. (2010) A comparison of azacitidine and decitabine activities in acute myeloid leukemia cell lines. PLoS One 5: e9001.

10 Palomino-Hernandez O, Jardinez-Vera A, Medina-Franco J (2017) Progress on the computational development of epigenetic modulators of dna methyltransferases 3A and 3B. J Mex Chem Soc 61: 266- 272.

11 Fernandez-de Gortari E, Medina-Franco JL (2015) Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. RSC Adv 5: 87465-87476.

12 Prieto-Martinez FD, Gortari EF, Mendez-Lucio O, Medina-Franco JL (2016) A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. RSC Adv 6: 56225-56239.

13 Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, et al. (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42: D1083-D1090.

14 Liu TQ, Lin YM, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucl Acids Res 35: D198-D201.

15 Huang Z, Jiang H, Liu X, Chen Y, Wong J, et al. (2012) HEMD: An integrated tool of human epigenetic enzymes and chemical modulators for therapeutics. PLoS One 7: e39917.

16 Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50: 1189-1204.

17 Molecular Operating Environment (MOE) Version 2014.08, Chemical Computing Group Inc., Montreal, Quebec, Canada.

18 R Development Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing V, Austria.

19 RStudio Team (2016) RStudio: Integrated development environment for R. RStudio I, Boston, MA, USA.

20 Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Delivery Rev 23: 3-25.

21 Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, et al. (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45: 2615-2623.

22 Medina-Franco JL, Maggiora GM (2013) Molecular similarity analysis: Chemoinformatics for drug Discovery. John Wiley & Sons, Inc. pp: 343-399.

23 Medina-Franco JL, Maggiora GM, Giulianotti MA, Pinilla C, Houghten RA (2007) A similarity-based data- fusion approach to the visual characterization and comparison of compound databases. Chem Biol Drug Des 70: 393-412.

24 Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39: 2887-2893.

25 Medina-Franco JL, Petit J, Maggiora GM (2006) Hierarchical strategy for identifying active chemotype classes in compound databases. Chem Biol Drug Des 67: 395-408.

26 Medina-Franco JL, Martínez-Mayorga K, Bender A, Scior T (2009) Scaffold diversity analysis of compound data sets using an entropy-based measure. QSAR Comb Sci 28: 1551-1560.