UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

CARACTERIZACIÓN DE LA DIVERSIDAD MOLECULAR, COBERTURA DEL ESPACIO QUÍMICO Y RECONOCIMIENTO MOLECULAR DE INHIBIDORES DE DNA METILTRANSFERASAS

TESIS

PARA OPTAR POR EL GRADO DE

DOCTOR EN CIENCIAS

PRESENTA

M. en C. ELI ANTONIO ALONSO FERNÁNDEZ DE GORTARI

DR. JOSÉ LUIS MEDINA FRANCO
Departamento de Farmacia, Facultad de Química, UNAM.

Facultad de Química, Diciembre de 2016

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

CARACTERIZACIÓN DE LA DIVERSIDAD MOLECULAR, COBERTURA DEL ESPACIO QUÍMICO Y RECONOCIMIENTO MOLECULAR DE INHIBIDORES DE DNA METILTRANSFERASAS

T E S I S
PARA OPTAR POR EL GRADO DE

DOCTOR EN CIENCIAS

P R E S E N T A

M. en C. ELI ANTONIO ALONSO FERNÁNDEZ DE GORTARI

DR. JOSÉ LUIS MEDINA FRANCO
Departamento de Farmacia, Facultad de Química, UNAM.

México, D. F. 2016

A mí querida madre. "Habremos de ser lo que hagamos, con aquello que hicieron de nosotros" (J.P. Sartre), "¡No llores!, siempre permanece en pie a pesar de todo" (J.A. Fernández)

Agradecimientos

Quiero agradecer especialmente el apoyo y dirección del Dr. Medina Franco para realizar este trabajo. Su incansable labor, ética de trabajo y su gran entendimiento y experiencia sobre nuestra rama, han sido el mejor aliado y ejemplo en mi desarrollo académico y personal durante mis estudios doctorales.

Mi más sincero agradecimiento al Dr. Castillo Bocanegra por todo el apoyo en los primeros pasos del proyecto. Sin él nada de ello me hubiera sido posible.

Al Dr. Ramón Garduño y al Dr. Fernando Cortés, ambos integrantes de mi Comité Tutor por sus valiosas observaciones y críticas.

También quiero agradecer profundamente a mis padres por todo el cariño y apoyo que me dieron a lo largo de la realización de este proyecto, a mi hermano por su insidiosa intelectualidad y a mis tíos por todo el apoyo que me han dado.

Un profundo reconocimiento a todos los integrantes del laboratorio 122 y equipo DIFACQUIM del departamento de Farmacia, por su entrega y su experiencia. En particular a las ideas y conocimientos aportados del M. en C. Aguayo Ortiz.

Agradezco también a todos los amigos y colegas que he conocido durante este tiempo, todos ellos han sido una gran fuente de inspiración y conocimiento. En particular quiero agradecer a la Q.F.B. Mariana Díaz por su incondicional apoyo y su cariñosa compañía durante todo mi proceso de formación doctoral.

Agradezco a la Universidad Nacional Autónoma de México y a la Facultad de Química UNAM por todos estos años de formación. Sin este insuperable espacio para los mexicanos nada de esto sería posible.

Agradezco al Posgrado de Ciencias Químicas de la UNAM por todo el apoyo y guía durante mi estancia en el doctorado.

Agradezco finalmente al Consejo Nacional de Ciencia y Tecnología (CONACyT) y a la dirección general de Posgrados por la beca otorgada para realizar mis estudios de doctorado con número No. 348291/240072, sin la cual no hubiera sido posible su realización.

Índice

1. Introducción 1

2. Antecedentes 2

2.1. DNMT 5

2.1.1. DNMT1: función y estructura 6

2.1.2. DNMT1: inhibidores 8

2.2 Métodos computacionales 11

2.2.1. Representaciones y similitud molecular 12

2.2.2. Tamizado molecular 21

2.2.2.1. Acoplamiento molecular 21

2.2.2.2. Modelo del farmacóforo 27

2.2.3. Identificación de sitios de unión 29

2.2.4. Entropía de Shannon 33

3. Objetivos 35

Capítulo 1. Quimioinformática 37

Metodología computacional 37

Resultados y discusión 44

Resumen de resultados 53

Capítulo 2. Farmacóforo 55

Metodología computacional 55

Resultados y discusión 56

Resumen de resultados 69

Capítulo 3. Identificación de sitios de unión de DNMT1

Metodología computacional 71

Resultados y discusión 72

Resumen de resultados 85

Capítulo 4. Huellas digitales de bases de datos moleculares (*Database Fingerprints,* DFP) 87

Metodología computacional 87

Resultados y discusión 89

Resumen de resultados 105

Capítulo 5. Estudios de acoplamiento molecular *a posteriori* de inhibidores de DNMT1 y DNMT3A 106

Metodología computacional 106

Resultados y discusión 107

Resumen de resultados 118

Conclusiones generales 120

Referencias 120

Anexo 130

Parte de los resultados de este trabajo se publicaron en:

- Gortari, E. and Medina-Franco, J. (2015). Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA-methyltransferases. *RSC Adv.*, 5(106), 87465-87476.
- Prieto-Martínez FD, Peña-Castillo A, Méndez-Lucio O, Fernández-de Gortari E, Medina-Franco JL. Molecular Modeling and Chemoinformatics to Advance the Development of Modulators of Epigenetic Targets: A Focus on DNA Methyltransferases. Adv Protein Chem Struct Biol. 2016; 105:1-26.
- Garella, D., Atlante, S., Borretto, E., Cocco, M., Giorgis, M., Costale, A., Stevanato, L., Miglio, G., Cencioni, C., Fernández-de Gortari, E., Medina-Franco, J. L., Spallotta, F., Gaetano, C. and Bertinaria, M. (2016), Design and synthesis of N-benzoyl amino acid derivatives as DNA methylation inhibitors. Chem Biol Drug Des, 88: 664–676.
- "Overview of Computer-Aided Drug Design for Epigenetic Targets", Rodrigo Aguayo-Ortiz & Eli Antonio Alonso Fernández-de Gortari. "Epi-informatics: Discovery and Development of Small Molecule Epigenetic Drugs and Probes Using Computational Approaches" Medina-Franco JL (Ed.). Elsevier (2016), ISBN: 978-0-12-802808-7**.**
- Fernando D. Prieto-Martínez, Eli Fernández-de Gortari, Oscar Méndez-Lucio, José L. Medina-Franco A Chemical Space Odyssey of Inhibitors of Histone Deacetylases and Bromodomains, RSC Adv, 2016,6,61, 56225-56239.
- Medina-Franco, JL.; Fernández-de Gortari, E.; Naveja, JJ. "DIFAC: Diseño de Fármacos por Computadora", *Educación Química*, 2015, 26, 180–186.
- "Developmental DNA methyltransferase inhibitors in the treatment of gynecologic cancer" junto con el Dr. Dueñas del Instituto de investigaciones Biomédicas, UNAM, del Instituto nacional de Cancerología. Dueñas-Gonzalez, A.; Medina-Franco, J. L.; Chavez-Blanco, A.; Dominguez-Gomez, G.; Fernández-de Gortari, E. Expert Opinion on Pharmacotherapy 2015, 17, 323–338.

Los avances del proyecto se presentaron en:

Póster del trabajo con título "Modelado del farmacóforo de inhibidores no nucleosídicos de DNA metiltransferasas basado en análisis quimioinformáticos" expuesto en Simposio en Química Medicinal y Farmacéutica, Congreso Estudiantil de Ciencia Sin Fronteras y Reunión Iberoamericana de la red CYTED en Nano vacunas para VIH. Realizado en la Escuela Superior de Medicina, IPN, en conjunto con el Centro de Tecnología Genómica y Red Iberoamericana CYYTED.

CONGRESO: (Eli Fernández-de Gortari & José Luis Medina-Franco, “Relevant epigenetic chemical space: a chemoinformatic characterization of small molecules – DNMT inhibitors”, The International Chemical Congress of PACIFIC BASIN SOCIETIES 2015,Honolulu, Hawaii, USA, Diciembre 15-20, 2015.

CONGRESO: (Eli Fernández-de Gortari & José Luis Medina-Franco, “ChemoinformaticBasedPharmacophoreModeling of Non-nucleoside Inhibitors of DNA methyltransferase1”, 251 ACS Meeting San Diego, California, USA, presentación de trabajo en sección de carteles. 14-19 Marzo, 2016.

Lista de Figuras

Figura A.1. 1: 5-Azacitidina R=OH, 1: Decitabina R=H, 2: Perthenolide, 3: Curcumina, 4: Nanaomicina A, 5: Epigallocateqina, 6: isoxazolina, 7: Procaína X=O, 7: Procainamida X=NH, 8: RG108, 9: NSC319745, 10: SGI-1027, 11: NSC14778, 12: NSC137546.

Figura A.2. Mecanismo de acción de los inhibidores nucleosídicos de DNMT1.

Figura A3. Algunos inhibidores representativos de DNMT1 clasificados según su fuente de obtención. A: Aprobados para uso clínico, B: Productos naturales, C: Reposicionamiento, D: Compuestos sintéticos producto de programas de optimización, E: Tamizado molecular de alta eficiencia.

Figura A.4. Representación molecular bidimensional de la cafeína.

Figura A.5.Reducción del espacio de propiedades a dos y tres dimensiones por medio de PCA.

Figura A.6.Representación conceptual de los métodos de representación 2D basados en diccionario.

Figura A.7.Representación conceptual de los métodos de representación 2D *extended connectivity*.

Figura A.8. Moléculas con el mismo quimiotipo según el núcleo base que comparten.

Figura A.9. Índice de Tanimoto.

Figura A.10. Archivo de parámetros de AutoDock4.

Figura A.11.Ecuación de Coulomb para interacciones electrostáticas de AutoDock4.

Figura A.12. Potencial de Lennard-Jones 12-6 de AutoDockLigand

Figura A.13. Ecuación para la energía torsional de AutoDock4, N es el número de enlaces rotables.

Figura A.14. Acoplamiento molecular de NSC319745 en sitio del cofactor de DNMT3A mediante el programa ICM de MolSoft.

Figura A.15. *Molecular grid* de un sitio presente en cruzaína (PDB ID: 3KKU). En rojo interacciones electrostáticas negativas, en blanco positivas, en otro color de naturaleza hidrofóbica.

Figura A.16. Alineamiento de proteasas de *Homo Sapiens* y la estructura de Cruzaína de *Tripanosoma Cruzi*. Alineamiento realizado en programa pymol académico.

Figura A.17. Entropía de Shannon.

Figura O.1. Diagrama de metodología general para lograr los objetivos.

Figura 1.1. Curvas del gráfico *Cyclic System Retrieval*.

Figura 1.2. Expresiones para obtener el factor de enriquecimiento.

Figura 1.3. Regiones del diagrama de frecuencia de núcleos base contra factor de enriquecimiento.

Figura 1.4. Valores estadísticos y representación mediante *box notch plots* de algunas de las propiedades y descriptores calculadas para los compuestos presentes en las bases de datos estudiadas.

Figura 1.5. Espacio de propiedades respecto a dos componentes principales.

Figura 1.6. Funciones acumulativas para los *fingerprints* MACCS keys y TGD.

Figura 1.7. Frecuencia del quimiotipo contra factor de enriquecimiento de los núcleos base más frecuentes en la base de datos de inhibidores de DNMT1.

Figura 2.1. Estructura cristalográfica de la enzima DNMT1.

Figura 2.2. Expresiones para especificidad (*true negative rate*) y sensibilidad (*true positive rate*) de los dos parámetros de ROC. P: positivos, N: negativos, PV: positivos verdaderos, FN: falsos negativos, FV: falsos verdaderos, FP: falsos positivos.

Figura 2.3. Metodología seguida para obtener el modelo del farmacóforo.

Figura 2.4. Ejemplo de conformación obtenida por acoplamiento molecular de un inhibidor no nucleosídico en el sitio catalítico de DNMT1.

Figura 2.5. Diagrama de frecuencias y código de barras representativos de PLIF.

Figura 2.6. Farmacóforo obtenido utilizando el quimiotipo RNDWX con el esquema PPCH_ALL. Las esferas verdes representan regiones hidrofóbicas planas, mientras que la esfera azul representa un aceptor de puente de hidrógeno.

Figura 2.7, Espacio ROC (*Receiver Operating Characteristic)*.

Figura 2.8. Farmacóforo obtenido utilizando el quimiotipo SU70D con el esquema PPCH_ALL. La esfera verde claro representa una región hidrofóbica plana, mientras que las esferas azules representan aceptores de puente de hidrógeno y la esfera verde una región hidrofóbica.

Figura 2.9. Compuesto **1**, **22** y **24**.

Figura 3.1. Diagrama de flujo para la campaña de tamizado virtual basado en estructura.

Figura 3.2. Sitios de unión de DNMT1 encontrados por AutoLigand. En rojo interacciones electrostáticas negativas, en blanco positivas y en otros colores hidrofóbicas.

Figura 3.3.Sitios identificados por PARS. Violeta: sitio de cofactor, Azul: sitio activo, Amarillo: sitios sin interés respecto a su grado de conservación y relación con la flexibilidad, Rojo: sitios poco conservados y con gran influencia en la flexibilidad.

Figura 3.4. Traducción de NMA a factor B para cada uno de los sitios identificados por PARS.

Figura 3.5. Matriz de conectividad entre sitios identificados (rojo mayor interacción, azul menor) y el correspondiente modelo de DNMT1 donde se muestra los sitios interconectados de DNMT1 SPACER.

Figura 3.6. Tamizado virtual ciego visualizado mediante pymol académico.

Figura 3.7. Metodología del tamizado. Vina: AutoDock Vina, AD4: AutoDock 4.2, HITS: candidatos, consenso clustering: agrupamiento consenso, I, II, III, IV: regiones del plano de agrupamiento y score consenso.

Figura 3.8. Gráficos de resultados del tamizado.

Figura 3.9. *Hits* consenso, frecuencia de interacciones según PLIF, mapa de potencial molecular y modelo del farmacóforo obtenido para el sitio 1Z.

Figura 4.1. A) Representación esquemática de representaciones moleculares 2D basadas en diccionario. B) Representación esquemática de DFP.

Figura 4.2. Ejemplo de un mapa de calor obtenido. Base de datos de inhibidores de DNMT1.

Figura 4.3. Diagrama de flujo de la metodología.

Figura 4.4. Distribuciones de probabilidad para algunas de las bibliotecas moleculares estudiadas.

Figura 4.5. Plano de entropía de Shannon contra similitud media.

Figura 4.6. Relaciones lineales entre DFP y BD inverso para las dos métricas de corte.

Figura 4.7. Mapa de calor de ACE, ACE *decoys*, MOR y MAO.

Figura 5.1. Estrategias de modulación sobre el compuesto NSC137546. A: modulación de la porción ácida, B: Conversión de amida a amina, C: substituciones sobre el anillo bencénico.

Figura 5.2. Análogos de NSC.

Figura 5.3. Pruebas de actividad residual contra concentración del compuesto para el lisado celular con sobreexpresión selectiva de DNMT1 y DNMT3A.

Figura 5.4. Modo de unión de **22** en DNMT1 y mapa 2D de interacciones.

Figura 5.5. Modo de unión de **22** en DNMT3A y mapa 2D de interacciones.

Lista de Tablas

Tabla 1.1. Fuente de los compuestos que constituyen la base de datos de inhibidores de DNMT1 construida dentro del grupo.

Tabla 1.2. Fuente y número de compuestos para cada una de las bases de datos de referencia.

Tabla 1.3. Contribuciones de los cuatro primeros componentes principales.

Tabla 1.4. Diversidad de núcleos base para inhibidores de DNMT1.

Tabla 2.1. Matriz de confusión. PV: positivos verdaderos, FP: falsos positivos, FN: falsos negativos, NV: negativos verdaderos.

Tabla 2.2. Resultados del tamizado basado en farmacóforo.

Tabla 3.1. Código de 4 letras, proteína, función y RMSD de las proteínas representativas para cada familia en Angstroms.

Tabla 4.1. Características generales de las bases de datos estudiadas, ES^a: entropía de Shannon.

Tabla 4.2. Valores de ES y SM.

Tabla 4.3. Matriz de valores de BD.

Tabla 5.1. Habilidad de inhibición de la metilación de DNA de los compuestos (1-27) expresado como metilación residual relativa de DNA.

Tabla 5.2. Resultados de pruebas de actividad inhibitoria selectiva.

Abstract

In this work we develop a characterization of the DNMT1 inhibitors chemical space reported in public sources. To achieve this goal we implement a variety of computational techniques that include chemioinformatic and molecular modeling studies.

Trough the chemical space analysis was possible to discriminate compounds regards its biological activity and chemical structure. The best compound clusters were selected to develop a pharmacophore model

The obtained pharmacophore models were subjected to internal validation through ROC space criteria. The external validation was carried out by virtual screening of independent active compounds databases synthetized by Dr. Massimo Bertinaria laboratory. This procedure gave rise to the selection of one pharmacophore model. The pharmacophore model is based in one common non-nucleosidic scaffold of a DNMT1 inhibitor.

The inhibitors information was also analyzed by molecular docking in flexible related binding sites of DNMT1 crystallographic structure. Two non-catalytic and non-cofactor flexible related binding sites were located, as well as a group of consensus hits against DNMT1. These hits may be a starting point of experimental optimization campaigns focused in allosteric inhibitors search.

Finally, we present the theoretical bases of a new molecular representation method for focused databases (Database Fingerprint). DFP is based in relative redundancies presented in binary representations to assess the general structural pattern of molecular collections. This work demonstrates the potential of DFP as a chemical space characterization metric with promising performance in virtual screening applications.

Resumen

En este trabajo se realizó la caracterización del espacio químico de inhibidores de DNMT1 reportados en fuentes públicas. Para alcanzar este objetivo se utilizaron una serie de técnicas computacionales que abarcan estudios quimioinformáticos y de modelado molecular.

A través de la caracterización del espacio químico fue posible hacer la discriminación de compuestos respecto a su actividad biológica y basada en estructura química. Los grupos de compuestos con mayores posibilidades de ser exitosos fueron seleccionados para la realización del modelado del farmacóforo.

Los modelos obtenidos fueron sujetos a validación interna mediante el uso del espacio ROC (*Receiver Operating Characteristic*). La validación externa fue realizada por medio de cribado virtual de bases de datos independientes al desarrollo del modelo y con compuestos activos obtenidos por el laboratorio del Dr. Massimo Bertinaria. Este procedimiento dio como resultado la selección de un modelo del farmacóforo basado en la estructura de un grupo de compuestos con un núcleo base de inhibidores de DNMT1 en común.

La información de inhibidores de DNMT1 también fue analizada por medio de una metodología de tamizado molecular basada en acoplamiento molecular en sitios de unión relacionados con la flexibilidad de DNMT1. Se localizaron dos sitios de unión distintos al sitio catalítico y del cofactor con una gran influencia sobre la flexibilidad de la proteína, así como una serie de candidatos consenso activos contra DNMT1 que pueden ser la base de programas de optimización de inhibidores alostéricos de dicha enzima.

Por último, se desarrolló un nuevo método de representación molecular para bases de datos enfocadas (***D****atabase* ***F****inger****p****rint, DPF*). Esta representación se basa en el número relativo de redundancias de representaciones binarias para obtener información general sobre el patrón estructural de los inhibidores estudiados. En este trabajo se muestra el potencial de DFP para ser utilizada como métrica en la caracterización del espacio químico, con expectativas prometedoras en su aplicación en campañas de cribado virtual a gran escala.

1. Introducción

En el diseño racional de fármacos confluyen una gran cantidad de áreas relacionadas por el mismo objetivo. A lo largo del tiempo, estos esfuerzos han dado como resultado la acumulación de grandes volúmenes de información química y biológica que en nuestros días es imposible analizar sin la ayuda de metodologías sistemáticas basadas en el uso de computadoras digitales. Ello impulsó a la formación de una nueva rama del conocimiento que se encarga del manejo, almacenamiento, análisis, representación y modelado de la información química llamada quimioinformática. Su presencia en el área farmacéutica ha ido en aumento, siendo ya parte integral de los procesos de desarrollo de sustancias bioactivas dentro del sector público y privado, como una de las herramientas fundamentales utilizadas en el diseño de estrategias que satisfagan la creciente demanda de insumos relacionados con la salud pública.

La DNA metiltransferasa 1 (DNMT1) es una de principales macromoléculas relacionadas con la regulación de la expresión genética mediante la metilación de regiones específicas del DNA. La desregulación de dicho mecanismo se encuentra directamente relacionado con enfermedades como cáncer, enfermedades del sistema nervioso y cardiovascular. En los últimos años han surgido terapias basadas en la inhibición de DNMT1, sin embargo, los fármacos disponibles en la actualidad, aun cuando son efectivos, presentan una gran toxicidad, baja selectividad y biodisponibilidad limitada. Por este motivo, es de suma importancia que se den esfuerzos dirigidos a la optimización y búsqueda de estructuras químicas alternativas que sean capaces de inhibir a DNMT1 sin la presencia de los efectos adversos presentes en los fármacos aprobados.

En este trabajo se presenta el uso y desarrollo de diversas metodologías quimioinformáticas y de modelado molecular que tienen por objetivo la exploración del espacio químico y el análisis de la información molecular de inhibidores de DNMT1 para la obtención de conocimiento clave que sea de utilidad para asistir a los esfuerzos multidisciplinarios relacionados en el diseño de nuevos inhibidores de DNMT1 no nucleosídicos.

2. Antecedentes

Mientras que la información genética se encuentra codificada en los genes, los procesos como transcripción, traducción y, en general, la expresión de la información contenida en el DNA se encuentra regulada por una serie de mecanismos englobados en el llamado epigenoma. Estos mecanismos son heredados mitóticamente y no afectan el contenido informacional del DNA.[1]

En términos generales, estos mecanismos pueden ser divididos en tres grandes etapas: generadora, iniciadora y de mantenimiento. El generador en un estímulo procedente del ambiente que desata una respuesta intracelular que desemboca en la aparición del iniciador.[2] Por su parte, el iniciador traduce la señal del generador para establecer un contexto de la cromatina en un sitio en específico. Es decir, el iniciador determina el lugar del cromosoma donde se llevará a cabo la regulación epigenética. La etapa de mantenimiento, sostiene el estado de la cromatina por medio de modificación de histonas, posicionamiento de nucleosomas, metilación de DNA, entre otros. [3]

Enzimas como la histona acetiltransferasas, histona deacetilasas y DNA metiltransferasas realizan modificaciones covalentes sobre la cromatina para mantener la regulación epigenética.

Actualmente se conoce que la metilación de DNA es un mecanismo de silenciamiento de genes supresores de tumores y genes responsables de la reproducción celular en células cancerígenas, donde el patrón de metilación se ve modificado respecto a células sanas. [2] Esto ha llevado a que este mecanismo cobre gran importancia como blanco terapéutico para el desarrollo de sustancias bioactivas. Lo que es de suma importancia como parte de las terapias enfocadas al tratamiento de diversas clases de cáncer, así como otros padecimientos vinculados con la desregulación epigenética como es el caso de desórdenes metabólicos, inflamación, enfermedades del sistema nervioso central e infecciones virales.[4, 5,6,7]

La metilación de DNA en mamíferos se encuentra mediada por una familia de enzimas llamadas DNA metiltransferasas (DNMT´s). Esta familia está

constituida por las enzimas DNMT1, DNMT3A, DNMT3B. Estas enzimas mantienen el patrón de metilación de la cromatina mediante la metilación selectiva en posiciones ricas en el dinucleótido CpG (islas CpG) hemimetilado o directamente mediante la metilación *de novo. En* células cancerígenas se ha detectado un patrón general hipometilado con presencia de hipermetilación en las regiones comprendidas como islas CpG. [3]

Actualmente la 5-aza y 5-aza-2´-deoxicitidina (respectivamente Azacitidina y Decitabina), dos inhibidores nucleosídicos de las DNMT, se encuentran aprobados por la FDA (*U.S.A. Food and Drug Administration* por sus siglas en inglés) para su uso clínico en el tratamiento del síndrome mielodisplásico, leucemia mielomonocítica crónica y leucemia mieloide aguda.[8] Ver **FiguraA.1**.

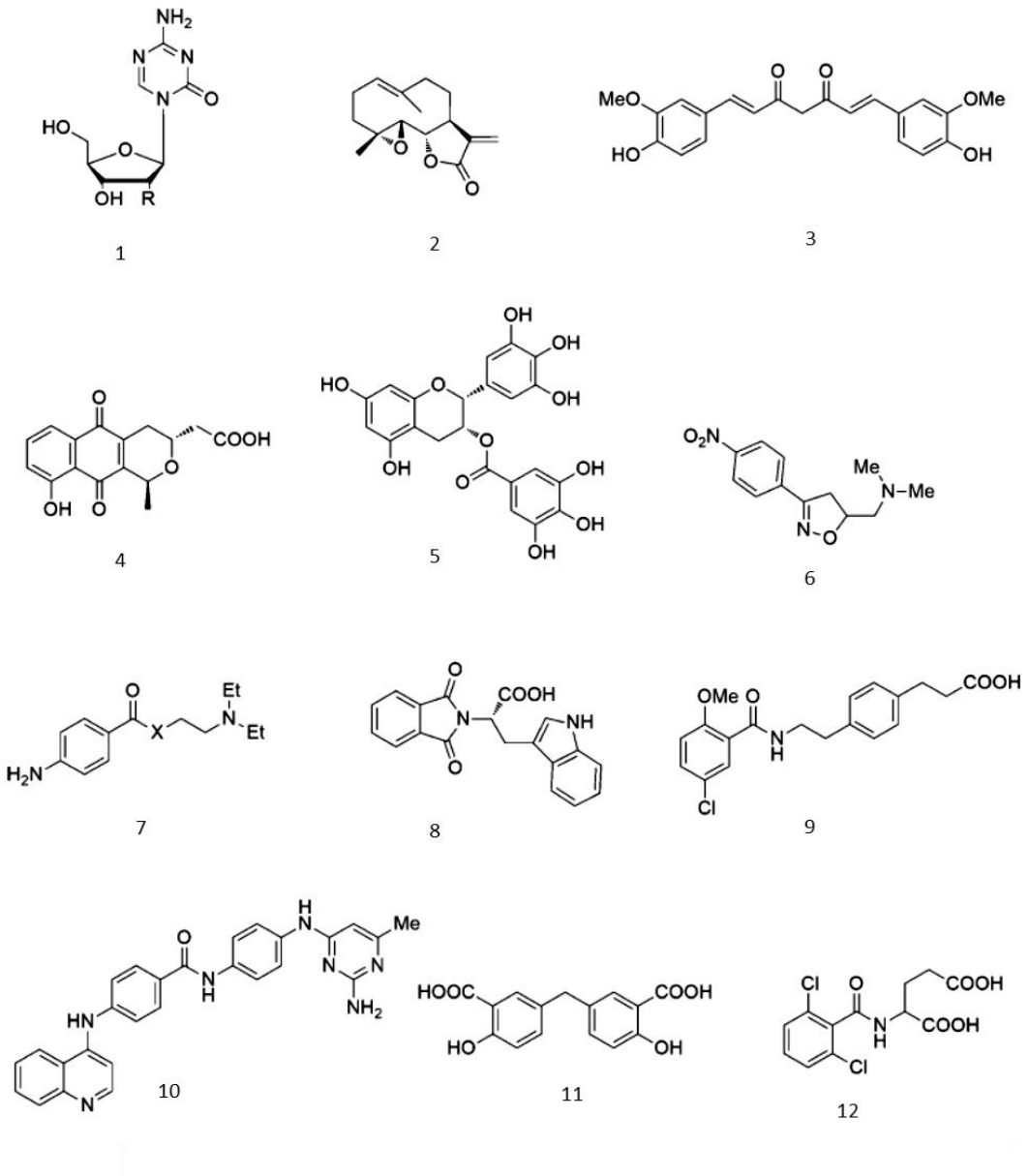


Figura. A.1. 5-Azacitidina R=OH, 1: Decitabina R=H, 2: Perthenolide, 3: Curcumina, 4: Nanaomicina A, 5: Epigallocateqina, 6: isoxazolina, 7: Procaína X=O, 7: Procainamida X=NH, 8: RG108, 9: NSC319745, 10: SGI-1027, 11: NSC14778, 12: NSC137546.[5]

Sin embargo, ambos presentan una gran toxicidad, baja selectividad y escasa estabilidad química. Esto apoya la hipótesis mecanicista que sostienen que su efecto terapéutico se da gracias a la incorporación de estas moléculas dentro de la estructura de DNA con el subsecuente agotamiento de las DNMT´s dado por la formación de un enlace covalente en el sitio catalítico. [9], [10]

Este trabajo presta especial atención a este mecanismo de mantenimiento epigenético como punto de partida para el desarrollo racional de *epiprobes* o epifármacos no nucleosídicos, por medio de metodologías computacionales enfocadas en la optimización del proceso de búsqueda de compuestos que presenten mayor selectividad y menor toxicidad como inhibidores de DNMT1 respecto a las moléculas disponibles actualmente en el mercado.

2.1. DNMT

La metilación de DNA es el mecanismo epigenético de mayor preponderancia en organismos eucariontes. Sin embargo, se ha demostrado la prevalencia de este mecanismo en distintos grupos de organismos como las bacterias o los virus[11,12]. La metilación de DNA se encuentra mediada por una familia de enzimas llamada DNMTs. En el ser humano esta familia incluye a las enzimas DNMT1, DNMT3A, DNMT3B y DNMT3L. Esta familia de enzimas facilitan la metilación de DNA en el carbono C-5 de la citosina en presencia del donador de metilo S-adenosil-L-metionina (SAM o AdoMet), dando como resultado a la base nitrogenada correspondiente en su forma metilada y la S-adenosil-L-homocisteína (SAH), dentro de regiones ricas en el dinucleótido CpG, llamadas islas CpG.[13]

DNMT1 es la DNMT más abundante en mamíferos y la primera en haber sido clonada y caracterizada bioquímicamente. Su función se relaciona con el mantenimiento del patrón de metilación en posiciones hemimetiladas de DNA, lo que estabiliza la estructura de la cromatina, silencia genes y protege la información genética de inserciones[13]. Por otro lado, DNMT3A y DNMT3B, son capaces de generar nuevos patrones de metilación en DNA desmetilado por medio de la metilación *de novo,* lo cual se encuentra estrechamente relacionado con la herencia mitótica de la información epigenética. Por último, no se ha identificado un sitio catalítico para DNMT3L, sin embargo, se sabe que actúa como regulador de la metilación *de novo* asociándose a las enzimas DNMT3A y DNMT3B.[7]

2.1.1. DNMT1: función y estructura

En los mamíferos, la metilación ocurre en el carbono C-5 de la citosina preferentemente en los dinucleótidos CpG. En general, sólo algunas de estas posiciones se encuentran metiladas, dando lugar a diferentes patrones de metilación en distintos tipos celulares y tejidos. La especificidad de las posiciones metiladas se puede ponderar si se toma en cuenta que aproximadamente del 60 al 80% de los 56 000 sitios CpG en mamíferos se encuentran metilados, lo que sólo representa del 4 al 6 % del total de citosinas presentes en el genoma humano. [11]

Esta especificidad se puede entender observando la estructura proteica de las DNMT's, las cuales cuentan con un gran multidominio N-terminal que es responsable de la localización internuclear y el reconocimiento proteína-proteína de estas enzimas. Las DNMT's también cuentan con una región C-terminal que corresponde al dominio catalítico. Este dominio contiene al sitio activo de las enzimas, que comprende diez motivos estructurales que se encuentra conservados entre las C5 DNMT's de organismos procariontes y eucariontes. En el dominio catalítico se encuentra la presencia de un sitio llamado *AdoMet-dependent Mtase fold,* formado por un conjunto de seis laminas β paralelas y siete laminas β antiparalelas intercaladas en la posición cinco y seis, que a su vez, se encuentran rodeadas de seis α hélices. [13]

Las variaciones presentes en el dominio catalítico, especialmente en una región no conservada responsable del reconocimiento del sustrato y la especificidad de la enzima (*Target Recognition Domain* TRD), son las principales causantes de las diferencias encontradas entre las distintas funciones realizadas por las DNMT's. Las diferentes funciones realizadas por las DNMT's en mamíferos, se pueden diferenciar en metilación *de novo (*DNMT3A y 3B) y mantenimiento del patrón de metilación del DNA, el cual se realiza en posiciones hemimetiladas del DNA (DNMT1). [13]

La DNMT1 humana está formada por 1616 residuos de aminoácidos. Aproximadamente tres cuartos de su secuencia forman al dominio de regulación. Dentro de esta región se encuentran los siguientes dominios:

- Dominio de reconocimiento de la unión de la proteína asociada a DNMT1 (DMAP1).
- Dominio RFTS (*Replication Foci Targeting Sequence*).
- Dominio de unión de Zinc CxxC.

El dominio DMAP es una región encargada de la unión de DNMT1 con DMAP1, lo que se ha asociado con la cosupresión de la transcripción al unirse con histona desacetilasa2 (HDAC2). La función de RFTS aún se desconoce pero se tiene la hipótesis de que puede estar relacionada con la localización de la enzima dentro del sitio de replicación durante la fase S del ciclo celular. El dominio CxxC participa en la actividad catalítica de DNMT1 al interactuar de forma específica con zonas del DNA hemimetilado.[13]

Se ha determinado que la unión proteolítica de la región N y C-terminal resulta en un aumento de la actividad metilante *de novo*. También existe evidencia que sostiene, que la interacción intermolecular entre la región CxxC del dominio de regulación y el dominio catalítico, lleva a la activación alostérica de la enzima una vez que esta se encuentra unida al sustrato metilado. [13]

La región C-terminal corresponde al dominio catalítico de la enzima. Este contiene una secuencia de trece residuos de Gly-Lys que se repite de forma alternada y que es encargada de la unión entre este dominio y el dominio de regulación.[13]

Cuando la metilación va a ser realizada, el dominio catalítico extrae a la base de la doble hélice (*base flipping*) para insertarla dentro de la cavidad catalítica junto al sitio AdoMet. La transferencia del grupo metilo se da de acuerdo a una reacción de adición de Michael. Un residuo conservado de cisteína del sitio catalítico inicia la reacción mediante un ataque nucleofílico sobre el carbono C-6 para formar un intermediario covalente entre la citosina y la enzima. Los siguientes

pasos incluyen la formación de los complejos enzima-sustrato y enzima-sustrato-cofactor lo que incluye el mecanismo del giro de la base. Enseguida se da las desprotonación del N3 endocíclico, lo que da lugar a la formación de un enlace covalente entre el grupo metilo del donador y el carbono C-5 de la citosina junto de la liberación del *S*-Adenosil L-homocisteína. El protón de la posición C-5 de citosina es extraído por algún residuo básico del sitio catalítico y la metilación es finaliza por medio de una β eliminación seguida del regreso de la base a su conformación original en la doble hélice (**Figura A.2.**). Toda la catálisis enzimática depende de la formación, acumulación y concentración de intermediarios de reacción así como de los factores que median la transferencia de protón durante el proceso de transferencia de metilo, lo que representa el paso limitante de la catálisis.[12,13]

2.1.2. DNMT1: Inhibidores

Los inhibidores de DNMT1 pueden ser divididos en dos grandes grupos: nucleosídicos y no nucleosídicos.

El primer grupo incluye compuestos que se incorporan dentro de la estructura del DNA durante la replicación. Estos compuestos actúan inhibiendo el mecanismo de acción de DNMT1 al generar un complejo covalente irreversible entre el DNA y la diana, dada la imposibilidad de la β eliminación. Los compuestos aprobados por la FDA para su uso clínico 5-aza citidina y 5-aza-2´-deoxicitidina (respectivamente Azacitidina (aprobado en 2004) y Decitabina (aprobado en 2006)) son dos análogos del nucleósido 2´-deoxicitidina sustituidos por un nitrógeno en el carbono C-5, lo que previne la disociación del complejo. La **Figura A.2.** muestra el mecanismo propuesto para la reacción de metilación sobre la citosina y la 5-azacitosina. [14]

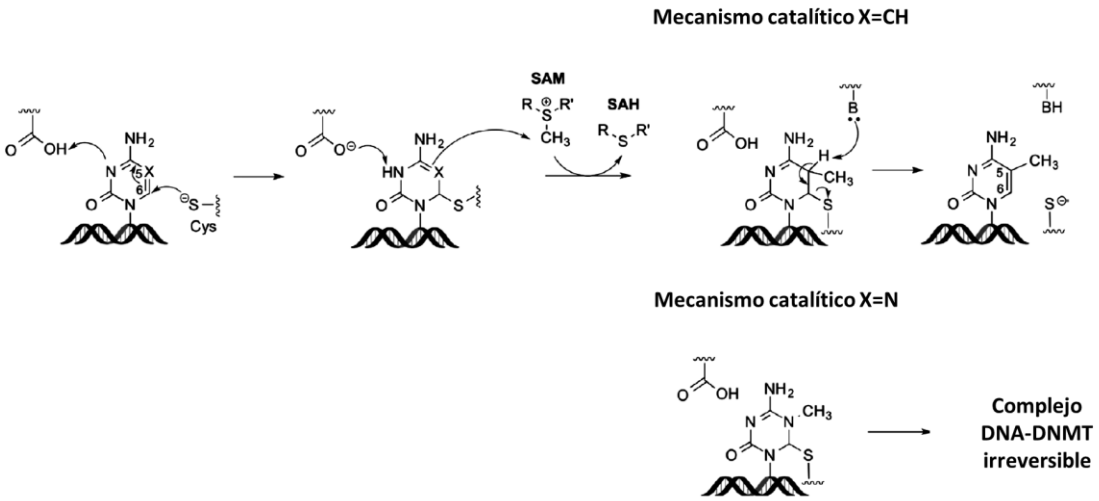


Figura A.2. Mecanismo de acción de los inhibidores nucleosídicos de DNMT1. [14]

Estos compuestos han demostrado una alta efectividad y sus efectos contra otras clases de cáncer se encuentran en fase clínica. Sin embargo, dado que su mecanismo de acción implica la integración de estas moléculas dentro de la estructura de DNA, su selectividad es muy baja. Se observan altos niveles de toxicidad dada su habilidad para integrarse en cualquier zona del DNA, además de presentar valores de vida media bajos cuando son expuestos a condiciones fisiológicas. [15]

Dadas estas condiciones, se ha vuelto cada vez más atractivo encontrar compuestos no nucleosídicos capaces de inhibir la metilación de DNA. Este conjunto de compuestos comprende una familia estructuralmente heterogénea que incluye compuestos de muy diversas fuentes: productos naturales, reposicionamiento de fármacos, productos sintéticos, producto de optimización experimental, tamizado de alto rendimiento, tamizado virtual,[16] entre otros.[17] En la **Figura A.3.** se muestran algunos ejemplos de compuestos que han demostrado actividad desmetilante por medio de la inhibición de DNMT1 clasificados de acuerdo a su fuente de obtención.

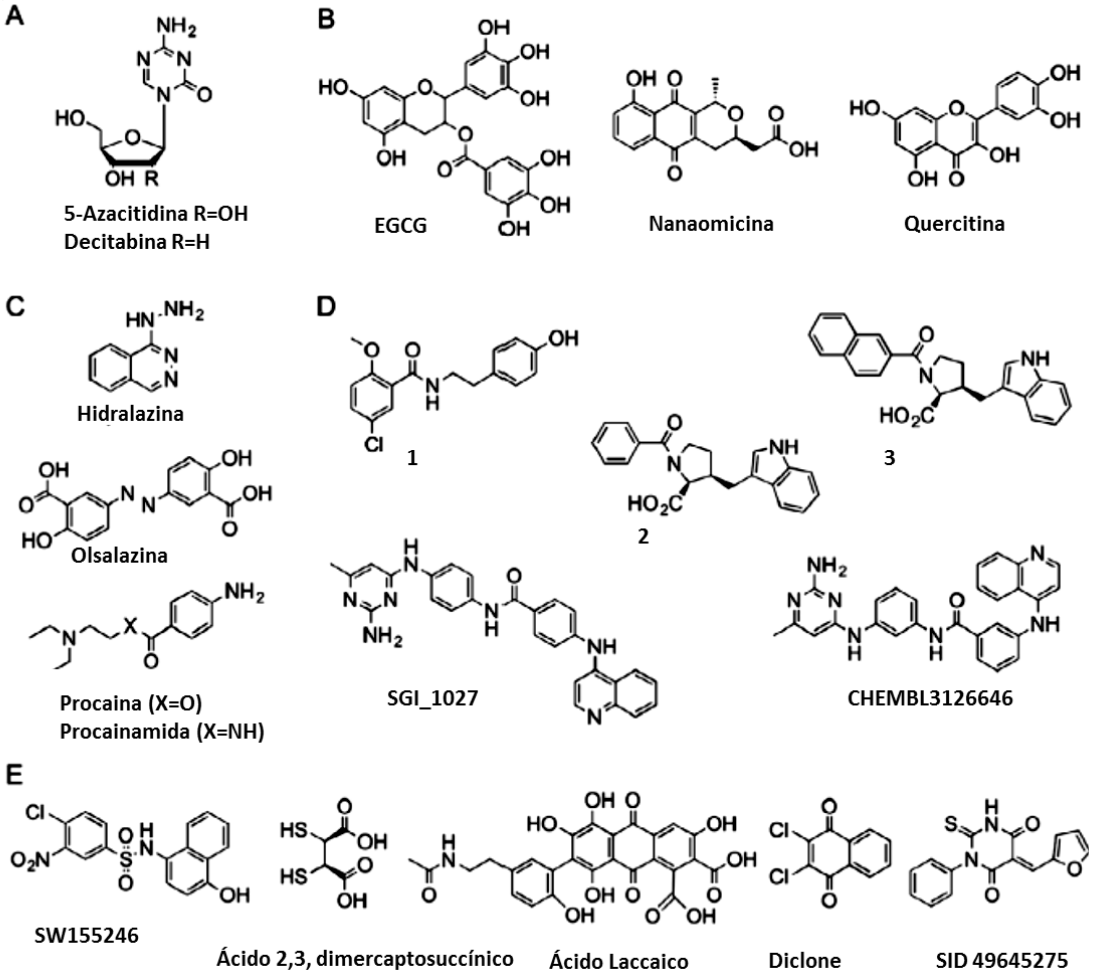


Figura A.3. Algunos inhibidores representativos de DNMT1 clasificados según su fuente de obtención.[17] A: Aprobados para uso clínico, B: Productos Naturales, C: Reposicionamiento, D: Compuestos sintéticos producto de programas de optimización, E: Tamizado molecular de alta eficiencia.

Estos compuestos prueban la existencia de moléculas capaces de inhibir a DNMT sin producir los efectos secundarios presentes en los compuestos nucleosídicos. Esto abre nuevas oportunidades en la investigación dedicada al diseño de epifármacos, que de ser exitosas se traduciría en quimioterapias menos agresivas para las personas que padecen enfermedades relacionadas con la desregulación epigenética.

2.2. Métodos computacionales

Actualmente, los métodos computacionales son utilizados dentro de instituciones académicas y privadas, para aumentar las posibilidades de éxito en las primeras fases del desarrollo y optimización de nuevos inhibidores de dianas relacionadas con la regulación epigenética. Estos métodos se pueden dividir en dos grandes grupos:

- Basados en estructura
- Basados en ligando

Los métodos basados en estructura son aquellos que son aplicados en caso de existir información experimental sobre la estructura de la diana de interés. Dichos estudios son comúnmente realizados por medio de difracción de rayos X y, en menor medida, por medio de resonancia nuclear magnética y modelado por homología [18]. Aunque es necesario recordar, que la aproximación por homología sólo es accesible en el caso de contar con información estructural de proteínas taxonómicamente relacionadas, lo que en muchos casos, pueden llevar a resultados cuestionables que deben ser posteriormente validados por medio de métodos experimentales.

Entre los métodos computacionales basados en estructura se pueden enumerar algunos de los más populares.[19]

- Acoplamiento molecular
- Dinámica molecular
- Acoplamiento ensamble
- Farmacóforo basado en estructura
- Métodos *ab initio*

Los métodos basados en ligante, generalmente utilizados en la ausencia de información estructural del blanco, hacen uso de la información de compuestos con actividades conocidas para inferir y proponer las características moleculares

generales que deben satisfacer los inhibidores prototipo que son propuestos dentro de campañas de desarrollo de sustancias bioactívas.[19] Entre los métodos basados en ligando se pueden mencionar:

- QSAR (*Quantitative Structure Activity Relationship*)
- SAS (*Structure Activity Similarity*), DAD (*Dual Activity Difference maps*) y TAT (*Triple Activity Difference maps*)
- Huellas digitales moleculares (*Molecular fingerprints*) y métricas de similitud
- Descriptores moleculares y propiedades fisicoquímicas
- Núcleos base (*scaffolds*)
- Farmacóforo basado en ligando

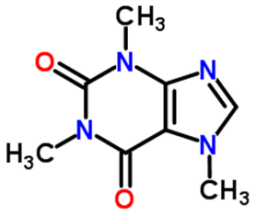
Aun cuando la cantidad de información estructural crece exponencialmente año con año, hasta el momento sigue siendo insuficiente respecto al número de biomoléculas implicadas en el metabolismo de enfermedades. Este hecho, aunado a los costos y tiempos de cálculo en los métodos estructurales, ha posicionado a los métodos basados en estructura del ligando dentro de los primeros puestos de popularidad en empresas farmacéuticas. Por su parte, los métodos basados en estructura generalmente se encuentran confinados a la investigación realizada en instituciones académicas.

En este trabajo se utilizaron metodologías que pueden ser incluidas en ambos grupos, así como algunas que se encuentran en la interface de esta clasificación.

2.2.1. Representación y similitud molecular

El concepto de similitud es ampliamente utilizado en cualquiera de las actividades abstractas que realiza el ser humano y la química no es la excepción. Dicho concepto es dependiente del observador que realiza el juicio de similitud entre dos o más entidades que pertenezcan a su foco de atención. Es decir, el concepto de similitud, incluso cuando es ampliamente usado en las distintas ramas del conocimiento humano, es subjetivo. [20]

La similitud molecular depende principalmente de dos conceptos: el modo de representación de las entidades comprendidas y el método de comparación, ambos a su vez interrelacionados. Sin reparar más en generalidades, dentro de la química se han utilizado cientos de formas de representación de los sistemas moleculares, siendo los más utilizados y representativos aquellos que se ilustran mediante grafos con pesos.[21] En la **Figura A.4.** se muestra a la molécula de cafeína mediante varios tipos de representación.



SMILES: Cn1cnc2c1c(=O)n(c(=O)n2C)C

$C_8H_{10}N_4O_2$

InChi: RYYVLZVUVIJVGH-UHFFFAOYSA-N

ChemSpider ID: 2424

Nombre sistemático1,3,7-Trimethyl-3,7-dihydro-1H-purine-2,6-dione

Nombre alternativo: 1-methyl-Theobromine

Nombre común: cafeína

Figura A.4. Representación molecular bidimensional de la cafeína.[22]

Como se puede observar en la **Figura A.4.**, todos estos métodos de representación apuntan a la misma entidad química, sin embargo, los elementos que la componen pueden diferir significativamente respecto al tipo y cantidad de información que contienen. Tanto el tipo como la cantidad de elementos incluidos en la representación dependerán de la cantidad de información con la que se cuenta, así como el objetivo de la misma. Sin tomar en cuenta las posibles relaciones de dependencia de los elementos que conforman a la representación, se puede decir que una condición necesaria para que esta sea efectiva es que contenga aquellos elementos susceptibles al análisis del sistema en cuestión.

A partir de la revolución informática que se dio algunas décadas atrás, iniciaron los esfuerzos por representar compuestos químicos de modo que fueran interpretables por computadoras. Este hecho daba la posibilidad de analizar grandes cantidades de información química en tiempos asequibles. Tomando en

cuenta que el espacio químico de moléculas pequeñas (30 átomos, C, H, O, N, F, tamaño promedio de una molécula *druglike*) se aproxima a un orden de 10^{30}, [23] la capacidad de procesamiento de la información molecular por medios computacionales se vuelve una necesidad de primer orden.

Los métodos de representación computacionales se pueden clasificar de la siguiente manera: unidimensionales (1D), bidimensionales (2D) y tridimensionales (3D). Entre los métodos de representación 1D se pueden encontrar propiedades macroscópicas y descriptores moleculares codificados con números continuos o discretos. Ejemplo de ello son las propiedades fisicoquímicas como punto de fusión, logP, PSA, TPSA, peso molecular, entre otras. Estas propiedades pueden ser obtenidas directamente de la experimentación o por medio de métodos computacionales que varían en su grado de exactitud respecto a la teoría en la cual se sustentan. En cuanto a los descriptores moleculares encontramos características ligadas con sistemas moleculares discretos, como es el caso de número de aceptores de hidrógeno, número de donadores de hidrógeno, enlaces rotables, número de enlaces sp^3, número de sistemas cíclicos o aromáticos, etc. [24]

Ambos métodos de representación permiten la visualización del espacio químico de propiedades y descriptores. Dicho espacio no tiene un punto de referencia fijo, por lo que es común utilizar bases de datos de compuestos, fragmentos o moléculas ampliamente caracterizadas como estándar para situarse dentro del mismo. Ya que generalmente esta clase de estudios utiliza grandes cantidades de información molecular con el propósito de correlacionar a los elementos utilizados para describir al sistema, el uso de métodos estadísticos y de ciencia de datos es una herramienta imprescindible para su estudio. Entre los métodos más utilizados para estudiar este tipo de sistemas multidimensionales se encuentran el aprendizaje de máquina, el agrupamiento (*clustering*), el análisis de componentes principales, la estadística inferencial, etc. [25]

Para ejemplificar estos conceptos, se puede tomar un conjunto de compuestos y obtener un n número de propiedades y/o descriptores, con lo que el tamaño de dicho espacio será de n dimensiones. Ya que no nos es posible visualizar espacios superiores a tres dimensiones, es común que se apliquen

métodos de reducción dimensional con el propósito de representar de forma gráfica al espacio de propiedades para así visualizar el comportamiento colectivo de los elementos del sistema. En la **Figura A.5.** se muestra la representación 2D y 3D de un grupo de moléculas que pertenecen a un espacio de propiedades de seis dimensiones mediante la reducción dimensional realizada por medio de análisis de los componentes principales.

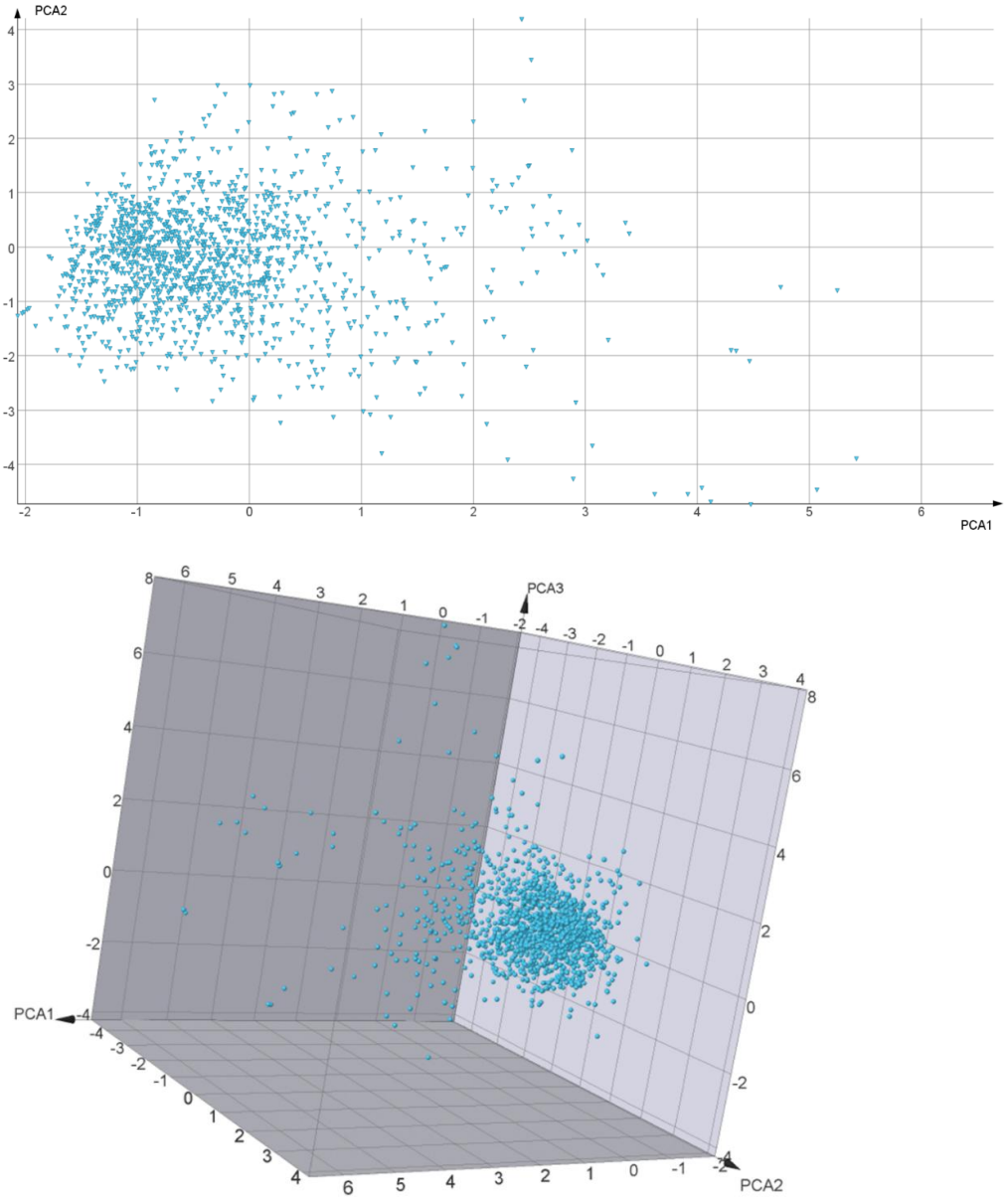


Figura A.5.Reducción del espacio de propiedades a dos y tres dimensiones por medio de PCA.

Este método realiza la reducción dimensional por medio del cálculo de los valores y vectores propios de la matriz de covarianza o matriz de coeficientes de correlación, la cual al ser simétrica contienen una base completa. Dichos vectores son tomados como nuevos sistemas de coordenadas. El número de dimensiones a reducir dependerá de la capacidad de cada uno de estos vectores para capturar la varianza de la distribución y, por lo tanto, de la dispersión de los datos en el espacio n dimensional que ocupan.

Los métodos de representación 2D, comúnmente conocidos como huellas digitales moleculares o *fingerprints*, hacen uso de notación lineal para almacenar información molecular. Uno de los métodos de representación 2D más utilizados son aquellos que se caracterizan por codificar sus componentes por medio de notación binaria. Siendo que, valores iguales a uno indican la presencia de algún elemento y cero su ausencia. Estos métodos se subdividen a su vez en: basados en un diccionario y calculados *in situ*. Los primeros tienen un diccionario en memoria que cuenta con elementos predefinidos para cada una de las posiciones del vector. La **Figura A.6.** muestra el concepto de representación basada en un diccionario.[25]

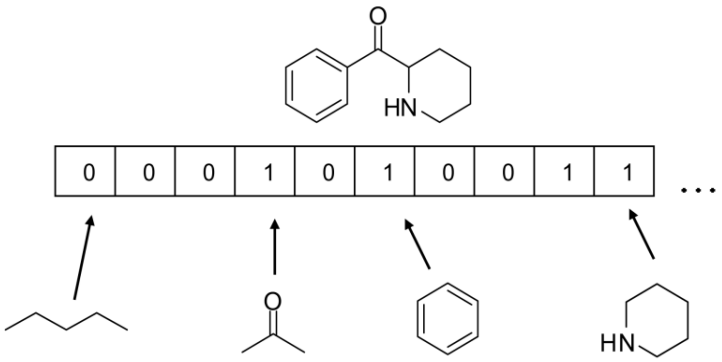


Figura A.6. Representación conceptual de los métodos de representación 2D basados en un diccionario.

Las huellas digitales binarias calculadas *in situ* o sobre la marcha, incluyen características singulares de cada molécula a representar. Entre estos métodos se pueden contar: basados en grafos, en puntos farmacofóricos y en conectividad. En

la **Figura A.7.** se muestra el concepto que se encuentra detrás del método de representación conocido como *extended connectivity*. [26]

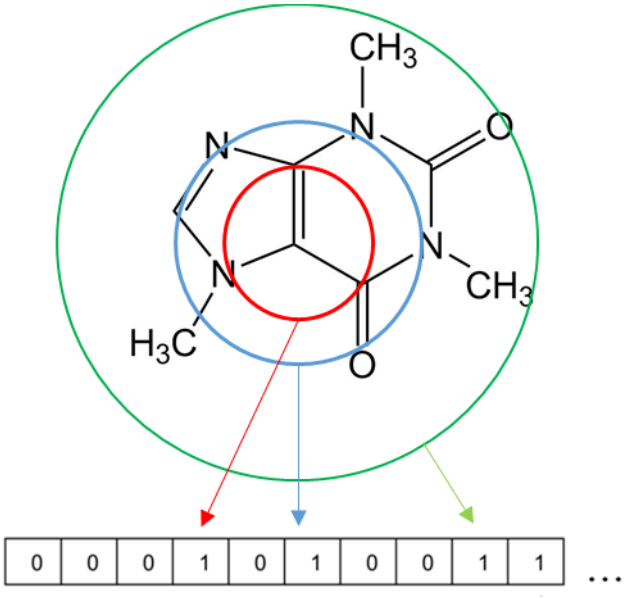


Figura A.7.Representación conceptual de los métodos de representación 2D *extended connectivity*.

Este método determina los átomos vecinos de cada uno de los átomos presentes en la molécula respecto a círculos concéntricos que difieren en diámetro.

Entre los métodos farmacofóricos se encuentra los llamados de n puntos. Estos métodos asocian distintas geometrías, dependientes del número de puntos seleccionados, a la vecindad de cada uno de los átomos de una molécula. Además de ello, pueden relacionar distintos descriptores moleculares a cada uno de los átomos identificados para generar patrones que son codificados dentro del vector resultante. [27]

Otro de los métodos de representación molecular 2D es el conocido como núcleos base o *scaffolds*. Este método retira cada uno de los elementos lineales de grafos moleculares para obtener sólo aquellos elementos de naturaleza cíclica; se retiran átomos que no pertenecen a sistemas cíclicos, a menos que participen en el enlace de dos de ellos.[28] Cada uno de los elementos encontrados se llama quimiotipos y pueden ser codificados por medio de códigos alfanuméricos. Los

quimiotipos permiten agrupar a las moléculas de un conjunto en diferentes grupos respecto al núcleo base compartido por dos o más entidades químicas. Este método de representación es de gran utilidad en estudios dirigidos a la investigación de sustancias bioactivas, ya que se sabe que diferentes elementos moleculares pueden ser intercambiados en un sistema sin perder la actividad deseada (bioisósteros), e incluso, existen estructuras de naturaleza cíclica que presentan actividad contra una gran variedad de dianas biológicas (estructuras privilegiadas). En la **Figura A.8.** se muestran ejemplos de moléculas que comparten un núcleo base y el quimiotipo asociado según la metodología utilizada por el programa MEQI. [28]

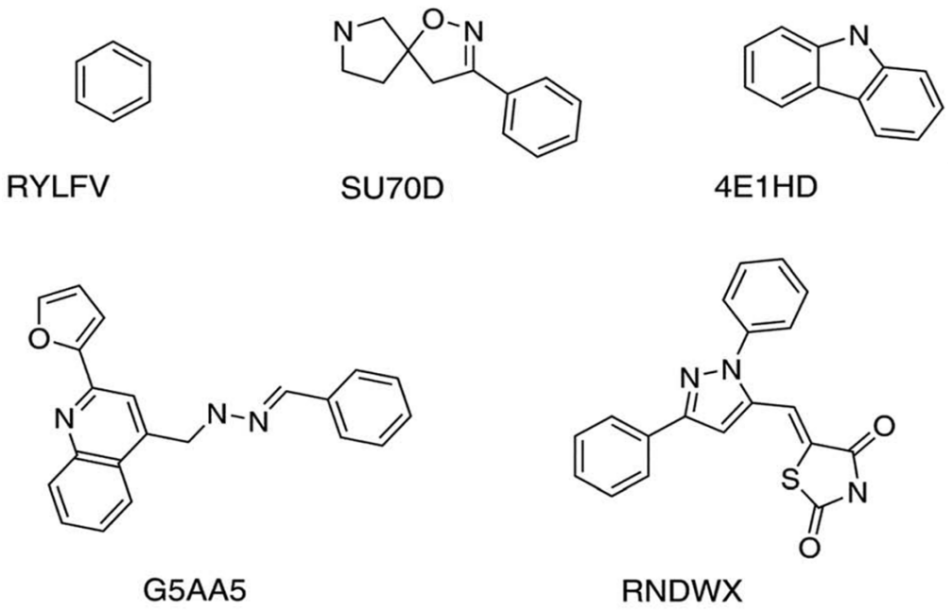


Figura A.8.Moléculas con el mismo quimiotipo según el núcleo base que comparten.[15]

Por último, los métodos de representación 3D tienen la información estructural y espacial de los elementos moleculares de los compuestos químicos. Aun cuando este método de representación parecería uno de los más completos para representar de forma adecuada a los sistemas moleculares, la dificultad que se presenta respecto a los grados de libertad del sistema han limitado en gran medida su verdadera utilidad, siendo que en muchas ocasiones, los métodos 2D

superan a estos últimos dentro de muchas aplicaciones quimioinformáticas. Este hecho se puede explicar, en parte, por la explosión combinatoria producto del posible número de conformaciones asociadas a los grados de libertad de una molécula, así como a los problemas asociados a la búsqueda de mínimos globales o de conformaciones bioactivas.

En el diseño racional de fármacos asistido por computadora estos métodos de representación son utilizados comúnmente para realizar comparaciones de dos o más moléculas. Uno de los principios centrales de esta rama del conocimiento postula que "moléculas similares tienen propiedades (actividad biológica) similares". Incluso cuando este postulado no se cumple en el 100% de los casos, años de pruebas experimentales lo validan como un primer acercamiento eficaz para el diseño de moléculas bioactivas. Dado este hecho, el concepto de similitud se coloca en uno de los lugares privilegiados de la investigación farmacéutica y, por ende, del diseño racional de fármacos.

Para que dos o más moléculas sean susceptibles de ser comparadas es necesario que se encuentren descritas por medio del mismo método de representación. Una vez que esto se encuentra satisfecho, el siguiente paso es aplicar algún tipo de métrica comparativa a cada uno de los elementos que se desea comparar. Dentro de la quimioinformática se utilizan una gran variedad de métricas de comparación que contemplan diferentes parámetros para determinar el grado de similitud entre dos entidades químicas. Entre las más simples y populares métricas de comparación se encuentra el llamado índice de Tanimoto. Este índice puede variar su forma respecto al método de representación, utilizando vectores o escalares según sea el caso, pero en esencia lo que intenta describir es la similitud de los elementos comparados por medio del número de elementos compartidos y no compartidos de dos entidades de interés. En la **Figura A.9.** se muestra la ecuación para calcular el índice de Tanimoto. [29]

$$T(a,b)=\frac{C}{A+B-C}$$

Figura A.9. Índice de Tanimoto.

Tomando dos moléculas representadas por medio de vectores binarios, el valor de C se determina como el número de posiciones binarias compartidas con valor de uno (presencia de algún elemento molecular), A como el número de elementos únicos en la molécula a y B como el número de elementos únicos en la molécula b. Los valores resultantes de este índice se encuentran en un intervalo decimal que adquiere valores entre cero y uno, donde cero representa la máxima disimilitud (S-1) y uno la máxima similitud. El espacio de valores formado por las comparaciones pareadas de las moléculas estudiadas conforma el llamado espacio de similitud. Este espacio puede ser estudiado por los mismos métodos utilizados para el estudio del espacio de propiedades, pero a diferencia de este, la dimensionalidad del sistema dependerá del número de elementos constituyentes del método de representación. Por ejemplo, MACCS keys representa a las moléculas por medio de un vector binario de 166 o 322 elementos, lo que da como resultado un espacio de 166 o 322 dimensiones respectivamente.[30] Otras métricas de comparación ampliamente utilizadas son: Euclidiana, *Block city distances,* Manhattan, índice de Dice, coeficiente coseno, distancias de Soergel, entre otras.

Entre las aplicaciones más relevantes que se pueden realizar por medio de la similitud molecular se encuentran la descripción del espacio químico y el tamizado molecular. Como se mencionó, ambas aplicaciones dependerán tanto del método de representación como de la métrica de comparación utilizada, lo que en el primer caso representa el cambio de la topografía del espacio químico y, en la segunda, el número y clase de *hits* obtenidos. [31]

Otro de los estudios más recurrentes del espacio químico utilizando similitud molecular, es la determinación de la diversidad química de bases de datos moleculares enfocadas a la búsqueda de sustancias bioactivas. En estos estudios se hace uso de uno o más métodos de representación para determinar la similitud interna de los compuestos presentes en la base de datos. Estos estudios permiten modular y/o diseñar el espacio químico de interés según el enfoque de la investigación. Por ejemplo, si lo que se desea es realizar tamizados moleculares, es importante que la porción del espacio químico estudiado sea lo más amplia posible para así aumentar las posibilidades de encontrar sustancias novedosas

contra una diana en particular. Si al contrario, lo que se desea es enriquecer una base de datos de compuestos con una actividad enfocada contra una diana biológica en particular, lo más conveniente es encontrar grupos de moléculas que difieran poco (baja diversidad) de las utilizadas como referencia.

2.2.2. Tamizado molecular

Como se mencionó en el apartado anterior, el tamizado molecular puede ser realizado por medio de similitud molecular. Sin embargo, esta técnica es más general e incluye una amplia diversidad de técnicas. En esencia, el tamizado molecular es la aplicación de cualquier método sistemático que permita distinguir y encontrar moléculas con propiedades particulares y separarlas de aquellas que no las presentan. Como en el caso de los métodos computacionales enfocados al diseño de fármacos, el tamizado molecular puede ser realizado basándose en la estructura de la diana biológica o en la estructura del ligante, como en el caso de la similitud molecular. A continuación se describen dos metodologías, acoplamiento molecular y modelo del farmacóforo, la primera de ellas totalmente basada en la información estructural, mientras que la segunda puede hacer uso de ambas fuentes de información según sea el caso.

2.2.2.1. Acoplamiento molecular

El acoplamiento molecular es una técnica de modelado molecular que hace uso de un campo de fuerza parametrizado para un grupo determinado de átomos, generalmente aquellos más frecuentes en sistemas orgánicos, y un algoritmo de optimización que puede variar de metodología a metodología.[32,33]

```
# $Id: AD4_parameters.dat,v 1.14 2007/04/27 06:01:47 garrett Exp $
#
# AutoDock
#
# Copyright (C) 1989-2007, Garrett M. Morris, David S. Goodsell, Ruth Huey, Arthur J. Olson,
# All Rights Reserved.
#
```

```
# AutoDock is a Trade Mark of The Scripps Research Institute.
#
# This program is free software; you can redistribute it and/or
# modify it under the terms of the GNU General Public License
# as published by the Free Software Foundation; either version 2
# of the License, or (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program; if not, write to the Free Software
# Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA.
# AutoDock Linear Free Energy Model Coefficients and Energetic Parameters
#           Version 1.0
#          $Revision: 1.14 $

# AutoDock 4 free energy coefficients with respect to original (AD2) energetic parameters
#
#              Free Energy Coefficient
#              ------
FE_coeff_vdW    0.1560
FE_coeff_hbond  0.0974
FE_coeff_estat  0.1465
FE_coeff_desolv 0.1159
FE_coeff_tors   0.2744

# AutoDock 4 Energy Parameters

# - Atomic solvation volumes and parameters
# - Unweighted vdW and Unweighted H-bond Well Depths#
# - Atom Types
# - Rii = sum of vdW radii of two like atoms (in Angstrom)
# - epsii = vdW well depth (in Kcal/mol)
# - vol = atomic solvation volume (in Angstrom^3)
# - solpar = atomic solvation parameter
# - Rij_hb = H-bond radius of the heteroatom in contact with a hydrogen (in Angstrom)
# - epsij_hb = well depth of H-bond (in Kcal/mol)
# - hbond = integer indicating type of H-bonding atom (0=no H-bond)
# - rec_index = initialised to -1, but later on holds count of how many of this atom type are in receptor
# - map_index = initialised to -1, but later on holds the index of the AutoGrid map
```

```
# - bond_index = used in AutoDock to detect bonds; see "mdist.h", enum {C,N,O,H,XX,P,S}
#
# - To obtain the Rij value for non H-bonding atoms, calculate the
#       arithmetic mean of the Rii values for the two atom types.
#       Rij = (Rii + Rjj) / 2
#
# - To obtain the epsij value for non H-bonding atoms, calculate the
#       geometric mean of the epsii values for the two atom types.
#       epsij = sqrt( epsii * epsjj )
#
# - Note that the Rij_hb value is non-zero for heteroatoms only, and zero for H atoms;
#       to obtain the length of an H-bond, look up Rij_hb for the heteroatom only;
#       this is combined with the Rii value for H in the receptor, in AutoGrid.
#       For example, the Rij_hb for OA-HD H-bonds will be (1.9 + 1.0) Angstrom,
#       and the weighted epsij_hb will be 5.0 kcal/mol * FE_coeff_hbond.
#
#       Atom   Rii                    Rij_hb      rec_index
#       Type       epsii        solpar       epsij_hb    map_index
#                    vol                    hbond     bond_index
#       --     ----  -----  -------  --------  ---  ---  -  --  -- --
atom_par H      2.00  0.020   0.0000   0.00051  0.0  0.0  0  -1  -1  3    # Non H-bonding Hydrogen
atom_par HD     2.00  0.020   0.0000   0.00051  0.0  0.0  2  -1  -1  3    # Donor 1 H-bond Hydrogen
atom_par HS     2.00  0.020   0.0000   0.00051  0.0  0.0  1  -1  -1  3    # Donor S Spherical Hydrogen
atom_par C      4.00  0.150  33.5103  -0.00143  0.0  0.0  0  -1  -1  0    # Non H-bonding Aliphatic Carbon…
```

Figura A.10. Archivo de parámetros de AutoDock4.[33]

El campo de fuerza contiene una lista de valores determinados de manera experimental o computacionalmente, que sirven como fuente de información para el algoritmo de optimización. Entre los elementos que se pueden encontrar en este tipo de archivos se encuentra distancias de enlace, ángulos de enlace, ángulos diedros, cargas totales o parciales, etc. El campo de fuerza incluye también ecuaciones derivadas de la teoría newtoniana que utilizan a los parámetros antes mencionados para describir a los diferentes tipos de interacciones moleculares que pueden existir entre un ligando y un receptor. Comúnmente estas ecuaciones

se pueden dividir en: interacciones electrostáticas, interacciones de Van der Waals, puentes de hidrógeno, ángulos y distancias de enlace.

El cálculo de las interacciones electrostáticas se aproxima mediante la ley de Coulomb. Esta ecuación relaciona la magnitud de la carga formal de los átomos implicados en una interacción con el inverso de la distancia que los separa, siendo posible la atracción en el caso de cargas opuestas o repulsión en el caso contrario. Ver **Figura A.11**.

$$\Delta Gelec = Welec \sum i,j \left(qi * qj \right) / \left(\varepsilon(rij) * rij \right)$$

Figura A.11.Ecuación de Coulomb para interacciones electrostáticas de AutoDock4.[32]

Las interacciones tipo Van der Waals son descritas como un potencial tipo Lennard-Jones 12-6, u otras combinaciones de exponentes. Esta ecuación da como resultado un pozo de energía con energía mínima igual a la energía de equilibrio de la interacción interatómica, lo que aspira no sólo a describir el comportamiento de las interacciones débiles sino que a su vez impide el colapso de dos átomos gracias a la repulsión presente a distancias cortas. Ver **Figura A.12**.

$$\Delta GvdW = WvdW \sum i,j \left(Aij / rij12 - Bij / rij6 \right)$$

Figura A.12. Potencial de Lennard-Jones 12-6 de AutoDockLigand.[32]

En algunas ocasiones las distancias interatómicas y ángulos de enlace covalentes se encuentran fijados a valores experimentales o calculados previamente para pares de átomos de distinta naturaleza. Sin embargo, también es posible obtener estos valores por medio de la ecuación de Hook, que es el producto de una constante asociada al muelle que conecta dos masas (átomos) y el cuadrado de la distancia o ángulo existentes entre ellas. En el caso de AutoDock4 se utiliza una

ecuación parametrizada para calcular la energía torsional asociada al cambio de estado acoplado y desacoplado.

$$\Delta Gtor = Wtor\, Ntor$$

Figura A.13. Ecuación para la energía torsional de AutoDock4, N es el número de enlaces rotables.[32]

Generalmente, en mecánica molecular, los puentes de hidrógeno se determinan según parámetros que toman en cuenta la electronegatividad de los posibles donadores y aceptores de puente de hidrógeno, así como la distancia y ángulo entre ellos. Si los requisitos contemplados en el campo se cumplen, un enlace de hidrógeno ideal se da entre un donador y aceptor de hidrógeno con grandes diferencias de electronegatividad, a una distancia de 2.5 Å y con un ángulo de 180°, el algoritmo determinara su existencia.

El siguiente elemento con que debe contar un programa de acoplamiento molecular es el algoritmo de optimización y muestreo. Este algoritmo puede estar basado en técnicas de optimización de la conformación de muestreo aleatorio o determinista. Dichos algoritmos exploran las diferentes combinaciones de valores que pueden adquirir las variables que forman el campo de fuerza para definir las interacciones ligando-receptor asociadas a la menor energía posible según una cota o línea de corte preestablecida. Todas estas metodologías de optimización aspiran a realizar exploraciones del espacio conformacional asociando una superficie de energía potencial definida por el campo de fuerza. La selección de este algoritmo puede variar según el programa que se utilice o ser definida según criterios de tiempo y costo. Algunos ejemplos de algoritmos aleatorios son: algoritmos genéticos, algoritmos Monte Carlo, recocido simulado (*simulated annealing*). Entre los métodos clásicos se encuentran algoritmos como: basados en gradiente, Newton-Raphson, método de bisección, etc. [34,35]

Ambas componentes, campo de fuerza y algoritmo de optimización, permiten realizar el acoplamiento de una o más moléculas sobre un sitio de una diana en particular. El resultado de este procedimiento será una serie de conformaciones asociadas a un valor conocido como *score* o puntuación que tipificará a las moléculas según la energía de enlace encontrada. En la **Figura 1.14.** se muestra la conformación e interacciones resultado del acoplamiento ligando-receptor.

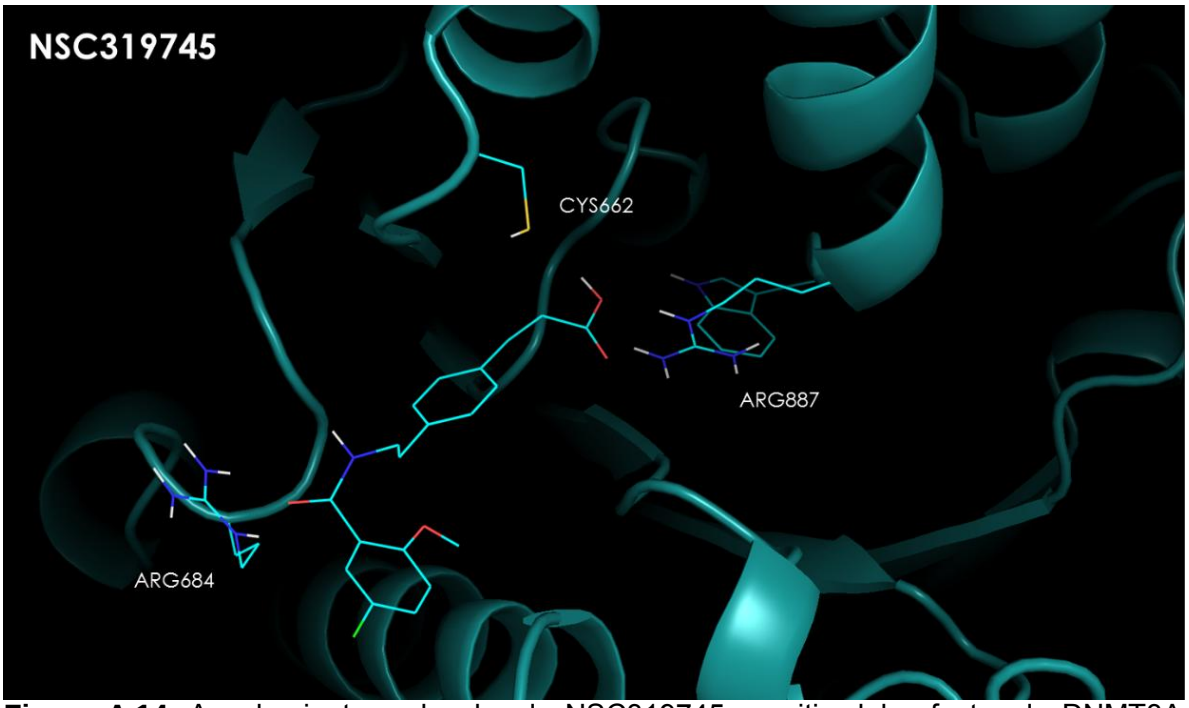


Figura A.14. Acoplamiento molecular de NSC319745 en sitio del cofactor de DNMT3A mediante el programa ICM de MolSoft. [36]

Algunos de los sesgos presentes en esta metodología computacional se enumeran a continuación:

- No se toma en cuenta el disolvente.
- No contempla factores entrópicos.
- El campo de fuerza no incluye factores importantes del comportamiento electrónico.
- Los algoritmos de optimización no garantizan la obtención de mínimos globales.

- No se puede simular la ruptura o formación de enlaces.
- No se contempla la movilidad del blanco.

Todos estos factores disminuyen la confiabilidad de esta metodología. Sin embargo, el *docking* es empleado por grupos y compañías con muchos recursos humanos y económicos. De hecho, las grandes empresas de software deben su existencia a las grandes cantidades económicas pagadas por empresas como Pfizer, Novartis, Merck, entre otras.[37] Ello es debido a que esta metodología representa una técnica de gran utilidad para la búsqueda de compuestos prometedores o para la descripción preliminar del tipo de interacciones presentes en compuestos de interés. Esto se hace más relevante cuando el volumen de información molecular es extenso, es decir, cuando se realiza la exploración de grandes zonas del espacio químico. Además de esto, los resultados obtenidos con esta técnica pueden ser tratados de tal manera que su rendimiento aumente de forma significativa, ejemplo de ello es el *clustering* de conformaciones, el uso de *scores* consenso, la restricción del espacio conformacional por medio de modelos del farmacóforo, el uso de información experimental del sistema estudiado como referencia interna, la combinación de distintas metodologías en flujos metodológicos, uso de residuos flexibles, uso de filtros o el uso de diferentes conformaciones del receptor, entre otras.[38]

2.2.2.2. Modelo del farmacóforo

El farmacóforo se encuentra definido como: “el conjunto de elementos estéricos y electrostáticos que garantizan la interacción supramolecular óptima con una diana biológica particular y así desatar o bloquear su respuesta biológica”. [39]

Los modelos computacionales del farmacóforo pueden basarse en información estructural de la diana, de ligantes o de ambas. Independientemente de la fuente de información, los diferentes modelos intentan localizar las regularidades o patrones formados por los elementos moleculares que constituyen el sistema. Una vez más, estos elementos dependerán del tipo de información

utilizado para representarlos así como de las posibles conformaciones accesibles para el sistema.

Los modelos del farmacóforo basados únicamente en la estructura del ligando hacen uso de algoritmos de mapeo que contemplan un número determinado de partículas sonda para la generación de *grids* o mallas moleculares. Esta técnica trata de determinar el potencial molecular formado por una serie de puntos separados por una distancia determinada. De esta manera, el resultado es un campo molecular donde se definen diferentes regiones respecto a un campo de fuerza preestablecido. En la **Figura A.15.** se muestra el resultado del mapeo de un sitio de unión proteico de acuerdo a la metodología AutoGrid del paquete de programas AutoDock Tools.[32]

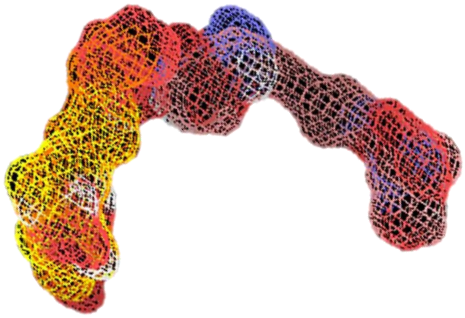


Figura A.15. *Molecular grid* de un sitio presente en cruzaína (PDB ID: 3KKU). En rojo interacciones electrostáticas negativas, en blanco positivas, en otro color de naturaleza hidrofóbica.

Este modelo del farmacóforo puede ser utilizado como molde de referencia para encontrar moléculas que cumplan con la distribución espacial de elementos electrostáticos y estéricos, lo que constituye un método de identificación de moléculas con posible actividad para esa diana en específico.

En caso de contar únicamente con información relacionada con las estructuras de ligantes que sean activos contra una diana, sea inferido por su efecto en un organismo o por la actividad específica demostrada

experimentalmente y/o por métodos computacionales, se puede realizar un modelo del farmacóforo que determine aquellos elementos moleculares y coordenadas espaciales compartidas por dos o más compuestos. Ya que generalmente no se conoce la conformación bioactiva de moléculas con actividad conocida, se puede aproximar esta a uno de sus mínimos de energía para después realizar un alineamiento rígido o flexible de las estructuras de interés. Otra estrategia, que implica la información estructural del blanco, consiste en utilizar la estructura cristalina de algún inhibidor cocristalizado como punto de referencia para el alineamiento. Aun cuando esta aproximación es razonable, su aplicación se encuentra limitada al número de moléculas cocristalizadas. [19]

La tercera estrategia que se puede seguir, es el uso de las huellas digitales llamadas SPLIF (*Structure Protein Ligand Interaction Fingerprints*). Estas huellas digitales codifican información espacial y electrostática de moléculas acopladas en un sitio en particular. La información con la que se alimenta a este algoritmo puede ser obtenida experimentalmente o por medio de acoplamiento molecular *in silico*. La información codificada en esta representación puede ser utilizada para realizar búsquedas por similitud o para construir el modelo del farmacóforo mediante distintos esquemas moleculares, lo cuales cuentan con diferentes elementos moleculares para describir el tipo de interacción contenida en los SPLIF. [40]

Cada uno de los modelos del farmacóforo mencionados, basados en estructura o en ligante, pueden ser utilizados como criterios para realizar tamizado molecular en bases de datos de diferente naturaleza. Su información también puede ser ingresada en algoritmos de acoplamiento molecular, como es el caso de RDock, como restricción conformacional, lo que puede aumentar la sensibilidad y selectividad del tamizado así como disminuir los tiempos de cálculo considerablemente. [41]

2.2.3. Identificación de sitios de unión

Los métodos que se enfocan en la búsqueda de sitios de unión de proteínas se pueden dividir en tres grandes grupos: conservación, adaptación del ligando y búsqueda geométrica.

El primer método de identificación se basa en la comparación de la secuencia de secuencias homólogas o de los residuos de sitios funcionales. La hipótesis en la que se sustentan sostiene que los residuos que son importantes para la unión con ligandos se encuentran conservados.

Esto encuentra eco en la teoría de la evolución molecular, ya que en esta se entiende a los mecanismos bioquímicos como resultado de miles de años de selección basados en pruebas y error de mecanismos moleculares, por lo que aquellos que podemos observar en la actualidad como parte constitutiva de los organismos vivos han sido lo suficientemente exitosos para ser conservados. La **Figura A.16.** muestra el resultado del alineamiento de secuencias encontradas en el PDB respecto a la secuencia de la enzima cruzaína de *Tripanosoma cruzi*. En esta imagen se puede observar la gran similitud de los residuos de aminoácido que constituyen el sitio activo de un grupo de proteasas de *Homo sapiens* y el sitio catalítico de la cruzaína, la cual también es una proteasa.[42]

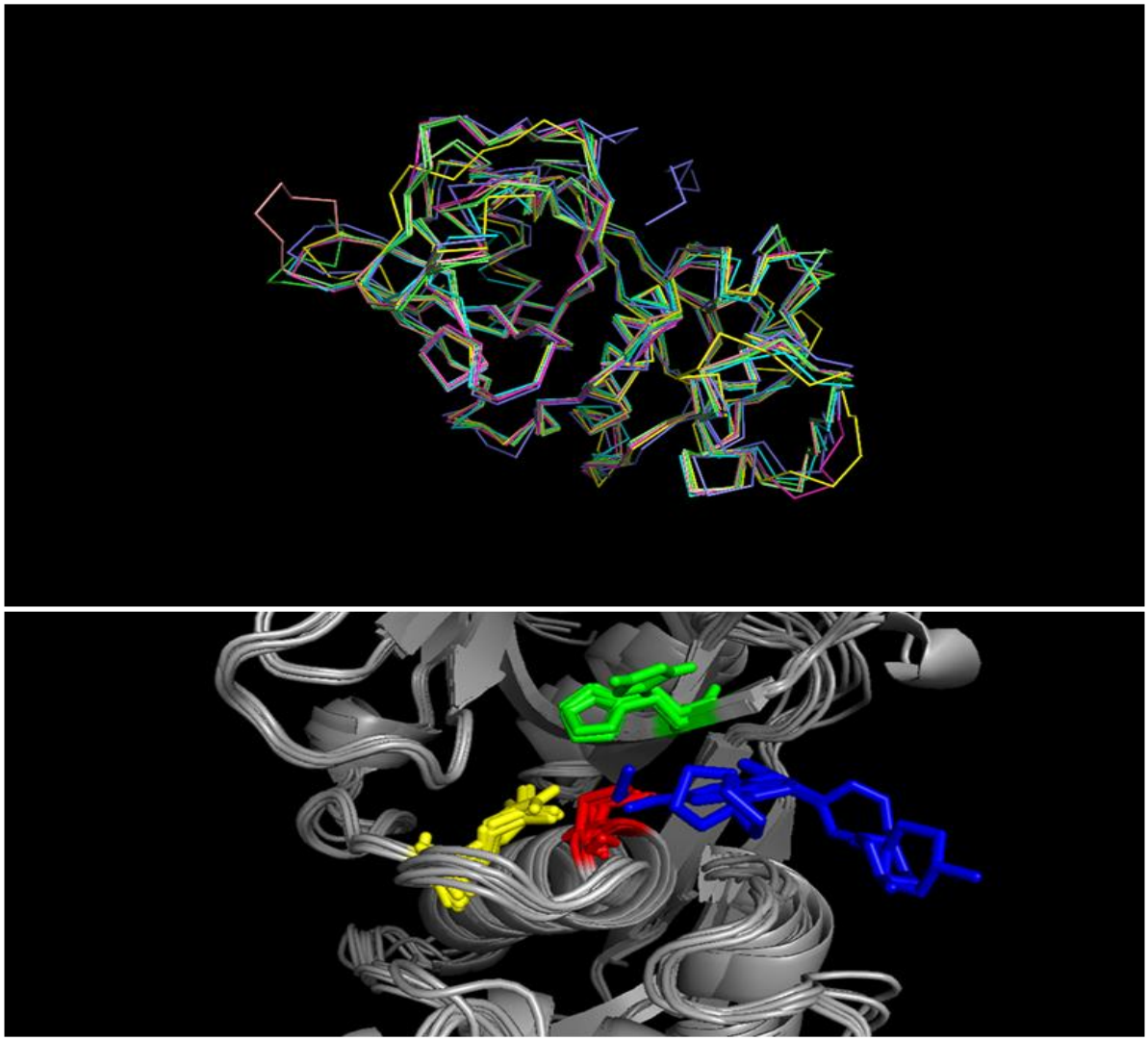


Figura A.16. Alineamiento de proteasas de *Homo sapiens* y la estructura de cruzaína de *Tripanosoma cruzi*. Alineamiento realizado en programa pymol académico.

Este ejemplo muestra claramente como este tipo de algoritmos es capaz de localizar sitios de unión implicados en la respuesta biológica. Sin embargo, estos métodos también pueden identificar secuencias que se encuentran relacionadas sólo con la estabilidad estructural de las macromoléculas biológicas, ya que muchas de estas también han sido conservadas por la selección natural.

Por otro lado, los métodos basados en la adaptación conformacional del ligando se fundamentan en las propiedades químicas de ligandos por medio de la construcción de plantillas tridimensionales o utilizando la información contenida en

bases de datos que representan una gran distribución de grupos funcionales donde se encuentran caracterizados algunos de sus elementos moleculares, como donadores y aceptores de puente de hidrógeno, hidrofobicidad, carga eléctrica, etc. Estos métodos también pueden ser catalogados como modelos farmacofóricos basados en estructura. [43]

El tercer grupo de métodos se puede dividir a su vez en tres grupos: puramente geométricos, geométricos con campo de energía y geométricos combinados con otros métodos. En todos los casos, la identificación se basa únicamente en la estructura del receptor. Los puramente geométricos utilizan la combinación de los diagramas de Voronoi y algoritmos de determinación de caminos mínimos como el de Dijkstra. Otra vertiente de estos métodos está basada en la caracterización geométrica por medio de esferas de diferentes tamaños que recorren la superficie del blanco para localizar cavidades que empaten en términos de volumen. [44]

Los métodos combinados, los más utilizados, son aquellos que mezclan los métodos geométricos con otros métodos de localización. Algunos de los algoritmos más comunes son, el algoritmo de *PocketFinder,* el cual construye una malla molecular por medio del cálculo de potenciales tipo Lennard-Jones. De acuerdo a estos resultados, se descartan aquellas regiones o sitios desfavorables, para después mapear el contorno de la superficie por métodos geométricos para identificar hendiduras con un volumen mayor a cierta cota. Otro de estos algoritmos es el *Q-SiteFinder,* que utiliza también una maya para determinar el potencial de van der Waals pero por medio de una sonda de metilo. Su segundo criterio de exclusión se basa en métodos geométricos que seleccionan a los mejores sitios de acuerdo a una línea de corte.[43]

Como herramientas complementarias a las antes mencionadas se pueden utilizar acoplamiento molecular ciego de bases de datos de moléculas bioactivas contra dianas en específico, el uso de acoplamiento de fragmentos junto al cálculo de modos normales y alineamiento de secuencias (servidor PARS) [45]o el uso de dinámica molecular en mezclas de disolventes para caracterizar el campo molecular de la superficie de proteínas (pyMDMix [46]), entre otros.

2.2.4. Entropía de Shannon

La teoría de la información es una rama de las matemáticas que trata sobre la transmisión y procesamiento de la información. Esta teoría fue desarrollada en parte por Claude E. Shannon y Warren Weaver a finales de los años cuarentas durante el periodo de la Segunda Guerra Mundial.[47] Esta teoría fue inaugurada por el trabajo de Claude E. Shannon titulado “Una teoría matemática de la comunicación”, donde en su intento por volver más eficiente la información a través de los canales de comunicación, plantea la ecuación de la entropía que lleva su nombre. En la **Figura A.17.** se muestra una de las formas matemáticas de la entropía de Shannon.

$$ES = -\sum p(x) \log_2 p(x)$$

Figura A.17. Entropía de Shannon.

Esta ecuación es el primer acercamiento teórico para cuantificar la cantidad de información contenida en un mensaje. Los datos contenidos en un mensaje, sea cual sea su naturaleza, se encuentran compuestos de dos partes, la información y el ruido. El ruido se puede entender como todas aquellas partes de un mensaje que no guardan ningún patrón de utilidad, siendo normalmente producto de fenómenos físicos de naturaleza aleatoria, mientras que la información se puede definir como aquellos datos transmitidos que son de utilidad.

Ya que los datos trasmitidos están compuestos por elementos finitos de acuerdo a la naturaleza de la señal, esta ecuación entiende que para que un grupo de datos sea considerado como información, debe tener un patrón definido formado por la presencia de redundancias. Es decir, la información es un grupo de señales que se encuentran acotadas por alguna estructura abstracta. En el caso contrario, el ruido, cualquier señal comprendida por un método de transmisión tiene la misma probabilidad de aparición.

En el área de quimioinformática este concepto ha cobrado gran relevancia ya que permite cuantificar de manera sistemática la cantidad de información contenida en un grupo de representaciones moleculares. La entropía de Shannon ha sido utilizada como fundamento para desarrollar un sinnúmero de métodos de representaciones moleculares, dentro de metodologías de tamizado virtual y en la caracterización del espacio químico.[25] Respecto a la caracterización del espacio químico, esta métrica permite catalogar a diferentes bases de datos respecto a su contenido de información, lo que se encuentra estrechamente ligado con el concepto de diversidad molecular. Por ejemplo, en el caso de una base de datos con un alto grado de diversidad molecular, estructuras químicas heterogéneas, la cantidad de información será baja. Mientras que una base de datos con baja diversidad, estructuras moleculares análogas, la cantidad de información será alta. Esto se debe al pequeño grado de redundancia encontrado dentro de una base de datos diversa, como es el caso de las diseñadas para realizar tamizado de alto rendimiento y que tiene como caso extremo una distribución aleatoria de señales. Caso contrario son las bases de datos enfocadas a dianas específicas. En este caso es alto el número de redundancias debido a la gran probabilidad de encontrar moléculas con un alto grado de similitud gracias al principio que relaciona la actividad con la similitud. [48]

Objetivos

Objetivo General. Cuantificar la diversidad estructural y cobertura en el espacio químico de inhibidores de DNA metiltransferasa (DNMT) evaluando el impacto que tiene esta diversidad en las bases estructurales que rigen el reconocimiento molecular y modo de unión de los inhibidores con la diana epigenética.

Objetivo específico 1. Estudio quimioinformático de inhibidores conocidos de DNMT

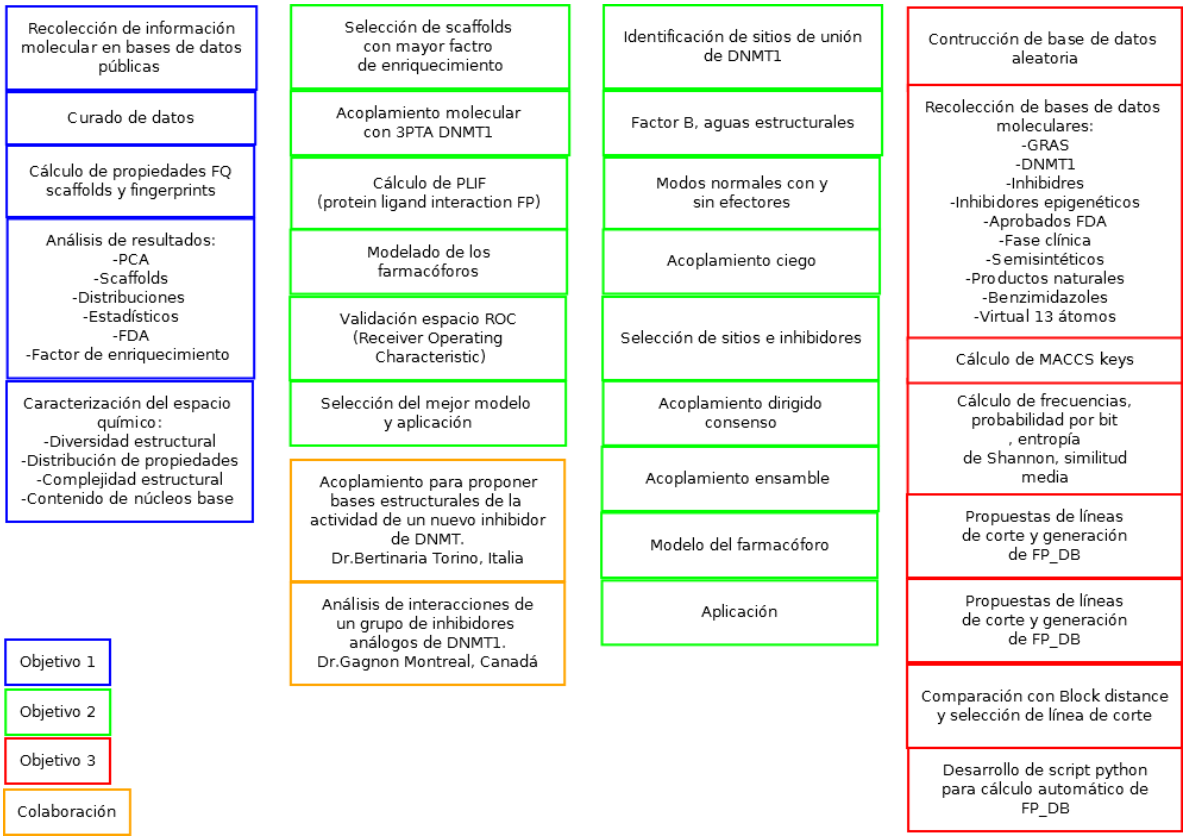
El objetivo es cuantificar la diversidad estructural y química para analizar la cobertura del espacio químico de compuestos que han sido identificados por nuestro grupo y otros grupos de investigación como inhibidores de DNMTs. Esta información permitirá conocer el grado de variación estructural y químico de inhibidores hasta ahora descritos contra esta diana epigenética.

Objetivo específico 2. Estudio computacional del reconocimiento molecular de inhibidores de DNMT con la diana epigenética

Por otro lado, existe una gran ausencia de información sobre los mecanismos de acción de los inhibidores de DNMT reportados. Esto también ha dado pie a incertidumbre respecto a los sitios de unión en los que actúan los diferentes inhibidores. Ambos vacíos de información han llevado a la generación de hipótesis que comprenden la unión en DNA, sitio del cofactor, sitio activo y posiblemente sitios que a larga distancia del sitio de unión (alostéricos o dinámicos). Por lo tanto, se planteó como objetivo establecer modelos que describan la interacción ligando-diana biológica de los inhibidores de DNA metiltransferasas descritos hasta la fecha para después generar hipótesis sobre el mecanismo de acción. Los resultados de este objetivo contribuirán a establecer las bases estructurales implicadas en el reconocimiento molecular de estos inhibidores y ayudarán a explicar su actividad biológica a nivel molecular.

Objetivo específico 3. Proponer una nueva representación molecular para la búsqueda de inhibidores de DNMT1 en bases de datos.

El objetivo es generar una representación total de bases de datos de compuestos químicos a través de la probabilidad asociada a los bits de representaciones binarias por su relación con la entropía de Shannon y el concepto de redundancia informacional. Se plantea utilizar estos criterios como una metodología novedosa en la búsqueda por similitud de compuestos activos contra DNMT1. Esto se encuentra de acuerdo con publicaciones donde se utiliza este tipo de métricas para realizar caracterización de bases de datos y búsquedas por similitud.[49]



FiguraO.1. Diagrama de metodología general para lograr los objetivos.

Capítulo 1. Quimioinformática

Metodología computacional

A)

Construcción y curado de una base de datos molecular de inhibidores de DNMTs a partir de bibliotecas públicas de compuestos químicos y de la literatura científica publicada en las principales revistas relacionadas con el desarrollo de fármacos. Esta base de datos fue analizada mediante una serie de herramientas quimioinformáticas establecidas[50] para cuantificar tres aspectos fundamentales y complementarios: diversidad estructural empleando descriptores de huellas digitales moleculares (*molecular fingerprints*), distribución de propiedades fisicoquímicas y métricas asociadas con la complejidad estructural, contenido y diversidad de núcleos base (*molecular scaffolds*). Ello se realizó por medio del paquete de programas *Molecular OperatingEnviroment* (**MOE**)[51] y el conjunto de programas *Mayachem Tools*.[52]

B)

Una vez identificados los compuestos de interés se realizaron estudios de acoplamiento molecular (*molecular docking*) [53] y modelado del farmacóforo a partir de superposiciones tridimensionales y/o caracterización de huellas digitales de la interacción ligando-proteína (*protein-ligand interaction fingerprints*). Aunado a ello, se aplicaron las estrategias utilizadas en el primer año de doctorado, lo que comprende: búsqueda de sitios de unión no conservados y su posible conexión con actividad alostérica y/o cambios dinámicos con impacto en la actividad biológica, acoplamiento molecular ensamble y técnicas que toman en cuenta la flexibilidad de la proteína blanco como estudio de modos normales.

Desarrollo

Se construyó una base de datos interna con inhibidores de DNMT1 a partir de cuatro fuentes principales que incluyen tres bases de datos públicas de compuestos químicos y búsqueda en literatura científica vigente. La búsqueda en bases de datos públicas dio como resultado 265 compuestos de *CheMBL* [54], 106 compuestos de *Human Epigenetic Enzyme Modulator Database*, *HEMD*[55] y 337 compuestos de *Binding Database.*[56] La búsqueda de compuestos adicionales no reportados en ninguna de las bases de datos públicas desde el año 2013 hasta el momento se realizó dentro de la literatura científica por medio del buscador Web of Science dando como resultado 47 compuestos con actividad inhibitoria contra alguna DNMT.

El curado de la información hallada se realizó de acuerdo con la metodología reportada por Fourches D. et al..[57] Para cada uno de los compuestos se realizó una homogenización de la actividad y se obtuvo la notación estructural lineal canónica de acuerdo a *Simplified Molecular Input Line-Entry system* (SMILE) antes de ser ingresada a la base de datos. Cada molécula fue preparada mediante la función *wash* incluida en el módulo contenido en el programa *Molecular Operating Environment* (MOE)[51], que desconecta sales de metales, remueve componentes simples y recalcula estados de protonación. Se eliminaron réplicas de compuestos idénticos con la misma actividad biológica. Cuando estas presentaban pequeñas diferencias en el valor de actividad reportado se obtuvo la media. Para el caso compuestos con actividades muy distantes fue necesario revisar el método experimental por el cual fueron obtenidos dichos valores de actividad y de acuerdo a ello se seleccionaron algunas de las posibilidades. En caso de que la información no fuera definitoria no se incluyó dentro de la base de datos. En la **Tabla 1.1.** se muestran algunas de las características de cada una de las fuentes de información utilizadas para este estudio.

Tipo de fuente	Fuente	Referencia	No.inhibidores
Bases de datos públicas	Binding Database	Ref 22 RCS	265
	ChEMBL	Ref 56 RCS	163
	HEMD	Ref 57 RCS	96
Literatura	Web of Science	https://isiknowledge.com	42
TOTAL	Conjunto curado y sin réplicas		566

Tabla 1.1. Fuente de los compuestos que constituyen la base de datos de inhibidores de DNMT1 construida dentro del grupo. [15]

El mismo proceso de curado fue realizado para cuatro bases de datos disponibles públicamente, entre las que se encuentran una base de datos de compuestos aprobados como fármacos (*Food and Drugs Administration* (FDA) de E.E.U.U.A), una de compuestos en fases clínicas, una colección general de inhibidores y una base de datos enfocada a inhibidores de blancos epigenéticos. En la **Tabla 1.2.** se presentan algunas de las características generales de las bases de datos utilizadas para la caracterización relativa del espacio químico de inhibidores de DNMT1.

Base de datos	Fuente	Número de compuestos
Fármacos aprobados	Drug Bank	1490
Colección para tamizado	Selleck	1100
En fase clínica	Therapeutic Target Database	837
Base de compuestos enfocada a inhibidores epigenéticos	Selleck	113

Tabla 1.2. Fuente y número de compuestos para cada una de las bases de datos de referencia. [15]

Representación estructural

Las bases de datos fueron comparadas y analizadas por medio del cálculo de propiedades fisicoquímicas, huellas digitales moleculares (*molecular fingerprints*) y análisis con el método de representación basado en núcleos base o *scaffolds*.

Propiedades fisicoquímicas y descriptores moleculares

Se calcularon seis propiedades fisicoquímicas con el paquete de programas MOE: coeficiente de partición octanol/agua (**SlogP**), enlaces rotables (**RB**), donadores de puente de hidrógeno (**HBD**), aceptores de puente de hidrógeno (**HBA**), área superficial topológica (*Topological polar Surface Area* **TPSA**) y peso molecular (**MW**).

Posteriormente se obtuvo la distribución de cada una de las propiedades así como el valor de su media, mediana, intervalo intercuartílico y desviación estándar. Para cada propiedad y por medio del programa estadístico RGui[58] se graficaron diagramas de caja y se calculó la divergencia de la normalidad, homocedasticidad, pruebas de hipótesis paramétricas o no paramétricas según fuera el caso, así como el correspondiente análisis post-hoc (Kruskal-Wallis y Nemenyi) utilizando PMCMR de RGui.[58]

Para poder generar una representación gráfica de las seis dimensiones en cuestión, se realizó un análisis de componentes principales (*Principal Component Analysis* **PCA**) utilizando el programa MOE[51] y *Osiris DataWarrior*. [59]

Huellas digitales moleculares (*molecular fingerprints*)

Las bases de datos también fueron estudiadas a partir de diferentes huellas digitales moleculares (en inglés, *molecular fingerprints)*. Por medio del programa MOE, se calcularon MACCS keys (166-bits) que originalmente fueron diseñadas para búsqueda de estructuras en bases de datos y donde cada bit describe una pequeña subestructura que cuenta con un máximo de diez átomos sin contar hidrógeno. GpiDAPH3 (*Pharmacophore Graph Triangle*) que consta de un farmacóforo de tres puntos calculado a partir del grafo en dos dimensiones de la estructura. TGD (*Typed Graph Distance*) que es un *fingerprint* del tipo distancia de

grafos, donde se representa cada *fingerprint* como un conjunto de tres elementos de la forma (u, v, d), donde u y v son tipos de átomos que incluyen la etiqueta ácido, base, donador y aceptor de puente de hidrógeno, etc. Mientras que d es la distancia mínima entre dos vértices del grafo. Y finalmente, *Extendended-Connectivity*, un *fingerprint* topológico circular, utilizado, en este caso, con un diámetro de cuatro y seis átomos de distancia. [26,30,27]

Seguido de ello se calculó la matriz de similitud utilizando el índice de Tanimoto (**Figura A.9.**), el cual es el cociente del conjunto de elementos que se comparten por dos *fingerprints* entre la suma de elementos únicos de cada estructura restada por el conjunto de elementos que se comparten.[29]

Para dichas matrices se tomaron los valores fuera de la diagonal y se realizó un muestreo aleatorio de 5000 valores para después calcular estadísticos como media, mediana, distancia intercuartílica, máximo, mínimo y desviación estándar. Finalmente, se determinó el tipo de distribución, su desviación respecto a una distribución normal, homocedasticidad, funciones de distribución acumulativa y cálculo de pruebas de hipótesis paramétricas y no paramétricas, así como el correspondiente análisis post-hoc según fuera el caso.

Núcleos base *(scaffolds)*

Otro de los criterios que se utilizó para la descripción de los inhibidores DNMT1 fue el conteo de *scaffolds* (también llamados núcleos base o sistemas cíclicos). El proceso de extracción de esta información se realizó por medio del programa *Molecular Equivalence Indices* (MEQI).[28] Este programa remueve de forma sistemática todos aquellos vértices que con grado uno, dando como resultado a los grafos cíclicos y en caso existir más de un ciclo, incluyendo a las cadenas que conectan a dichos sistemas.

Estos sistemas cíclicos son parte de los quimiotipos definidos en la metodología propuesta por Johnson y Xu. [28] Para cada uno de los sistemas cíclicos, esta metodología asigna un único identificador de cinco caracteres de longitud (código quimiotípico). [60, 61]

Por medio de estos resultados es posible determinar la frecuencia para los diferentes sistemas cíclicos que conforman la base de datos y de esta manera realizar una comparación directa con el contenido y diversidad en otras bases de datos. Para lograr dicho objetivo, se realizó el estudio de los gráficos *Cyclic System Retrieval* (CSR) (**Figura1.1.**). En dichas gráficas se presenta la fracción de sistemas cíclicos contra la fracción acumulativa de la base de datos. La información contenida en este método de visualización se puede explicar por medio de sus casos extremos: de contener un sistema cíclico diferente para cada uno de los compuestos en la biblioteca dará como resultado una línea recta con pendiente de 45°, lo que correspondería a la máxima diversidad de *scaffolds*. Mientras de encontrar un solo sistema cíclico que englobe a todos los compuestos se obtendrá un escalón con su valor máximo en uno, es decir, el mínimo valor de diversidad posible.

Para poder medir de forma exacta todas las variaciones existentes entre estos dos posibles extremos es recomendable calcular el área debajo de la curva de la función resultante. Este valor rondará entre 0.5 para el caso de máxima diversidad y uno para el caso opuesto.

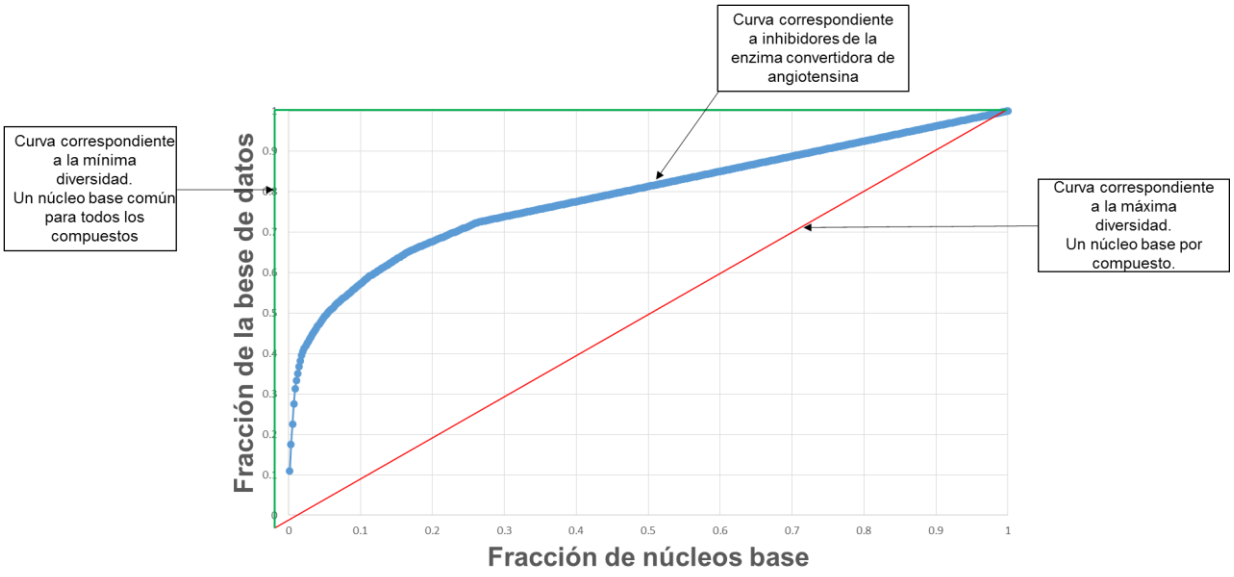


Figura1.1. Curvas del gráfico *Cyclic System Retrieval*.

Otras métricas comúnmente utilizadas y que son paralelas a la anterior son: valor F_{50} (valor de la fracción de sistemas cíclicos cuando se recupera el 50% de la fracción de compuestos en la biblioteca), número y fracción de "núcleos únicos" (*singletons*: sistemas cíclicos que contienen una sólo molécula) respecto al número de compuestos y al número de sistemas cíclicos, fracción de sistemas cíclicos, entre otras. [62]

Ya que se cuenta con los valores de actividad para los inhibidores de DNMT1, también es posible construir los gráficos de factor de enriquecimiento respecto a frecuencia de sistemas cíclicos. Para realizarlo se eligieron a los sistemas cíclicos más representativos, los más frecuentes, y se calculó para cada uno de ellos el factor de enriquecimiento (**Figura 1.2.**)

$$Act(C) = [C^*]/[C]$$
$$Act(C_\lambda) = [C_\lambda^*]/[C_\lambda]$$
$$FE(C_\lambda) = Act(C_\lambda)/Act(C)$$

Figura 1.2. Expresiones para obtener el factor de enriquecimiento.

Estos gráficos son de gran utilidad ya que nos permiten tener una idea general sobre la cantidad de análogos reportados para cada clase de inhibidor, así como su actividad promedio respecto a los otros sistemas presentes (**Figura 1.3**). Esta información puede convertirse en información clave en la toma de decisiones al proponer estrategias para la síntesis de nuevos compuestos bioactivos.

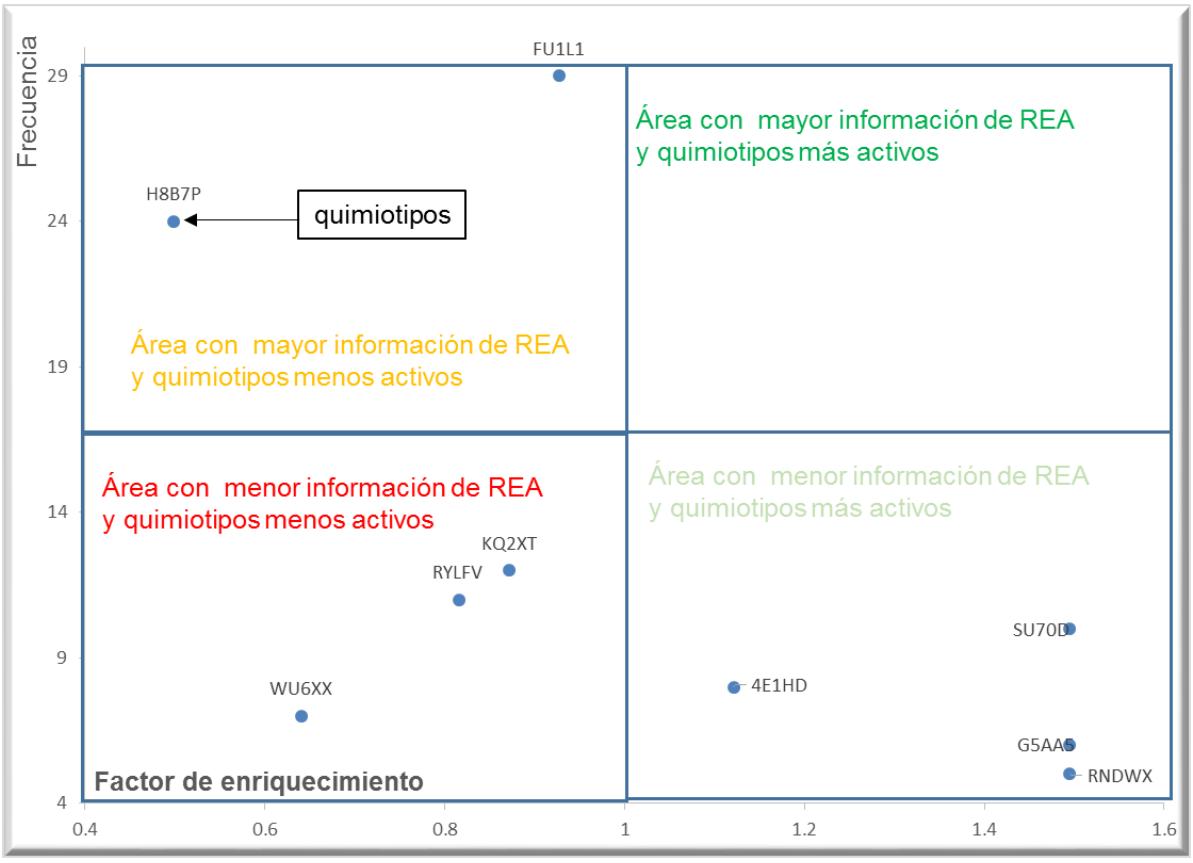


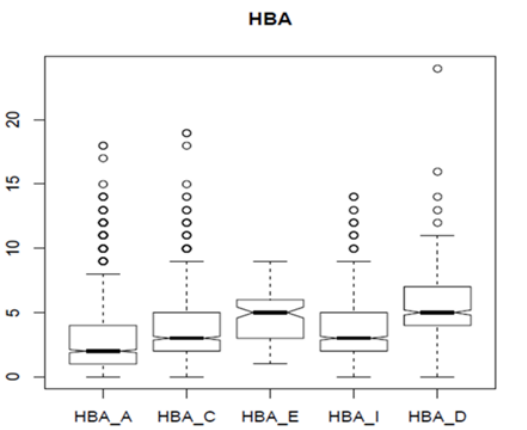
Figura 1.3. Regiones del diagrama de frecuencia de núcleos base contra factor de enriquecimiento.

Resultados y Discusión

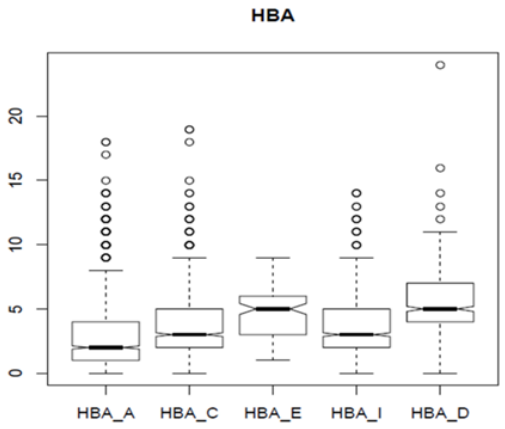
Los resultados se presentarán en tres secciones que representan las diferentes métricas utilizadas para la caracterización de los inhibidores de DNMT1.

Propiedades fisicoquímicas y descriptores moleculares

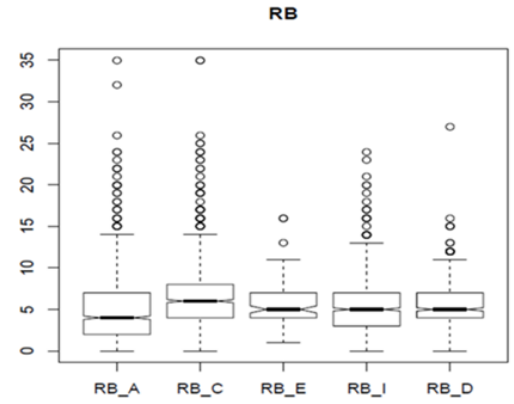
En la **Figura 1.4.** se puede encontrar el resumen de los estadísticos calculados y los diagramas de caja con muesca (*notch box plots*) para algunas de las propiedades y descriptores calculados con la paquetería de programas MOE, realizados a través del lenguaje de programación R.[58]



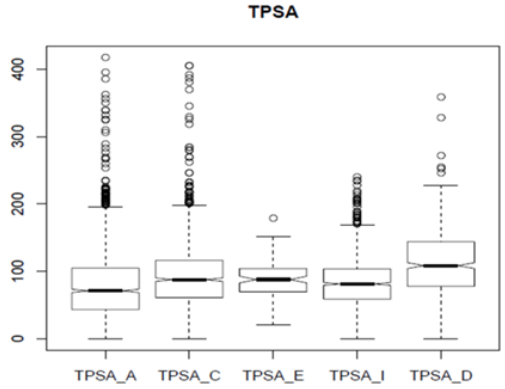
HBA	Approved	Clinical	Epigenetic	Inhibitors	DNMT
Min	0.0	0.0	1.0	0.0	0.0
1Q	1.0	2.0	3.0	2.0	4.0
Median	2.0	3.0	5.0	3.0	5.0
Mean	2.9	3.6	4.5	3.4	5.3
3Q	4.0	5.0	6.0	5.0	7.0
Max	18.0	19.0	9.0	14.0	24.0
SD	2.6	2.5	1.6	2.1	2.5



HBA	Approved	Clinical	Epigenetic	Inhibitors	DNMT
Min	0.0	0.0	1.0	0.0	0.0
1Q	1.0	2.0	3.0	2.0	4.0
Median	2.0	3.0	5.0	3.0	5.0
Mean	2.9	3.6	4.5	3.4	5.3
3Q	4.0	5.0	6.0	5.0	7.0
Max	18.0	19.0	9.0	14.0	24.0
SD	2.6	2.5	1.6	2.1	2.5



RB	Approved	Clinical	Epigenetic	Inhibitors	DNMT
Min	0.0	0.0	1.0	0.0	0.0
1Q	2.0	4.0	4.0	3.0	4.0
Median	4.0	6.0	5.0	5.0	5.00
Mean	5.0	6.3	5.5	5.3	5.6
3Q	7.0	8.0	7.0	7.0	7.0
Max	35.0	35.0	16.0	24.0	27.0
SD	3.8	4.2	2.8	3.3	2.9



TPSA	Approved	Clinical	Epigenetic	Inhibitors	DNMT
Min	0.0	0.0	20.3	0.0	0.0
1Q	43.4	61.1	69.6	59.1	78.1
Median	71.2	88.0	88.0	81.4	108.6
Mean	81.9	95.0	89.0	84.1	111.7
3Q	104.9	116.3	104.3	104.0	143.9
Max	417.3	405.4	178.3	240.5	358.2
SD	55.9	52.5	29.2	37.3	47.1

Figura1.4 Valores estadísticos y representación mediante *box notch plots* de algunas de las propiedades y descriptores calculadas para los compuestos presentes en las bases de datos estudiadas. [15]

Este método de visualización de las distribuciones de los descriptores presenta una caja que incluye a todos los valores contenidos entre el primer y tercer cuartil, la media de la distribución como una línea horizontal obscura y círculos vacíos en la parte superior e inferior representado los valores atípicos. Las muescas que se encuentran a cada lado de la caja representan un intervalo de confianza del 95% respecto a la media. De acuerdo al análisis realizado con el examen estadístico de

Shapiro-Wilk, ninguna de las distribuciones presenta un comportamiento normal, por lo que fue necesario aplicar la prueba no paramétrica Kruskal-Wallis y el análisis *post hoc* de Nemenyi implementada en el módulo PMCMR[58] de R para determinar la diferencias estadísticas entre ellas.

De las diferencias encontradas se puede decir que los inhibidores de DNMT1 muestra, en conjunto, un mayor número de donadores y aceptores de puentes de hidrógeno que los compuestos aprobados, en fase clínica y la colección general de inhibidores. Al contrario, para estos descriptores se encontró una gran similitud con la colección de inhibidores epigenéticos, con una media de 5 y 2 respectivamente y un valor de p de 0.35 y 0.79. En general se observan valores más altos para TPSA, lo que se puede traducir en una tendencia de estos inhibidores a ser más polares que los compuestos encontrados en otras bases de datos.

Los valores de peso molecular y SlogP de los inhibidores de DNMT1 son similares a las colecciones de referencia, con excepción de los compuestos aprobados, que a su vez muestra, para estos descriptores, valores menores respecto a todas las colecciones estudiadas.

Por último, el número de enlaces rotables muestra una distribución muy similar respecto a todas las colecciones estudiadas y especialmente con los inhibidores epigenéticos, con un valor p=0.99, lo que indica una flexibilidad molecular compartida.

Representación visual del espacio de propiedades

Para generar una representación visual del espacio de propiedades (6D), se aplicó una reducción dimensional por medio del análisis de componentes principales basado en covarianza. La **Tabla 1.3.** muestra las contribuciones porcentuales de los tres primeros componentes. Siendo que los dos primeros componentes recuperan un 77.3% de la covarianza, y los tres primeros un 89.4%, utilizar cualquiera de estas combinaciones para representar el comportamiento general del espacio de propiedades es razonable respecto a la cantidad de información perdida.

	PC1	PC2	PC3	PC4
Eigenvalores (EV)	-2.8	-1.27	-1.63	-9.15
Contribución de EV	53.86	77.31	89.35	95.11
HBA	0.106	-0.023	-0.134	0.155
HBD	0.134	-0.111	-0.280	-0.619
RB	0.051	0.083	0.220	-0.216
SlogP	-0.011	0.236	-0.213	-0.086
TPSA	0.005	-0.004	0.002	0.007
MW	0.002	0.002	0.000	0.001

Tabla 1.3. Contribuciones de los cuatro primeros componentes principales. [15]

En esta tabla también se incluyen las contribuciones de cada uno de los descriptores para cada una de las componentes. Siendo aceptores y donadores de hidrógeno los que más contribuyen en CP1, SlogP para CP2 y enlaces rotables para CP3.

La visualización del espacio de propiedades en dos dimensiones se puede observar en la **Figura1.5**.

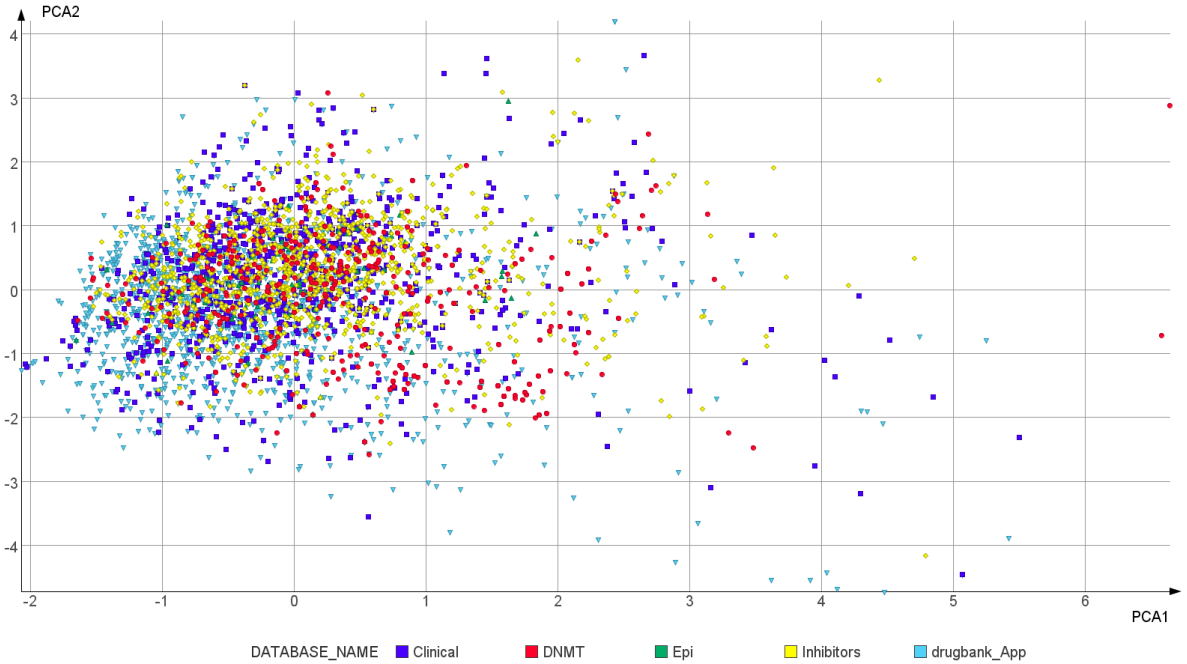


Figura 1.5. Espacio de propiedades respecto a dos componentes principales.

Después de un examen visual, se puede concluir que los inhibidores de DNMT1 ocupan y son contenidos mayoritariamente por el espacio ocupado por las colecciones de referencia.

Las dos colecciones que cubren mayoritariamente el espacio son la de compuestos aprobados y en fase clínica, además, y como era esperado, estas dos colecciones son las más densamente pobladas dentro de este espacio. La colección de inhibidores de DNMT1 y la biblioteca enfocada a inhibidores ocupan un espacio más restringido que se encuentra incluido en el anterior, sin embargo, DNMT1 también se expande a regiones no ocupadas por las colecciones de referencia. Tomando en cuenta las limitaciones y pérdida de información implícita en la reducción dimensional, se puede decir que la colección de inhibidores de DNMT1 ocupa el espacio de propiedades de química farmacéutica tradicional y algunas de sus moléculas exploran nuevas regiones no incluidas en el anterior, lo cual es coherente con los valores divergentes para algunas de las propiedades.

También se realizó una subdivisión entre los compuestos activos e inactivos de DNMT1 respecto a una cota de 10 μM. Obteniendo 378 compuestos por debajo de esta cota y 32 moléculas con actividades menores a 1 μM. El espacio que ocupan dichos compuestos respecto a los denominados inactivos es similar. Sin embargo, los compuestos activos presentan mayor peso molecular y SlogP, sugiriendo que son de mayor tamaño y más hidrofóbicos que los inactivos. En esta población también se observaron valores mayores para el número de enlaces rotables, indicando que estos son, en términos generales, más flexibles que los inactivos.

Huellas digitales moleculares *(molecular fingerprints)*

Como se mencionó anteriormente, para determinar la variación del espacio químico ocupado respecto a la representación molecular y la diversidad de cada una de las bases de datos estudiadas, se calcularon cinco *fingerprints* diferentes que difieren en su metodología y naturaleza. La similitud interna de cada base de

datos fue calculada por medio del índice de Tanimoto con la paquetería de programas MayaChem Tools.[52]

Los resultados muestran una gran diferencia de la similitud media respecto al método de representación utilizado. TGD presenta los valores de similitud más altos, con una media entre 0.54 y 0.64, seguido de MACCS keys, con media entre 0.31 y 0.41, GpiDAPH3, con media de entre 0.13 y 0.26, y ECFP con media entre los valores 0.06 y 0.07. Esta tendencia también puede ser observada en otros trabajos donde se hace uso de diferentes métodos de representación. [31] Esta tendencia se puede explicar parcialmente por el grado de resolución con el que cuenta cada una de las representaciones.

Mientras que MACCS keys está basado en un diccionario preestablecido de fragmentos, ECFP es calculado al vuelo y depende del número y tipo de vecinos que compartan las moléculas representadas.

En conclusión, se puede decir que la base de datos de inhibidores de DNMT1 es diversa, y que es comparable con la de la biblioteca enfocada a inhibidores epigenéticos según TGD, MACCS keys y GpiDAPH3. En la **Figura 1.6** se puede observar la comparación de las diferentes bases de datos representadas por medio de un *fingerprint* determinado en su forma de funciones de distribución acumulada.

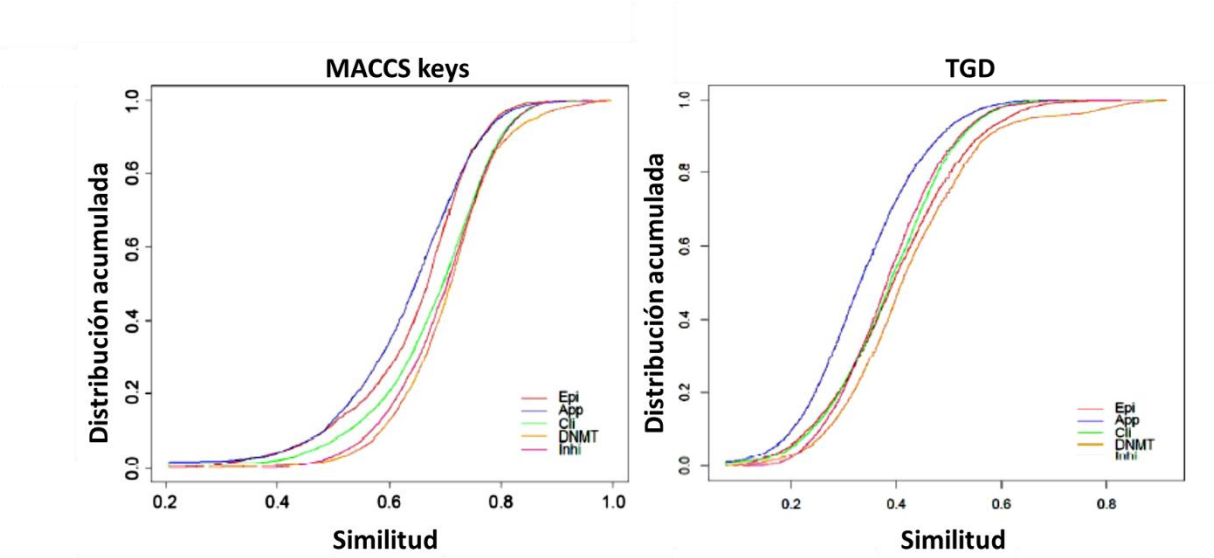


Figura 1.6. Funciones acumulativas para los *fingerprints* MACCS keys y TGD.[15]

Al comparar las diferentes bibliotecas se puede afirmar que la base de datos de compuestos aprobados es la que presenta mayor diversidad, seguida de los compuestos en fase clínica y la biblioteca de inhibidores. Esto no nos sorprende ya que dichas bases no son enfocadas en ninguna diana particular, sino que incluyen inhibidores de diversas dianas. Sin embargo, los resultados apuntan a validar el uso de al menos tres tipos de representación molecular para la comparación y estudio de la diversidad. Estos son: TGD, GpiDAPH3 y MACCS keys.

Aun cuando en otros estudios ECFP ha dado buenos resultados para el estudio de las relaciones estructura actividad, en nuestro caso los valores que se obtienen son demasiado pequeños para poder compararlos con los resultados obtenidos con otras métricas de representación.

Estudio de núcleos base

El estudio de la diversidad núcleos base de la biblioteca de inhibidores de DNMT1 se realizó por medio del conteo de frecuencia y de la curva CSR. Los valores asociados a los núcleos base como es el número de núcleos base, el número de núcleos únicos, y distintas fracciones se resumen en la **Tabla 1.4.**

	Conjunto DNMT1i
N	291
N/M	0.515
N_{sing}	170
N_{sing}/N	0.584
N_{sing}/M	0.300

Tabla 1.4. Diversidad de núcleos base para inhibidores de DNMT1.[15]

Estos resultados muestran una gran diversidad de núcleos base, esto se ve reflejado en proporción de núcleos únicos respecto al número de sistemas cíclicos (58%), así como al número de compuestos en la base de datos (30%). Esto

también se puede verificar mediante el área debajo de la curva y fracción 50, con valores de 0.67 y 0.23, respectivamente.

Los núcleos base encontrados pueden dividirse en dos grandes grupos: nucleosídicos y no nucleosídicos. Los primeros están relacionados con análogos del cofactor *S*-adenosil-L-metionina, mientras que el segundo grupo incluye estructuras de diversa naturaleza. Dentro del conjunto de compuestos con mayor actividad se pueden encontrar quimiotipos relacionados con ambos grupos. Entre los núcleos base más frecuente de naturaleza nucleosídicos se encuentran los siguientes quimiotipos expresados por su identificador de cinco caracteres FUIL1, H8B7P, KQ2XT y WU6XX con frecuencias de alrededor del 5% comprendiendo el 12.7% de la base de datos con un total de 72 compuestos. Por el otro lado, los núcleos base con identificador RYLFV, SU70D, 4E1HD, G5AA5 y RNDWX en total representan sólo el 7.1% de la base de datos con un total de 40 compuestos.

Estos resultados expresan las tendencias históricas que se han dado en el desarrollo de inhibidores de esta diana, las cuales están relacionadas con la síntesis de moléculas similares los compuestos análogos al cofactor SAM aprobados para su uso clínico.

Enriquecimiento de quimiotipos

Para calcular a los quimiotipos más relevantes de la base de datos de DNMT1 se calculó la frecuencia y factor de enriquecimiento como se muestra en la sección de metodología.

El factor de enriquecimiento da la proporción de compuestos activos para determinado núcleo base, ello combinado con la información de frecuencias permite identificar a aquellos núcleos que son más atractivos por la información accesible para generar relaciones de estructura-actividad. En segundo término, también posibilita la selección de núcleos novedosos donde la información de sus análogos aún no se encuentra explorada. Esta selección dependerá directamente del enfoque experimental, de factores económicos y de viabilidad sintética.

La **Figura 1.7.** muestra a los nueve núcleos base más frecuentes junto con su factor de enriquecimiento. Cuatro de ellos presentan valores de enriquecimiento

mayores a 1.0, de ellos, tres pertenecientes al grupo de compuestos no nucleosídicos, con un factor de 1.4 (SU70D, G5AA5 y RNDWX). La fuente experimental de estos núcleos proviene de experimentos de tamizado molecular de alto rendimiento realizados por medio de ensayos de *molecular beacon probes* por Sanford-Burnham Center for Chemical Genomics. [63]

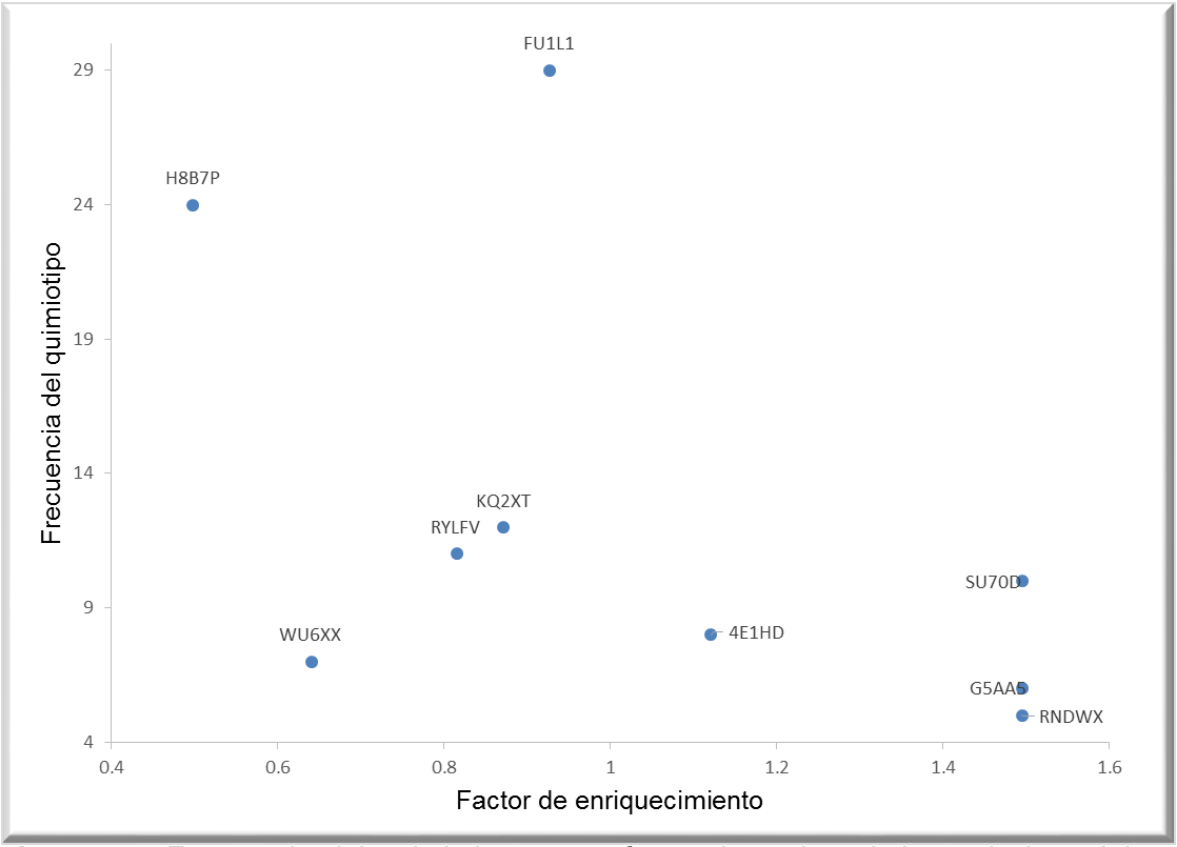


Figura 1.7. Frecuencia del quimiotipo contra factor de enriquecimiento de los núcleos base más frecuentes en la base de datos de inhibidores de DNMT1.[15]

En especial, en núcleo base con identificador SU70D presenta una de las mejores opciones al contar con mayor información para un análisis de estructura-actividad dado su alto valor de frecuencia. Los análisis realizados indican que una buena estrategia inicial para encontrar nuevos agentes quimioterapéuticos que tengan como diana molecular a DNMT1 sería la exploración de las relaciones estructura actividad de estos tres núcleos (SU70D, G5AA5 y RNDWX).

Uno de los quimiotipos más interesantes desde la perspectiva computacional es el 4E1HD. Esta familia de compuestos fue encontrada mediante

el tamizado virtual basado en acoplamiento molecular y agrupamiento de una base de datos de 111,121 moléculas por el grupo de investigación de Chen et al..[64] De estos resultados se obtuvieron 50 *hits*, de los cuales de compraron análogos para realizar análisis de actividad contra DNMT1 y explorar así las relaciones estructura actividad.

Por último, el núcleo base con mayor frecuencia (29 compuestos) pertenece al quimiotipo FU1L1 con un factor de enriquecimiento de 0.98. Este núcleo pertenece a análogos del cofactor SAM, lo que desde nuestra perspectiva caen en segundo término de importancia dados los efectos adversos potenciales que un inhibidor de esta naturaleza pueda presentar.

Es importante recalcar que además de las posibles estrategias experimentales que pudieran surgir gracias a esta información, en términos computacionales sería posible utilizar cualquiera de los núcleos antes mencionados como punto de partida de metodologías de tamizado molecular basadas en similitud o en otra técnica de búsqueda virtual.

Resumen de resultados

Esta parte de la tesis doctoral representa uno de los primeros esfuerzos por caracterizar el espacio químico de los inhibidores de DNMT1. Esto se realizó por medio de tres métricas complementarias que contemplan: diversidad estructural por medio de similitud, núcleos base y estudio de la distribución de descriptores moleculares y propiedades de los inhibidores de DNMT1.

A raíz del estudio sobre el espacio de propiedades fisicoquímicas se puede concluir que estos inhibidores presentan una polaridad promedio mayor que la molecular aprobadas para uso clínico, en fase clínica y que la colección general de inhibidores, reflejado en los valores encontrados para donadores y aceptores de puente de hidrógeno y TPSA. También se encontró que los inhibidores de DNMT1 presentan una flexibilidad comparable a estas bases de datos, lo cual se refleja en los valores de enlaces rotables.

La representación visual por medio de la reducción dimensional efectuada por análisis de componentes principales, muestra que la base de datos de

inhibidores de DNMT1 ocupa regiones contenidas en el espacio químico farmacéutico tradicional, con algunas moléculas que exploran regiones distantes.

Estas moléculas podría también ser punto de partida de estudios de estructura actividad en busca de inhibidores novedosos, aun cuando a su vez corren mayor riesgo de presentar efectos secundarios o baja biodisponibilidad.

La diversidad medida por medio de distintas métodos de representación molecular y con el índice de Tanimoto, demuestra que la colección inhibidores de DNMT1 presenta diversidad estructural por debajo de la encontrada para la colección de compuestos aprobados para uso clínico. El análisis de núcleos base plantea la posibilidad de encontrar quimiotipos que resulten en estructuras privilegiadas epigenéticas, lo que representa una oportunidad para la exploración de la relaciones estructura actividad de sus análogos.

Además estos resultados están de acuerdo con las tendencias actuales sobre el desarrollo de inhibidores de DNMT1 no basados en la estructura del cofactor SAM, lo que es posible apreciar en los cuatro quimiotipo no nucleosídicos mencionados en el texto. Sobre estos núcleos es conveniente realizar otro tipo de estudios computacionales basados en la estructura cristalográfica de DNMT1 como es el caso de acoplamiento molecular rígido y flexible-flexible, lo que puede sentar las bases atomísticas de la interacción ligando proteína y por lo tanto guiar campañas experimentales de optimización que pretendan explorar más a fondo las relaciones estructura-actividad de estos inhibidores prometedores.

Para obtener información más detallada de este trabajo se puede consultar la siguiente referencia:

Fernández-de Gortari, E., Medina-Franco, J.L. et al. “Relevant epigenetic chemical space: a chemoinformatic characterization of small molecules: DNMT inhibitors”. Publicado en RCS. Gortari, E. and Medina-Franco, J. (2015). Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. RSC Adv., 5(106), 87465-87476.

Capítulo 2. Farmacóforo

Metodología computacional

A)

Utilizando el gráfico de factor de enriquecimiento contra frecuencia de núcleos base como criterio, se seleccionó a los compuestos pertenecientes a quimiotipos no nucleosídico más sobresalientes como punto de partida para realizar acoplamiento molecular según la metodología presente en el programa *Internal Coordinate Mechanics* (ICM).[36] El acoplamiento molecular se realizó sobre la estructura cristalizada de DNMT1 con código PDB 3PTA en su forma activa dentro de un volumen que comprende el sitio del sustrato y del cofactor en ausencia del cofactor SAM. [65]

B)

Para cada uno de los modos de unión encontrados se calcularon las huellas digitales de interacción ligando receptor llamados *Protein Ligand Interaction Fingerprints* (PLIF) implementados en *Molecular Operating Environment* (MOE).[51] Los PLIF guardan información del tipo de interacción y residuos de aminoácido involucrados en el reconocimiento ligando-proteína. Los resultados se pueden analizar por medio de histogramas y gráficos de barras, incluyendo el tipo de interacción y su frecuencia respecto a los residuos con los que interactúa.

C)

Se seleccionaron a aquellos núcleos base que presentaron los compuestos con mayor frecuencia de interacción para generar modelos del farmacóforo. Este análisis se realizó tanto para el total de inhibidores seleccionados, como para cada quimiotipo por separado.

D)

Cada uno de los modelos del farmacóforo obtenidos fue probado mediante la implementación de una búsqueda basada en dichos modelos. Los *hits* obtenidos fueron clasificados con falsos positivos, positivos verdaderos, falsos negativos y negativos verdaderos para construir la matriz de confusión y obtener la

sensibilidad y selectividad de cada uno de ellos. Con estos valores se construyó el espacio ROC[66] correspondiente para la validación y selección de los mejores modelos que servirán como base para posteriores campañas de tamizado virtual en bases de datos externas a la utilizada para la generación de los modelos.

Resultados y discusión

El acoplamiento molecular para cada uno de los compuestos pertenecientes a los quimiotipos seleccionados fue realizado en por medio del programa ICM con dos metodologías diferentes: acoplamiento flexible-flexible y acoplamiento rígido-flexible. La metodología computacional que usa este paquete de programas está basada en la optimización de las coordenadas internas del ligando dentro de un mapa de potenciales mediante algoritmos Monte Carlo. [67]

El sitio de acoplamiento fue definido alrededor de la cisteína catalítica (Cys1226) de la DNMT1 incluyendo al sitio del substrato y cofactor en ausencia de SAM (**Figura 2.1.**). Esta estructura cristalina está reportada en el PDB con el código 3PTA. La estructura fue preparada y equilibrada, además de ser sujeta a estudios de dinámica molecular para obtener una posible conformación activa.[68] La conformación activa se caracteriza por la presencia un asa relacionada con la interacción y reconocimiento de DNA en su forma abierta.

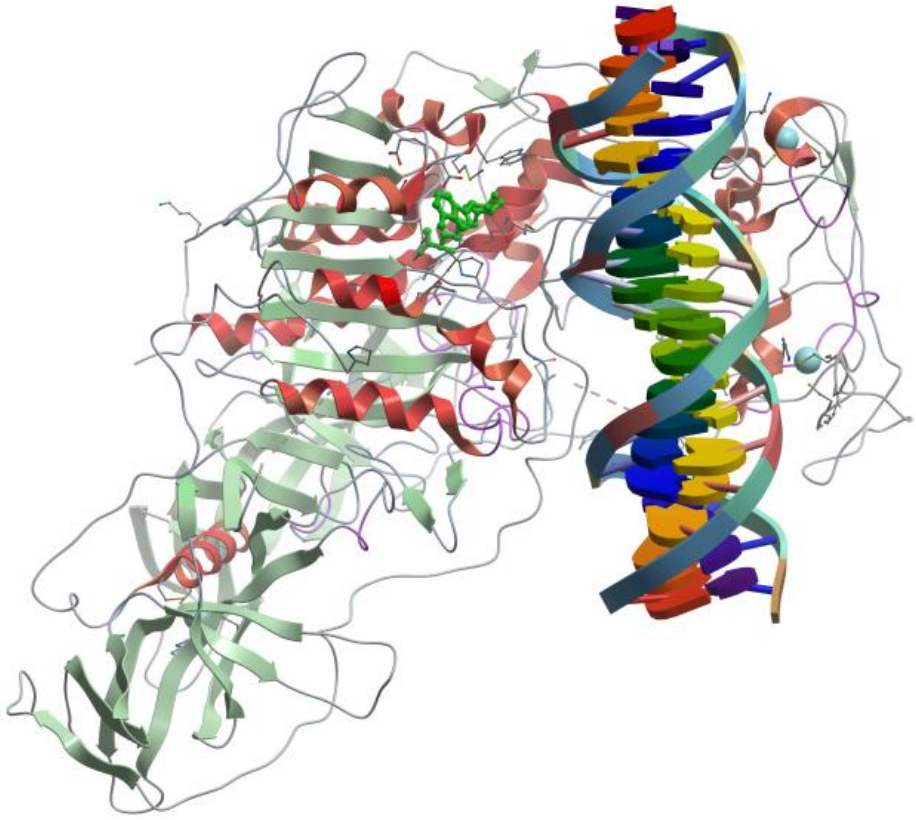


Figura2.1. Estructura cristalográfica de la enzima DNMT1 (PDB ID: 3PTA).

Para asegurar la convergencia de los resultados obtenidos por el algoritmo de optimización, se realizaron tres ciclos de acoplamiento y se seleccionaron moléculas de acuerdo a los valores de la función de puntuación y conformación.

Los compuestos seleccionados sirvieron de base para calcular los PLIF implementados en MOE. Este método describe las interacciones entre el ligando y el receptor utilizando una huella digital molecular que consta de seis tipos de interacciones: donador y aceptor de hidrógeno de cadena lateral, aceptor y donador del esqueleto de la proteína, interacciones iónicas y de superficie. En este método de representación los puentes de hidrógeno se determinan a través de la media de distribuciones obtenidas del análisis de este tipo de interacciones en sistemas proteicos, donde se determina un porcentaje de probabilidad de acuerdo a los valores de longitud y ángulo de enlace de los átomos que participan en la

interacción. Las interacciones iónicas se aproximan con la fuerza de Coulomb de átomos que presentan cargas formales contrarias, mientras que las interacciones dependientes de la superficie se obtienen por medio del cálculo de la exposición de los residuos frente al disolvente.[69]

Los PLIF contienen la información necesaria para generar referentes de búsqueda basados en el farmacóforo de cada uno de los núcleos base seleccionados. Para realizarlo se utilizaron dos esquemas diferentes llamados *Unified* y *PPCH_All*, respectivamente. Los esquemas son un grupo de reglas que determinan el número y tipo de puntos farmacofóricos que satisfacen la interacción observada entre el ligando y su diana. Cada uno de estos esquemas contiene una serie de elementos basados en la interacción ligando proteína.

Estos dos esquemas se pueden separar en: 1) anotaciones atómicas, que describen directamente a un átomo según el tipo de interacción que presenta en el sistema, 2) anotaciones del centroide, que determina el centro geométrico de un conjunto de átomos, 3) anotaciones proyectadas, que determinan las posibles interacciones entre dos átomos. El esquema llamado *Unified* incluye también proyecciones para aceptores y donadores de puentes de hidrógeno, distinciones generales entre sistemas π y no π, ligandos metálicos, aniones, cationes, bioisósteros NCN +, bioisósteros COO-, centroides aromáticos y grupos R, mientras que *PPCH_All* sólo incluye algunas de ellas más anotaciones hidrofóbicas.[69]

El grupo de modelos así obtenidos para cada uno de los quimiotipos fueron utilizados como criterio de búsqueda dentro de una base de datos de confórmeros de los inhibidores presentes en la base de datos de DNMT1. Los resultados se analizaron para determinar el número de falsos positivos, falsos negativos y el número de positivos y negativos verdaderos según un límite de actividad de 10 μM. Para cada modelo se construyó la matriz de confusión que contiene la información necesaria para determinar la especificidad y la sensibilidad de cada hipótesis (**Figura2.2**).

$$Sensibilidad = \frac{PV}{P} = \frac{PV}{PV + FN}$$

$$Especificidad = \frac{FV}{N} = \frac{FV}{FV + FP}$$

Figura2.2. Expresiones para especificidad (*true negative rate*) y sensibilidad (*true positive rate*) de los dos parámetros de ROC. P: positivos, N: negativos, PV: positivos verdaderos, FN: falsos negativos, FV: falsos verdaderos, FP: falsos positivos.

Con dichos elementos se determinó el espacio ROC, el cual fue interpretado para seleccionar la mejor hipótesis para implementarla en metodologías de búsqueda de moléculas bioactivas.

En la **Figura 2.3.** se resumen los pasos realizados dentro de la metodología:

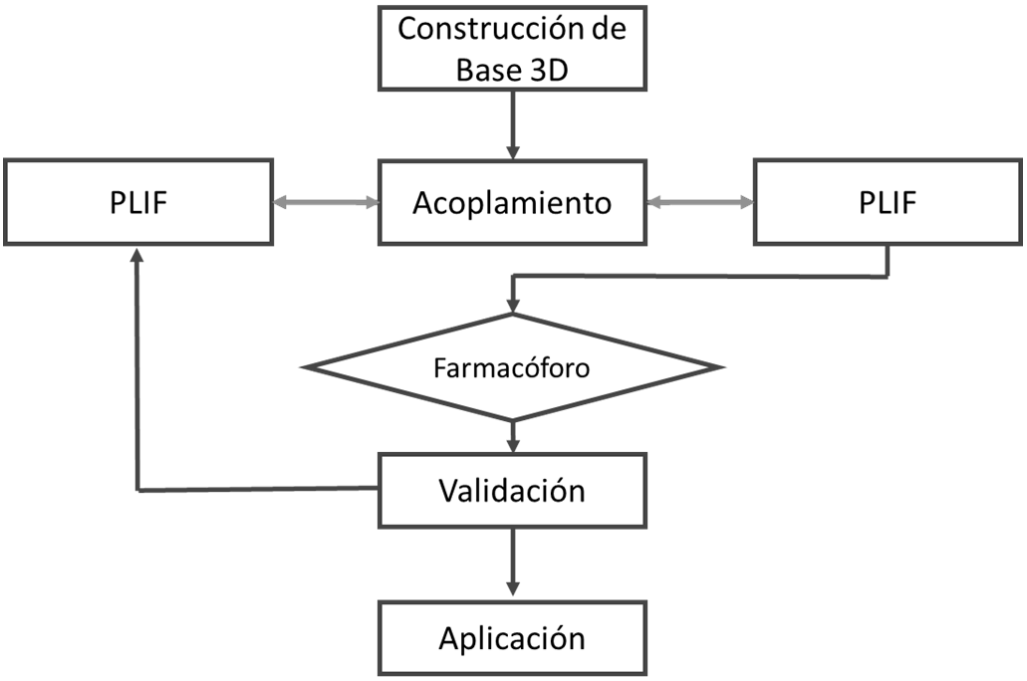


Figura 2.3. Metodología seguida para obtener el modelo del farmacóforo.

Acoplamiento molecular

Después de depurar la información de los compuestos con un núcleo base común, se realizó el acoplamiento molecular según la metodología implementada en el programa ICM. Las estructuras con una conformación de interacción acorde a lo reportado (**Figura 2.4.**) y con los mejores valores para la función de puntuación fueron seleccionadas como punto de partida para la siguiente fase del modelado.

Esta etapa es de suma importancia, ya que no existe evidencia experimental sobre el mecanismo de acción de los inhibidores. Por ejemplo, en los casos donde el análisis de actividad es funcional, es probable que la actividad desmetilante esté dada por mecanismos donde DNMT1 no participa.

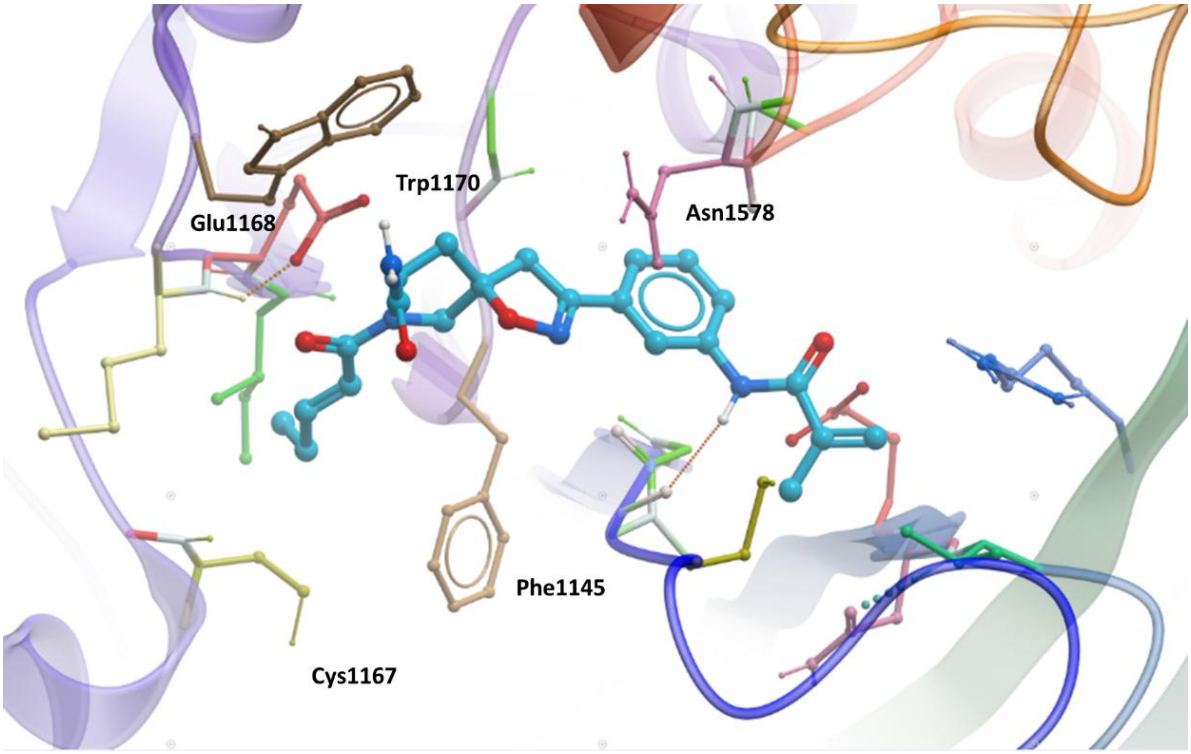


Figura 2.4. Ejemplo de conformación obtenida por acoplamiento molecular de un inhibidor no nucleosídico en el sitio catalítico de DNMT1.

El acoplamiento molecular se llevó a cabo utilizando la estructura cristalográfica reportada en el *Protein Data Bank* con el código 3PTA, lo que representa una diferencia sustancial respecto a otros trabajos donde se realizaron esfuerzos para desarrollar modelos del farmacóforo de los inhibidores de DNMT1.[70,68] Otras de las

grandes diferencias respecto a trabajos similares, es el uso de información molecular actualizada y su tratamiento quimioinformático para la descripción del espacio químico. Esto último, utilizando descriptores moleculares, métricas de similitud y análisis de núcleos bases. Además de que la metodología utilizada para desarrollar los modelos del farmacóforo difieren. En este trabajo se hizo uso de representaciones moleculares que caracterizan el número y tipo de interacciones entre los diferentes compuesto y su diana biológica para extraer aquellas que son preponderantes para la actividad biológica de los inhibidores. A su vez se utilizaron estos modelos para determinar las variables que conforman el espacio ROC para ser utilizadas como criterio de evaluación y selección de los modelos que serán utilizados para futuras campañas de cribado virtual.

PLIF y generación de los modelos del farmacóforo

Con la información contenida en los PLIF para cada uno de los compuestos con un núcleo base en común y con un factor de enriquecimiento alto, respecto al conjunto de los núcleos base considerados, se analizaron las frecuencias y el código de barras de interacciones para determinar aquellas que son comunes dentro de cada familia de compuestos y en el total de ellas, así como los aminoácidos del sitio de unión relacionados con dicha asociación. (Ver **Figura 2.5**)

Una vez que dichas relaciones fueron determinadas se generaron los modelos del farmacóforo por medio de los esquemas *Uniffied* y *PPCH_All*. La idea inicial constaba de realizar un solo modelo a través de la mezcla de los elementos presentes en la poses en todos los compuestos de diferente quimiotipo. Sin embargo, esto no fue posible ya que la diferencia estructural entre los análogos que pertenecen a diferentes núcleos base difiere significativamente. Es decir, la diversidad estructural de la colección es grande y por lo tanto la probabilidad de encontrar conformaciones que presenten interacciones comunes disminuye considerablemente.

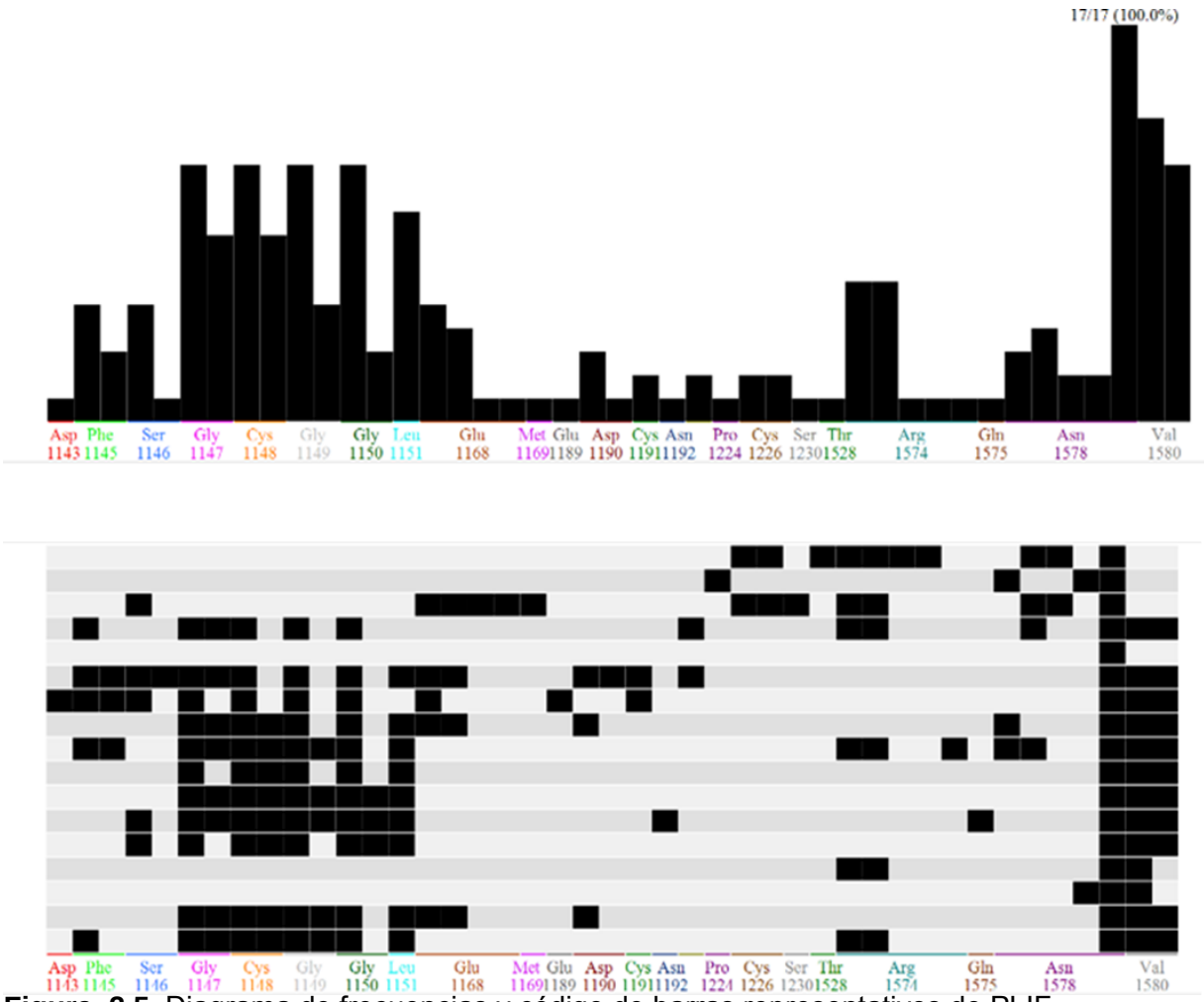


Figura. 2.5. Diagrama de frecuencias y código de barras representativos de PLIF.

Debido a la gran diversidad estructural de los compuestos reportados, se procedió a realizar un modelo del farmacóforo para cada uno de los conjuntos de compuestos con quimiotipo en común. (**Figura 2.6.**) De ellos se obtuvo un modelo con el máximo número de elementos farmacofóricos, así como un modelo para cada detrimento secuencial de elementos. Esto se realizó para determinar en fases posteriores cuáles o cuál de estos elementos tiene la mayor contribución al momento de discriminar compuestos inactivos de activos.

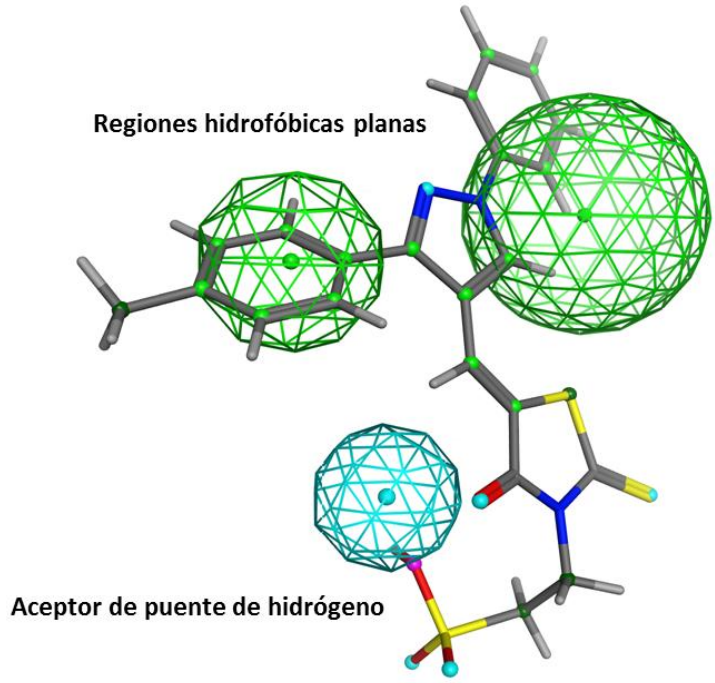


Figura 2.6. Farmacóforo obtenido utilizando el quimiotipo RNDWX con el esquema PPCH_ALL. Las esferas verdes representan regiones hidrofóbicas planas, mientras que la esfera azul representa un aceptor de puente de hidrógeno.

Análisis del modelo del farmacóforo

Todos los modelos generados en el paso anterior fueron utilizados como criterio de búsqueda en dos bases de datos distintas. La primera base de datos es la biblioteca de inhibidores original de donde se seleccionaron los compuestos base para la modelación.[15] La segunda consta de 25 confórmeros de baja energía para cada una de las moléculas presentes en la base de datos original. Esto es necesario, ya que el modelo del farmacóforo toma en cuenta la disposición espacial de sus elementos para determinar si existe similitud con la el modo de unión del compuesto con el que se está comparando. Es decir, pueden existir conformaciones de compuestos activos que no sean identificados por el modelo al no presentar un arreglo tridimensional adecuado.

Los resultados obtenidos de las diferentes búsquedas fueron catalogados en cuatro grupos: falsos positivos, falsos negativos, negativos verdaderos y positivos verdaderos.

La proporción existente entre estos resultados permitió obtener los parámetros de sensibilidad y selectividad de cada uno de los modelos propuestos. Los valores de selectividad o especificidad dan cuenta de la capacidad del modelo para distinguir compuestos activos en bases de datos, mientras que la sensibilidad indica la capacidad del modelo para discriminar a compuestos inactivos o negativos verdaderos. Es importante recordar que la diferenciación entre activos e inactivos depende del límite de actividad seleccionado (10μM en este trabajo), lo cual a su vez se ve determinado por la distribución de actividades de la base de datos, el objetivo de la investigación y los recursos disponibles.

Modelo	Hits	Activos	PV	FP	FN	NV	Sensibilidad	Selectividad	Precisión
SU1Uni	0	0	0	0	314	251	0.00	1.00	0.00
SU2Uni	178	80	80	98	234	153	0.25	0.61	0.39
SU3Uni	191	90	90	101	224	150	0.29	0.60	0.40
SU4Uni	181	89	89	92	225	159	0.28	0.63	0.37
SU5Uni	310	184	184	126	130	125	0.59	0.50	0.50
SU6Uni	0	0	0	0	314	251	0.00	1.00	0.00
SU1All	1	2	2	0	312	251	0.01	1.00	1.00
SU2All	51	18	18	33	296	218	0.06	0.87	0.35
SU3All	185	118	118	67	196	184	0.38	0.73	0.64
SU4All	2	2	2	0	312	251	0.01	1.00	1.00
SU5All	21	9	9	12	305	239	0.03	0.95	0.43
SU6All	15	2	2	13	312	238	0.01	0.95	0.13

Tabla. 2.1. Matriz de confusión. PV: positivos verdaderos, FP: falsos positivos, FN: falsos negativos, NV: negativos verdaderos.

La combinación de estos valores genera el llamado espacio ROC. Este espacio está dividido por una línea de 45° que delimita a aquellos modelos que tienen mayor capacidad de discriminar moléculas activas respecto a una selección

aleatoria. De esta manera, los modelos que se encuentran la parte superior serán aquellos con mayor capacidad para distinguir a compuestos inactivos, mientras que los que encuentran hacia valores situados a la derecha del espacio presentarán mayor capacidad para identificar compuestos activos. Cualquier modelo que caiga debajo de la diagonal que divide al espacio ROC tendrá una capacidad igual o menor respecto a una búsqueda aleatoria.

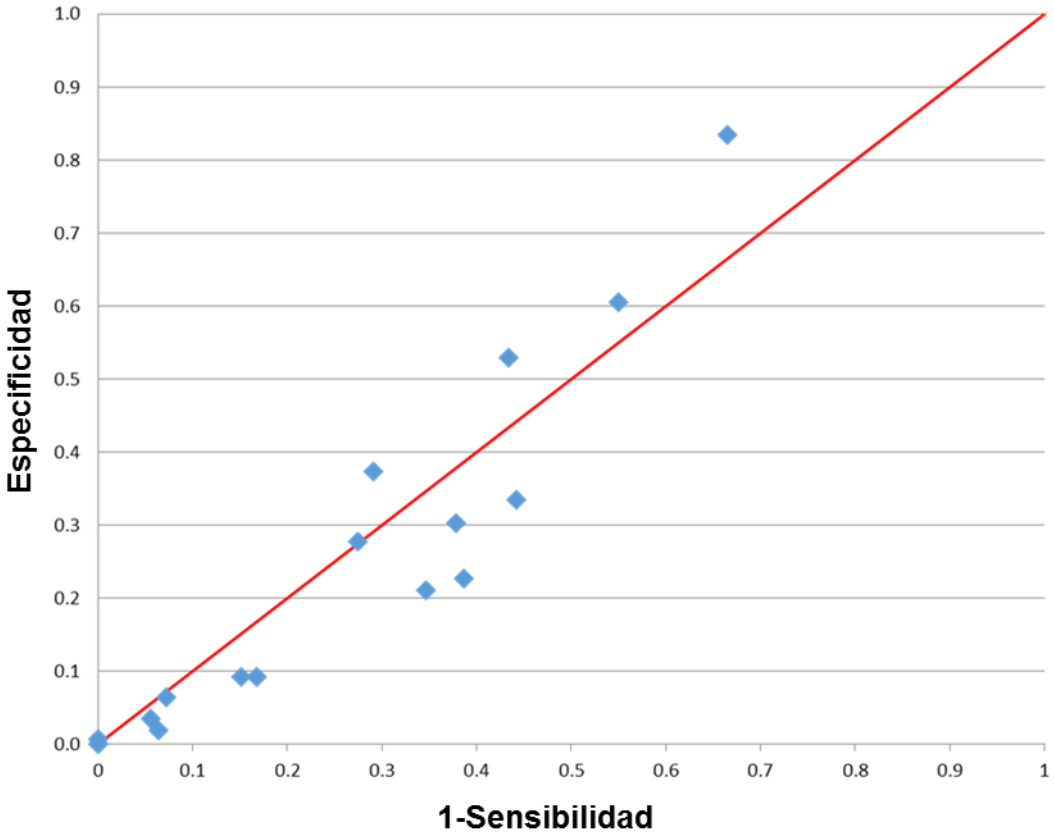


Figura. 2.7. Espacio ROC (*Receiver Operating Characteristic*).

En nuestro caso se decidió seleccionar a los modelos dentro del área superior izquierda (mayor sensibilidad) de la **Figura 2.8**. Seguido de ello se realizó una caracterización del tipo de compuesto (nucleosídico o no nucleosídico) así como el número de análogos, favoreciendo a aquellos que identifican a la mayor cantidad de compuestos con estructura no nucleosídica.

El mejor de los modelos encontrados (mayor sensibilidad) fue realizado con el esquema *PPCH_All*, con un radio máximo de 3 Å y una cobertura de 50%. El modelo fue extraído de la familia de compuestos con código quimitípico RNDWX. Debe mencionarse que con el esquema *Uniffied* ningún elemento farmacofórico consenso fue identificado.

El modelo seleccionado consta de tres puntos farmacofóricos: una región hidrofóbica plana con un radio promedio de 1.8 Å, un aceptor de puente de hidrógeno y ligando metálico con radio promedio de 1.4 Å, y otra región hidrofóbica plana con radio promedio de 2.6 Å (**Figura. 2.9.**).

También se seleccionó la hipótesis con la mejor selectividad, el cual fue propuesto utilizando a la familia de compuestos con quimiotipo SU70D. El modelo también fue realizado con el esquema *PPCH_All* con un radio máximo de 3 Å y una cobertura del 50%. Este modelo consiste en cuatro puntos farmacofóricos: una región hidrofóbica plana con radio promedio de 2.6 Å, una región no planar de radio promedio de 1.4 Å y dos regiones de aceptores de puente de hidrógeno planos con radio promedio de 1.3 Å y 0.8 Å respectivamente.

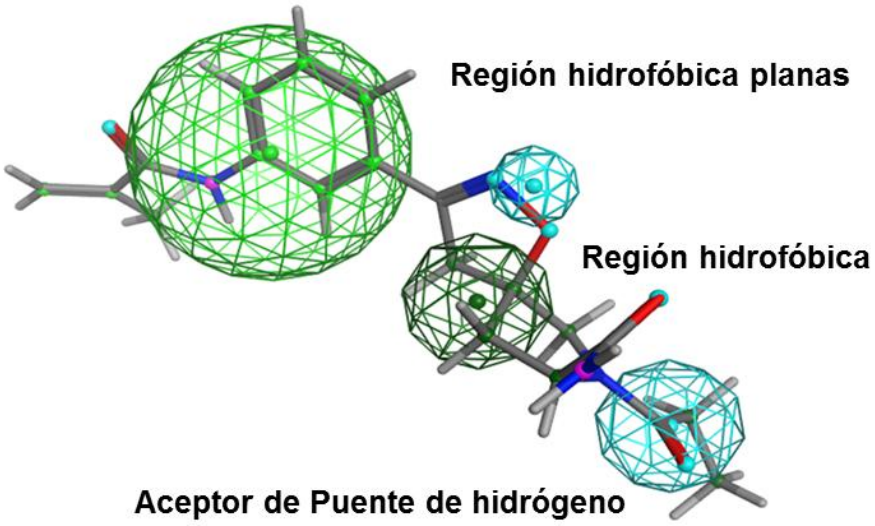
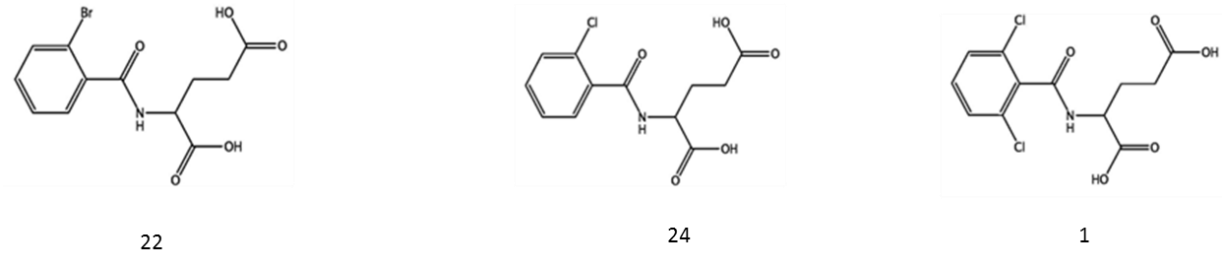


Figura. 2.8. Farmacóforo obtenido utilizando el quimiotipo SU70D con el esquema PPCH_ALL. La esfera verde claro representa una región hidrofóbica plana, mientras que las esferas azules representan aceptores de puente de hidrógeno y la esfera verde una región hidrofóbica.

Validación externa

Recientemente el laboratorio del Dr. Massimo Bertinaria de la Universidad de Torino en colaboración con la Universidad Goethe probó una serie de análogos del compuesto NSC137546, el cual fue originalmente detectado como inhibidor de DNMT por medio un tamizado virtual de una base de datos molecular realizado en el National Cancer Institute (NCI) de E.E.U.U.A. Este inhibidor fue propuesto como un inhibidor reversible del sitio activo de DNMT1, presentando una selectividad moderada contra DNMT1 frente a DNMT3B en una concentración de 100μM.[71]

De los análogos sintetizados por el laboratorio del Dr. Massimo Bertinaria, dos compuestos fueron encontrados con mayor actividad que NSC137546. En particular, de los análogos sintetizados el compuesto referido como **22** (**Figura 2.10.**) fue identificado como el más activo, presenta inhibición de la metilación de DNA mediada por DNMT1 y DNMT3A en una concentración dentro del intervalo de concentración de micro molar. También se caracterizó la inhibición aislada de DNMT1 y DNMT 3A por inmunoprecipitación. Además el compuesto **22** presenta estabilidad dentro de condiciones fisiológicas en suero humano. [72]



22 24 1

Figura 2.9. Compuesto **1**, **22** y **24**.[72]

Basados en esta información se hizo la validación externa de los dos modelos del farmacóforo seleccionados en el espacio ROC para probar su capacidad de reconocimiento y discriminación de compuestos activos contra DNMT1 en un conjunto de compuestos externo a la modelación. Para realizarlo utilizamos una base de datos que incluía análogos de NSC137546 (**Figura 5.1.**) reportados por el Dr. Bertinaria. Fue grato encontrar que los modelos fueron capaces de identificar a los análogos de NSC137546 así como al propio NSC137546 y al compuesto **22**.

También se identificó al compuesto **24**, el cual es un inhibidor de DNMT1 con una actividad intermedia entre el compuesto **22** y NSC137546.

Estos resultados muestran de forma preliminar la capacidad de los modelos como un criterio de búsqueda suficientemente robusto para realizar campañas de tamizado molecular en diferentes bases de datos moleculares para la identificación de inhibidores no nucleosídicos de DNMT1.

Tamizado preliminar basado en farmacóforo

Se realizaron dos búsquedas basadas en los modelos seleccionados en los pasos anteriores sobre dos bases de datos moleculares diferentes: una colección de compuestos aprobados por la FDA de E.E.U.U.A. con 1490 compuestos obtenidos del repositorio público *Drug Bank* y una base de datos formada de tres colecciones de compuestos anticancerígenos aprobados también por la FDA para su uso clínico. Este conjunto de compuestos se encuentra disponible como platos para tamizado de la *Division of Cancer Treatment & Diagnosis of the National Cancer Institute U.S.A*. con los identificadores ID: AOD6_Plate4825, AODIV_Plate1 y 2, y ApprovedOncDrugs_5_PlateMap con 119, 114 y 95 compuestos respectivamente.[73] En la **Tabla 2.1.** se puede revisar los resultados de la búsqueda sobre estas colecciones moleculares.

Base de datos	RN_All Hits	SU3_All Hits
AOD_Plate4825	72	34
AODIV_Plate1 and 2	52	28
ApprovedOncDrugs_5_PlateMap_AOD	67	33

Tabla 2.2. Resultados del tamizado basado en farmacóforo.

Para la búsqueda realizada sobre el conjunto de compuestos oncológicos se identificaron moléculas como el Celecobix, Femara, Sprycel y Xallkori como consenso para ambos modelos. En especial se puede decir que la cantidad de *hits* encontrados por el modelo SU3_All refleja su selectividad frente al modelo RN_All. Sin embargo no es evidente la relación estructural y espacial entre las moléculas identificadas y el núcleo base en el que fueron basados. En contraste, las

moléculas identificadas por el modelo RN_All muestran en muchos casos relaciones estructurales claras con el quimiotipo de referencia. Esto puede deberse a la novedad del sistema cíclico en el cual están basados.

En cuanto a las moléculas identificadas sobre la base de datos de compuestos aprobados, se pretende utilizarlas como punto de partida en investigaciones que tengan como objetivo el reposicionamiento de fármacos.

Resumen de resultados

En este capítulo de la tesis se presenta el desarrollo de dos modelos del farmacóforo para la identificación de inhibidores de DNMT1 basado en la caracterización inicial del espacio químico. Dicha caracterización se realizó por medio de tres métricas: propiedades fisicoquímicas y descriptores moleculares (espacio de propiedades), métricas de similitud con huellas dactilares moleculares y el contenido y análisis de núcleos base. Todo ello dentro de la base de datos construida mediante la información recabada en bases de datos públicas y literatura científica actual de inhibidores de DNMT1. Por medio de la caracterización de los núcleos, su frecuencia y factor de enriquecimiento, se seleccionaron a aquellos sistemas cíclicos con mejores características (actividad y estructura no nucleosídica) para ser punto de partida para la obtención del modelo farmacofórico.

Las familias de compuestos seleccionadas fueron sujetas a acoplamiento molecular sobre el sitio activo y del cofactor en ausencia de SAM de una estructura cristalina de DNMT1 en una posible conformación activa como una primera aproximación para encontrar la conformación activa de estos inhibidores. De todas las moléculas seleccionadas y acopladas al receptor se seleccionaron a aquellas con conformaciones razonables y mejores valores de score como punto de partida para los pasos subsecuentes de la metodología.

Para caracterizar el tipo de interacción presente en los compuestos seleccionados en el paso anterior, se calcularon los PLIF de cada una de estas conformaciones.

La información de interacción fue utilizada para generar modelos del farmacóforo con diferente número de elementos. Los modelos se utilizaron como

referencia para búsquedas dentro de la base de datos original y en una base de datos de confórmeros de los inhibidores de DNMT1, favoreciendo a aquellos que identificaban hits no nucleosídicos y con alta diversidad. Estos modelos fueron caracterizados y seleccionados por medio del espacio ROC calculado con un límite de corte sobre la actividad de 10μM y posterior inspección visual de los compuestos obtenidos.

El mejor modelo obtenido fue el RN_All que fue desarrollado bajo el esquema *PPCH_All* con la familia de compuestos con quimiotipo RNDWX con un radio de 3Å y una cobertura del 50%. Este modelo consiste en tres elementos: una región hidrofóbica plana con un radio promedio de 1.8 Å, un aceptor de puente de hidrógeno y ligando metálico con radio promedio de 1.4 Å, y otra región hidrofóbica plana con radio promedio de 2.6 Å.

Como se demostró, este modelo puede ser utilizado en investigaciones posteriores como parte integral de metodologías de tamizado virtual o como criterio de búsqueda en campañas de *scaffold hopping* para la identificación de nuevos inhibidores de DNMT1 no nucleosídicos. Hasta donde sabemos, este es el modelo de farmacóforo más reciente basado en la caracterización del espacio químico de inhibidores de DNMT1 por medio de métodos quimioinformáticos.

Capítulo 3. Identificación de sitios de unión de DNMT1

Metodología computacional

Esta metodología tiene como objetivo facilitar la búsqueda de *hits* computacionales con potencial afinidad sobre sitios alternativos de DNMT1 relacionados con la flexibilidad de la proteína. La metodología que se describe en este capítulo busca resolver la baja selectividad y alta toxicidad presente en los fármacos aprobados localizando sitios de unión alternativos relacionados con la respuesta biológica.

A)

Se determinó el grado de conservación de la secuencia de aminoácidos así como las diferencias de RMSD respecto a la estructura tridimensional PDB ID: 3PTA, utilizando las estructuras contenidas en el PDB por medio de servidor Dali.[42]

B)

Distintas metodologías computacionales fueron utilizadas para la identificación de sitios de unión. Por medio de AutoDock Ligand,[32] SPACER[74] y NMA PARS,[45] se localizaron y caracterizaron cavidades con características apropiadas para permitir la interacción con moléculas pequeñas. Los resultados consenso fueron seleccionados para realizar estudios de tamizado virtual basado en estructura.

C)

Se realizó cribado virtual ciego con las bases de datos de inhibidores de DNMT1 con los programas AutoDock Vina[53] y AutoDock 4.2.[32] Los sitios consenso identificados con ambos motores fueron seleccionados para realizar tamizado virtual consenso dirigido.

D)

Por medio de un criterio de selección basado en agrupamiento (*clustering)* y en valores de la función de puntuación consenso se determinaron las moléculas *hits*. Como segundo criterio de selección se identificaron a aquellas moléculas

presentes en el mismo sitio de unión al comparar los resultados de las dos metodologías de acoplamiento molecular aplicadas. (**Figura3.1**)

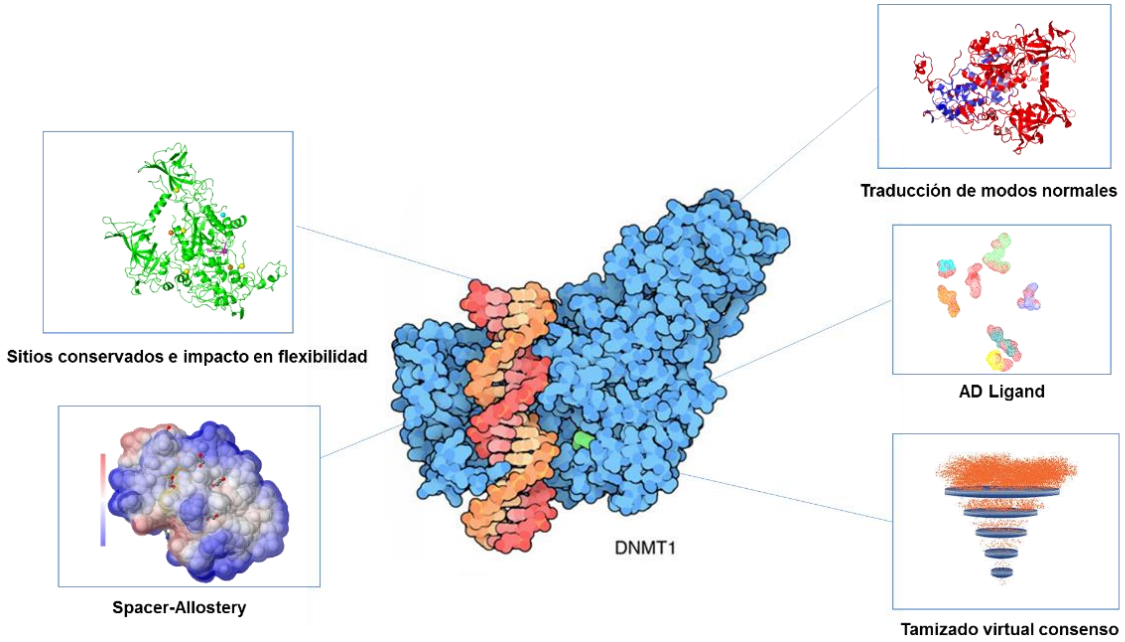


Figura3.1. Diagrama de flujo para la campaña de tamizado virtual basado en estructura.

Resultados y Discusión

Análisis estructural

Se analizó la estructura cristalográfica de la enzima DNMT1 (PDB ID: 3PTA). Por medio del servidor Dali se realizó la comparación de la estructura tridimensional de DNMT1 contra las estructuras reportadas en el PDB. Esta comparación se hizo con el propósito de encontrar proteínas similares que pudieran en principio ser afectados por inhibidores conocidos de DNMT1. La comparación de estructuras en tres dimensiones se hace más relevante cuando se encuentran valores de conservación de la secuencia menores al 40%, ya que cambios en la secuencia pueden resultar en arreglos tridimensionales similares y por lo tanto a funciones biológicas relacionadas.

De un grupo de 800 salidas se eligieron aquellas proteínas que presentaron valores de RMSD menor a 2.0 Å. Como segundo elemento de exclusión se eligieron a las estructuras exclusivas de *Homo sapiens*. En la **Tabla 3.1.** se

muestran los resultados después de aplicar los criterios de selección mencionados.

Código PDB	Proteína	Función	RMSD
3PTA	DNMT1	Metilación	0
4DA4	DNMT1	Metilación	1.6
4H0N	DNMT2	Metilación	2.9
4DOW	Reconocimiento	Reconocimiento de histona H4 en Lys20	2.2
1VI5	DNMT	Metilación en Bacillus	2.7
2FVU	Dom. SIR3/BAH	Silenciamiento de transcripción	2.7

Tabla 3.1. Código de 4 letras, proteína, función y RMSD de las proteínas representativas para cada familia en Angstroms.

Al mismo tiempo, es importante obtener información sobre la actividad biológica reportada. Esta información puede sugerir interacciones entre los inhibidores del blanco de interés y dianas que se encuentran relacionadas con el metabolismo de fármacos, lo cual anticipa posibles efectos adversos.

Afortunadamente las diferencias estructurales entre DNMT1 y las estructuras reportadas en el PDB son suficientemente grandes. Aunque esta información no es concluyente aumenta las posibilidades de encontrar moléculas que interacciones con sitios únicos de DNMT1. Esto cobra importancia si se toma en cuenta la gran similitud entre el sitio catalítico y del cofactor de las diferentes enzimas desmetilantes.

Identificación de sitios de unión

Una vez determinada esta información se realizó una búsqueda de sitios sobre la estructura cristalográfica de DNMT1 utilizando distintas metodologías computacionales.

El programa computacional AutoLigand utiliza AutoGrid[75] para generar mapas de afinidad por medio del barrido de átomos sonda (carbono, oxígeno e hidrógeno) que recolectan la información de puntuación para cada punto de la

superficie de la proteína. A su vez, AutoLigand utiliza un algoritmo de relleno por difusión el cual rastrea todos los puntos de la superficie que no son parte de la estructura para determinar los volúmenes que no se traslapan. Una vez determinada la energía y volumen de los sitios encontrados selecciona a los que presentan valores de menor energía y volúmenes superiores a los 1000 $Å^3$. Este procedimiento da como resultado una colección de posibles sitios de unión caracterizados por regiones que sugieren la complementariedad con elementos moleculares. En la **Figura 3.2.** se muestran los sitios identificados para DNMT1 incluido el sitio catalítico y el del cofactor.

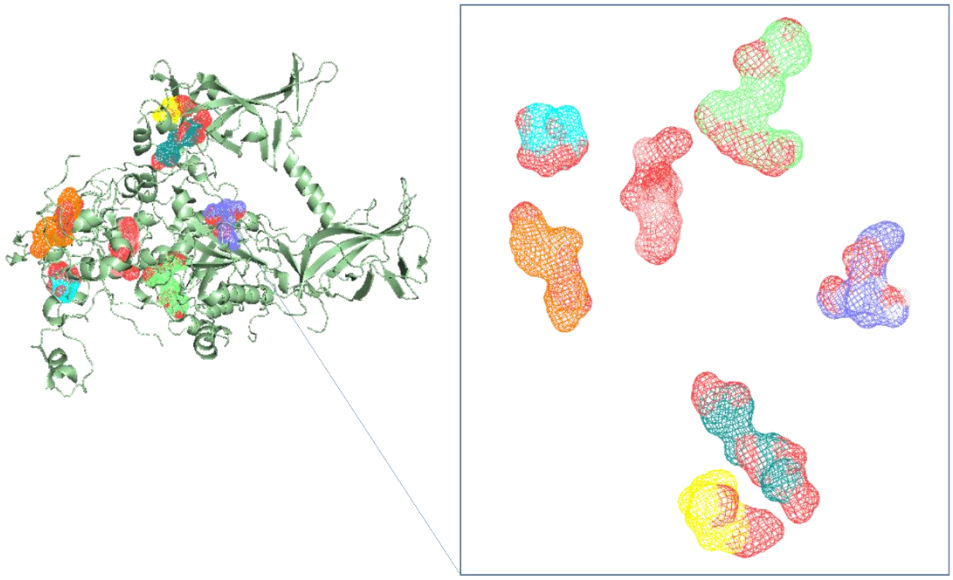


Figura 3.2. Sitios de unión de DNMT1 encontrados por AutoLigand. En rojo interacciones electrostáticas negativas, en blanco positivas y en otros colores hidrofóbicas.

Otra de las metodologías utilizadas fue PARS (*Protein Allosteric and Regulatory Sites*) desarrollada por Alejando Panjkovich y Xavier Daura[76,77] el cual se encuentra disponible en el sitio http://bioinf.uab.cat/cgi-bin/pars-cgi/pars.pl. A partir de información de 91 proteínas con actividad alostérica, esta metodología sostiene que los métodos convencionales de búsqueda de sitios de unión fallan en la identificación de sitios alostéricos ya que estos presentan formas planas respecto a los sitios primarios de interacción. Por este motivo el algoritmo PARS realiza una caracterización de los sitios identificados de acuerdo a la conservación estructural

y sus efectos en la flexibilidad de la proteína. Esto se logra introduciendo átomos de prueba (*dummy*) sobre los sitios identificados. La introducción de átomos de prueba también se puede hacer manualmente en caso de tener información sobre sitios de unión introduciendo ligandos en la estructura de entrada. Paso seguido se realiza un análisis de modos normales (NMA) de la forma holo y apo de la proteína. Los resultados obtenidos se transforman a factor B de los cuales se determina las diferencias entre las dos formas (holo y apo). A partir ellas se calcula el valor p, el cual a valores menores a 0.05 indica la presencia de perturbaciones grandes en la flexibilidad de la proteína.[45,77] (**Figura3.3.**)

RANK & SITE ID	FLEXIBILITY P-VALUE	STRUCTURAL CONSERVATION
1. CAV_1_Z	0.00	33.30
2. ZN_2_A	0.03	0.00
3. CAV_3_Z	0.03	0.00
4. CAV_5_Z	0.63	50.00
5. ZN_1_A	0.30	0.00
6. ZN_3_A	0.08	0.00
7. ZN_5_A	0.11	0.00
8. CAV_2_Z	0.21	0.00
9. CAV_6_Z	0.89	0.00
10. CAV_7_Z	0.29	0.00
11. CAV_8_Z	0.36	0.00
12. SAH_1601_A (CAT)	0.33	83.30
13. CAV_4_Z (CAT)	0.12	83.30

P<0.05

Figura 3.3. Sitios identificados por PARS. Violeta: sitio de cofactor, Azul: sitio activo, Amarillo: sitios sin interés respecto a su grado de conservación y relación con la flexibilidad, Rojo: sitios poco conservados y con gran influencia en la flexibilidad. [78]

Como se ha dicho esta metodología favorece a aquellos sitios con valores p bajos y un alto grado de conservación. En efecto, este criterio determinaría sitios

de unión alostéricos potenciales, ya que una de sus características es la conservación estructural de los elementos que llevan a cabo esta función. Sin embargo, el enfoque de este trabajo es la identificación de sitios que afecten la flexibilidad de la proteína, lo cual no necesariamente se encuentra previsto evolutivamente. Es decir, según el criterio seguido se deben favorecer aquellos sitios que tienen menor grado de conservación y valores p bajos para garantizar la selectividad y a su vez cambios en la flexibilidad que puedan dar como resultado variaciones en la actividad biológica.

A partir de los resultados se determinó que los sitios de mayor interés son los nombrados CAV_1_Z y CAV_3_Z (en adelante nombrados 1Z y 3Z). Como se observa en la **Figura 3.4.** estos sitios tienen un gran efecto en la flexibilidad (valores de p bajos) y un nivel de conservación bajo, lo cual los convierte en excelentes candidatos para el desarrollo de moléculas selectivas contra DNMT1.

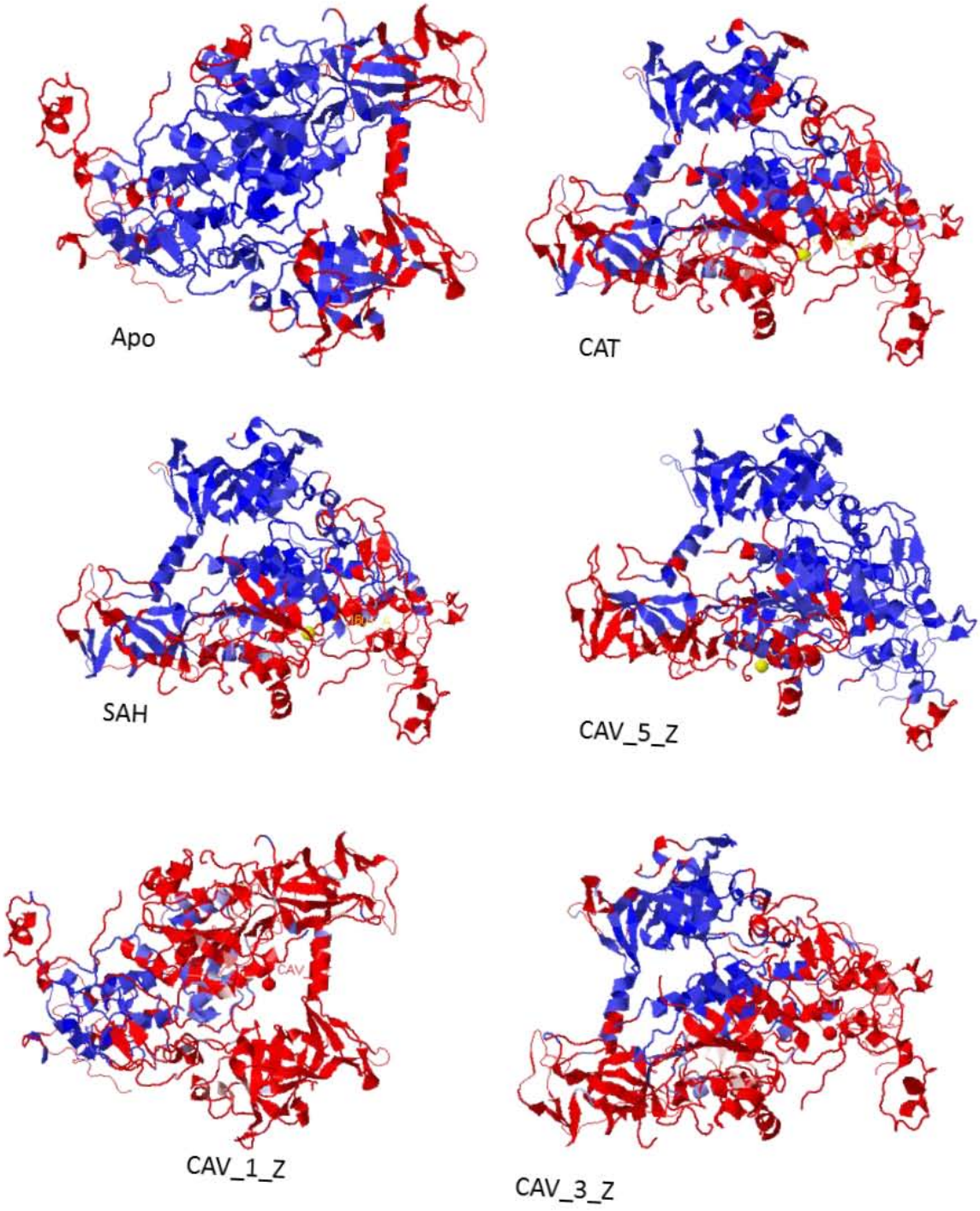


Figura 3.4. Traducción de NMA a factor B para cada uno de los sitios identificados por PARS.[78]

A la par de este análisis se utilizó el servidor SPACER (*Server for Predicting Allosteric Communication and Effects of Regulation*)[74] el cual permite analizar

posibles sitios de unión y determinar la existencia de comunicación estructural entre ellos. La metodología se basa en la aseveración de que cada proteína tiene un número determinado de grados de libertad que son capaces de describir las fluctuaciones conformacionales alrededor de la conformación nativa de la misma. De esta manera los movimientos que se dan sobre algún de estos grados de libertad deben estar interrelacionados con movimientos en otras zonas de la proteína. Se dice que existirá acoplamiento molecular en caso de que el efecto de la unión de un ligando sobre un sitio de la estructura modifique la afinidad de un segundo ligando en otra región de la estructura.[74]

Los resultados obtenidos para este análisis demuestran la existencia de un sitio de unión en la misma región donde AutoLigand y PARS lo identifican (sitio 1Z) mientras que el sitio 3Z no es reconocido. En la **Figura 3.5.** se muestra un modelo de la enzima con el sitio identificado. Este sitio (en color verde), aun cuando no tiene valores de correlación altos, se encuentra acoplado con el sitio del cofactor SAM (en color naranja). Este hecho aumenta las posibilidades de encontrar ligandos específicos para el sitio 1Z que tengan como resultado la desestabilización de la estructura funcional de la proteína y por lo tanto en su repuesta biológica.

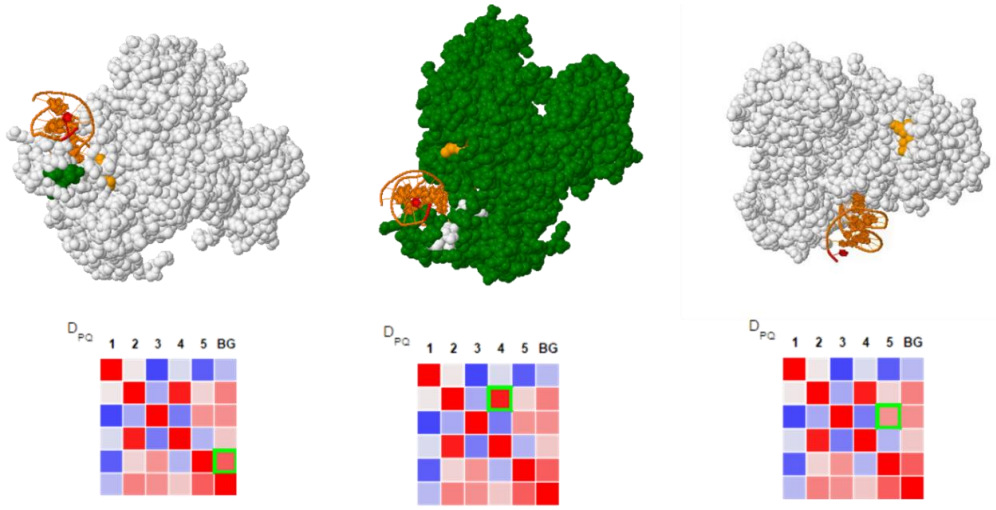


Figura 3.5. Matriz de conectividad entre sitios identificados (rojo mayor interacción, azul menor) y el correspondiente modelo de DNMT1 donde se muestra los sitios interactuantes de DNMT1 SPACER.[79]

Tamizado molecular

Como siguiente paso en la identificación de sitios de unión se utilizó la base de datos de inhibidores de DNMT1 como colección para realizar tamizado virtual basado en acoplamiento molecular utilizando AutoDock 4.2[32] y AutoDock Vina.[53] Como primer acercamiento se realizó el acoplamiento en serie sin favorecer ningún de los sitios antes identificados (tamizado virtual ciego). Esto se realizó para determinar si algunas de las moléculas presentes en la base de datos tienen una afinidad por alguno de los sitios identificados frente al sitio del catalítico o del cofactor.

En la **Figura 3.6.** se muestra que algunas moléculas presentan afinidad tanto para 1Z como para 3Z utilizando ambas metodologías de acoplamiento molecular.

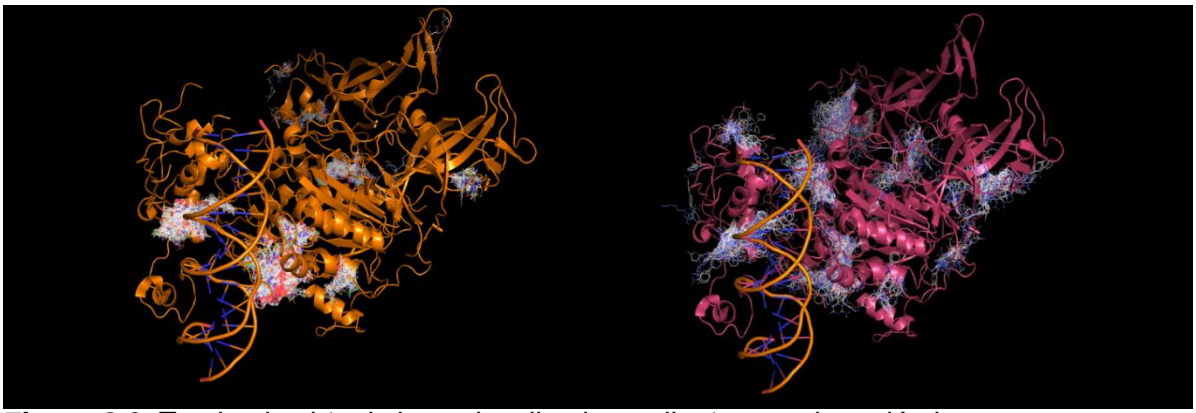


Figura 3.6. Tamizado virtual ciego visualizado mediante pymol académico.

Dichas moléculas fueron posteriormente seleccionadas para realizar tamizado virtual enfocado en los dos sitios de interés por medio de la combinación de ambas metodologías.

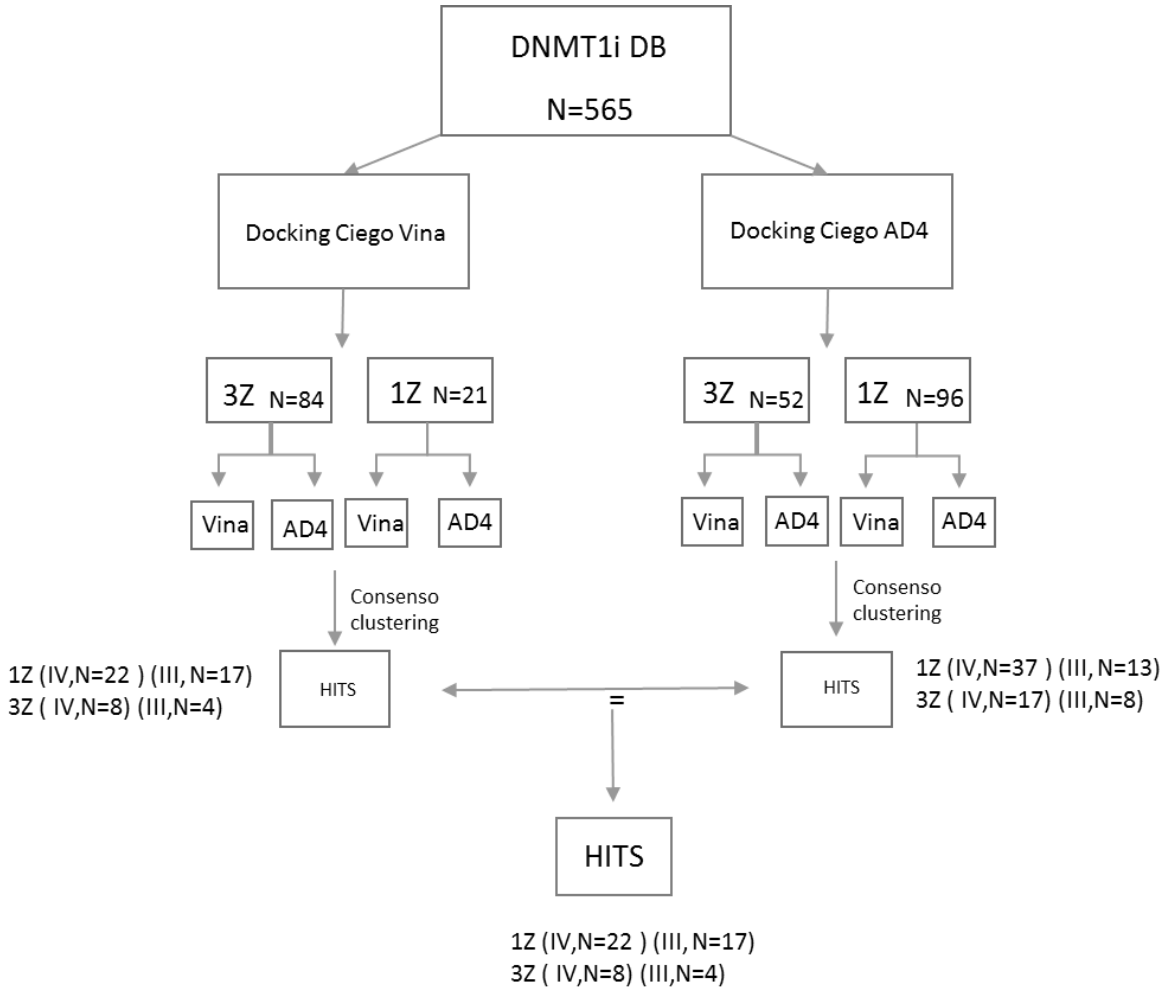
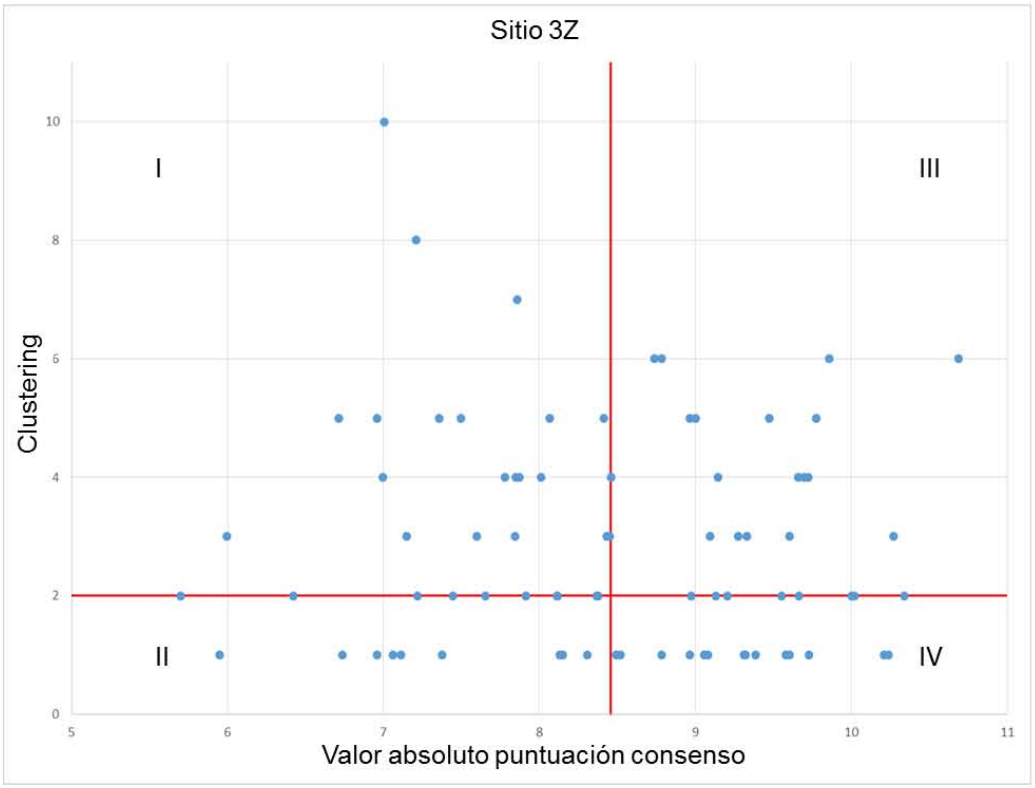


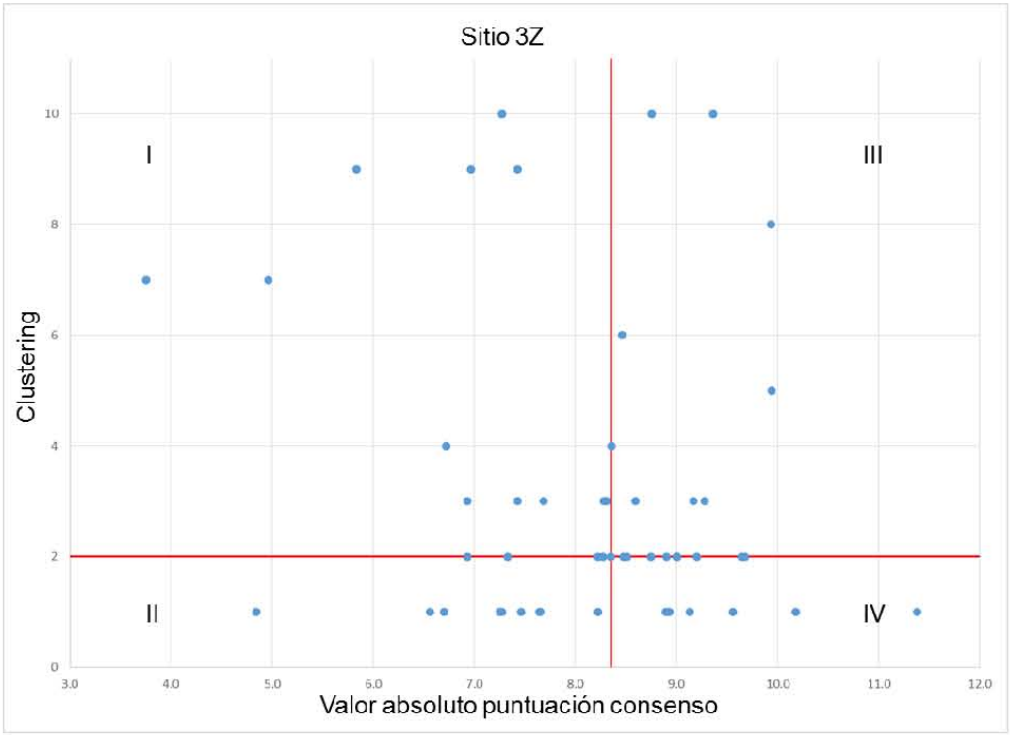
Figura 3.7. Metodología del tamizado virtual. Vina: AutoDock Vina, AD4: AutoDock 4.2, HITS: candidatos, consenso clustering: agrupamiento consenso, I, II, III, IV: regiones del plano de agrupamiento y score consenso.

La **Figura 3.7.** muestra 84 moléculas fueron identificadas para el sitio 3Z y 21 moléculas para el sitio 1Z utilizando el motor AutoDock Vina. Mientras que se obtienen 52 moléculas y 96 moléculas para cada sitio utilizando AutoDock 4.2. Cada grupo de moléculas fue acoplado utilizando ambos motores dentro del sitio donde se demostró afinidad previamente. Los resultados fueron analizados utilizando dos métricas: el promedio de los valores absolutos de la función de puntuación y el agrupamiento de las conformaciones de acuerdo a su RMSD, ver **Figura3.8**.

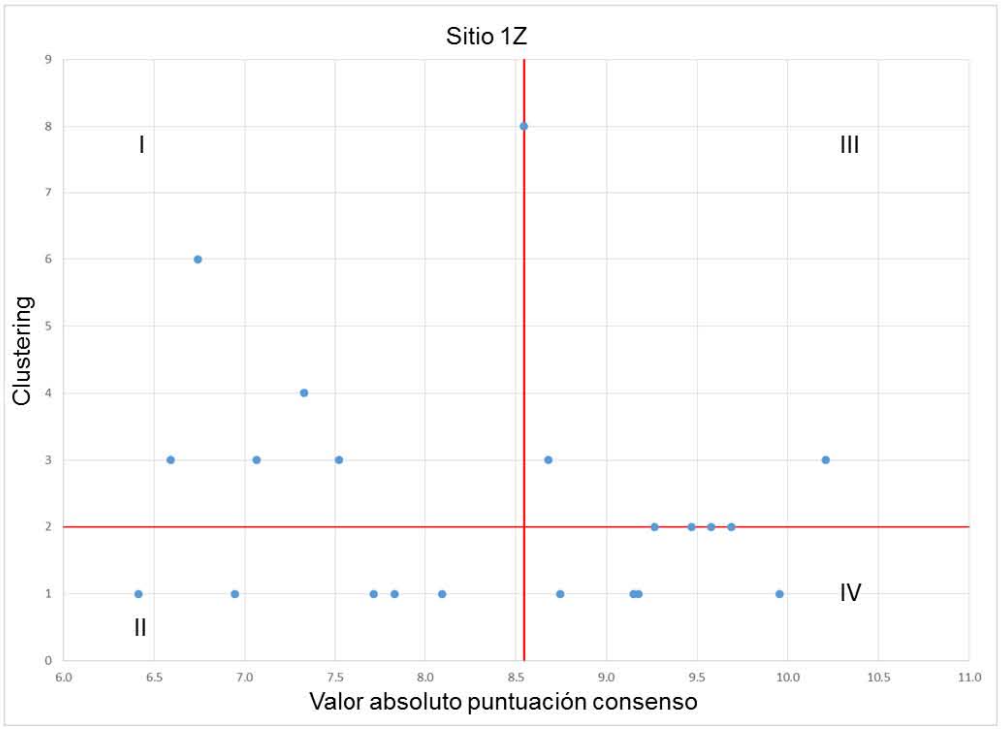
A)


Sitio 3Z
I
III
II
IV
Clustering
Valor absoluto puntuación consenso
0
2
4
6
8
10
5
6
7
8
9
10
11

B)


Sitio 3Z
I
III
II
IV
Clustering
Valor absoluto puntuación consenso
0
2
4
6
8
10
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0

C)



D)

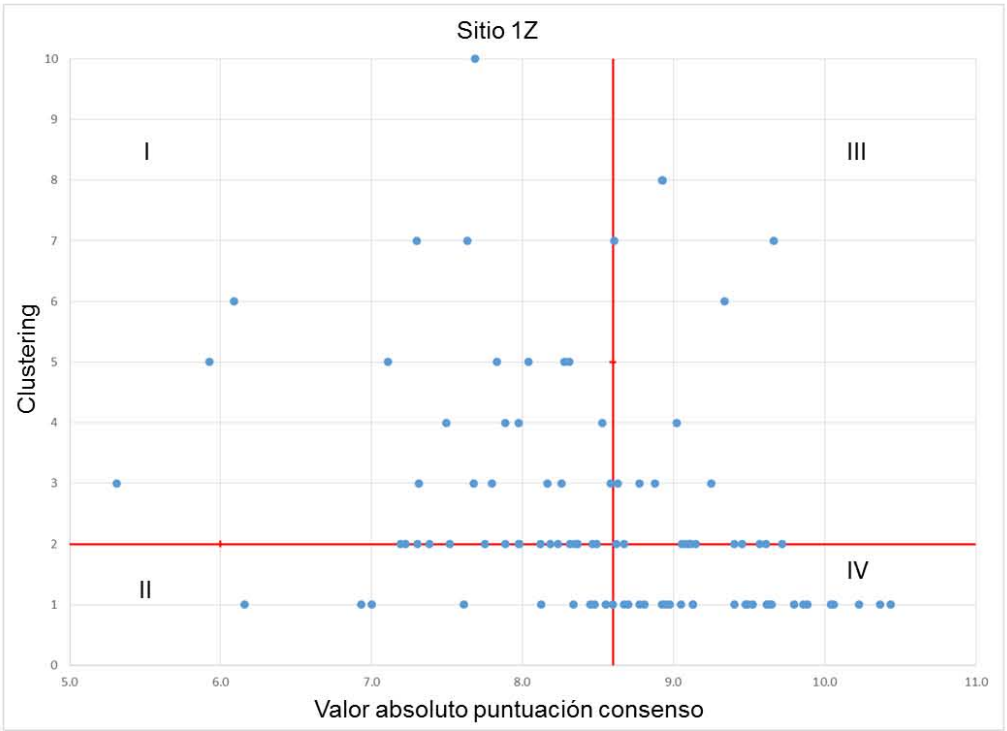


Figura 3.8. Gráficos de resultados del tamizado. Resultados obtenidos a partir de A: Vina ciego, B: AutoDock4.2 ciego, C: Vina ciego, D: AutoDock4.2 ciego.

Estos gráficos fueron divididos en cuatro regiones utilizando la mediana de las distribuciones formadas por ambas variables. La región I contiene moléculas con valores bajos de puntuación y con altos valores agrupamiento, lo que hace referencia a moléculas con pocos grados de libertad y pequeño tamaño, mientras que la región II contiene moléculas que presentan valores bajos de afinidad y agrupamiento por lo que no presentan interés. La región IV contiene a moléculas con un alto valor de puntuación y niveles bajos de agrupamiento, lo que normalmente está asociado a moléculas con gran número de grados de libertad y alto peso molecular, lo que resulta en posibles artefactos al aumentar el número de posibles interacciones. Por último, la región III contiene a las moléculas de mayor interés al presentar valores altos para ambas variables. El número de *hits* para cada una de las regiones puede ser consultado en la **Figura 3.7**.

Se realizó un análisis visual de las estructuras que se encontraban dentro de la región III y IV para identificar como *hits* a aquellas presentes en ambas ramas de la metodología. Ver **Figura 3.9.**

Para el sitio 1Z se encontraron 41 hits de los cuales cinco se repiten para ambas ramas de la metodología.

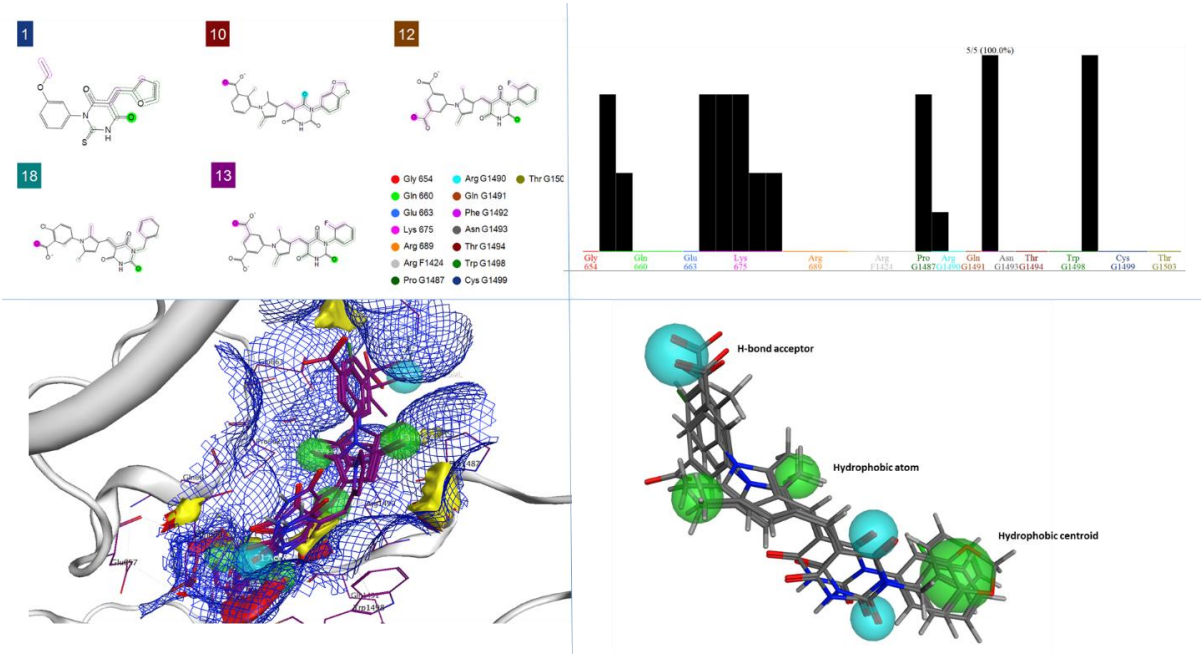


Figura 3.9 Hits consenso, frecuencia de interacciones según PLIF, mapa de potencial molecular y modelo del farmacóforo obtenido para el sitio 1Z.

Como se puede observar en la **Figura 3.9.** todos ellos contienen un núcleo base común, una base pirimidínica con una insaturación exocíclica con alto carácter aromático dado el sistema conjugado. Todo ello hace al carbono insaturado de la doble ligadura exocíclica un excelente aceptor 1-4 de Michael, por lo que este grupo de inhibidores probablemente deban su actividad a la adición inespecífica de nucleófilos en dicha posición. Se puede pensar que esta doble ligadura fue integrada al diseño de estos inhibidores para fijar la posición relativa de los dos anillos que forman al núcleo base. Sin embargo, para evitar la alta reactividad de esta posición podrían proponerse otras modificaciones estructurales que cumplieran este propósito como parte de una campaña de optimización.

En la **Figura 3.9.** se puede observar el campo molecular obtenido con átomos sonda implementados en el programa MOE. En azul se representa la superficie de la proteína que se encuentra expuesta directamente al disolvente, las regiones en color amarillo representan zonas de carácter hidrofóbico, mientras que las regiones en color rojo son de carácter electrofílico. Si se observan los elementos

presentes en el modelo del farmacóforo basado en PLIF que se muestra en la **Figura 3.9.** todas las regiones que aparecen en el campo molecular coinciden con los elementos de este modelo (Phe1492, Trp1498, Gln660, Glu663). Se prevé que el extremo de los inhibidores que se encuentra expuesto al disolvente no jugará un papel de aceptor de puente de hidrógeno como muestra el modelo farmacofórico, ya que es más probable que se encuentre rodeado por moléculas del disolvente. Además de todo esto es importante subrayar que el sitio de unión 1Z se encuentra dentro de la estructura del asa de reconocimiento de DNA, lo cual está directamente relacionado con la forma inactiva y activa de la enzima.[68]

Por el otro lado, para el sitio 3Z no se encuentran hits consenso o núcleos base frecuentes. Una de las características de los inhibidores que presentaron afinidad por este sitio es la presencia de un gran número de anillos aromáticos y regiones hidrofóbicas, lo cual tiene congruencia con el carácter hidrofóbico de este sitio. Sin embargo, el sitio 3Z se encuentra en una región plana de la proteína que se encuentra totalmente expuesta al disolvente. Estas características no permiten observar interacciones específicas entre los *hits* y el sitio de unión y por lo tanto no fue posible proponer un modelo del farmacóforo o algún tipo de estrategia para desarrollar inhibidores dirigidos a 3Z.

Resumen de resultados

Aplicando la metodología planteada en este trabajo sobre la estructura cristalina de DNMT1 fue posible la identificación de los sitios consenso 1Z y 3Z, los cuales están potencialmente relacionados con la flexibilidad de la proteína. Tanto la información estructural de estos sitios, como la información relacionada con los *hits* identificados, posibilitan el uso de una gran variedad de herramientas computacionales y experimentales para guiar la búsqueda y/o optimización de inhibidores de DNMT1.

La información estérica y electrónica de los sitios permite el uso de farmacóforo basado en estructura, el uso de fragmentos moleculares, o la aplicación PLIF. Por su parte, la información estructural de los *hits* consenso

puede ser un punto de partida en campañas de tamizado basado en similitud así como en modelación del farmacóforo basado en el ligante. Estos resultados también permiten generar modelos QSAR u otras técnicas de aprendizaje estadístico para la selección de compuestos selectivos contra esta diana, además de proponer mecanismos de reacción razonables para esta colección de inhibidores.

Los resultados obtenidos demuestran el potencial de la metodología planteada como herramienta para la determinación sitios de unión alternativos relacionados con la flexibilidad de dianas de interés farmacéutico para la identificación y caracterización molecular de inhibidores selectivos.

Capítulo 4. Huellas digitales de bases de datos moleculares (*Database Fingerprints,* DFP)

Metodología computacional

A)

Selección de bases de datos con diferentes grados de diversidad molecular. Para este problema se consideró la construcción de una base de datos formada de cadenas de 166 dígitos binarios, emulando las características de MACCS keys. De Ello se obtuvo a partir del ruido atmosférico, es decir, de forma estrictamente aleatoria. El siguiente paso constó de la obtención de información referente a compuestos con un quimiotipo en común, en este proyecto el núcleo base elegido fue el bencimidazol. También se realizó el proceso de curado para los siguientes conjuntos de compuestos, los cuales presentan valores de diversidad molecular intermedios respecto a los dos bases de datos mencionadas anteriormente: biblioteca de inhibidores enfocada a dianas epigenéticas, base de datos interna de inhibidores de DNMT1, compuestos en fase clínica, conjunto de inhibidores para pruebas de tamizado, productos naturales, semi-sintéticos, moléculas probadas para uso clínico, moléculas GRAS (*Generally Recognized as Safe*) y una base de compuestos generados computacionalmente siguiendo la reglas de valencia y con un número igual o menor a trece átomos.

B)

Para cada una de estas bases se calcularon los *MACCS keys* [50] por medio del conjunto de *scripts* de *Perl* contenidos en *MayaChem Tools*.[52] A través del índice de Tanimoto[29] se determinó la inter e intra similitud molecular promedio de cada una de las bases de datos, la entropía de Shannon total y la distancia de bloque (*city block distances*).[80]

C)

Para cada una de las matrices formadas por *MACCS keys* se calcularon la frecuencia y probabilidad por posición binaria. A estos valores se les aplicó dos

líneas de corte diferentes. La primera consta de la media más una desviación estándar de la base de datos de interés, mientras que la segunda es la media de la probabilidad para la distribución homogénea obtenida a través del ruido atmosférico.

D)

Las representaciones moleculares así obtenidas fueron comparadas con los valores de distancia de bloque. De acuerdo al comportamiento de dichos valores se seleccionó la línea de corte para realizar los pasos posteriores de la metodología.

E)

Una vez determinado el proceso a seguir para la construcción del DFP, se programó, con el lenguaje interpretado Python 3.5, el método para automatizar la obtención de las huellas digitales DFP. Este programa también incluye la posibilidad de comparar bases de datos con el DFP calculado, utilizando el índice de Tanimoto para realizar tamizado virtual y comparación de base de datos por medio de la visualización de los resultados en mapas de calor.

F)

Prueba preliminar de DFP como criterio para realizar tamizado molecular para la identificación inhibidores selectivos o polifármarcos. En este caso se eligieron a las siguientes bases de datos: inhibidores de enzima convertidora de angiotensina (ACE), ACE decoys, inhibidores del receptor opioide mu (MOR), inhibidores de MAO, sulfotransferasa humana (SULTS), inhibidores de P450, y receptor X de pregnano (PXR). Las últimas, pertenecientes a los llamados *anti-targets,* con el propósito de identificar compuestos con mayor probabilidad de producir efectos adversos. Para cada biblioteca se calculó el DFP correspondiente para luego ser comparado con cada uno de los compuestos contenidos en una base de datos de compuestos aceptados para uso clínico. Los resultados de dicha comparación fueron expresados como mapas de calor de similitud. Todo el procedimiento fue automatizado dentro de un *script* antes mencionado.

Resultados y discusión

El primer paso de exploración del concepto DFP se basó en la selección de una serie de bases de datos que contemplaran una gran variedad de características generales como la diversidad molecular. Para poder cuantificar esta característica se decidió hacer uso de dos métricas diferentes: entropía de Shannon y similitud media, obtenida por medio de representaciones moleculares binarias e índice de Tanimoto. Ambas métricas han sido utilizadas para este propósito en investigaciones anteriores a este trabajo.[15,31]

Las siguiente colección de bases de datos públicas (**Tabla 4.1.**) fue elegida a raíz de sus diferentes valores de diversidad molecular: colección de compuestos con núcleo base único, bencimidazoles,[81] biblioteca de inhibidores enfocada a dianas epigenéticas (Selleck),[82] base de datos interna de inhibidores de DNMT1,[15] compuestos en fase clínica,[83] conjunto de inhibidores para pruebas de tamizado,[84] productos naturales,[85] compuestos semi-sintéticos,[86] moléculas aprobadas para uso clínico,[87] moléculas GRAS (*Generally Recognized As Safe*)[88] y una base de compuestos generados computacionalmente siguiendo la reglas de valencia y con un número igual a trece átomos (GDB13).[89]

Base de datos	Tipo / fuente	Tamaño	Media MACCS keys/Tanimoto	ES[a]
Bencimidazol	Laboratorio 122 Farmacia	92	0.61	32.37
Epigenética	Comercial	113	0.45	49.36
DNMT1	DIFACQUIM	566	0.46	48.72
Clínica	Base de dianas terapéuticas	837	0.43	52.83
Tamizado	Comercial (website)	1100	0.43	51.91
Productos Naturales	PNat	1498	0.64	33.71
Semi-sintéticos	Relacionados con PNat	1498	0.60	29.19
Fármacos	Aprobados para uso clínico http://www.drugbank.ca	1490	0.37	54.20
GRAS	Aprobados para industria de alimentos	1500	0.38	31.40
GDB13	http://gdb.unibe.ch/downloads/	1500	0.44	49.04

Tabla 4.1. Características generales de las bases de datos estudiadas, ES[a]: entropía de Shannon.

La hipótesis seguida para selección de los conjuntos moleculares es la selección de un grupo de bases de datos que abarquen una gran gama de valores de diversidad.

Dicho ello, se espera que la base de bencimidazoles sea la de menor diversidad molecular al contener un núcleo base único. Seguida de ella lo valores aumentaran respecto al número y naturaleza de los compuestos contenidos en las

bases de datos seleccionadas. Por ejemplo, la base de datos enfocada a inhibidores de DNMT1 podría ocupar la siguiente posición dentro del orden relativo a la diversidad de conjunto de librerías, esto por supuesto, siguiendo la regla que postula que compuestos similares deben presentar actividades similares. Siguiendo esta lógica se puede prever que el orden creciente de diversidad molecular para las bases estudiadas es el siguiente: biblioteca de inhibidores enfocada a dianas epigenéticas, moléculas aprobadas para uso clínico, compuestos en fase clínica, conjunto de inhibidores para pruebas de tamizado, compuestos semi-sintéticos, productos naturales, moléculas GRAS y GDB13. Aquí el contenido de información de GDB13 sólo estaría limitado por la redundancia informacional generada por las reglas de valencia y por supuesto, por el número de átomos que componen a sus moléculas.

Para poder afirmar que GDB13 puede representar un límite superior sería necesario contar con moléculas dentro de un intervalo de tamaños mucho mayor. Por esta razón, se propuso la construcción de una base de datos formada únicamente por cadenas de 166 dígitos binarios generados de forma aleatoria. Para evitar los posibles sesgos de los algoritmos semi-aleatorios se recurrió al servidor accesible en random.org,[90] el cual cuenta con distintos generadores de números aleatorios basados en ruido atmosférico. Esta base de datos, aun cuando no necesariamente contiene cadenas binarias válidas, presenta una distribución aleatoria homogénea. Esto permite afirmar que contiene el máximo grado de entropía informacional posible, es decir, no contienen redundancias, patrones definidos o información, ya que el ruido predomina. Está característica es de gran utilidad, ya que permitió plantear la existencia de un límite superior en cuento a diversidad o entropía se refiere.

Una vez planteadas la barreras conceptuales del sistema se calcularon *MACCS keys* para cada uno de los compuestos seleccionados. A partir de ellas se construyó la matriz de *fingerprints* para cada base de datos. Para cada columna, que corresponde una posición binaria determinada para el total de compuestos, se calculó la frecuencia de cifras binarias iguales a uno, que representa la presencia

de algún elemento definido en el diccionario previsto dentro del diseño de *MACCS keys*.

A partir de los valores de frecuencia es trivial calcular la probabilidad asociada a cada una de estas posiciones si se divide la frecuencia contra el número de compuestos totales de una base de datos en particular. En este punto, se encuentra el paso importante para la construcción del DFP. Ya que se cuenta con los valores de probabilidad media de los *fingerprints* contenidos en la base de datos aleatoria (probabilidad media=0.55) es posible plantear que cualquier valor que se encuentre por debajo de dicha cota es producto de algún fenómeno relacionado con el azar y por lo tanto su contenido informacional no es relevante para describir al conjunto de compuestos contenidos en la base de datos correspondiente.

Para contrastar este último punto se planteó el uso de otra línea de corte basada en la probabilidad media y desviación estándar interna. Esta métrica es usada en el tratamiento se señales digitales como un filtro del ruido asociado a la trasmisión y componentes del conjunto emisor-receptor. [91]

Este procedimiento dio como resultado dos cadenas de 166 dígitos binarios para cada base de datos. Para obtenerlas se generó el siguiente criterio de selección: si el valor de la probabilidad media de una posición dada estaba por encima del valor elegido como límite de corte, se asignó el valor de uno a dicha posición dentro de una nueva cadena de dígitos binarios. Para el caso contrario, el valor era menor al límite de corte, se asignó cero.

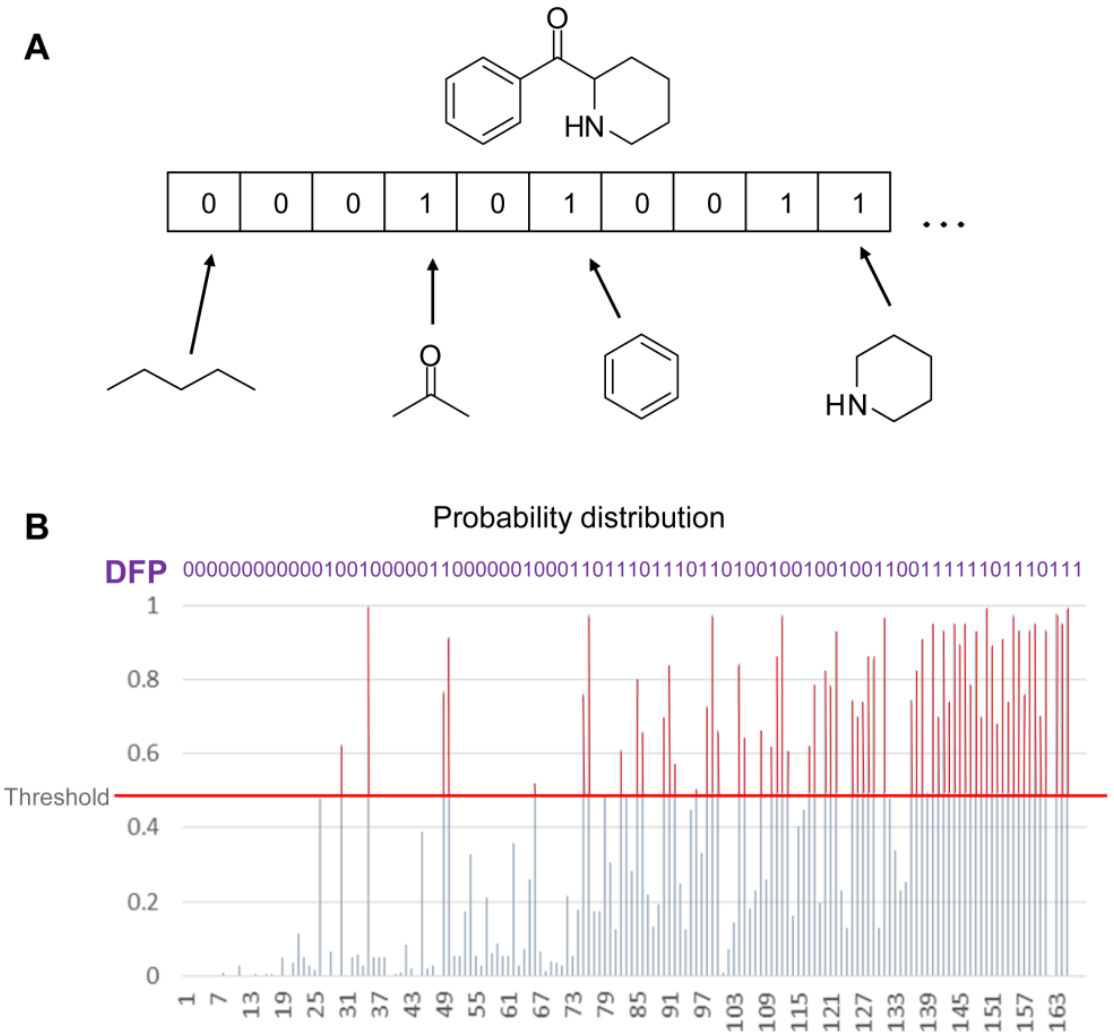


Figura 4.1. A) Representación esquemática de representaciones moleculares 2D basadas en diccionario. B) Representación esquemática de DFP.

Ambas representaciones binarias fueron comparadas respecto a los valores de similitud media interna (*MACCS keys*/Tanimoto) y al valor de distancia de bloque de la comparación pareada. Estos resultados también fueron contrastados con la relación que guardan los valores de entropía y similitud media. Esto es, el comportamiento de la bases de datos dentro del plano formada por ambas variables.

Afortunadamente se encontró que los valores obtenidos por las métricas tradicionales y DFP son linealmente comparables. Esto muestra la capacidad de

DFP para ser utilizado como una métrica para la descripción del espacio químico. Esto sin obviar que es necesario realizar pruebas de validación más exhaustivas utilizando mayor número de datos. Una vez seguros de la validez del método se procedió a la escritura de un *script* basado en el lenguaje de programación Python 3.5[92] para automatizar el cálculo de la representación molecular de bases de datos para cualquier conjunto de compuestos (Código disponible en **Anexo**).

Ya que los resultados hasta este punto fueron satisfactorios, el paso natural fue probar la metodología para realizar estudios de tamizado molecular. En dicho sentido, este método permite comparar el patrón general de una colección de compuestos con otras bibliotecas de interés por medio de una única cadena de elementos binarios.

Para probar dicha hipótesis, se eligieron una serie de colecciones enfocadas a dianas de relevancia clínica (inhibidores de enzima combatidora de angiotensina (ACE), ACE *decoys*, inhibidores del receptor opioide mu (MOR), inhibidores de MAO, sulfotransferasa humana (SULTS), inhibidores de P450, y receptor X de pregnano (PXR)[93] para ser comparadas con una base de datos de compuestos aprobados para uso clínico. Dicho conjunto incluye colecciones de inhibidores de dianas relacionadas con el metabolismo de fármacos, como es el caso de P450. Esto último con el objetivo de identificar compuestos presentes en bases de datos de dianas interés clínico que puedan interferir simultáneamente con los llamados *anti-targets*, lo cual puede estar asociado con una mayor probabilidad de producir efectos no deseados. Sobra decir que se espera utilizar este tipo de comparaciones para identificar moléculas con posible actividad polifarmacológica.

Se utilizaron mapas de calor como método de visualización de resultados para facilitar el análisis de resultados y la selección de moléculas prometedoras o no deseadas. Nuevamente, esto se automatizó dentro de un *script* escrito en el lenguaje Python 3.5 que es capaz de calcular el DFP de la base o bases de interés para ser comparadas con los MAACs keys de alguna biblioteca de interés. Los mapas de calor pueden ser modificados según el número de bibliotecas

elegidas dentro de una escala de color que comprende el intervalo de similitud unitario. (Código disponible en **Anexo**)

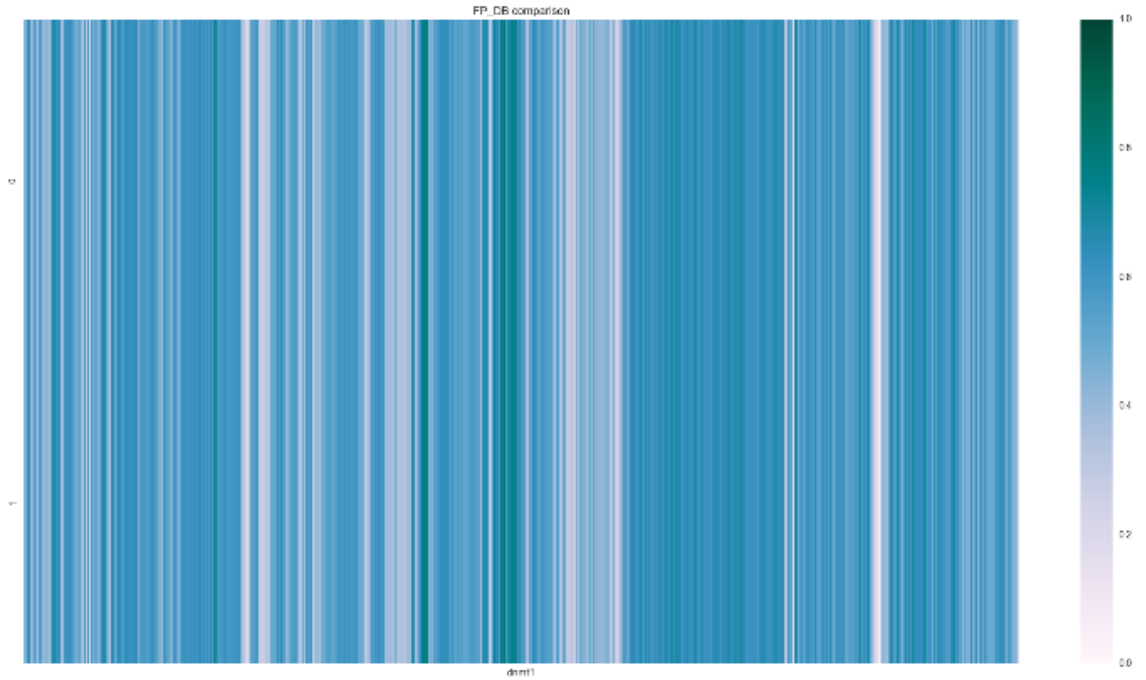
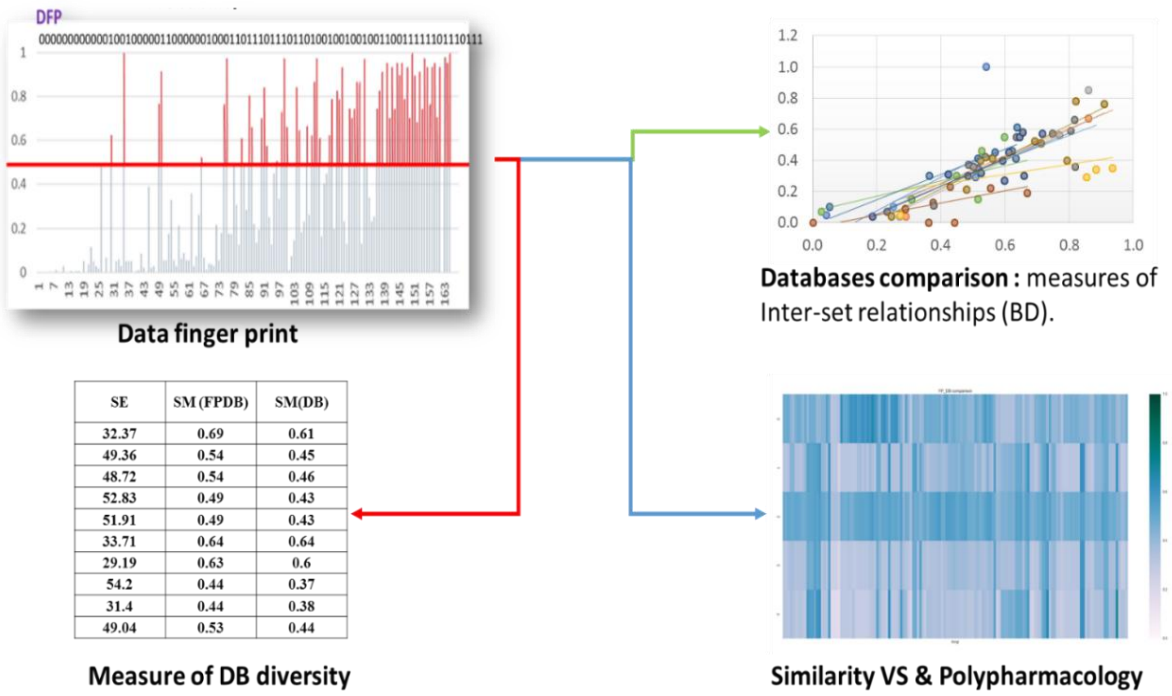


Figura 4.2. Ejemplo de un mapa de calor obtenido. Base de datos de inhibidores de DNMT1.

El proceso de tamizado debe ser sujeto a un estudio exhaustivo de validación. Para ello, se inició el estudio de tamizado de bases de datos que contienen compuestos reconocidos como activos y 36 de sus señuelos.[94] Este procedimiento de validación será continuado en trabajos posteriores a la presentación de este trabajo.



SE	SM (FPDB)	SM(DB)
32.37	0.69	0.61
49.36	0.54	0.45
48.72	0.54	0.46
52.83	0.49	0.43
51.91	0.49	0.43
33.71	0.64	0.64
29.19	0.63	0.6
54.2	0.44	0.37
31.4	0.44	0.38
49.04	0.53	0.44

Figura 4.3. Diagrama de flujo de la metodología.

Distribución de *fingerprints* binarios

La **Figura 4.4.** muestra la distribución de probabilidades los *MACCS keys* para las diferentes bases de datos estudiadas en este trabajo. Los valores correspondientes a la entropía de Shannon y la similitud media MACCS keys/Tanimoto de cada uno de los conjuntos se reportan en la **Tabla 4.2**. Tanto la tabla como la **Figura 4.5.** muestran que los valores de entropía de Shannon como la similitud media difieren para cada una de las librerías. Esto también puede ser observado de forma visual mediante el patrón generado por las redundancias presentes en cada una de las distribuciones de probabilidad.

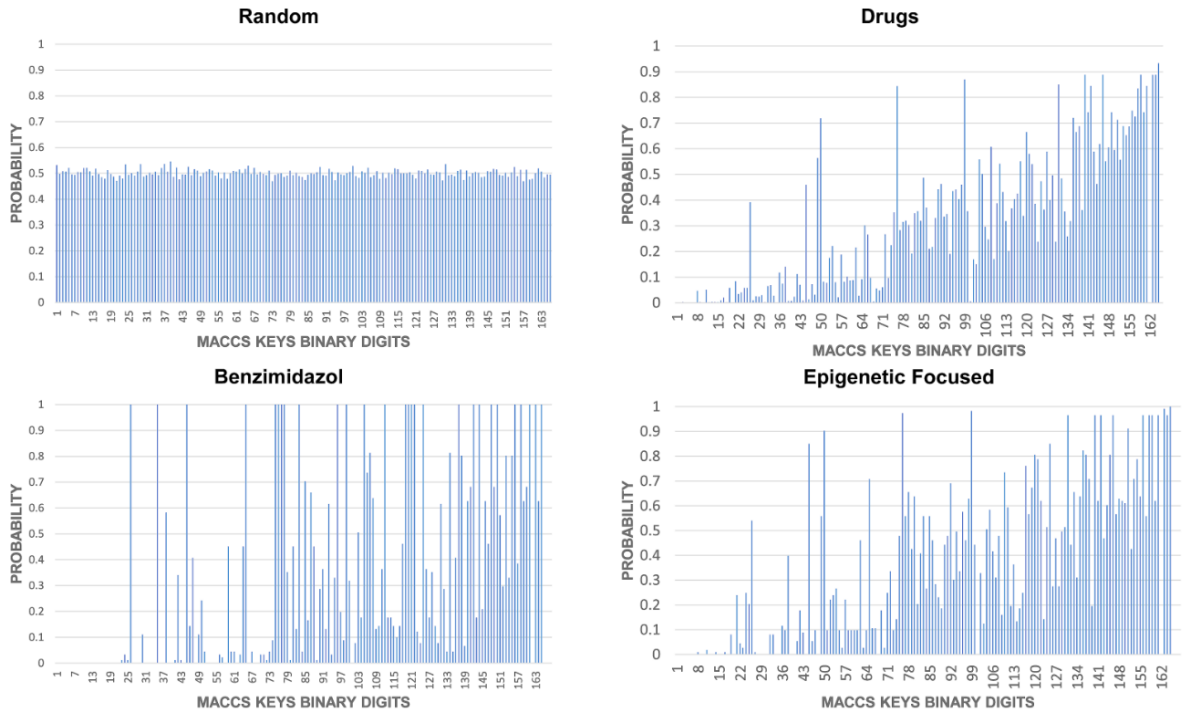


Figura4.4. Distribuciones de probabilidad para algunas de las bibliotecas moleculares estudiadas.

Base datos	Tipo / fuente	N	Media	ES[a]
Bencimidazoles	Laboratorio 122 Farmacia	92	0.61	32.37
Epigenética enfocada	Comercial	113	0.45	49.36
DNMT1	DIFACQUIM	566	0.46	48.72
Clínica	Base de dianas terapéuticas	837	0.43	52.83
Tamizado general	Comercial (website)	1100	0.43	51.91
Productos Naturales (PNat)	PNat	1498	0.64	33.71
Semi-sintéticos	Relacionados con PNat	1498	0.60	29.19
Fármacos	Aprobados para uso clínico http://www.drugbank.ca	1490	0.37	54.20
GRAS	Aprobados para industria de alimentos	1500	0.38	31.40
GDB13	http://gdb.unibe.ch/downloads/	1500	0.44	49.04

ES[a]: Entropía de Shannon

Tabla 4.2. Valores de ES y SM.

La **Tabla. 4.2.** muestra la relación existente entre los valores de similitud MACCS keys/Tanimoto y la entropía de Shannon. Se observa que valores de entropía altos están asociados con una intra-diversidad alta o baja similitud, mientras que los valores de entropía bajos se relacionan con una baja diversidad o alta similitud media. Es decir, a valores altos de entropía en menos probable encontrar dos compuestos con secuencias binarias iguales. Esto también se aplica elemento a elemento, entre mayor es el valor entrópico, menor es el número de redundancias y por lo tanto menor el número de elementos compartidos. Ya que esto es así, no es de sorprender que la similitud de Tanimoto, que está basada en el número de unidades compartidas, siga una proporción directa con la entropía de Shannon. Una excepción encontrada en el conjunto de bases fue el caso de GRAS. Esta presenta una entropía relativamente baja pero una gran diversidad según el índice de Tanimoto. Esto puede ser interpretado tomando en cuenta que el índice de Tanimoto no sólo utiliza el número de redundancias como factor de comparación, además de ello, sopesa la cantidad de elementos no compartidos, únicos, de cada una de las representaciones. A diferencia del índice de Tanimoto, la entropía de Shannon sólo se ve modificada por el patrón general de redundancias presentes en el conjunto de compuestos.

Otra de las bases de datos que se distinguen por sus características únicas es la generada de forma aleatoria. Como era de esperarse, la entropía de esta base de datos es mayor que la del resto del conjunto, dado que cada una de las posiciones binarias tienen la misma probabilidad de presentar cualquiera de los dos valores posibles, lo que se expresa directamente en la media de probabilidades (0.5).

La **Figura 4.5.** muestra la distribución de las diferentes bases de datos en el plano de entropía contra similitud media.

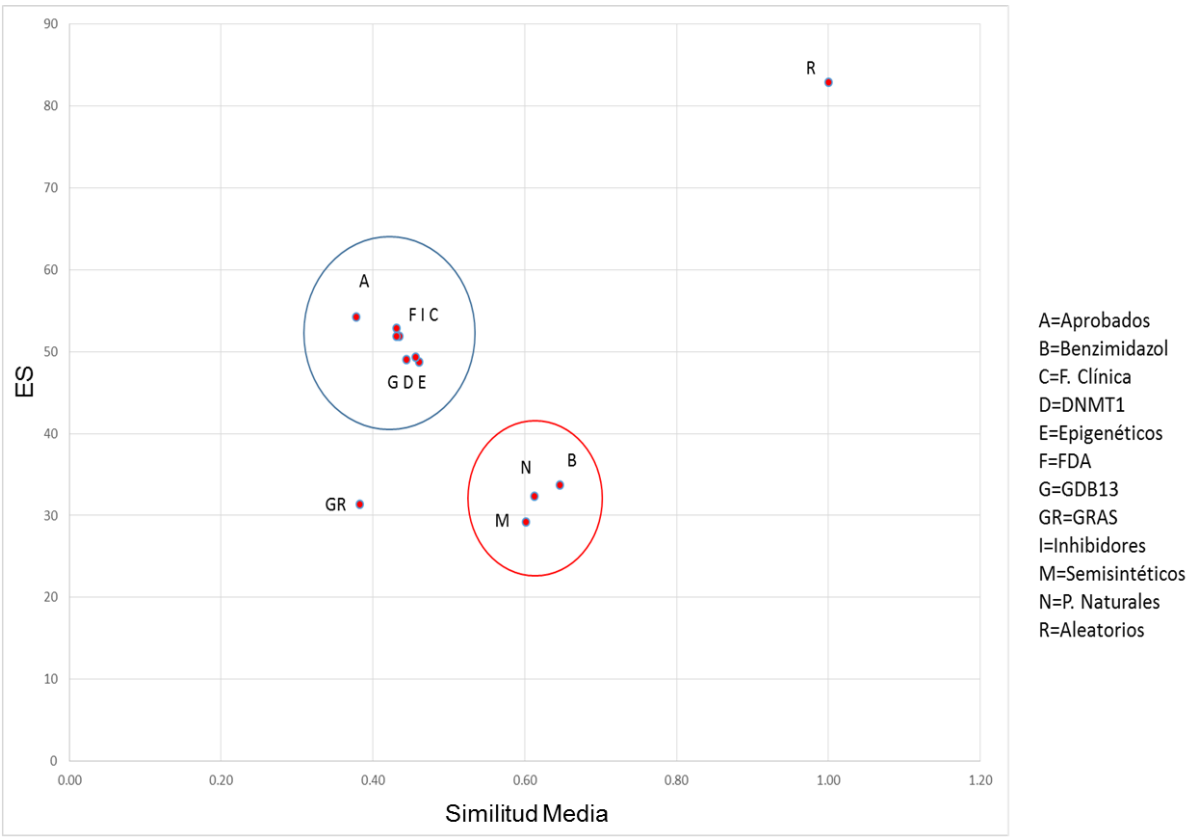


Figura 4.5. Plano de entropía de Shannon contra similitud media.

Como se puede observar, las bases de datos se agrupan en regiones (*clusters*) sin tomar en cuenta a las dos excepciones antes mencionadas. De alguna manera dichas bases de datos se encuentran organizadas respecto a la naturaleza de sus compuestos. En el grupo más poblado, todas las bases de datos, con excepción de GDB13, están relacionadas implícitamente al contener moléculas bioactivas en distintos estadios de desarrollo clínico (compuestos aprobados para uso clínico, compuestos en fase clínica, colección para tamizado molecular, inhibidores epigenéticos e inhibidores de DNMT1). Mientras que el grupo con menor población incluye compuestos que se encuentran presentes en organismos vivos. Esto sugiere que la entropía de Shannon obtenida a partir de las redundancias presentes en conjuntos *MACCS keys* pueden ser utilizado como un criterio complementario (aunado con diversidad de núcleos base, diversidad de *fingerprints*, etc.), para caracterizar la diversidad de un grupo de bases de datos o como un método de visualización de la "topografía" del espacio químico en una

nivel de abstracción mayor (meta espacio químico) a la comúnmente estudiada (espacio químico molecular).

El concepto de entropía de Shannon, originado en la teoría informacional para medir el contenido de información en un mensaje, ha sido utilizado anteriormente dentro de la quimioinformática como una métrica de diversidad estructural basada en el contenido de núcleos base, como pilar conceptual para la generación de *fingerprints* moleculares y recientemente como una herramienta indirecta para enriquecer bases de datos por medio de tamizado molecular. En especial, Wang et al. establecieron que la diferencia de entropía de una base de datos enfocada antes y después de la inclusión de nuevas moléculas puede ser utilizada como criterio para seleccionar compuestos similares a los contenidos en dicho conjunto.[95]

DFP

Como ya se mencionó, se calcularon los *MACCs keys* de 166 unidades binarias para cada una de las bases de datos seleccionadas. Sobre la distribución de probabilidades obtenida para cada uno de los conjuntos se aplicaron dos métricas de corte diferentes: la media de la distribución de probabilidades de la base de datos generada aleatoriamente, y una que depende de la base de datos estudiada con un valor variable correspondiente al valor de su media más una distribución estándar.

Resultado de estas métricas de corte fue la generación de dos DFP´s diferentes para cada una de las bases estudiadas. Para verificar la capacidad de representación de ambas metodologías se realizó una comparación pareada de los valores de distancia de bloque con los valores del índice de Tanimoto aplicado a los DFP´s. En la **Figura 4.6.** se muestra la relación lineal que existe entre estos dos valores para las dos metodologías propuestas. Para obtener relaciones dentro de la unidad, se realizó una normalización y se obtuvo el inverso de los valores de distancia de bloque.

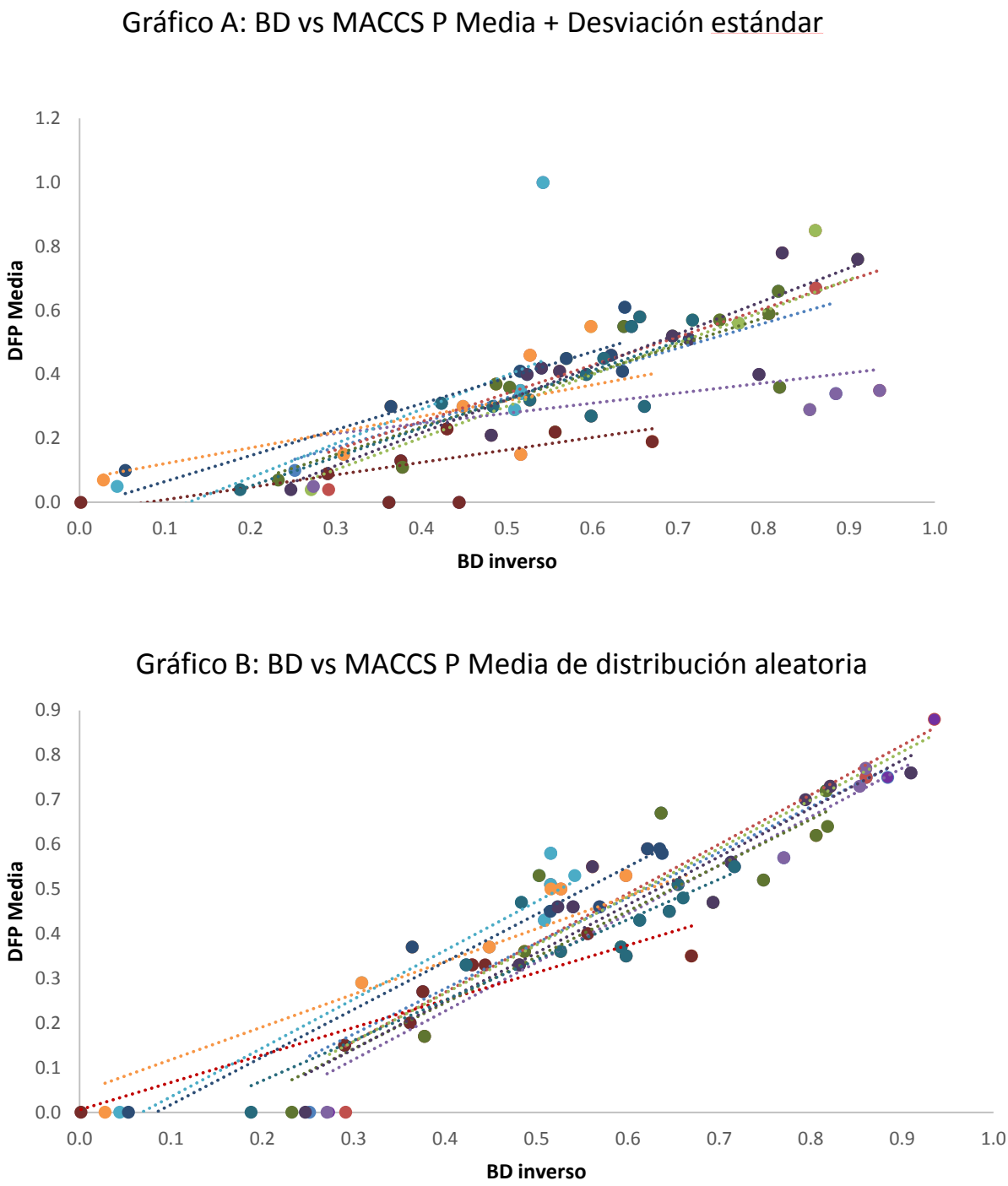


Figura 4.6. Relaciones lineales entre DFP y BD inverso para las dos métricas de corte.

La **Figura 4.6.** muestra que la primera métrica de corte (valor de la media de la distribución aleatoria, gráfica B) resulta en una mejor relación lineal con los valores de distancia de bloque. La **Tabla 4.3.** muestran los valores de DFP/Tanimoto y

distancia de bloque para cada comparación pareada. Estos resultados están de acuerdo con publicaciones anteriores[15,31] donde se demuestra que existe una distancia pequeña entre los compuestos aprobados y en fase clínica. De la misma manera, se presenta pequeñas distancias entre los inhibidores epigenéticos y los compuestos para tamizado molecular. De hecho, se puede esperar que estos compuestos compartan regiones amplias del espacio químico cuando son representadas con MACCS keys, que en este caso, es la representación binaria elegida para obtener a DFP. [15]

Para la base de datos aleatoria se obtuvieron las distancias más grandes. También GRAS presenta este comportamiento, lo que es consistente con otros trabajos que demuestran que estos compuestos son disimilares a los moléculas contenidas en bases de datos comúnmente utilizadas en el descubrimiento de fármacos.

Aleatoria	0										
GDB13	54	0									
DNMT1	51	27	0								
GRAS	67	27	42	0							
PNat	63	39	24	48	0						
SS	65	32	34	22	43	0					
Benci	64	35	33	43	32	46	0				
TG	49	23	12	37	24	32	31	0			
Aprobados	49	19	17	30	29	27	33	10	0		
Clínica	47	23	13	38	25	32	32	4	9	0	
Epi	50	26	12	42	24	37	32	8	15	9	0

Tabla 4.3. Matriz de valores de BD.

De acuerdo a estos resultados, se puede afirmar que DFP es una representación razonable para ser utilizada como un descriptor del espacio químico al ser capaz contener el patrón general de una base de datos molecular. Estos resultados deben ser ampliados para definir los límites de esta metodología de representación molecular. Se puede prever que DFP puede ser un método óptimo para enriquecimiento de bases de datos enfocadas con un tamaño y diversidad media, ya que para el caso de conjuntos con alta diversidad, el número de redundancias será muy bajo y por lo tanto no existirá un patrón yacente. Esta limitante también puede verse como una ventaja en el caso del diseño bases de datos diversas para tamizado de alta y medio rendimiento, ya que indicará la presencia de conjuntos moleculares donde sus componentes no guardan relación estructural.

Tamizado virtual

Como se menciona en la sección anterior, uno de los objetivos de DFP es su implementación en tamizado virtual. Para avanzar hacia esta dirección, se realizó un script en lenguaje Python 3.5 que es capaz de calcular de forma automática los valores de entropía, similitud media de MACCs keys/Tanimoto y DFP para cualquier conjunto molecular seleccionado. El *script* se ilustra en la en el **Anexo**. Además de ello, tienen la posibilidad de comparar un número n de DFP´s con un conjunto molecular de interés. Los resultados del este tamizado virtual son arrojados como mapas de calor que utilizan escalas de color para indicar el grado de similitud entre el patrón general de un conjunto dado y los compuestos de la base de datos a comparar.

Este *script* fue aplicado a las siguientes bases de datos: inhibidores de enzima convertidora de angiotensina (ACE), ACE *decoys*, inhibidores del receptor opioide μ (MOR), inhibidores de P450, inhibidores de MAO, sulfotransferasa humana (SULTS) y receptor X de pregnano (PXR). Sus correspondientes DFP´s se compararon con un conjunto de aproximadamente 1500 compuestos aprobados para uso clínico (FDA). Se incluyeron *anti-targets* con el propósito de corroborar la posible interacción de estos compuestos con enzimas relacionadas con el metabolismo de degradación de fármacos. Estos mapas presentan la

ventaja de localizar compuestos con actividad polifarmacológico. Para todas las dianas farmacológicas se localizaron compuestos activos. Dentro de los primeros 20 lugares: ACE (3 compuestos activos), ACE *decoys* (7 compuestos activos), MOR (8 compuestos activos) y MAO (un compuesto activo). La diferencia encontrada entre ACE y ACE *decoys* es seguramente debida al aumento de redundancias gracias a las moléculas señuelo, la cuales son moléculas muy similares pero sin actividad prevista. En la **Figura 4.9.** se muestran los resultados tamizado virtual en forma de mapas de calor.

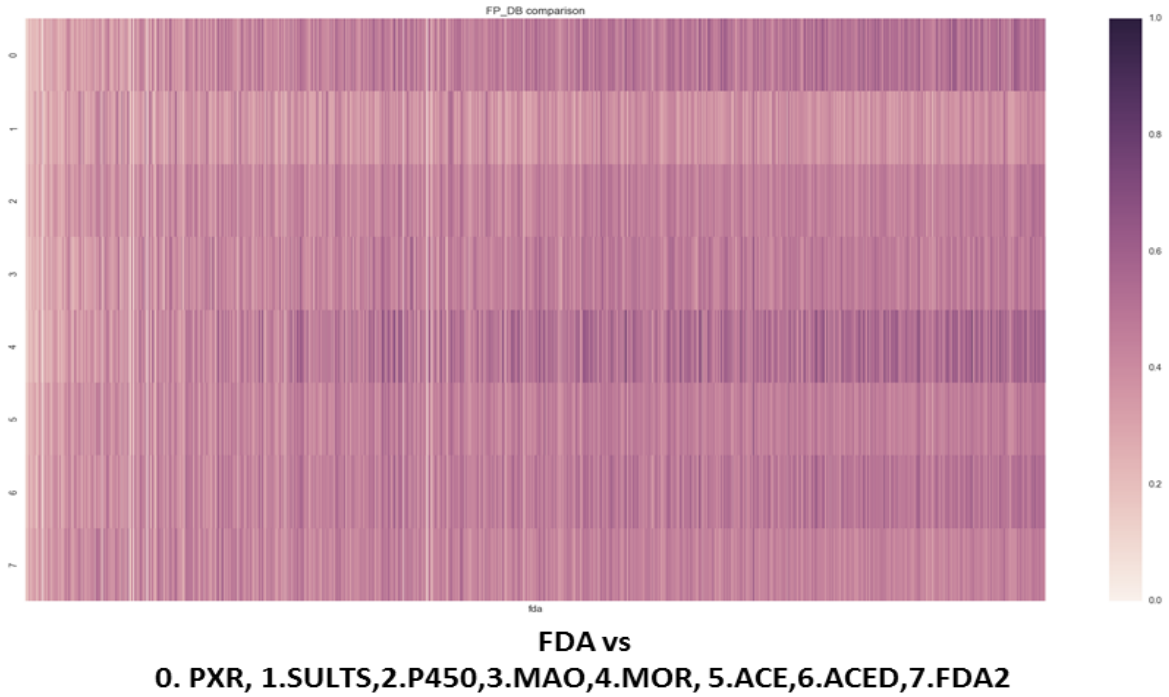


Figura 4.7. Mapa de calor de ACE, ACE *decoys*, MOR y MAO.

Estos resultados se muestran como una ejemplificación de las posibles aplicaciones que tiene el método desarrollado, así como la capacidad de automatización del programa diseñado para tamizado virtual basado en DFP.

Aun cuando este método es potencialmente útil para realizar tamizado virtual, es necesario realizar un estudio de validación, así como campañas de tamizado virtual sobre sistemas con distintas propiedades informacionales.

Resumen de resultados

En este capítulo de la tesis se presenta a DFP como un método de representación de bases intuitivo basado en la redundancia informacional presente en bases de datos formadas por representaciones moleculares binarias. Este *fingerprint* tiene la capacidad de incluir el patrón general de un conjunto de compuestos, lo que es de suma utilidad dentro del desarrollo de fármacos si se tiene en cuenta el postulado de similitud-actividad.

DFP está construido mediante la aplicación de dos métricas de corte sobre la distribución de probabilidades de MACCs keys de 166 unidades binarias. En este trabajo se aplicó este método a diez bases de datos con características de diversidad y entropía variadas, así como su comparación con los valores de similitud media MACCS keys/Tanimoto, DFP/Tanimoto y distancia de bloque. Mediante estos resultados se determinó una métrica final igual a la media de probabilidad de una distribución aleatoria

También se concluyó que la entropía de Shannon basada en la distribución de representaciones moleculares binarias es una métrica adecuada para medir la diversidad de bases de datos moleculares. El plano formado por la entropía y la similitud media de DFP/Tanimoto o *MACCs keys*/Tanimoto, es un método de visualización que permite la caracterización de la diversidad y relación entre bases de datos moleculares.

DFP es un método razonable para el estudio de la inter e intra relación de bases de datos dentro del espacio químico. Independiente de la necesidad de un estudio más extenso sobre el comportamiento de DFP frente a variaciones de la entropía, es posible prever su aplicación en campañas de tamizado molecular y diseño de bases de datos moleculares. En especial, las bases de datos enfocadas contienen una diversidad molecular y contenido de información ideal para ser representadas por medio de este método, lo que hace de DFP una excelente alternativa para la búsqueda de inhibidores de DNMT1.

Capítulo 5. Estudios de acoplamiento molecular *a posteriori* de inhibidores de DNMT1 y DNMT3A

Metodología

A partir de los resultados de pruebas actividad biológica obtenidos para los análogos del compuesto NSC137546 sintetizados en el grupo del Dr. Massimo Bertinaria de la Universidad de Torino, Italia, se seleccionaron a aquellas moléculas con los valores más altos de inhibición contra DNMT1 y DNMT3A para realizar estudios de acoplamiento molecular en el sitio catalítico de la estructura cristalográfica de ambas enzimas. Estos compuestos fueron denominados **22** y **24**, ambos N-benzoyl aminoácidos. Ambos compuestos fueron obtenidos mediante síntesis en un programa de optimización del compuesto NSC137546 (**Figura. 5.1**). Este compuesto, a su vez, fue identificado como inhibidor de DNMT utilizando técnicas de tamizado virtual dentro de una de base de datos del *National Cancer Institute* (NCI). [96]

A)

La estrategia aplicada por el grupo de investigación del Dr. Bertinaria constó de la variación de tres regiones moleculares de NSC137546: sustituciones sobre el anillo bencénico, conversión de amina por amida y modulación del fragmento ácido. De este procedimiento se obtuvieron 27 análogos que fueron probados con diferentes técnicas *in vitro* e *in situ* para medir la actividad contra DNMT1 y DNMT3A, así como pruebas de estabilidad en condiciones fisiológicas.

B)

De acuerdo a los resultados de actividad inhibitoria se seleccionó un compuestos (**22**) para realizarle estudios de acoplamiento molecular. Todos los acoplamientos y cálculos de la función de puntuación se realizaron con el programa *Internal Coordinates Mechanics* (ICM-Pro versión 3.8-4), dentro de la cavidad catalítica de las estructuras cristalográficas de DNMT1 (PDB ID: 3PTA)[65] y DNMT3A (PDB ID: 2QRV)[97] preparadas con las opciones preestablecidas de ICM. Todos los

acoplamientos se realizaron por triplicado para asegurar la convergencia del algoritmo de optimización.

C)

Se seleccionaron e inspeccionaron visualmente a las diez conformaciones de menor energía. Para cada una de ellas se generaron los mapas bidimensionales de interacción con el programa MOE versión 2014.09. A partir de estos resultados se describieron aquellas interacciones que favorecen la interacción del complejo proteína-ligando.

Resultados y Discusión

Fase experimental

Como se ha mencionado, los inhibidores no nucleosídicos de DNMT representan una gran oportunidad para el desarrollo de compuestos activos con mayor selectividad. Dentro de estos esfuerzos y gracias a campañas de tamizado virtual en la base de datos del NCI se pudo identificar a un derivado del ácido glutámico llamado NSC137546 con actividad moderada contra DNMT1 (inhibición selectiva de DNMT1 contra DNMT3B a una concentración de 100μM).[72]

Este compuesto fue seleccionado por el laboratorio del Dr. Massimo Bertinaria de la Universidad de Torino, Italia, como punto de partida para realizar un programa de optimización sintética para identificar compuestos activos contra DNMT y explorar su RAE. Es importante mencionar que hasta el momento no se cuenta con información estereoquímica del NSC137546, ni con resultados cristalográficos que debelen el modo de unión de este compuesto en ninguna de sus posibles dianas. Dadas estas condiciones y a raíz de estudios de acoplamiento molecular, los cuales no logran distinguir diferencias significativas entre los enantiómeros (R) y (S), el grupo decidió utilizar la forma (S) para así favorecer al estereoisómero con mayor abundancia natural.[72]

En la **Figura 5.1.** se muestra la estructura del NSC137546 así como las regiones moleculares que fueron elegidas para realizar modificaciones estructurales.

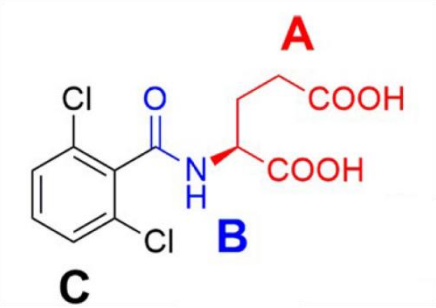


Figura 5.1. Estrategias de modulación sobre el compuesto NSC137546. A: modulación de la porción ácida, B: Conversión de amida a amina, C: substituciones sobre el anillo bencénico.[98]

Como muestra la **Figura 5.2.**, tres regiones de la molécula fueron modificadas para tratar de modular la actividad del núcleo base N-benzoyl aminoácido. De la modificación de la porción ácida terminal se obtuvieron los primeros ocho compuestos (**1-8**). Estas modificaciones se realizaron para probar el efecto de variaciones estéricas homogéneas. El análogo **9** fue sintetizado para verificar el papel de la amida en la actividad, mientras que los compuestos **10** a **27** presentan modificaciones en el fragmento aromático con el propósito de obtener información preliminar de las RAE relacionadas con el anillo aromático.[72]

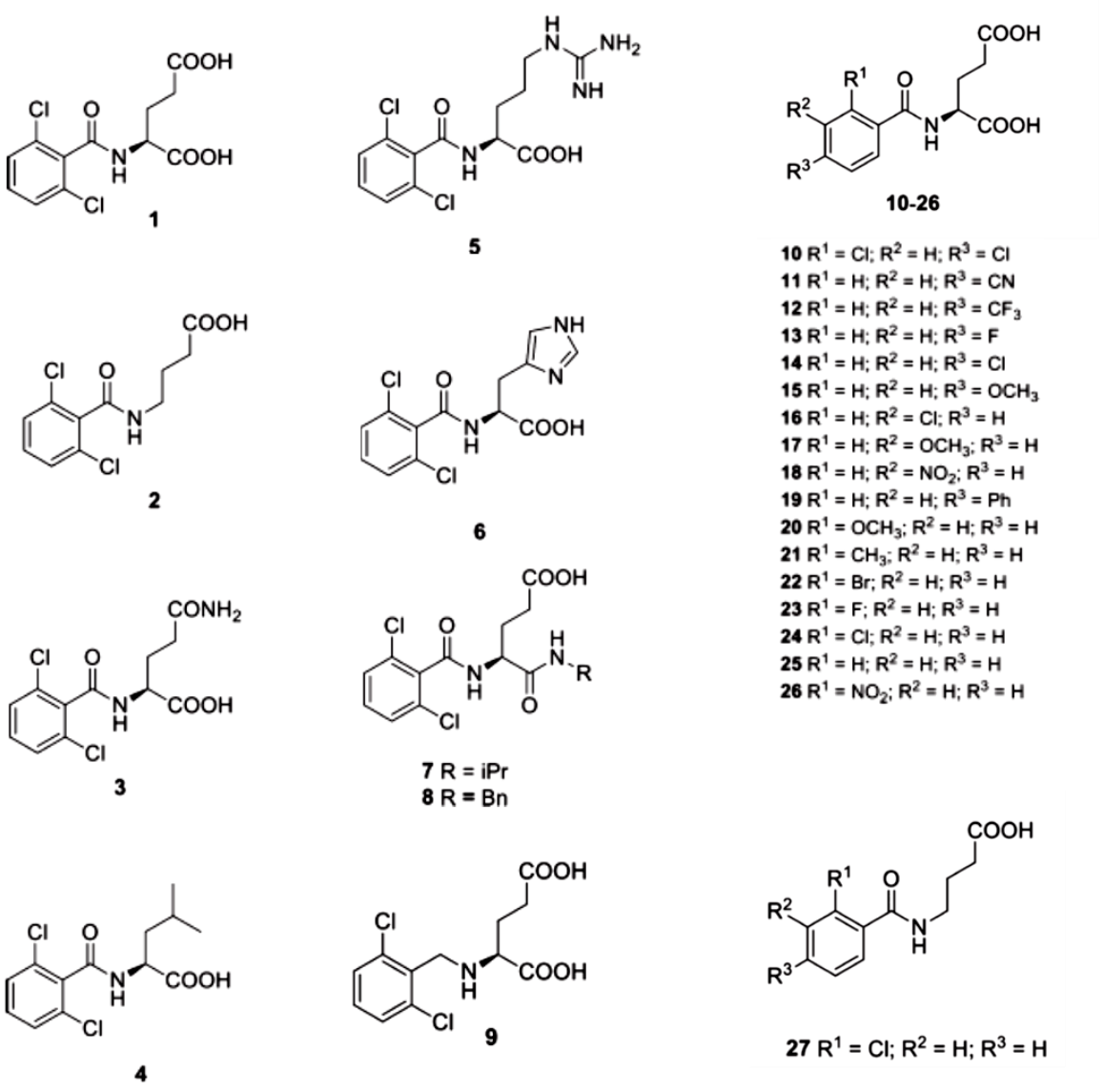


Figura 5.2. Análogos de NSC137546 sintetizados. [72]

La capacidad de inhibición de la metilación de DNA total de cada uno de los compuestos sintetizados fue probada *in vitro* para dos concentraciones diferentes (50 y 100 μM) incubando junto con un lisado celular recién obtenido durante 2 horas a 37°C (**Tabla 5.1.**).

A partir de estos resultados, el laboratorio del Dr. Bertinaria determinó las siguientes RAE para los análogos sintetizados (ANOVA y pruebas *post hoc* Bonferroni): [72]

- El compuesto **1**, compuesto de partida, presenta una actividad de 39 ± 12% ($p < 0.05$). La remoción del α-carboxilo, compuesto **2**, produce una actividad de 40 ± 2% ($p < 0.05$) lo que es comparable con el compuesto **1**.
- Bajo las condiciones del ensayo, la conversión del α-carboxilo a una amida sustituida, presente en el compuesto **7** y **8**, no muestra una habilidad desmetilante significativa, lo que demuestra que la sustitución de grupos voluminosos en esta posición no representa una estrategia prometedora. Por otro lado, la importancia del α-carboxilo en la actividad de los análogos aún debe ser estudiada a mayor profundidad.
- El remplazo de una amida o la remoción del grupo γ-carboxilo, compuesto **3** y **4** respectivamente, produce una pérdida total de la actividad desmetilante.
- Los derivados de imidazol, compuesto **6**, conserva la actividad inhibitoria (42 ± 7% ($p < 0.05$)) mientras que la sustitución altamente básica presente en **5**, guanidina, presenta total ausencia actividad.
- La reducción del enlace amida a una amina, compuesto **9**, a una concentración de 100 μM mantiene una actividad comparable al compuesto **1** (37 ± 1% ($p < 0.05$)).

Compuesto	100 µM	50 µM	Compuesto	100 µM	50 µM
1	61±12[c]	96±6	**15**	72±7	82±6
2	60±2[c]	87±15	**16**	75±2	89±5
3	99±11	122±26	**17**	69±3	97±2
4	126±29[c]	96±2	**18**	58±3[c]	92±4
5	90±1	98±22	**19**	64±11	93±6
6	58±7[c]	104±15	**20**	81±2	85±5
7	82±16	118±1[c]	**21**	60±5[d]	86±18
8	94±39	90±17	**22**	45±1[c]	63±16[d]
9	63±1[d]	111±2	**23**	51±9[c]	68±3[d]
10	116±9	108±36	**24**	68±21[d]	65±1[d]
11	85±20	122±16	**25**	74±4	95±10
12	72±23	100±1	**26**	107±10	103±11
13	57±4[c]	106±13	**27**	99±14	93±5
14	70±6[c]	100±16			

ANOVA y pruebas post hoc con el test de Bonferroni.

[a] Determinado en un lisado de células HaCat utilizando el ensayo actividad-inhibición de metilación de DNA Epiquik.

[b] Datos Expresados como el porcentaje de metilación residual relativa con un vehículo de DMSO al 1%. Los resultados son la media de triplicados del ensayo.

[c] $p < 0.05$ contra vehículo.

[d] $p < 0.01$ contra vehículo.

Tabla 5.1. [a] Habilidad de inhibición de la metilación de DNA de los compuestos (1-27) expresado como metilación residual relativa de DNA.

De la exploración del efecto de las sustituciones dentro de las diferentes posiciones del anillo aromático se concluyó:

- La disustitución en la posición 2 y 4 del anillo bencénico, compuesto **10**, elimina la actividad.
- La sustitución de un átomo de cloro en la posición orto, meta y para (compuesto **24**, **16** y **14**), disminuye la actividad a 68 ± 21% ,75 ± 2% y 70 ± 6% respectivamente. Mientras que **24** en concentraciones menores tiene mayor actividad que **1**, sustitución de un átomo de cloro en la posición orto (35 ± 1%).
- La sustitución de un electrodonador (grupo metoxi) no demuestran un aumento de la actividad respecto a **24** a una concentración de 50 μM.

Estos resultados preliminares sugieren que la sustitución de un halógeno en posición orto del anillo aromático puede jugar un papel fundamental en la modulación de la actividad de estos compuestos. Por esta razón realizaron una serie de sustituciones dentro del anillo aromático contemplando la variación de diferentes propiedades estéricas y electrónicas, compuestos **11-13**, **18**, **19**, **21-23** y **26**. De estas sustituciones sólo **13**, **22** y **23** presentan aumentos en la capacidad desmetilante total.

Para obtener información detallada sobre la metodología experimental que se utilizó en este trabajo se puede consultar el trabajo de D. Garella, S. Atlante, *et al*.[72]

A raíz de estos resultados preliminares se seleccionaron al compuesto **22** y **24** como candidatos para realizar pruebas de actividad más específicas. Estos compuestos fueron sujetos a pruebas de actividad en un lisado de células que sobre expresan DNMT1, DNMT3A y DNMT3B de forma selectiva. Las actividades de los compuestos seleccionados fueron comparadas con **1** y RG108 (inhibidor de referencia), a siete concentraciones diferentes dentro del intervalo de 1 a 150 μM. Los resultados se muestran en la **Figura 5.3.**[72]

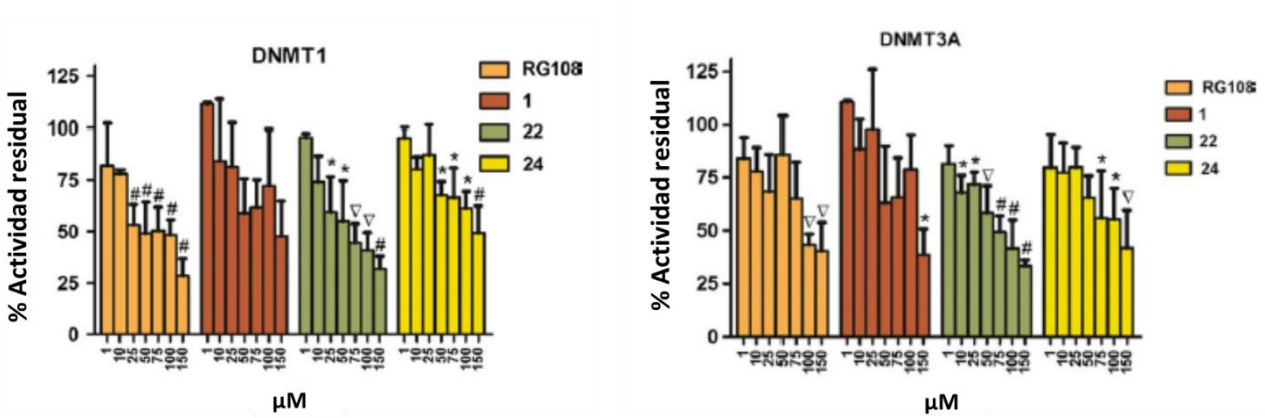


Figura5.3. Pruebas de actividad residual contra concentración del compuesto para el lisado celular con sobreexpresión selectiva de DNMT1 y DNMT3A.[72]

Todos los compuestos presentaron actividad inhibitoria contra DNMT1 y DNMT3A, en un intervalo del 28 al 49% para el caso de DNMT1 y de 39 a 41% para DNMT3A de actividad residual a la máxima concentración utilizada. Es notorio que el compuesto **22** presenta una actividad tan alta como el compuesto de referencia RG108, con valores de inhibición de 25 µM (41 ± 18% frente a 47 ± 10% para RG108 en el caso de DNMT1 y de 38 ± 6%. frente 32 ± 17% de RG108 para DNMT3A). [72]

Para ambas enzimas, tanto el compuesto **1** como el **24** demostraron una actividad inhibitoria menor a los antes mencionados. Por último, ninguno de los compuestos presentó actividad significativa contra DNMT3B a una concentración mayor a 100 µM.[72]

Para evaluar la capacidad inhibitoria del compuesto **22** de manera más exhaustiva, se realizaron pruebas *in situ* directamente sobre la enzima DNMT1 y DBMT3A utilizando como compuestos de referencia al RG108 y al compuesto **1**. Los resultados se resumen en la **Tabla 5.2**.

Compuesto	DNMT1(media+DS[b])	DNMT1(media+DS[b])
1	66±15[c]	61±17[c]
22	58±11[d]	51±13[c]
RG108	65±8[c]	79±3

ANOVA y pruebas post hoc con el test de Bonferroni.

[a] Las enzimas inmunoprecipitadas fueron incubadas con el compuesto 1 y 22 a 100 μM(1% de DMSO) a 37°C por 2hrs. La inhibición de fue medida mediante el ensayo de actividad-inhibición de DNMT del kit.

[b] Datos Expresados como el porcentaje de metilación residual relativa con un vehículo de DMSO al 1%. Los resultados son la media de triplicados del ensayo.

[c] $p < 0.1$

[d] $p < 0.05$ contra vehículo.

Tabla 5.2. [a]Actividad enzimática residual relativa selectiva.[72]

Las enzimas fueron tratadas a una concentración de 100 μM para cada uno de los compuestos. Todas las moléculas fueron capaces de inhibir a las dos dianas. En especial **22** presenta actividad inhibitoria contra DNMT1 y DNMT3A de 42 y 49% respectivamente. Estos resultados sugieren que la química de los derivados de este N-benzoyl-aminoácidos pueden ser explorados en el futuro para generar inhibidores polifarmacológicos contra DNMT1 y 3A.[72]

Estudios de acoplamiento molecular del compuesto 22 sobre DNMT1 y DNMT3A

Las técnicas de simulación molecular asistidas por computadora han demostrado ser una herramienta invaluable para la generar hipótesis estructurales que expliquen la interacción ligando-receptor.[99,100,101,102,103] En muchas ocasiones estas técnicas también son utilizadas para sentar las bases estructurales que dan razón a los posibles modos de unión de compuestos que han sido probados experimentalmente como inhibidores de una diana en específico. Generalmente, los estudios *a posteriori* se realizan cuando no se cuenta información experimental suficiente sobre la estructura e interacción de ligando sobre una diana

determinada. La ausencia de dicha información se encuentra habitualmente relacionada con las dificultades experimentales implicadas en la determinación estructural de sistemas de alta complejidad.

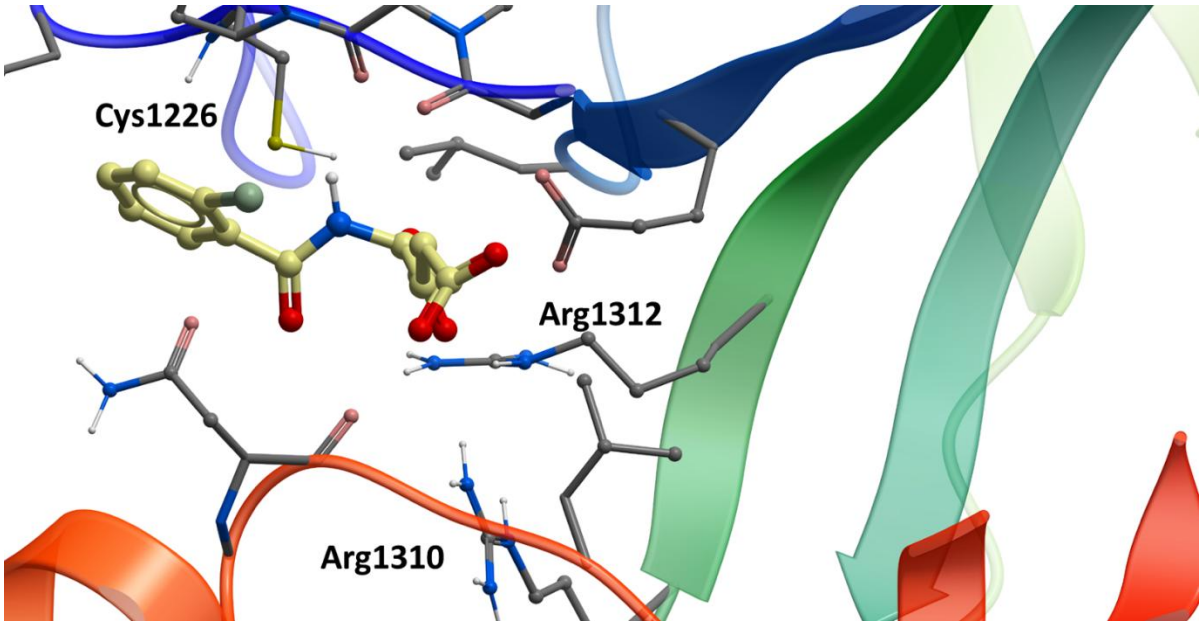
Estos hechos justifican el uso acoplamiento molecular como una herramienta eficaz dentro de los pasos preliminares de campañas experimentales de optimización de compuestos bioactivos para la modulación de los sustituyentes que deben satisfacer al farmacóforo. Para el caso de DNMT1 y DNMT3A esto se encuentra plasmado en el trabajo el trabajo de Garella, Atlante, Borretto, et al. [72]

Para el presente trajo de colaboración, se utilizó esta técnica como medio para elucidar los posibles modos de unión del compuesto **22** sobre el sitio catalítico de las estructuras cristalográfica de DNMT1 (PDB ID: 3PTA) y DNMT3A (PDB ID: 2QRV) curadas por medio de las opciones preestablecidas del programa ICM. El acoplamiento molecular flexible fue realizado mediante el mismo paquete de programas. Este, utiliza un algoritmo de optimización Monte Carlo, por lo que cada uno de los experimentos de simulación fue realizado por triplicado para garantizar la convergencia de los resultados.

Se seleccionaron a los diez primeros modos de unión de menor energía utilizando inspección visual dentro de la interface de ICM como criterio de exclusión de artefactos. Para cada uno de los modos de unión se obtuvo el mapa bidimensional de interacciones ligando-proteína mediante el paquete de programas MOE 2014.09.

La **Figura 5.4.** muestra la representación gráfica tridimensional de modo de unión del compuesto **22** sobre el sitio catalítico de DNMT1 en usencia del cofactor SAM. Esta imagen también muestra el mapa bidimensional de las interacciones ligando-receptor.

A



B

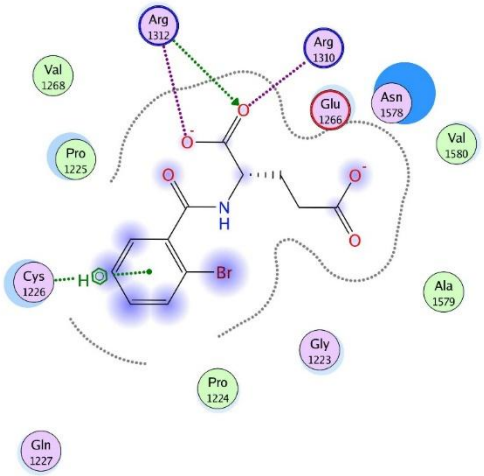


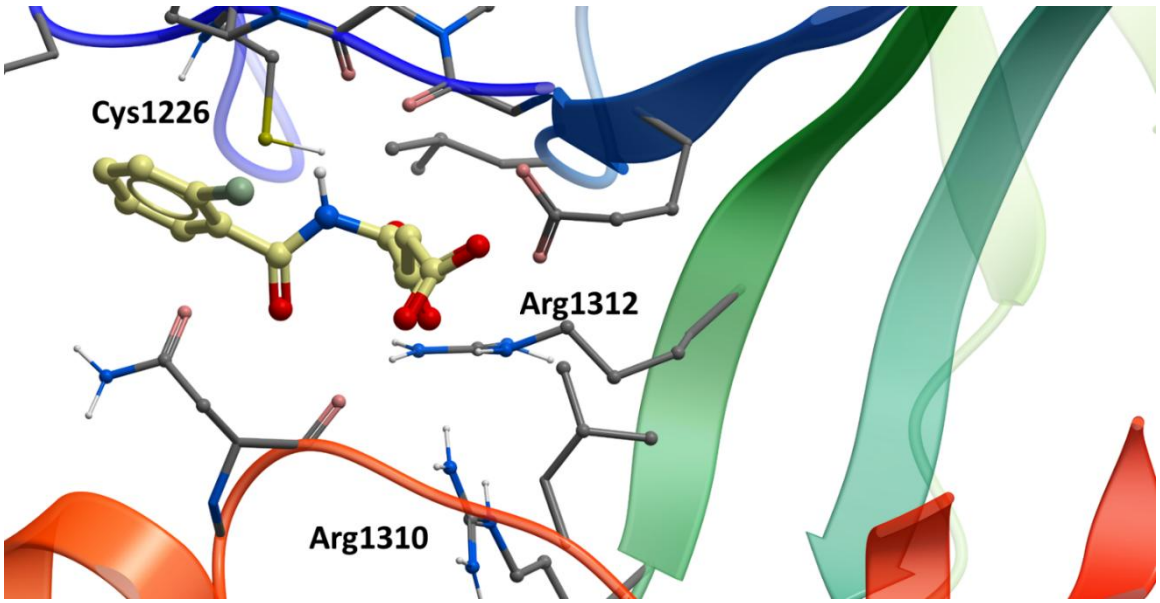
Figura 5.4. A) Modo de unión de compuesto **22** en DNMT1 y B) mapa 2D de interacciones.[72]

De acuerdo con este modelo, el grupo γ-carboxílico del compuesto **22** participa en dos interacciones tipo puente de hidrógeno con el residuo Arg1312 y Arg1310. También se observa una interacción entre el hidrógeno del tiol de la Cys1226 y el areno de ligando. Tanto el modo de unión que presenta el compuesto **22** dentro de

esta cavidad, como la interacción con las cisteína catalítica, sugieren que este ligando puede realizar la inhibición bloqueando el sitio catalítico de DNMT1.[72]

La **Figura 5.5.** muestra el modelo de acoplamiento molecular obtenido para la interacción del compuesto **22** con el sitio catalítico de DNMT3A, incluyendo la representación gráfica del modo de unión y el mapa de interacciones bidimensional.

A



B

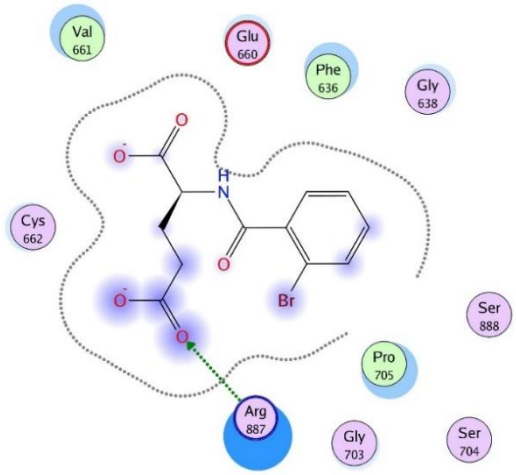


Figura 5.5. A) Modo de unión de compuesto **22** en DNMT3A y B) mapa 2D de interacciones. [72]

El modo de unión presenta una interacción entre el grupo γ-carboxilo de **22** y la Arg887, sin embargo, a diferencia del modelo propuesto para el sitio catalítico de DNMT1, no se encuentra ninguna interacción con la cisteína catalítica (Cys662). Aun cuando se podría suponer que la inhibición realizada por el compuesto **22** sobre DNMT3A se realiza bloqueando en sitio catalítico, esta hipótesis es menos verosímil que el caso anterior al contar con un número menor de interacciones que estabilicen la interacción proteína-ligando.[72]

Resumen de resultados

En este trabajo se realizaron estudios de acoplamiento molecular *a posteriori* para elucidar los modos de unión de un inhibidor de DNMT1 y DNMT3B sintetizado dentro de una campaña experimental de optimización de compuestos N-benzoyl aminoácidos inspirados en el inhibidor de DNMT NSC137546 en el laboratorio del Dr. Massimo Bertinaria (Universidad de Torino). El compuesto NSC137546 fue identificado anteriormente por medio de tamizado molecular de una base de datos del NCI.

Para realizar la optimización del compuesto, se modificaron de forma sistemática tres regiones moleculares del inhibidor: fragmento aromático, enlace amida y grupo ácido terminal. Ello dio como resultado la producción de 27 análogos que fueron sometidos a pruebas de actividad *in vitro* e *in situ* para probar su capacidad desmetilante. De los resultados de las pruebas actividad biológicas se encontró que el compuesto denominado **22** (en la publicación original), el cual presenta la sustitución de un átomo de bromo en la posición orto del fragmento aromático, exhibe los mejores valores de actividad inhibitoria de la serie, presentando un porcentaje de inhibición *in situ* contra DNMT1 y DNMT3A de 42 y 49% respectivamente.

A partir de los resultados experimentales se seleccionó al compuesto **22** para proponer los posibles modos de unión sobre el sitio catalítico en ambas dianas DNMT1 (PDB ID: 3PTA) y DNMT3A (PDB ID: 2QRV) curadas por medio de las opciones preestablecidas del programa ICM. El acoplamiento molecular flexible fue realizado por triplicando para asegurar la convergencia del algoritmo Monte

Carlo del programa ICM-Pro versión 3.8-4. Los resultados obtenidos sugieren que la actividad inhibitoria del compuesto **22** puede realizarse mediante el bloqueo de la cavidad catalítica de DNMT1 y 3A. En el primer caso se encuentran dos interacciones tipo puente de hidrógeno entre el grupo γ-carboxílico y la Arg1310 y Arg1312 además de un enlace areno-hidrógeno con la cisteína catalítica 1226. Para el caso de DNMT3A el modelo propone sólo una interacción entre la Arg887 y el grupo γ-carboxílico del compuesto **22**.

De manera general, los resultados de la simulación se encuentran de acuerdo con las RAE dilucidadas experimentalmente y pueden ser parte de las evidencias que sustentan dichas relaciones. Estos resultados también pueden ser de gran utilidad para futuras campañas de optimización de los N-benzoyl aminoácidos, ya que permiten generar estrategias de optimización precisas basadas en la modulación estructural del ligando para favorecer la formación de interacciones específicas que lleven a un aumento de la actividad inhibitoria.

Es importante mencionar que el papel de los halógenos sustituidos en el fragmento aromático de estos análogos también puede ser gran relevancia en el RAE. Ya que el acoplamiento molecular está basado en mecánica molecular, no considera factores relevantes para la simulación de interacción halógeno receptor. De concretarse evidencia experimental sobre la relevancia de las RAE de dichas sustituciones en el grupo aromático de los N-benzoyl aminoácidos, sería necesaria la aplicación de métodos *ab initio* para determinar sus causas a un nivel de descripción electrónico.

Para obtener información detallada sobre la metodología de síntesis de los análogos de NSC137546, así como de las pruebas de actividad inhibitoria *in situ* e *in vito* se puede consultar el trabajo Garella, Atlante, Borretto, et al.[72]

Conclusiones generales

Por medio de la aplicación de una serie heterogénea de modelos computacionales que hacen uso tanto la información estructural de DNMT1, como de la información disponible públicamente de inhibidores probados biológicamente, fue posible determinar su diversidad estructural y cobertura del espacio químico, así como las bases estructurales que rigen su reconocimiento molecular y modo de unión sobre esta diana biológica.

Estos resultados pueden ser de gran utilidad en campañas de optimización de candidatos moleculares como una alternativa al diseño racional clásico. Al acotar el espacio químico de los inhibidores de DNMT1, estas técnicas permiten disminuir los costos de investigación, el uso de modelos animales y los tiempos de desarrollo.

Referencias

(1) Yoo, C. B.; Jones, P. A. Epigenetic Therapy of Cancer: Past, Present and Future. *Nat Rev Drug Discov* **2006**, *5* (1), 37–50.

(2) Auclair, G.; Weber, M. Mechanisms of DNA Methylation and Demethylation in Mammals. *Biochimie* **2012**, *94* (11), 2202–2211.

(3) Berger, S. L.; Kouzarides, T.; Shiekhattar, R.; Shilatifard, A. An Operational Definition of Epigenetics An Operational Definition of Epigenetics. *Genes Dev.* **2009**, *23* (7), 781–783.

(4) Rius, M.; Lyko, F. Epigenetic Cancer Therapy: Rationales, Targets and Drugs. *Oncogene* **2012**, *31* (39), 4257–4265.

(5) Medina-Franco, J. L.; Méndez-Lucio, O.; Dueñas-González, A.; Yoo, J. Discovery and Development of DNA Methyltransferase Inhibitors Using in Silico Approaches. *Drug Discov. Today* **2015**, *20* (5), 569–577.

(6) Stefanska, B.; Karlic, H.; Varga, F.; Fabianowska-Majewska, K.; Haslberger, A. G. Epigenetic Mechanisms in Anti-Cancer Actions of Bioactive Food Components - The Implications in Cancer Prevention. *Br. J. Pharmacol.* **2012**, *167* (2), 279–297.

(7) Gros, C.; Fahy, J.; Halby, L.; Dufau, I.; Erdmann, A.; Gregoire, J. M.; Ausseil, F.; Vispé, S.; Arimondo, P. B. DNA Methylation Inhibitors in Cancer: Recent and Future Approaches. *Biochimie* **2012**, *94* (11), 2280–2296.

(8) Fahy, J.; Jeltsch, A.; Arimondo, P. B. DNA Methyltransferase Inhibitors in Cancer: A Chemical and Therapeutic Patent Overview and Selected Clinical Studies. *Expert Opin. Ther. Pat.* **2012**, *22* (12), 1–16.

(9) Medina-Franco, J. Y. and J. L. Inhibitors of DNA Methyltransferases: Insights from Computational Studies. *Current Medicinal Chemistry*. 2012, pp 3475–3487.

(10) Méndez-Lucio, O.; Romo-Mancillas, A.; Medina-Franco, J. L.; Castillo, R. Computational Study on the Inhibition Mechanism of Cruzain by Nitrile-Containing Molecules. *J. Mol. Graph. Model.* **2012**, *35*, 28–35.

(11) Jeltsch, A. Beyond Watson and Crick: DNA Methylation and Molecular Enzymology of DNA Methyltransferases. *Chembiochem* **2002**, *3* (4), 274–293.

(12) Bestor, T. H. The DNA Methyltransferases of Mammals. *Hum. Mol. Genet.* **2000**, *9* (16), 2395–2402.

(13) Dhe-Paganon, S.; Syeda, F.; Park, L. DNA Methyl Transferase 1: Regulatory Mechanisms and Implications in Health and Disease. *Int. J. Biochem. Mol. Biol.* **2011**, *2* (1), 58–66.

(14) Erdmann, A.; Arimondo, P. B.; Guianvarc'h, D. *Structure-Guided Optimization of DNA Methyltransferase Inhibitors*; Elsevier Inc., 2016.

(15) Gortari, E. F.; Medina-Franco, J. L. Epigenetic Relevant Chemical Space: A Chemoinformatic Characterization of Inhibitors of DNA Methyltransferases. *RSC Adv.* **2015**, *5* (106), 87465–87476.

(16) Ceccaldi, A.; Rajavelu, A.; Ragozin, S.; Sénamaud-Beaufort, C.; Bashtrykov, P.; Testa, N.; Dali-Ali, H.; Maulay-Bailly, C.; Amand, S.; Guianvarc'H, D.; et al. Identification of Novel Inhibitors of Dna Methylation by Screening of a Chemical Library. *ACS Chem. Biol.* **2013**, *8* (3), 543–548.

(17) Prieto-Martínez, F. D.; Peña-Castillo, A.; Méndez-Lucio, O.; Fernández-De Gortari, E.; Medina-Franco, J. L.; Endez-Lucio, O. M. Molecular Modeling

and Chemoinformatics to Advance the Development of Modulators of Epigenetic Targets: A Focus on DNA Methyltransferases. *Adv. Protein Chem. Struct. Biol.* **2016**, *105* (105), 1–26.

(18) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(19) Medina-Franco, J. L. *Epi-Informatics*, 1st ed.; Medina-Franco, Ed.; Elsevier: Mexico City, 2016; Vol. 1.

(20) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures BT - Chemoinformatics and Computational Chemical Biology; Bajorath, J., Ed.; Humana Press: Totowa, NJ, 2011; pp 39–100.

(21) Barnard, J. M. Representation of Molecular Structures-Overview. In *Handbook of Chemoinformatics*; Wiley-VCH Verlag GmbH, 2008; pp 27–50.

(22) Www.chemspider.com.

(23) Shoichet, B. K. Drug Discovery: Nature's Pieces. *Nat Chem* **2013**, *5* (1), 9–10.

(24) Maggiora, G. M. Introduction to Molecular Similarity and Chemical Space BT - Foodinformatics: Applications of Chemical Information to Food Chemistry; Martinez-Mayorga, K., Medina-Franco, J. L., Eds.; Springer International Publishing: Cham, 2014; pp 1–81.

(25) Guha, R. A Handbook of Cheminformatics Algorithms; Faulon, J. L., Bender, A., Eds.; Wiley: New York, NY, 2009.

(26) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(27) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (3), 569–574.

(28) Xu, Y.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (1), 181–185.

(29) Jaccard, P. Étude Comparative de La Distribution Florale Dans Une Portion

Des Alpes et Des Jura. *Bull. del la Société Vaudoise des Sci. Nat.* **1901**, *37*, 547–579.

(30) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.

(31) Prieto-Martínez, F. D.; Gortari, E. F.; Méndez-Lucio, O.; Medina-Franco, J. L. A Chemical Space Odyssey of Inhibitors of Histone Deacetylases and Bromodomains. *RSC Adv.* **2016**, *6* (61), 56225–56239.

(32) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791.

(33) Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, *28* (6), 1145–1152.

(34) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: MEthods and Applications. *Nat. Rev. Drug Disc.* **2004**, *3* (11), 935–949.

(35) Chaudhary, K. K.; Mishra, N. A Review on Molecular Docking : Novel Tool for Drug Discovery. *JSM Chem.* **2016**, *4*, 1–4.

(36) Www.molsoft.com.

(37) Wang, Y.; Law, W.-K.; Hu, J.-S.; Lin, H.-Q.; Ip, T.-M.; Wan, D. C.-C. Discovery of FDA-Approved Drugs as Inhibitors of Fatty Acid Binding Protein 4 Using Molecular Docking Screening. *J. Chem. Inf. Model.* **2014**, *54* (11), 3046–3050.

(38) Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. Docking: Successes and Challenges. *Curr. Pharm. Des.* **2005**, *11* (c), 323–333.

(39) Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998) . *Pure and Applied Chemistry* . 1998, p 1129.

(40) Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF)

for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* **2014**, *54* (9), 2555–2561.

(41) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput Biol* **2014**, *10* (4), e1003571.

(42) Holm, L.; Sander, C. Dali: A Network Tool for Prtoein Structure Comparison. *Trends Biochem Sci* **1995**, *20* (June), 478–480.

(43) Harris, R.; Olson, A. J.; Goodsell, D. S. Automated Prediction of Ligand-Binding Sites in Proteins. *Proteins* **2008**, *70* (4), 1506–1517.

(44) Brady, G. P. J.; Stouten, P. F. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *J. Comput. Aided. Mol. Des.* **2000**, *14* (4), 383–401.

(45) Panjkovich, A.; Daura, X. Exploiting Protein Flexibility to Predict the Location of Allosteric Sites. *BMC Bioinformatics* **2012**, *13* (1), 273.

(46) Alvarez-Garcia, D.; Barril, X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J. Med. Chem.* **2014**, *57* (20), 8530–8539.

(47) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, 1963.

(48) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28* (11–12), 1551–1560.

(49) Vogt, M.; Wassermann, A. M.; Bajorath, J. Application of Information—Theoretic Concepts in Chemoinformatics. *Information* **2010**, *1* (2), 60–73.

(50) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; 2004.

(51) Molecular Operating Environment (MOE), version 2010, Chemical Computing Group Inc., Montreal, Quebec, Canada. Available at http://www.chemcomp.com.

(52) Www.mayachemtools.org.

(53) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of

Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455–461.

(54) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100-7.

(55) Huang, Z.; Jiang, H.; Liu, X.; Chen, Y.; Wong, J.; Wang, Q.; Huang, W.; Shi, T.; Zhang, J. HEMD: An Integrated Tool of Human Epigenetic Enzymes and Chemical Modulators for Therapeutics. *PLoS One* **2012**, *7* (6), e39917.

(56) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein–ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35* (Database issue), D198–D201.

(57) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204.

(58) Www.r-Project.org.

(59) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473.

(60) Medina-Franco, J. L.; Edwards, B. S.; Pinilla, C.; Appel, J. R.; Giulianotti, M. A.; Santos, R. G.; Yongye, A. B.; Sklar, L. A.; Houghten, R. A. Rapid Scanning Structure-Activity Relationships in Combinatorial Data Sets: Identification of Activity Switches. *J. Chem. Inf. Model.* **2013**, *53* (6), 1475–1485.

(61) Pérez-Villanueva, J.; Méndez-Lucio, O.; Soria-Arteche, O.; Medina-Franco, J. L. Activity Cliffs and Activity Cliff Generators Based on Chemotype-Related Activity Landscapes. *Mol. Divers.* **2015**, *19* (4), 1021–1035.

(62) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28* (11–12), 1551–1560.

(63) Www.sbpdiscovery.org.

(64) Chen, S.; Wang, Y.; Zhou, W.; Li, S.; Peng, J.; Shi, Z.; Hu, J.; Liu, Y. C.; Ding, H.; Lin, Y.; et al. Identifying Novel Selective Non-Nucleoside DNA Methyltransferase 1 Inhibitors through Docking-Based Virtual Screening. *J. Med. Chem.* **2014**, *57* (21), 9028–9041.

(65) Song, J.; Rechkoblit, O.; Bestor, T. H.; Patel, D. J. Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. *Science* **2011**, *331* (6020), 1036–1040.

(66) Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation . *Casp. J. Intern. Med.* **2013**, *4* (2), 627–635.

(67) Morris, G. M.; Lim-Wilby, M. Molecular Docking. *Methods Mol. Biol.* **2008**, *443*, 365–382.

(68) Yoo, J.; Kim, J. H.; Robertson, K. D.; Medina-Franco, J. L. *Molecular Modeling of Inhibitors of Human DNa Methyltransferase with a Crystal Structure: Discovery of a Novel dnmt1 Inhibitor*, Eighth Edi.; Elsevier Inc., 2012; Vol. 87.

(69) Www.chemcomp.com/journal/cstat.htm.

(70) Yoo, J.; Medina-Franco, J. L. Homology Modeling, Docking and Structure-Based Pharmacophore of Inhibitors of DNA Methyltransferase. *J. Comput. Aided. Mol. Des.* **2011**, *25* (6), 555–567.

(71) Mendez-Lucio, O.; Tran, J.; Medina-Franco, J. L.; Meurice, N.; Muller, M. Toward Drug Repurposing in Epigenetics: Olsalazine as a Hypomethylating Compound Active in a Cellular Context. *ChemMedChem* **2014**, *9* (3), 560–565.

(72) Garella, D.; Atlante, S.; Borretto, E.; Cocco, M.; Giorgis, M.; Costale, A.; Stevanato, L.; Miglio, G.; Cencioni, C.; Fernández-de Gortari, E.; et al. Design and Synthesis of N-Benzoyl Amino Acid Derivatives as DNA Methylation Inhibitors. *Chem. Biol. Drug Des.* **2016**, No. May, 664–676.

(73) Https://dtp.cancer.gov.

(74) Goncearenco, A.; Mitternacht, S.; Yong, T.; Eisenhaber, B.; Eisenhaber, F.; Berezovsky, I. N. SPACER: Server for Predicting Allosteric Communication

and Effects of Regulation. *Nucleic Acids Res.* **2013**.

(75) Sallem, M. A. S.; Sousa, S. A. jo de. AutoGrid: Towards an Autonomic Grid Middleware. In *Proceedings of the 16th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*; WETICE '07; IEEE Computer Society: Washington, DC, USA, 2007; pp 223–228.

(76) Mitternacht, S.; Berezovsky, I. N. A Geometry-Based Generic Predictor for Catalytic and Allosteric Sites. *Protein Eng. Des. Sel.* **2011**, *24* (4), 405–409.

(77) Panjkovich, A.; Daura, X. Assessing the Structural Conservation of Protein Pockets to Study Functional and Allosteric Sites: Implications for Drug Discovery. *BMC Struct. Biol.* **2010**, *10*, 9.

(78) Pars @ Bioinf.uab.cat.

(79) Allostery.bii.a-Star.edu.sg.

(80) Boyd, D. B. *Reviews in Computational Chemistry Volume 17 Reviews in Computational Chemistry Volume 16*; 2002; Vol. 16.

(81) Aguayo-Ortiz, R.; Pérez-Villanueva, J.; Hernández-Campos, A.; Castillo, R.; Meurice, N.; Medina-Franco, J. L. Chemoinformatic Characterization of Activity and Selectivity Switches of Antiprotozoal Compounds. *Future Med. Chem.* **2013**, *6* (3), 281–294.

(82) Mei, F.; Fancy, S. P. J.; Shen, Y.-A. A.; Niu, J.; Zhao, C.; Presley, B.; Miao, E.; Lee, S.; Mayoral, S. R.; Redmond, S. A.; et al. Micropillar Arrays as a High-Throughput Screening Platform for Therapeutics in Multiple Sclerosis. *Nat. Med.* **2014**, *20* (8), 954–960.

(83) Qin, C.; Zhang, C.; Zhu, F.; Xu, F.; Chen, S. Y.; Zhang, P.; Li, Y. H.; Yang, S. Y.; Wei, Y. Q.; Tao, L.; et al. Therapeutic Target Database Update 2014: A Resource for Targeted Therapeutics. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1118–D1123.

(84) Medina-Franco, J. L.; Yoo, J.; Dueñas-González, A. Chapter 13 - DNA Methyltransferase Inhibitors for Cancer Therapy A2 - Zheng, Y. George BT - Epigenetic Technological Applications; Academic Press: Boston, 2015; pp 265–290.

(85) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-

Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49* (4), 1010–1024.

(86) Gonzalez-Medina, M.; Prieto-Martinez, F. D.; Naveja, J. J.; Mendez-Lucio, O.; El-Elimat, T.; Pearce, C. J.; Oberlies, N. H.; Figueroa, M.; Medina-Franco, J. L. Chemoinformatic Expedition of the Chemical Space of Fungal Products. *Future Med. Chem.* **2016**, *8* (12), 1399–1412.

(87) Http://www.drugbank.ca.

(88) Burdock, G. A.; Carabin, I. G.; Griffiths, J. C. The Importance of GRAS to the Functional Food and Nutraceutical Industries. *Toxicology* **2006**, *221* (1), 17–27.

(89) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730.

(90) Www.random.org.

(91) Smith, S. W. Digital Signal Processors. *Sci. Eng. Guid. to Digit. Signal Process.* **1997**, No. 1, 503–534.

(92) Python. Www.Python.Org. 2015.

(93) Www.ebi.ac.uk.

(94) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594.

(95) Wang, Y.; Geppert, H.; Bajorath, J. Shannon Entropy-Based Fingerprint Similarity Search Strategy. *J. Chem. Inf. Model.* **2009**, *49* (7), 1687–1691.

(96) Kuck, D.; Singh, N.; Lyko, F.; Medina-Franco, J. L. Novel and Selective DNA Methyltransferase Inhibitors: Docking-Based Virtual Screening and Experimental Evaluation. *Bioorg. Med. Chem.* **2010**, *18* (2), 822–829.

(97) Jia, D.; Jurkowska, R. Z.; Zhang, X.; Jeltsch, A.; Cheng, X. Structure of Dnmt3a Bound to Dnmt3L Suggests a Model for de Novo DNA Methylation. *Nature* **2007**, *449* (7159), 248–251.

(98) Garella, D.; Atlante, S.; Borretto, E.; Cocco, M.; Giorgis, M.; Costale, A.; Stevanato, L.; Miglio, G.; Cencioni, C.; Fernández-de Gortari, E.; et al.

Design and Synthesis of N-Benzoyl Amino Acid Derivatives as DNA Methylation Inhibitors. *Chem. Biol. Drug Des.* **2016**.

(99) Scharfe, M.; Maurer, B.; Aktories, K.; Jung, M.; Sippl, W. Docking and Virtual Screening of Novel Inhibitors for Mono-ADP-Ribosylating Toxins. *J. Cheminform.* **2011**, *3* (SUPPL. 1), 2011.

(100) Schneider, G.; Böhm, H. J. Virtual Screening and Fast Automated Docking Methods. *Drug Discov. Today* **2002**, *7* (1), 64–70.

(101) dos Santos Filho, J. M.; Leite, A. C. L.; Oliveira, B. G. de; Moreira, D. R. M.; Lima, M. S.; Soares, M. B. P.; Leite, L. F. C. C. Design, Synthesis and Cruzain Docking of 3-(4-Substituted-Aryl)-1,2,4-Oxadiazole-N-Acylhydrazones as Anti-Trypanosoma Cruzi Agents. *Bioorganic Med. Chem.* **2009**, *17* (18), 6682–6691.

(102) Leite, A. C. L.; Moreira, D. R. D. M.; Cardoso, M. V. D. O.; Hernandes, M. Z.; Alves Pereira, V. R.; Silva, R. O.; Kiperstok, A. C.; Lima, M. D. S.; Soares, M. B. P. Synthesis, Cruzain Docking, and in Vitro Studies of Aryl-4-Oxothiazolylhydrazones against Trypanosoma Cruzi. *ChemMedChem* **2007**, *2* (9), 1339–1345.

(103) Chen, Y.; Pohlhaus, D. T. In Silico Docking and Scoring of Fragments. *Drug Discov. Today Technol.* **2010**, *7* (3), e149–e156.

Anexo

```
#Función de comparación entre el FP_DB que se calcula a partir de una Base de Datos con otra de referencia
import os
import csv
import pandas as pd
import numpy as np
from numpy import reshape
from numpy import sum as sumnp
import itertools
import math
def DB_FPVS(e):
	print("\n")
	print("##########################################################################################")
	print("#### Antes de usar instalar MayaChem Tools e incluirlo en las variables de entorno ###")
	print("##########################################################################################")
	#Cambiar carpeta de trabajo
	cd=os.getcwd()
	os.chdir('%s'%(cd))
	archivoSDF=input("Dame el nombre del archivo .sdf para calcular su DFP: ")
	# Calcular MACCS key con script de MayaChem Tools
	os.system ("MACCSKeysFingerprints.pl -r MACCSFP_%s -o %s.sdf" %(archivoSDF,archivoSDF))
	m=("MACCSFP_%s.csv" %(archivoSDF))
	##Recibir archivo .csv y edición
	g = []
	with open('%s'%(m)) as csvarchivo:
		entrada = csv.DictReader(csvarchivo)
		for i in entrada:
			i = i['MACCSKeysFingerprints']
			i = i[-168:]
			g.append(i)
	#Convertir a enteros
	f = []
	g = str.join('',g)
	for i in g:
```

```
            i = into(i)
            f.append(i)
        #Frecuencia
        j = []
        y = reshape(f, ((len(f)/168), 168) )
        h = sumnp(y,axis=0)
        #Probabilidad
        for i in h:
            i = i/(len(f)/168)
            j.append(i)
        #Entropía
        s=[]
        for i in j:
            if i==0:
                s.append(i)
            else:
                i=i*(math.log2(i))
                s.append(i)
        s=-(sum(s))
        #FP_de base de datos
        FP=[]
        for i in j:
            if i<0.55:
                FP.append("0")
            else:
                FP.append("1")
        #del FP[-2:]
        FP_DB = ''.join(FP)
        print(FP_DB)
        #Integrar el fingerprint en lista de MACCS
        with open('MACCSFP_%sVS.csv'%(e),"r") as f:
            reader = csv.reader(f,delimiter = ",")
            data = list(reader)
            row_count = len(data)
        fd = open('MACCSFP_%sVS.csv'%(e),'a')
        fd.write('"Cmpd%s","FingerprintsBitVector;MACCSKeyBits;166;BinaryString;Ascen
ding;%s"\n'%(row_count,FP_DB))
        fd.close()
        # Similitud Media:
```

```
        data=os.system ("SimilarityMatricesFingerprints.pl -o --InputDataMode LoadInMemory --OutMatrixFormat RowsAndColumns --OutMatrixType LowerTriangularMatrix MACCSFP_%sVS.csv" %(e))
        pd.set_option('max_columns', 10000000000000000000)#aumentar número de columnas explicitas en print
        print(e)
        data=pd.read_csv("MACCSFP_%sVSTanimotoSimilarity.csv"%(e),header=None,chunksize=1000000)#borrar col y raw no deseadas
        data.drop([0],axis=0,inplace=True)
        data.drop([0],axis=1,inplace=True)
        data.drop([row_count],axis=1,inplace=True)
        a1=data.iloc[[row_count-1]]#Seleccionar ultimo renglon
        a1=a1.values.tolist()#convertir a lista
        merged = list(itertools.chain.from_iterable(a1))#quitar lista de listas
        Sim=[]
        for i in merged:
                i = float(i)
                Sim.append(i)
        b1=np.array(Sim)
        SD=np.std(Sim)
        mean=np.mean(Sim)
        corte=(mean+(2*SD))
        #Guardar resultados en csv
        fd = open('Similitud_FP%s_vs_%sDB.csv'%(archivoSDF,e),'a')
        fd.write('Sim vector FP%s vs %sDB= %s \n'%(archivoSDF,e,Sim))
        fd.write('SD= %s, mean= %s, SE= %s,corte= %s '%(SD,mean,s,corte))
        fd.close()
        #Borrar última línea de csv
        f = open('MACCSFP_%sVS.csv'%(e),"r")
        lineas = f.readlines()
        lineas =lineas[:-1]# slice en lineas
        f.close()
        f = open('MACCSFP_%sVS.csv'%(e),"w")
        for linea in lineas:
                f.write(linea)
        f.close()
        return b1
```

***Script* de Python para calcular DFP.**

```
#Comparación de DFP con bases de datos de referencia y generación de Heatmaps
from FpDb_singleDB_comp2 import *
import os
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
#Cambiar dirección de carpeta de trabajo
cd=os.getcwd()
os.chdir('%s'%(cd))
print("calcular MACCS keys previamente, nombreArchivo + VS")
e="ACE_deco_act"
# e=input("Dame el nombre de la BD (.sdf) para compararla (Tanimoto) con FP_DB calculado: ")
# os.system ("MACCSKeysFingerprints.pl -r MACCSFP_%sVS -o %s.sdf" %(e,e))
#Llamada a la función DB_FPVS para calcular FP_DB y comparar con base de datos
DB1=DB_FPVS(e)
#Heatmap
data=np.array([DB1,DB1])
sns.set()
sns.heatmap(data, vmin=0, vmax=1, xticklabels=False)#cmap="PuBuGn"
ax=plt.axes()
##Leyenda
# red_patch = mpatches.Patch(color='red', label='The red data')
# plt.legend(handles=[red_patch])
##Ejes y nombres
#ax.set_ylabel('')
ax.set_xlabel('%s'%(e))
ax.set_title('FP_DB comparison')
#sns.heatmap(data, annot=True, fmt="d")
sns.plt.show()
```

***Script* python para comparar bases de datos con DFP determinados. Generación de HeatMaps.**

RSC Advances

PAPER

Cite this: RSC Adv., 2015, 5, 87465

Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases†

Eli Fernández-de Gortari and José L. Medina-Franco*

Received 23rd September 2015
Accepted 5th October 2015

DOI: 10.1039/c5ra19611f

www.rsc.org/advances

DNA methylation is an epigenetic mechanism mediated by a family of proteins called DNA methyltransferases (DNMTs). The misregulation of the covalent modification of DNA through the addition of a methyl group at the carbon-5 position of cytosine residues is common in many diseases including cancer. Recent advances in synthetic and screening technologies for DNMT inhibitors (DNMTi) have made significant contributions to uncover promising candidates for epigenetic drug discovery. The structure–activity information, not available a few years ago, is being collected in public molecular databases. However, no systematic chemoinformatic studies that analyze the structural diversity and coverage of the chemical space of DNMTi with experimental activity have been discussed thus far. Herein, we report the assembly and curation of a molecular database of small-molecule DNMTi with a special focus on inhibitors of DNMT1. The compound collection was characterized using a comprehensive chemoinformatic approach that involved physicochemical properties, structural fingerprints, and molecular scaffolds. The availability of activity information enabled us to conduct chemotype enrichment analysis and suggest potential privileged epigenetic scaffolds. The structures of inhibitors of DNMT1 were compared to drugs approved for clinical use, compounds in clinical trials, a commercial screening library focused on epigenetic targets, and a general screening collection. The results of this work provided key insights to start characterizing the epigenetic relevant chemical space.

1. Introduction

DNA methylation is an epigenetic modification involving the addition of methyl group at the C-5 position of a cytosine residue. This process plays a key role in mammal development and in cancer cell growth. The methylation process is mediated by an enzymatic family called DNA methyltransferases (DNMTs). In humans, this family includes DNMT1, DNMT2, DNMT3A and DNMT3B.[1] DNMT1 and DNMT3B exhibit higher activity, which can be inferred from the strong reduction in DNA methylation in cell lines with double knock-out. DNA methylation represents one of the main mediations of epigenetic regulation. Therefore, the identification of novel DNMT inhibitors (DNMTi) is a promising research avenue to develop novel therapies against cancer and other diseases associated with epigenetic alterations.[2–4]

Currently, 5-aza and 5-aza-2′-deoxycytidine are two drugs approved for clinical use for the treatment of myelodysplasia (Fig. 1). 5-Aza and 5-aza-2′-deoxycytidine are nucleoside analogues which, after its incorporation into DNA, cause depletion of the DNMTs. However, these drugs have high toxicity, low bioavailability and low chemical stability, coupled with an uncertain mechanism of antitumor activity.[5] For this reason, research efforts to discover non-nucleoside DNMTi with greater specificity and lower toxicity are needed.

One of the main advantages of non-nucleoside DNMTi is that they do not need to be incorporated into the DNA. This characteristic contributes to the possible development of selective inhibitors against different DNMTs with the consequent decrease of unwanted side effects. Thus far, several non-nucleoside inhibitors have been identified such as SGI-1027, procainamide, tea polyphenol (−)-epigallocatechin 3-gallate, genistein, NSC401077, hydralazine, among others.[6,7] The first-generation of inhibitors showed low activity and selectivity against DNMTs. However, new generations of inhibitors with increased activity and selectivity profile have been developed, such as analogs of SGI-1027 (Fig. 1).[8] Nevertheless, these compounds have low potency especially in cells and lack of selectivity towards different DNMTs. Fig. 1 shows representative inhibitors of DNMT1 and molecules associated with demethylating properties.

The increased research efforts to develop more potent and specific DNMTi have augmented notoriously the number of screening data. Compounds from different sources including

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

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5ra19611f

Received: 1 February 2016 | Revised: 2 May 2016 | Accepted: 14 May 2016

DOI: 10.1111/cbdd.12794

RESEARCH ARTICLE

WILEY

Design and synthesis of *N*-benzoyl amino acid derivatives as DNA methylation inhibitors

Davide Garella[1] | Sandra Atlante[2] | Emily Borretto[1] | Mattia Cocco[1] | Marta Giorgis[1] | Annalisa Costale[1] | Livio Stevanato[1] | Gianluca Miglio[1] | Chiara Cencioni[2] | Eli Fernández-de Gortari[3] | José L. Medina-Franco[3] | Francesco Spallotta[2] | Carlo Gaetano[2] | Massimo Bertinaria[1]

[1]Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Torino, Italy

[2]Division of Cardiovascular Epigenetics Department of Cardiology, Goethe University, Frankfurt am Main, Germany

[3]Facultad de Química Departamento de Farmacia, Universidad Nacional Autónoma de México, Mexico City, México

Correspondence

Davide Garella, davide.garella@unito.it

and

Carlo Gaetano, gaetano@em.uni-frankfurt.de

The inhibition of human DNA Methyl Transferases (DNMT) is a novel promising approach to address the epigenetic dysregulation of gene expression in different diseases. Inspired by the validated virtual screening hit NSC137546, a series of *N*-benzoyl amino acid analogues was synthesized and obtained compounds were assessed for their ability to inhibit DNMT-dependent DNA methylation *in vitro*. The biological screening allowed the definition of a set of preliminary structure–activity relationships and the identification of compounds promising for further development. Among the synthesized compounds, *L*-glutamic acid derivatives **22**, **23**, and **24** showed the highest ability to prevent DNA methylation in a total cell lysate. Compound **22** inhibited DNMT1 and DNMT3A activity in a concentration-dependent manner in the micromolar range. In addition, compound **22** proved to be stable in human serum and it was thus selected as a starting point for further biological studies.

KEYWORDS

DNA methylation, DNMT inhibitors, docking, epigenetics, structure–activity relationships

Epigenetic modifications play an essential role in the establishment and regulation of cellular differentiation and gene expression.[1,2] DNA methylation is the most stable epigenetic mark in humans.[3] The DNA methylation occurs at the C5 position of the cytosine ring, particularly in a CpG dinucleotide context, through the action of three active DNA methyltransferases (DNMTs): DNMT1, DNMT3A, and DNMT3B. These enzymes catalyze the transfer of a methyl group from *S*-adenosyl-*L*-methionine (SAM) to the C5-cytosine.[4] DNMT1 is responsible for DNA methylation maintenance during cell replication by methylation of newly synthesized DNA strands; however, it was hypothesized that this enzyme can also participate in the *de novo* methylation process.[5] DNMT3A and DNMT3B are responsible for *de novo* DNA hemimethylated DNA strands.[6,7] Another protein, lacking enzymatic activity, namely DNMT3L, is capable of interacting with DNMT3A and DNMT3B with the consequence of stimulating their catalytic activity.[8]

In human genome, CpG dinucleotides are typically clustered in regions called CpG islands, which are located in the proximal promoter of more than half of all human genes.[9] When promoter CpG islands are methylated, the corresponding gene is repressed because of poor recognition by transcription factors and by other methyl-binding proteins (MBDs) involved in chromatin remodeling and reorganization.[10]

Aberrant DNA methylation, or the failure to maintain the appropriate DNA methylation status, results in the expression of non-optimal level of gene-associated proteins, which

RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: RSC Adv., 2016, 6, 56225

A chemical space odyssey of inhibitors of histone deacetylases and bromodomains†

Fernando D. Prieto-Martínez, Eli Fernández-de Gortari, Oscar Méndez-Lucio and José L. Medina-Franco*

Received 18th March 2016
Accepted 4th June 2016

DOI: 10.1039/c6ra07224k

www.rsc.org/advances

The interest in epigenetic drug and probe discovery is growing as reflected in the large amount of structure-epigenetic activity information available. Therefore, the significance of understanding the entire or fractions of the epigenetic relevant chemical space is increasing. Major epigenetic targets are histone lysine deacetylases (HDACs), bromodomains (BRDs), and DNA methyltransferases (DNMTs). However, with the exception of DNMTs, characterization of the chemical space of these epi-targets is limited. This work is the first chemoinformatic analysis of the physicochemical properties, structural diversity, and coverage of the chemical space of compounds screened as inhibitors of HDACs and BRDs. The chemical space was compared to DNMTIs, approved drugs, commercial screening compounds, and generally recognized as safe (GRAS) molecules. The structural complexity of compounds directed towards epigenetic targets was also addressed. The outcome of this analysis indicated that it is required to increase the structural diversity and molecular complexity of screening libraries tested as modulators of DNMTs, HDACs and BRDs. Results also suggested that it is feasible to develop dual inhibitors targeting HDACS and BRDs. This work has implications in repurposing of food chemicals with potential epigenetic activity and design of poly-epigenetic compounds.

1. Introduction

Every living being has the ability to inherit its genetic material. However this process is not flawless. After a few decades, the study of DNA repair lead to the discovery of higher order mechanisms and the term 'epigenetics' was coined.[1] Initially, inhibition of epigenetic targets was considered a novel alternative for the treatment of cancer. While this approach may be true, current research showed that environmental factors such as radiation exposure, nutritional history, dietary intake, reproductive factors, among others, also play a key role on the expression of specific epigenetic modifications.[1–3] Nowadays it is accepted that epi-modulation can act as a link between genotype and environment stimuli[4] and may be used as a Rosetta Stone to better understand, prevent and/or cure diseases. While this is yet to be proved, many researchers consider epigenetics as the missing link on the biogenesis of chronic diseases like Alzheimer's, dementia, schizophrenia, diabetes, metabolic syndrome, to name few examples.[5,6]

Chemical modifications are key features in epigenetics. Although the number of reactions and enzymes involved are different and comprise more than one hundred, it is possible to distinguish three functions: writers, erasers and readers. Writers add chemical groups that can be labile or stable. Erasers remove the groups added by writing enzymes. Readers are 'effector proteins' that identify specific chemical groups associated with epigenetic modifications and produce large scale changes such as chromatin remodeling or recruitment of other enzymes involved in DNA replication or gene expression.[7]

The correlation between epigenetic changes and carcinogenesis attracted attention to histone deacetylases (HDACs). Acetylation on lysine residues is one of the most common processes on epigenetics.[3] Eighteen different HDACs have been identified, characterized and classified in three classes. Class I comprises HDACs 1, 2, 3 and 8, that are located on the nucleus with involvement on development of numerous cancer types.[8] Class III gathers seven HDACs that are NAD^+ dependent and sirtuin constituted and are known as SIRT 1–7. This class has been mainly involved with pancreas and breast cancer, nevertheless some of them (*e.g.* SIRT1) may be involved with type II diabetes.[9] Although the removal of acetate groups from histone tails may be conceived as the first step towards transcriptional repression, it has been shown that regulation by HDACs goes beyond histones acting in a plethora of cellular pathways.[10] Despite major efforts from industry and academia and the baffling amount of chemical and biochemical studies towards these enzymes, the Food and Drug Administration (FDA) of the United States has approved only four drugs so far for clinical

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

CHAPTER 2

c0002 Overview of Computer-Aided Drug Design for Epigenetic Targets

Rodrigo Aguayo-Ortiz, Eli Antonio Alonso Fernández de Gortari
Departamento de Farmacia, Universidad Nacional Autónoma de México, México City, México

s0010 1. INTRODUCTION

p0015 Computational approaches have become indispensable tools to accelerate the development of epigenetic inhibitors helping in the selection, design, and lead identification of novel compounds. Computational methods are also useful for optimization phases and to decrease the costs of industrial drug development.

p0020 During the past few years, several in silico approaches such as similarity searching, molecular docking, virtual screening, pharmacophore modeling, and molecular dynamics (MD) have been applied to help understand the activity of known compounds and to design novel epigenetic inhibitors. Chemoinformatics tools are used to select compounds that are already characterized to identify compounds that have an increased opportunity of having activity against a target during the experimental tests. Following these processes, and depending on the information available, more detailed (and usually costly) techniques are further employed to identify relationships between structure and activity, the binding modes and affinities with docking and MD or to perform searches based in pharmacophore modeling.

p0025 This chapter provides an overview of the main techniques used in computational drug design to date. Special emphasis is made on the application of computational approaches within the efforts for the design of bioactive molecules against targets related to epigenetic regulation, which are known to be involved in diseases such as cancer, heart disease, psychiatric, among others. In order to facilitate the understanding of computer-aided drug design (CADD) techniques, the discussion in this chapter has been divided into two principal approaches: ligand-based drug design (LBDD) and structure-based drug design (SBDD) methods. Table 1 summarizes computational methods employed for the design of epigenetic inhibitors.

CONTENTS

1. Introduction 1
2. Ligand-Based Drug Design 2
   2.1 Chemoinformatics Analysis and Structure Similarity 3
   2.2 Pharmacophore Modeling 5
   2.3 Quantitative Structure–Activity Relationships ... 7
3. Structure-Based Drug Design ... 10
   3.1 Structural Data of Epigenetic Targets 10
   3.2 Docking 11
      3.2.1 Filtering Methods ... 14
      3.2.2 Docking Considering Protein Flexibility . 18
   3.3 Molecular Dynamics 21
4. Combining Methods 24
   4.1 Docking: 3D-QSAR 24

Epi-Informatics. http://dx.doi.org/10.1016/B978-0-12-802808-7.00002-2

10002-MEDINA-FRANCO-9780128028087

Educación Química (2015) 26, 180–186

educación Química

www.educacionquimica.info

PROFESORES AL DÍA

Avances en el diseño de fármacos asistido por computadora

José L. Medina-Franco[a,*], Eli Fernández-de Gortari[a] y J. Jesús Naveja[a,b]

[a] Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, México D.F., México
[b] Facultad de Medicina, PECEM, Universidad Nacional Autónoma de México, México D.F., México

Recibido el 5 de marzo de 2015; aceptado el 15 de abril de 2015
Disponible en Internet el 11 de junio de 2015

PALABRAS CLAVE
Quimiogenómica;
Quimioinformática;
Modelado molecular;
Relaciones estructura-actividad;
Cribado virtual

Resumen El diseño de fármacos asistido por computadora (DIFAC) tiene como objetivos el diseño, optimización y selección de compuestos con actividad biológica. El DIFAC forma parte de un esfuerzo multidisciplinario y tiene numerosas aplicaciones específicas durante el proceso de desarrollo de fármacos. A la fecha ha tenido contribuciones significativas en el diseño de fármacos que se encuentran en uso clínico. Es por esto que DIFAC cobra cada vez mayor importancia en la investigación que se hace en la industria farmacéutica, en universidades y centros de investigación. Métodos empleados en DIFAC pueden aplicarse a otras áreas, por ejemplo, productos naturales, bioquímica, química en alimentos, orgánica y teórica. En este artículo se discuten ejemplos de proyectos de diseño de fármacos realizados por un grupo de investigación enfocado en el DIFAC.

KEYWORDS
Chemogenomics;
Chemoinformatics;
Molecular modeling;
Structure-activity relationships;
Virtual screening

Progress in computer-aided drug design

Abstract The goals of computer-aided drug design (CADD) are the design, optimization and selection of compounds with biological activity. CADD is part of a multidisciplinary effort and has several specific applications during the drug development process. So far, this discipline has made significant contributions to the development of drugs that are currently in clinical use. Therefore, CADD has an increasing relevance in the research performed at the pharmaceutical industry, universities and research centers. Methods used in CADD can be used in other research areas such as natural products, biochemistry, food, organic, and theoretical chemistry. Herein, we discuss examples of drug design projects performed by an academic group focused on CADD.

* Autor para correspondencia.
Correo electrónico: medinaj@unam.mx (J.L. Medina-Franco).

La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

http://dx.doi.org/10.1016/j.eq.2015.05.002

CHAPTER ONE

Molecular Modeling and Chemoinformatics to Advance the Development of Modulators of Epigenetic Targets: A Focus on DNA Methyltransferases

F.D. Prieto-Martínez, A. Peña-Castillo, O. Méndez-Lucio, E. Fernández-de Gortari, J.L. Medina-Franco[1]

Facultad de Química, Universidad Nacional Autónoma de México, México City, México
[1]Corresponding author: e-mail addresses: medinajl@unam.mx; jose.medina.franco@gmail.com

Contents

1. Introduction 2
2. Progress on Chemical Information 4
3. Chemoinformatic Studies of DNMTs 9
   3.1 Characterization of Chemical Space: ERCS 9
   3.2 Chemoinformatic-Based Pharmacophore Model 12
   3.3 Activity Landscape Modeling 14
   3.4 Quantitative Structure–Activity Relationships 15
4. VS: Hit Identification and Optimization 16
   4.1 Novel VS Hits 16
   4.2 Follow-Up of VS Hits 18
5. Computer-Assisted Drug Repurposing 19
6. Food Chemicals as Potential Modulators of DNMTs and Other Epigenetic Targets 20
7. Concluding Remarks 21
Acknowledgments 22
References 23

Abstract

In light of the emerging field of *Epi-informatics*, ie, computational methods applied to epigenetic research, molecular docking, and dynamics, pharmacophore and activity landscape modeling and QSAR play a key role in the development of modulators of DNA methyltransferases (DNMTs), one of the major epigenetic target families. The increased chemical information available for modulators of DNMTs has opened up the avenue to explore the epigenetic relevant chemical space (ERCS). Herein, we discuss

Taylor & Francis

Expert Opinion on Pharmacotherapy

ISSN: 1465-6566 (Print) 1744-7666 (Online) Journal homepage: http://www.tandfonline.com/loi/ieop20

Developmental DNA methyltransferase inhibitors in the treatment of gynecologic cancers

Duenas-Gonzalez Alfonso, L. Medina-Franco José, Chavez-Blanco Alma, Dominguez-Gomez Guadalupe & Fernández-de Gortari Eli

To cite this article: Duenas-Gonzalez Alfonso, L. Medina-Franco José, Chavez-Blanco Alma, Dominguez-Gomez Guadalupe & Fernández-de Gortari Eli (2015): Developmental DNA methyltransferase inhibitors in the treatment of gynecologic cancers, Expert Opinion on Pharmacotherapy, DOI: 10.1517/14656566.2016.1118053

To link to this article: http://dx.doi.org/10.1517/14656566.2016.1118053

Accepted author version posted online: 11 Nov 2015.

Preliminary communication **En Revisión dentro del Journal of Chemoinformatics**

Database fingerprint (DFP): An approach to the representation of molecular databases

Eli Fernández-de Gortari[*], José L. Medina-Franco[*]

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

[*]Corresponding authors: hidragyrum@gmail.com; jose.medina.franco@gmail.com

ABSTRACT

Background: Molecular fingerprints are widely used in several areas of chemoinformatics including diversity analysis and similarity searching. The fingerprint-based analysis of chemical libraries, in particular of large collections, usually requires the representation of each compound in the library leading and may lead to issues such as storage space and redundant calculations. However, in several cases there are information redundancies and not all of the binary digit positions in the fingerprint contain significant information. **Results:** Herein is proposed a general approach to approximate the representation of an entire compound library with a binary fingerprint. The development of the so-called database fingerprint (DFP) is illustrated using a well-known fingerprint (MACCS keys) but other fingerprints can be used. In this work a DPF was developed for 10 representative data sets of general interest in chemistry covering a broad range of size from ca. 100 to 1500 compounds. The DFP can be developed for other compound libraries. **Conclusions:** The DFP is designed to capture key information of the compound collection and can be used to compare compound libraries and assess the diversity of libraries. In this work is shown the potential of the novel fingerprint to conduct inter-library relationships. A major perspective is to apply the DFP for virtual screening. **Keywords:** compound databases, diversity, information content, molecular fingerprints, similarity, Shannon entropy.