VNiVERSiDAD
D SALAMANCA

UNIVERSIDAD DE SALAMANCA

Departamento de Informática y Automática

ANALÍTICA VISUAL APLICADA AL DISEÑO DE NUEVOS FÁRMACOS

TESIS DOCTORAL

REALIZADA POR D. CARLOS ARMANDO GARCÍA PÉREZ

Directores:

DR. D. ROBERTO THERÓN SÁNCHEZ

DR. D. JOSÉ LUIS LÓPEZ PÉREZ

DR. D. RAFAEL PELÁEZ LAMAMIE DE CLAIRAC ARROYO

Junio 2013

VNiVERSiDAD
D SALAMANCA

UNIVERSIDAD DE SALAMANCA
Departamento de Informática y Automática

ANALÍTICA VISUAL APLICADA AL DISEÑO DE NUEVOS FÁRMACOS

TESIS DOCTORAL PRESENTADA POR:
D. CARLOS ARMANDO GARCÍA PÉREZ

Dirigida por:
DR. D. ROBERTO THERÓN SÁNCHEZ
DR. D. JOSÉ LUIS LÓPEZ PÉREZ
DR. D. RAFAEL PELÁEZ LAMAMIE DE CLAIRAC ARROYO

El doctorando

Salamanca, Junio de 2013

Roberto Therón Sánchez, *Profesor titular del Departamento de Informática y Automática de la Universidad de Salamanca*

José Luis López Pérez, *Catedrático del Departamento de Química Farmacéutica de la Universidad de Salamanca*

Rafael Peláez Lamamie de Clairac Arroyo, *Profesor titular de del Departamento de Química Farmacéutica de la Universidad de Salamanca*

HACEN CONSTAR: *Que D. Carlos Armando García Pérez, ha realizado bajo nuestra dirección la Memoria de la tesis doctoral que lleva por título* Analítica Visual Aplicada al Diseño de Nuevos Fármacos, *con el fin de obtener el grado de Doctor por la Universidad de Salamanca*.

Y para que surta los efectos oportunos firman en Salamanca, a trece de Junio de dos mil trece.

Fdo.: Roberto Therón Sánchez Fdo.: José Luis López Pérez

Fdo.: Rafael Peláez Lamamie de Clairac Arroyo

Agradecimientos y Dedicatoria

Al Gran Constructor del Universo, Dios.
A mi prometida Katja Sonnenberg, por ser el ángel de mi vida, porque con su amor y paciencia ha recorrido conmigo este camino hasta su desenlace.
A mi madre, María Alicia Pérez, por ser un valioso ejemplo de moral, lucha y fortaleza. Por guiarme con amor y cariño en todo los momentos importantes de mi vida. Gracias mamá.
A mis hermanas, Belinda García y Estrella García, por apoyarme e infundirme ánimos en momentos de flaqueza.
A mis cuñados, Raúl Cantú y Francisco Ríos.
A mis sobrinas, Dora Cantú, Cielo cantú y Lucero Cantú, porque con sus sonrisas me hacen sonreír también.
A mi familia, tíos y primos, por sus palabras de ánimo.
A mis amigos y segunda familia en mi paso por Salamanca, Juan García, Diego Gómez, José Castellanos, Antonio González, Vadim paz, Jannine Nieto, Odette Maldonado, Hannah Spungin y Olga Mikhina.
A mis directores, Roberto Therón, por ofrecerme la oportunidad de comenzar una vida en la ciencia e investigación, a Rafael Peláez y José Luis López, por acercarme al hermoso mundo de la química y bioinformática.

A todos, ¡Gracias!

Reflexión sobre el naufragio de Simónides.
Todo hombre culto lleva siempre en sí mismo todas sus riquezas.

- ¿Y tú Simónides, no te llevas ninguna de tus riquezas?
- Todas ellas -contestó- las llevo encima.
Cuando Simónides los encontró por azar, les dijo: Os advertía que yo llevaba encima de mí todos mis bienes.

Resumen

Actualmente se genera una gran cantidad de datos a tal ritmo que en la que práctica, más allá de almacenarla, no es posible recuperarla, transportar, analizar y, sobretodo extraer información útil para los usuarios finales. Este fenómeno conocido como *Big Data*, es el nuevo reto al que se enfrentan diariamente diferentes áreas y usuarios que van desde las redes sociales, educación, ciencia, economía, seguridad, química, biología, etc. Dos de estas áreas son la bioinformática y quimioinformática, donde se ha producido un gran cambio en la manera en la que se diseñan y desarrollan fármacos. La disponibilidad de un gran número de estructuras tridimensionales de macromoléculas biológicas, dianas potenciales o de facto de fármacos, ha permitido que el diseño de fármacos basado en la estructura de la diana se incorpore de lleno al arsenal de herramientas utilizadas. Entre los métodos computacionales disponibles, el *docking* y el cribado virtual son los métodos de selección de compuestos más utilizados en la búsqueda de nuevas moléculas bioactivas. Estas metodologías generan una gran cantidad de información en cada etapa del diseño. La analítica visual, a través de visualizaciones interactivas, afrontan este enorme reto de extraer, analizar y presentar información de manera sencilla al químico. Más aun, la analítica visual puede ser una pieza fundamental en el desarrollo de fármacos. En el presente trabajo de tesis, se aprovecha esta oportunidad mediante el desarrollo de una herramienta para analizar resultados de *docking* y propiciar el descubrimiento de puntos de mejora para futuros experimentos de *docking*. Empleando diferentes visualizaciones enlazadas a un visualizador molecular tridimensional, es posible presentar la información relevante al químico que le ayude en la toma de decisiones. Para la consecución de este objetivo, ha sido necesaria la implementación de algoritmos de *clustering* que ayudan a reducir la enorme cantidad de datos, de forma que sea posible la presentación de información de manera sencilla y configurable. Por otro lado, se debe resaltar que al corregir los puntos débiles de los resultados de *docking*, se logra utilizar el programa Autodock como si se tratase de una herramienta para la búsqueda por farmacóforos, lo que resulta en una significativa aceleración del proceso de diseño de fármacos.

Abstract

Currently a large amount of data is generated at such a pace that, in practice, beyond the storage, it is not possible to retrieve, transport, analyze and extract useful information for end users. This phenomenon, known as Big Data, has become the new challenge that different areas, ranging from social networks, education, science, economy, safety, chemistry, biology, etc.., are facing daily. Two of these areas are bioinformatics and chemoinformatics, where a big change has been introduced in the way we design and develop drugs. The availability of a large number of three-dimensional structures of biological macromolecules, potential or de facto targets of drugs, allowed drug design based on the structure of the target to be incorporated fully into the list of used tools. Among the available computational methods, docking and virtual screening methods are most commonly used for the screening of compounds in the search for new bioactive molecules. These methods generate a large amount of information at each stage of the design. Visual analytics through interactive displays, face the enormous challenge of extracting, analyzing and presenting information in a simple manner to the chemist. Moreover, visual analytics can be a cornerstone in the development of drugs. In this thesis work, we take advantage of this opportunity by developing a tool to analyze docking results and facilitate the discovery of areas of improvement for future docking experiments. Using different visualizations linked to a three-dimensional molecular display, relevant information can be presented to the chemist in order to support decision-making. To achieve this goal, it was necessary to implement clustering algorithms that help reduce the huge amount of data, so that it is possible to present information in a simple and configurable way. Furthermore, it should be noted that by correcting the weaknesses of the docking results, the use of the AutoDock program, as if it were a tool for searching by pharmacophore, is enabled, resulting in a significant acceleration of the drug design process.

Índice general

Índice de figuras

Índice de tablas

Parte I

Introducción

Capítulo 1

Introducción

1.1. Diseño de Fármacos

El diseño y desarrollo de nuevos fármacos es un proceso largo que con frecuencia no acaba con éxito. De las dificultades que entraña el proceso habla el escaso número de fármacos nuevos introducidos en el mercado en los últimos años [85] (ver figura 1.1). Para intentar mejorar el porcentaje de éxito, las empresas, centros de investigación y universidades realizan grandes inversiones en el desarrollo de nuevas herramientas y métodos para facilitar y acelerar el proceso de diseño de nuevos fármacos. Muchos son los métodos que se aplican en diferentes etapas para diseñar y desarrollar un fármaco. La selección de ellos dependerá de la información disponible, las singularidades del problema y los recursos humanos y tecnológicos [93]. El diseño de un fármaco comienza con la selección de un objetivo terapéutico (tratamiento de una enfermedad, desarrollo de un sistema de diagnostico, modificación de una situación fisiológica, etc.). Una vez establecido éste, es necesario identificar la o las dianas biológicas importantes para el proceso que se desea modificar o estudiar, y establecer medios para evaluar la actividad de las nuevas moléculas frente a ellas (ensayos). El siguiente paso es identificar compuestos que presenten actividad en el ensayo (*hit*), para posteriormente mejorarlos hasta conseguir líderes (compuestos activos más potentes y con propiedades adecuadas). A partir de los líderes, mediante nuevas etapas de optimización, se obtienen los candidatos a fármaco (hasta aquí el proceso se denomina fase preclínica) que, si superan las fases de evaluación en humanos (fase clínica), se convierten en fármacos.

La búsqueda de los compuestos activos iniciales (*hits*) se hace a partir de bibliotecas químicas que contienen un gran número de compuestos. En situaciones favorables, esta misma estrategia puede aplicarse con líderes más avanzados, utilizando bibliotecas de compuestos más específicos. Una técnica frecuentemente utilizada para la identificación de estos nuevos activos o líderes es el ensayo -de estas enormes bibliotecas químicas- frente a la diana biológica; es lo que se conoce como cribado masivo o HTS (*High Throughput Screening*) [115]. Dado que identificar físicamente estos compuestos líderes en las enormes bibliotecas químicas conlleva un enorme tiempo y, además, no todos los centros de investigación cuentan con la tecnología ni los medios económicos para llevar a cabo esta tarea, se opta por otras alternativas como es el cribado virtual.

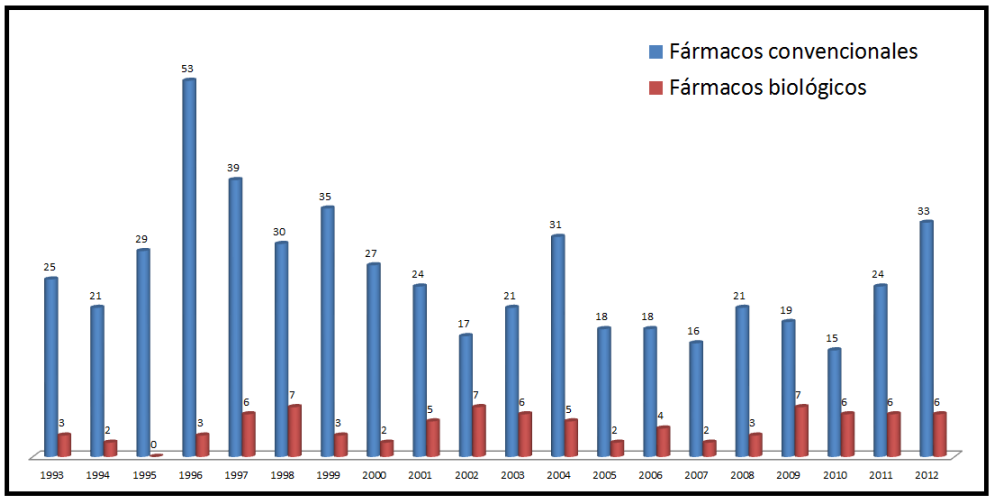


Figura 1.1: Fármacos aprobados por la FDA (*Food and Drug Administration*), la agencia encargada de autorizar los nuevos fármacos en los Estados Unidos, entre 1993 y 2012.

El cribado virtual, consiste en evaluar *in silico* grandes bibliotecas de compuestos químicos virtuales frente a modelos de las dianas biológicas, con el fin de seleccionar aquellos que potencialmente se unirán a ellas con mayor afinidad (por lo general, estas bibliotecas son más grandes que cualquier biblioteca utilizada en la evaluación empírica). Su objetivo principal es la predicción de la actividad biológica, reduciendo el número de compuestos reales que se emplearán en etapas posteriores.

El cribado virtual abarca una gran variedad de técnicas por ordenador que permiten a los químicos reducir las enormes bibliotecas virtuales a otras de tamaño manejable. Como resultado de este proceso se selecciona un cierto número de posibles nuevos ligandos -compuestos con afinidad por la diana terapéutica- muchos de los cuales pueden ser preparados o adquiridos y posteriormente ensayados. Cuando no se conoce la estructura tridimensional de las dianas biológicas el cribado virtual se basa en definir las características estructurales comunes de los compuestos que presentan la actividad deseada y utilizarlas para la selección de aquellos compuestos que las presentan en detrimento de otros que carezcan de ellas. Un tipo de cribado de este tipo es la búsqueda por elementos farmacofóricos -un farmacóforo es el conjunto de elementos que son necesarios para asegurar la interacción óptima con una diana biológica y producir o bloquear su actividad biológica-. Cuando se conoce la estructura de la diana biológica es posible llevar a cabo búsquedas en las que las interacciones con la diana biológica se evalúan átomo a átomo, siendo uno de los ejemplos más conocidos el *docking*.

El modelado molecular es un conjunto de técnicas encaminadas al estudio de la estructura tridimensional de las moléculas y sus propiedades, utilizando los ordenadores y técnicas de visualización gráfica. Los estudios de *docking* son estudios de modelado molecular dirigidos a obtener la estructura tridimensional de complejos formados por la asociación de un fármaco con su diana, utilizando como punto de partida modelos de las estructuras de los fármacos y sus dianas. Se trata de una herramienta por ordenador para el diseño de fármacos basado en la estructura de la diana, ya que predice la geometría de interacción de la proteína y el ligando y su afinidad de unión [132]. Esto implica ubicar los compuestos en el sitio de la diana (pose) para después hacer una jerarquización de los

resultados de los diferentes compuestos (puntuación). Como los modelos no son exactos, en lugar de obtenerse una única pose, lo habitual es obtener un número de ellas para cada ligando.

Posteriormente, es necesaria una inspección visual de todas las poses de los compuestos seleccionados. La inspección visual se torna en una tarea totalmente inabordable, dada la enorme cantidad de estructuras generadas durante el proceso del cribado virtual, por ejemplo, si se generan 10 conformaciones para cada compuesto de una biblioteca de 700000 moléculas, se obtendrían 7×10^6 conformaciones.

Este cuello de botella puede suponer un campo fértil para aplicar un enfoque de analítica visual, ya que el objetivo de esta ciencia es proporcionar una mejor percepción de grandes cantidades de datos de fuentes heterogéneas, como es el caso del diseño de un fármaco, a través de la conjugación de la potencia actual de cálculo de los ordenadores con las habilidades cognitivas humanas, francamente poderosas. Así, la aplicación de técnicas procedentes de la analítica visual contribuirá al diseño de fármacos abordando el problema desde otro enfoque y generando nuevo conocimiento a partir de los datos disponibles, facilitando la toma de decisiones.

1.2. Analítica Visual

La sobrecarga de información es un problema actualmente conocido como *Big data*. Kaisler *et al.* [58], definen el termino *Big data* como: "El total de datos que superan la capacidad de almacenar, manejar y analizarlos eficientemente". No cabe duda que los componentes informáticos de hardware y software tradicionales, han sido superados por la velocidad en que se generan nuevos datos para analizar y procesar. Esta avalancha de datos puede llegar a ser del orden de exabytes (1018). En consecuencia, ante esta abrumadora avalancha de datos, se hace necesaria la propuesta de técnicas y métodos para desentrañarlos, como es el caso de la visualización de información. Ante un problema de análisis de datos, la visualización de información, a través de un mecanismo de exploración visual de los datos, puede ayudar al usuario a extraer nuevo conocimiento a partir de los datos y a generar nuevas hipótesis. Estas hipótesis se pueden verificar a través de la exploración visual, con técnicas estadísticas o de aprendizaje automático. Sin embargo, la visualización de información se quedaría corta en el momento de examinar los datos representados si no se aplicasen técnicas de análisis de datos como minería de datos.

La analítica visual es la ciencia del razonamiento analítico facilitado por interfaces visuales altamente interactivas [127], que representan la información visualmente, permitiendo al usuario interactuar con los datos para comprenderlos mejor, sacar conclusiones y finalmente tomar mejores decisiones basadas en conjuntos de datos extensos y complejos [64] [60]. Aplicando técnicas procedentes de la analítica visual es posible combinar las capacidades de exploración visual del ser humano con el poder de procesamiento de los ordenadores para generar un entorno de conocimiento. Por otra parte, el usuario dejará de ser un mero espectador pasivo que sólo interpreta datos; en su lugar se convertirá en el primer actor de todo el proceso [139].

La analítica visual es más que una visualización; puede verse como un enfoque en la toma de decisiones, combinando la visualización de información, factores humanos y análisis de datos. El reto es identificar el algoritmo de análisis más adecuado para el dominio de los datos, identificar sus limitaciones (que ya no puedan ser automatizados), y a continuación desarrollar una solución integrada que combine el algoritmo de análisis del problema con técnicas de visualización e interacción

apropiadas [60].

La visualización es un componente importante en el proceso de diseño de fármacos. Las representaciones visuales son la forma más efectiva para presentar y comprender enormes volúmenes de datos. Sin embargo, la utilización correcta de la visualización de información es un factor crítico, dado que con volúmenes grandes de información, las gráficas y representaciones tradicionales -incluso cuando sean completamente interactivas- no son suficientes para comprender claramente el comportamiento de la información. En lugar de eso, es necesario implementar métodos y técnicas para comunicar claramente el valor que aportan los datos analizados.

Finalmente, es preciso considerar la importancia de comprender los múltiples tipos de datos y sus interrelaciones durante todo el proceso de diseño de fármacos tales como datos: numéricos, categóricos, las estructuras químicas, documentos de texto, etc. Por eso la analítica visual tiene un valor particular en el área del diseño de fármacos, ya que el uso de sistemas de visualización permite cruzar dominios y tipos de datos para ofrecer la posibilidad de integrar los análisis y ofrecer un rápido soporte a la toma de decisiones eficaces [106].

1.3. Problema de Investigación

Como se ha indicado anteriormente, el *docking* es una estrategia de cribado virtual basada en el conocimiento de la estructura tridimensional de la diana. En dichos experimentos, los ligandos virtuales que se desea evaluar se enfrentan a la diana, permitiendo que se coloquen (unan) en distintas posiciones y disposiciones (poses) con respecto a la diana. En el llamado *docking* flexible, se permite además una cierta flexibilidad de la diana, lo que es importante para el descubrimiento de compuestos estructuralmente diversos. Las poses generadas para cada compuesto se evalúan con algoritmos específicos que valoran la bondad de la interacción con la diana, permitiendo una ordenación de las mismas. Sin embargo, estas predicciones no son demasiado precisas, ya que esto requeriría mucho tiempo de cálculo, incompatible con la evaluación de un número elevado de candidatos, por lo que no es posible considerar exclusivamente la pose mejor valorada.

Durante el proceso de *docking* de grandes números de compuestos virtuales se genera gran cantidad de información. Sin embargo, actualmente no existe una técnica que agrupe y seleccione las mejores poses en un proceso automático para una gran cantidad de compuestos candidatos; hasta ahora gran parte de este proceso se realiza de forma manual. La reducción del número de poses de un mismo compuesto a un número manejable de representantes se realiza con facilidad, mediante procesos de agrupación y clasificación. Sin embargo, la comparación de estructuras diversas es un proceso no trivial que con frecuencia incorpora una notable subjetividad. Un ejemplo de esta carencia se puede encontrar en nuestro trabajo previo [93], donde se utilizó un proceso semiautomatizado para la selección de las mejores poses de cada compuesto evaluado (cerca de 700000 compuestos). La agrupación de las poses de un mismo compuesto pudo hacerse fácilmente de forma automática o semiautomática, pero para la comparación de compuestos diferentes, a pesar de utilizar un algoritmo de clasificación, fue necesario realizar una clasificación en forma manual, esto es, mediante una inspección visual de los representantes (estructuras con la mejor puntuación dadas por el programa de *docking*). Para llevar a cabo de forma automática esta segunda fase de comparación, agrupación y selección de estructuras, en este trabajo se parte del postulado de que poses similares deberían colocar átomos similares en posiciones cercanas.

Otro factor a considerar en el momento de selección de un representante es disponer de la mayor cantidad posible de información de ese compuesto; entiéndase como el tener el mayor número posible de variables a considerar. Este hecho es factible siempre y cuando el número de compuestos sea pequeño, pero deja de serlo cuando se manejan enormes cantidades, como suele ser el caso.

Las técnicas actuales de diseño de fármacos, no proporcionan un entorno de análisis visual, dado que la mayoría de ellas emplean sistemas de visualización simples que no ofrecen una aportación suficiente de conocimiento como para poder elaborar un discurso analítico sobre lo que se está explorando. Habitualmente, se limitan a representar los resultados o están diseñadas para realizar representaciones gráficas, dejando a un lado la generación de nuevo conocimiento. Muchos de estos sistemas de visualización son estáticos y cerrados y la interacción con otros sistemas de representación es limitada, impidiendo la incorporación de nueva información al análisis en curso. Por todo lo expuesto, existe la necesidad de contar con un método que incorpore la experiencia del químico junto con los procedimientos analíticos para maximizar el análisis de información generada por el *docking*, que derivará en la selección de las mejores poses del conjunto de compuestos químicos. De esta manera se incorporará la suficiente información para agrupar las estructuras no sólo por su similitud de forma, sino que, además, tendrá en cuenta el resto de información que surge durante el proceso de *docking*, por consiguiente esta información se mantendrá a lo largo del resto del proceso de diseño del fármaco.

1.4. Objetivos y Preguntas de Investigación

Tomando como punto de partida los resultados del *docking* de muchos compuestos químicos, este trabajo de investigación pretende desarrollar un método para seleccionar las mejores poses de cada uno de los compuestos candidatos, llevar a cabo agrupaciones de las distintas poses de los diferentes compuestos candidatos para continuar con el proceso de diseño del fármaco y, por último, posibilitar, mediante esta información, la búsqueda por farmacóforos. Por lo que los objetivos son los siguientes:

- Determinar e implementar un método automático de agrupación de *n* poses por similitud geométrica para cada compuesto candidato y posteriormente seleccionar las mejores poses de un conjunto de compuestos químicos diferentes.
    1. Automatizar la selección de representantes de los grupos por cada compuesto candidato.
    2. Automatizar la comparación de distintos compuestos candidatos.
- Definir e implementar un método para realizar búsquedas por farmacóforos.
    1. Diseñar e implementar una técnica que ayude al usuario a definir búsquedas por farmacóforos.
    2. Crear nuevas formas de retroalimentación en el proceso de *docking* que faciliten la incorporación de más información al diseño de fármacos.

A partir de los objetivos anteriores, surgen las siguientes preguntas de investigación:

- ¿Cómo pueden contribuir las técnicas y métodos de analítica visual a mejorar el diseño de fármacos durante el proceso de comparación de compuestos químicos diferentes para la obtención de estructuras químicas candidatas similares?

    1. ¿Cómo los métodos y técnicas de analítica visual son aplicables durante el proceso de comparación de estructuras diferentes, para permitir el discurso analítico en el diseño de fármacos?
    2. ¿Cómo contribuyen estas técnicas y métodos a la generación y obtención de información en el proceso de diseño de fármacos?

- ¿Cómo se puede diseñar e implementar un método para la creación de búsquedas por farmacóforos, aplicando analítica visual?

1.5. Organización de la Tesis

Una vez establecidos los objetivos y las preguntas de investigación, a continuación se describen los capítulos en los que está organizado este trabajo de tesis.

En la segunda parte de la tesis se tratan los "Conceptos Generales". El capítulo "Diseño de Fármacos" basado en la estructura introduce al lector en los conceptos básicos del diseño de fármacos, así como le ofrece una breve explicación tanto del ciclo de vida del diseño de fármacos, como de las diferentes etapas que lo constituyen. Posteriormente en el capítulo de "Visualización de Información" se hace una revisión de aquellas técnicas y métodos que han sido fundamentales para la realización del trabajo de tesis, se explican conceptos generales de visualización de información necesarios para la interacción y comprensión de grandes volúmenes de datos. Para finalizar con los conceptos generales, en el capítulo de "Analítica Visual" se introduce, y explica detalladamente al usuario, el papel que juega la analítica visual en el proceso de razonamiento analítico, así como las ventajas que se pueden obtener al aplicar este enfoque al diseño de fármacos.

La tercera parte de la tesis abarca el "Estado del Arte". Mediante los capítulos expuestos aquí, el lector podrá conocer el estado del arte en relación a los diferentes tipos de algoritmos que se han empleado y que se emplean actualmente para analizar la enorme cantidad de datos generada en las diferentes etapas del diseño de fármacos, así como las propuestas visuales y analíticas para afrontar de la mejor manera el problema de aprehender el conocimiento que se esconde tras el ingente volumen de datos. Para ello el lector encontrará en el capítulo de "*Clustering* en el Diseño de Fármacos" los diferentes algoritmos y estrategias de clasificación en el diseño de fármacos. Posteriormente en el capítulo de "Visualización de Información en el Diseño de Fármacos" se exponen las diversas propuestas tanto para visualizar los resultados de clasificación y visualización que apoyan el análisis en bioinformática. Finalmente, el capítulo "Analítica Visual en el Diseño de Fármacos", recoge lo expuesto en capítulos anteriores y se discute la necesidad de un enfoque de analítica visual, y se ofrece una revisión de los trabajos que han explorado la posibilidad de emplear la analítica visual como discurso analítico para el diseño de fármacos.

En la cuarta parte se presenta detalladamente el problema a solucionar, así como la quinta parte explica detalladamente la solución propuesta a través de una herramienta que aplica la analítica visual, que constituye la principal aportación de este trabajo de tesis. Se presenta una herramienta que emplea las ventajas de la analítica visual a través de diversas visualizaciones que expresamente

se emplean para el análisis de grandes cantidades de moléculas de diferente número y tipo de átomo, así como en el propio diseño de farmacóforos.

Finalmente se presentan los capítulos de “Resultados” y “Conclusiones” de este trabajo de tesis.

Parte II

Conceptos Generales

Capítulo 2

Diseño de Fármacos

El diseño de fármacos es un proceso largo en el que pueden destacarse una serie de etapas (figura 2.1). Estas etapas no siempre ocurren de forma explícita y con frecuencia coexisten en el tiempo, pero para su descripción es razonable ordenarlas de forma secuencial de acuerdo con la lógica subyacente en el proceso. En él participan profesionales de distintas áreas, desde economistas (sobre todo en los programas empresariales) hasta investigadores básicos y, por supuesto, gran cantidad de especialistas en áreas concretas de clínica, biología, química, farmacia e informática, entre otras.

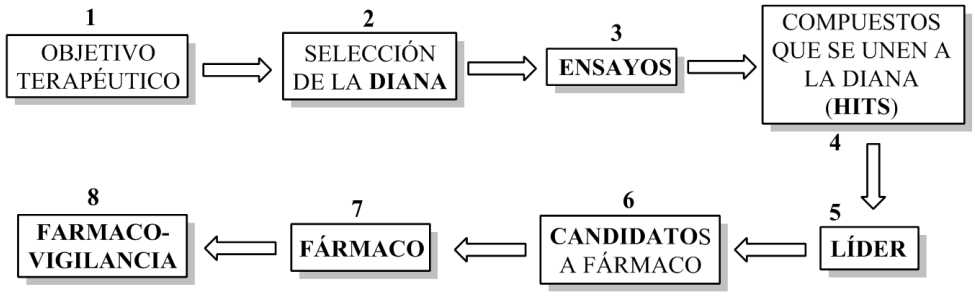


Figura 2.1: Esquema de las etapas en el desarrollo de fármacos.

El primer paso en el diseño de un fármaco es seleccionar una enfermedad -**objetivo terapéutico**(1)-, para la que se quiere encontrar una cura o un tratamiento o, incluso, un método de diagnóstico. En ocasiones, el objetivo no es propiamente una enfermedad sino una situación fisiológica no deseada, como por ejemplo un embarazo.

Una vez decidido el objetivo, el siguiente paso es identificar una diana biológica (2) para el nuevo fármaco (un receptor, enzima, un ácido nucleico). En este paso es importante que la diana seleccionada sea relevante (que su alteración produzca el efecto deseado) y que pueda ser utilizada como diana (no todas las moléculas implicadas en un proceso concreto son necesariamente buenas dianas). En el pasado, sólo se podía establecer la existencia de una diana si existía previamente un compuesto (no necesariamente un fármaco) que producía un efecto biológico, lo cual demostraría

que existe una molécula diana. De esta forma, el encontrar una diana dependía principalmente de encontrar primero un compuesto activo. En la química médica moderna se ha superado esta fase de aproximación farmacológica a la búsqueda de dianas, y gracias a los avances de la biología y la bioquímica y es posible encontrar y validar dianas potenciales antes de disponer de moléculas que modulen su actividad. Sin embargo, este proceso dista de ser trivial y son numerosos los casos en que las dianas elegidas resultan no ser apropiadas para el desarrollo de fármacos.

Una vez seleccionada la diana, es necesario establecer un **bioensayo** (3) que permita diferenciar los compuestos químicos capaces de modificar la actividad de la diana de los compuestos inactivos. Además, los ensayos deben permitir clasificar los compuestos en función de su potencia. Obviamente, aunque el objetivo terapéutico pertenezca a la medicina humana, no es posible hacer pruebas en humanos y tienen que realizarse por otros medios, que deben permitir analizar un gran número de compuestos. Habitualmente se habla de pruebas *in vitro* o *in vivo*. Las pruebas *in vitro* son estudios efectuados con tejidos, células o dianas aisladas en los que se mide el crecimiento, el metabolismo, la aparición o desaparición de un efecto, la producción o cese de una reacción, o, simplemente, la unión de un compuesto -en este caso se suelen denominar ligandos- a una diana. Las pruebas *in vitro* pueden hacerse a gran escala, proceso que se conoce como cribado masivo (*High throughput screening, HTS*). Las pruebas *in vivo* se realizan en animales a los cuales se les induce una condición clínica (enfermedad) para ver si el tratamiento con los compuestos elimina dichos síntomas. Las pruebas *in vivo* presentan problemas para estadios iniciales del desarrollo, como son la lentitud del proceso, el sufrimiento que causa a los animales y su elevado coste económico. Por el contrario, presentan la ventaja de ser más representativos de la situación en humanos que los ensayos *in vitro*, por lo que se suelen realizar una vez confirmada la actividad *in vitro* de los compuestos.

Una vez que se dispone de un sistema biológico (enfermedad-diana-ensayo) comienza la búsqueda de compuestos químicos capaces de modificar su comportamiento. Normalmente esto implica descubrir sustancias que se unen a la diana (ligandos), que esta unión haga que la diana cambie su comportamiento (actividad específica) y que este cambio se traduzca en una modificación de la situación clínica. Aunque a primera vista las tres acciones parecen una consecuencia necesaria una de la siguiente, con frecuencia se observa una desconexión entre un efecto y el siguiente, por lo que suelen ser necesarios ciclos repetidos de diseño-obtención-evaluación de compuestos hasta tener un candidato con propiedades adecuadas. En cada uno de estos ciclos, el primer paso es encontrar compuestos que se unan a la diana (*hits*, 4) y lo que se consigue en los pasos sucesivos es mejorar sus propiedades y el comportamiento de los mismos hasta llegar a conseguir compuestos que presenten ciertas garantías de éxito en su aplicación clínica -candidatos a fármaco-. En las etapas intermedias de este proceso se usa con frecuencia la denominación de líderes (5) para compuestos activos de un cierto tipo estructural, destacados porque presentan ya unas ciertas propiedades que se semejan a lo que finalmente se espera para los candidatos a fármacos (6).

Uno de los problemas en las fases iniciales del proceso es encontrar compuestos activos para modificarlos. Como el porcentaje de compuestos que presentan actividad frente a una diana suele ser muy pequeño, a menudo es necesario ensayar muchos para encontrar unos pocos activos. Esto traslada el problema a encontrar suficientes sustancias para ensayar. Tradicionalmente, las fuentes de compuestos han sido fármacos preexistentes o ligandos o moduladores descritos, la naturaleza, las grandes bibliotecas de compuestos que las compañías han ido acumulando a lo largo de los años, bibliotecas de compuestos sintetizadas expresamente para la ocasión o compuestos generados por unión de fragmentos más pequeños que se sabe se unen a la diana. Sea como fuere la generación de las moléculas a ensayar, el proceso finaliza con el ensayo de los compuestos frente a la diana

mediante un bioensayo adecuado. Sin embargo, el ensayo de grandes números de compuestos es lento y caro, por lo que se han diseñado numerosas estrategias para reducir el número de compuestos a ensayar.

Una de estas metodologías auxiliares en este proceso es el uso de herramientas informáticas que intentan predecir los resultados del bioensayo, por lo que con frecuencia se refiere a ellos como ensayos *in silico*. Se reduce así el número de compuestos a ensayar a sólo aquellos para los que las posibilidades de que sean activos son altas (positivos). Esto permite un notable abaratamiento de los ensayos y una agilización del proceso, al reducirse el número de compuestos que deben ser obtenidos, ensayados y analizados, siempre y cuando la herramienta enriquezca la selección en compuestos activos (verdaderos positivos) frente a inactivos (falsos positivos). Una importante desventaja de estos métodos es que, debido a sus limitaciones intrínsecas y a la necesidad de hacerlos muy rápidamente, con frecuencia no se seleccionan compuestos que podrían haber resultado activos (falsos negativos). Pese a estas limitaciones, las aplicaciones de la informática en el diseño de fármacos son múltiples y cada vez más variadas, extendiéndose desde la bioinformática a la quimioinformática. De forma genérica, las aproximaciones a la búsqueda de compuestos activos asistidas por ordenador pueden clasificarse en dos grandes categorías, que no se utilizan de forma excluyente sino complementaria:

a) aproximaciones basadas en información indirecta sobre la interacción de los ligandos con la diana: Cuando no se conoce la estructura tridimensional de la diana, los requerimientos estructurales de los ligandos se deducen de la estructura de los compuestos que poseen la actividad deseada. Un farmacóforo es el resumen de los elementos estructurales que son importantes y que se requieren para la actividad biológica. Esta información puede utilizarse en búsquedas de bases de datos de compuestos, permitiendo encontrar compuestos que satisfagan dichos requisitos. Así, por ejemplo, es posible buscar compuestos que tengan dos grupos básicos separados 7Å y una zona hidrofóbica a 4Å de uno de dichos grupos. Estas búsquedas por farmacóforo permiten reducir los compuestos a ensayar de una biblioteca de posibles candidatos. Cuando se dispone de información sobre la actividad frente a una diana de una serie de compuestos de estructura relacionada es posible establecer relaciones entre la estructura y la potencia de los mismos, lo que permite diseñar compuestos más potentes (estudios de relación estructura-actividad).

b) aproximaciones basadas en el conocimiento de la estructura tridimensional de la diana. Existen distintos métodos para determinar la estructura tridimensional de las dianas, pero entre los más utilizados están los métodos difractométricos (como la difracción de rayos X de cristales) y la Resonancia Magnética Nuclear (RMN). Cuando se conoce la estructura tridimensional de la diana, es posible evaluar y cuantificar a escala atómica las interacciones entre la diana y los posibles ligandos, mediante lo que se conoce como experimentos de *docking*. Estas aproximaciones evalúan de forma aproximada la energía de la interacción (expresada en forma de constantes de asociación o energías de unión) y permiten clasificar los ligandos, para seleccionar aquellos mejor valorados.

El término de cribado virtual (*Virtual screening, VS*) se utiliza para describir el proceso de analizar por ordenador enormes cantidades de compuestos, bien sea utilizando métodos de *docking* o búsquedas por farmacóforo u otras [132]. Para poder realizar VS es necesario construir las moléculas a evaluar, para lo que se suele recurrir a sistemas de construcción automatizados, a bases de datos de

moléculas o de fragmentos de moléculas (como por ejemplo Zinc[1]) o a la construcción de conjuntos específicos de moléculas. En general se suele limitar los compuestos generados a aquellos que serán accesibles sintética o comercialmente, en lo que se utilizan también los ordenadores.

Tras numerosos ciclos de diseño, obtención y evaluación bioquímica de compuestos, es frecuente conseguir compuestos con alta potencia frente a la diana seleccionada. Estos deben ser modificados posteriormente para conseguir que alcancen en el organismo los niveles adecuados, no sean tóxicos en sí mismos o metabolizados a compuestos tóxicos, su acción dure el mayor tiempo posible, etc. A estas propiedades importantes para obtener un candidato a fármaco se las conoce como de la fase farmacocinética, por oposición a la fase farmacodinámica de interacción con la diana. Antes de que los fármacos (7) sean ensayados en humanos (ensayos clínicos) se tienen que realizar pruebas de toxicidad *in vitro* e *in vivo*, tanto de toxicidad aguda como de toxicidad crónica, para de terminar su posible efecto carcinógeno.

Al concluir estas pruebas, se procede a abordar el último paso en el desarrollo de un fármaco: los ensayos clínicos en humanos (8). Estos ensayos se dividen en cuatro fases:

I. Pruebas en voluntarios sanos para evaluar la efectividad, la potencia, la farmacocinética, y los posibles efectos secundarios.

II. Pruebas en un grupo reducido de pacientes voluntarios para comprobar los efectos y la dosis a administrar.

III. Comparación con otros tratamientos en un grupo mayor de pacientes y una comparación con un placebo. En caso de superar con éxito la fase III, las agencias de evaluación de los medicamentos (FDA en EEUU y EMEA en Europa) autorizan la comercialización del mismo.

IV. Continúa el seguimiento del fármaco en el mercado para confirmar la efectividad y detectar posibles efectos secundarios de baja frecuencia.

El porcentaje de compuestos que supera todas estas fases es muy reducido, y el coste del mismo es elevadísimo, habiéndose estimado que el precio de poner un nuevo fármaco en el mercado es de unos 800 millones de dólares.

[1] http://zinc.docking.org/

Capítulo 3

Visualización de Información

Stuart Card y John Mackinlay definen la Visualización de Información como "El proceso visual asistido por ordenador para obtener conocimiento" [13]. Otra posible definición es "La Visualización de Información busca representar eficientemente todas o la mayoría de las variables representándolas en sus posibles dimensiones" [36]. En otras palabras, estas representaciones gráficas se generan a partir del contenido de los datos (numéricos, caracteres, lógicos o abstractos), y cuando se ofrecen a un observador son procesados a través de la vista (el sentido que mejor nos permite captar una gran cantidad de información); la capacidad humana de reconocer patrones subconscientemente [6] es un factor determinante, y todo ello permite al observador acceder a conocimiento nuevo en solo un vistazo. Así, el propósito de la visualización de información es hacer posible para el usuario obtener modelos mentales internos del contenido de la información de todos los conjuntos de datos involucrados en la representación visual; modelos que subsecuentemente son utilizados para predecir y/o tomar decisiones [86].

El progreso logrado en el almacenamiento de datos permite hoy en día almacenar una gran cantidad de información [61], lo cual dificulta el análisis de la misma. A esta limitación se suma el problema de diseñar representaciones interactivas que estén fuertemente relacionadas con el número de atributos involucrados (variables y sus dimensiones) [121]; aquí es donde la visualización de información ofrece su gran valor ya que, por medio de diferentes técnicas, permite proporcionar un mejor análisis y presentación de la información sin perder un gran número de datos significativos o cruciales. En otras palabras, el número de dimensiones que se puedan representar así como el tamaño del dispositivo (monitor de PC, portátil, móvil, tableta, etc.) en el cual se esté proyectando la gráfica, no sean un impedimento para representar el número de variables (dimensiones) involucradas en el análisis. Por lo tanto, podemos finalizar diciendo que la visualización de información tiene como objetivo reducir la complejidad en el examen y comprensión de información, diseñando técnicas apropiadas para la representación visual de datos [22].

Unos de los campos de interés en donde se aplican métodos visuales de análisis de datos masivos es la bioinformática, y una rama en concreto es el diseño de fármacos basados en la estructura de la diana. De las diferentes etapas que conforman el diseño de fármacos, una de las que genera gran cantidad de datos, es el análisis de resultados de ubicación (en el espacio) de compuestos en relación con el sitio activo de la diana (*docking*) -por lo general estos resultados suelen ser masivos, alrededor de cien por cada compuesto-, los cuales se deben examinar visualmente. Es, por tanto, imperativo

que exista un método visual e interactivo que ayude a explorar, analizar y sobre todo a interpretar de una manera sencilla los resultados obtenidos. La visualización de información con sus diferentes técnicas, puede ayudar a los químicos a seleccionar (o desechar) de manera más sencilla y rápida aquellos compuestos que satisfagan los requisitos de interacción del sitio activo de la diana, sin tener que inspeccionar cada resultado por separado.

3.1. Visualización de Información y Visualización Científica

Si buscáramos en la literatura científica técnicas para visualizar nuestros datos nos toparíamos con que éstas pertenecen fundamentalmente a dos categorías: técnicas de visualización de información y técnicas de visualización científica [15]. Aunque hoy en día la frontera entre estas dos categorías es muy difusa, podemos acudir a la definición original de la más antigua. Card [15] define la visualización científica como: "El uso de representaciones visuales interactivas de información científica, datos físicos, para ampliar el conocimiento". Esto es, la visualización científica cubrirá datos del mundo real para representarlos en gráficos interactivos. En contraposición, la visualización de información se encarga de representar datos abstractos mediante gráficos (interactivos o no). Un ejemplo de visualización de información sería el representar mediante una gráfica en forma de red la interacción entre proteínas de diferentes organismos, en cambio un ejemplo de visualización científica sería la representación de isosuperficies de esas proteínas. Esto no quiere decir que la visualización de información no aborde problemas científicos ni trate con datos reales; la diferencia fundamental es que mientras la visualización científica intentará hacer una representación fidedigna de un fenómeno real, la visualización de información tratará de encontrar la forma de representación de los datos de un problema para que se ajusten mejor al modelo mental que el observador inspeccionará durante el proceso de análisis o comprensión de ese conjunto de datos complejo.

Así, Rhyne [103] sugiere que no debe haber una diferencia entre las dos disciplinas, ya que en muchas ocasiones se emplean las dos y defiende que su separación crea confusión en lugar de ayudar a los investigadores y desarrolladores a crear herramientas o aplicar eficazmente técnicas de visualización para sus datos. Podemos finalizar diciendo sobre este debate acerca de la diferencia entre la visualización de información y científica, que actualmente son numerosos los ejemplos en los que para proporcionar una herramienta eficaz se emplean de forma combinada técnicas procedentes de ambas categorías, o incluso ubicadas directamente en la frontera, pues se exhiben simultáneamente características de ambas. Éste será el caso de nuestra propuesta, pues con la intención de ofrecer al químico la mejor herramienta, se combinan representaciones que reflejan estructuras físicas en 3 dimensiones, con representaciones que permiten entender cómo estas estructuras se agrupan según sus distintas propiedades.

3.2. Interfaces Gráficas de Usuario

La forma en que interactuamos con los ordenadores en estos días es a través de interfaces gráficas -ya que en los inicios de los ordenadores personales solo existía una línea de comandos en donde escribíamos los comandos para realizar alguna acción determinada-, esto es, por medio de ventanas o íconos invocamos alguna tarea en específico; este método permite una comunicación entre el ordenador y el usuario, haciendo la tarea más sencilla. Por otro lado, la velocidad de cálculo y la resolución

gráfica que ofrecen los sistemas informáticos actuales han incrementado, mejorando así las posibles formas de visualización de información. Por eso, las herramientas con un enfoque de visualización de información deben proporcionar una vía de comunicación (interacción) entre la representación gráfica de los datos, otras visualizaciones de la herramienta y el usuario. Para nuestra propuesta es imprescindible que el químico pueda interactuar con los resultados del *docking* de manera visual, permitiéndole seleccionar regiones de la visualización o facilitándole la retroalimentación cuando ocurra algún cambio en la repersentación o representaciones. Más aun, la visualización científica le permitirá al químico estudiar y comprender la interacción entre los compuestos y el sitio de la diana -Por ejemplo, le puede mostrar qué aminoácidos participan en la interacción-. De esa manera, le proporcionan información valiosa sobre la función y el comportamiento de la macromolécula [21]. Por otro lado, la visualización de información abstracta puede revelar patrones, grupos, vacíos o valores atípicos de datos estadísticos de los resultados del *docking*. Todo esto adquiere su máxima potencia cuando se aprovecha la gran capacidad perceptiva que tenemos los seres humanos para escanear, reconocer, y recordar imágenes rápidamente, así como para detectar cambios en tamaño, color, forma, movimiento o textura [114].

3.2.1. Proceso de Visualización

A partir de un conjunto de datos, el diseñador decide qué técnicas de visualización son las más apropiadas para que el usuario final explore de la forma más eficiente posible ese conjunto de datos. El usuario posteriormente ha de experimentar -a través de la interacción- con los diferentes tipos de opciones para personalizar la representación gráfica tales como el estilo, diseño, las vistas, colores, funciones, hasta que obtenga una o varias representaciones que satisfagan sus necesidades. Dependiendo del tipo de visualización seleccionada y tal vez personalizada, ésta proporcionará los resultados esperados. A través de estas ayudas visuales (representaciones gráficas) interactivas -adaptación de las representaciones para que éstas respondan a la pregunta que el usuario se está haciendo sobre los datos-, el usuario podrá obtener suficiente información o conocimiento del conjunto de datos, o incluso obtener un panorama completo acerca de los datos y de esa forma ayudar a otros en el proceso de adquisición de conocimiento [20].

Por ejemplo, un investigador necesita filtrar una red de interacción proteína-proteína con 2700 nodos para encontrar posibles blancos farmacológicos para el tratamiento de alguna enfermedad. Lo primero que él desearía ver es qué nodos de la red contienen el mayor número de interacciones, por lo cual emplea un programa que visualice dicha red. El programa le proporciona diferentes tipos de disposición de los nodos de la red: grafo dirigido por fuerzas (*force directed*), malla (*grid*) y circular. Supongamos que selecciona la de malla. El resultado son una serie de círculos (nodos) interconectados por líneas y, al ser demasiados, resulta imposible hacerse una idea de las conexiones entre los nodos, por lo tanto, de este tipo de representación no puede sacar mucho provecho. Como segunda opción selecciona la circular y obtiene un resultado similar; finalmente selecciona el grafo dirigido por fuerzas (en el que los nodos siguen un modelo de atracción/repulsión similar a las moléculas o los astros) el cual le ofrece una mejor comprensión de la interacción entre las proteínas. Ahora el investigador, realiza un filtrado por aquellas proteínas que no sean semejantes a las de humano (no ortólogos a humano) para determinar los posibles blancos farmacológicos. Con el resultado del filtrado, el investigador obtiene una idea general de las interacciones de proteínas en función del filtrado. Posteriormente, el investigador resaltará de alguna forma aquellas proteínas que sean de su interés (blancos farmacológicos) o empleará algún método para personalizar la visualización actual, para continuar con el análisis. Por lo general el proceso de análisis visual de este tipo de redes in-

volucra una mayor cantidad de datos, por lo que, la visualización debe ofrecer la mayor cantidad de información posible, para que a medida que el investigador haga uso de las distintas funciones de la visualización tenga una idea clara del contexto de lo que está viendo. En otras palabras, el proceso de análisis visual ocurre mientras se va navegando/explorando sobre la visualización. En cada interacción que se tenga con la visualización, nos debe revelar nueva información que se sumará al conocimiento previo del problema, y por consiguiente se encontraran respuestas o más preguntas -en el sentido de que ha respondido una parte del problema-.

El proceso de visualización -aprehender un gráfico generado a partir de datos abstractos- se puede comparar con el proceso que se realiza con un motor de búsqueda de internet [20], excepto que, en términos generales, es mucho más complejo que concatenar unas cuantas palabras clave en el motor de búsqueda. En la visualización, las acciones equivalentes al botón de "buscar" son por lo general llevadas a cabo mediante técnicas específicas. Algunos ejemplos puden ser: *zoom*, barrido -interacción con una representación que dispara un filtrado sobre los datos originalmente representados-, foco+contexto -aunque el foco de exploración esté en unos determinados detalles de la representación, se mantiene el contexto global de los datos, para disminuir la carga cognitiva del usuario-, etc.; estas técnicas, junto a otras, se explican en detalle en la sección 3.3. Dado que el número de parámetros que se emplean para realizar la "búsqueda" es grande, el proceso de búsqueda se traducirá en explorar varias vistas o realizar diferentes funciones de filtrado; por ejemplo, si se deseara analizar los resultados de la expresión de genes en un *microarray* de 300 genes para 20 pruebas, podría, inicialmente, representarse visualmente el *microarray* y posteriormente, una vez que se encontrara computacionalmnete realizarse un agrupamiento de los datos, a continuación visualizarlos en un diagrama de dispersión. Sin embargo, para poder representar gráficamente esos datos habría que aplicar otro método, en este caso uno de reducción de dimensionalidad para pasar de 20 a tres o dos dimensiones. Por otro lado, la interacción con las representaciones es vital para la "búsqueda", ya que permitirá al usuario explorar la información de tal manera que pueda obtener un mayor conocimiento o entendimiento de la información, o, puesto en otras palabras, adaptar la representación para que ésta responda a la pregunta que el usuario se está haciendo sobre los datos o el problema en sí mismo. Una dificultad que puede surgir, sin embargo, es que la interacción puede ser lenta -especialmente cuando se visualizan conjuntos de datos enormes-. Así, es lógico que, en las décadas pasadas, se haya puesto énfasis en mejorar la velocidad de las herramientas de visualización, para que de esta forma puedan tener una interacción rápida a la hora de realizar las "búsquedas".

3.2.2. Visualización para Asistir en la Adquisición de Conocimiento

Se requiere en la visualización de datos que el usuario tenga un dominio amplio sobre los datos que está explorando, ya que forma parte del propio proceso. Por ejemplo, el usuario, al codificar con colores específicos para diferentes objetos en la visualización, estará dándole un significado específico de acuerdo al dominio de su conocimiento sobre los datos que está explorando. Retomando el ejemplo del diagrama de dispersión, el usuario podría interactivamente cambiar la forma, el color o la textura de los grupos formados con el fin de poder diferenciar aquellos genes que se han expresado más que otros, así, cuando el usuario abandone temporalmente el proceso de análisis (tal vez unas horas, días o meses) y vuelva a ver la gráfica podrá retomar el análisis en el punto en el que la dejó sin tener que repetir todo el proceso. También, podría variar las vistas que utiliza (por ejemplo, una representación jerárquica de los grupos encontrados), y así, obtendría otro tipo de información significativa o conseguiría desvelar otros escenarios que requieran posterior investigación con más detalle. Por otro lado, en muchas ocasiones, la carencia de cierto conocimiento por parte del usuario

sobre la visualización o herramienta desarrollada, es un obstáculo para la explotación de las técnicas de visualización. Así, si el usuario no recibe el entrenamiento adecuado sobre cómo realizar ciertas funciones o no dispone de tiempo para explorar todas las posibles vistas, la eficacia de las soluciones visuales desarrolladas se ve comprometida. Estas dos situaciones mencionadas son indicadores de la necesidad de poner especial cuidado en desarrollar herramientas y técnicas que sean intuitivas para los usuarios poco experimentados y así ayudar a la adquisición del conocimiento a través de las visualizaciones. De igual forma, al compartir el conocimiento generado (gracias a las visualizaciones) entre diferentes usuarios, mediante técnicas de visualización orientadas a reducir la carga cognitiva, otros usuarios menos experimentados también podrán obtener conocimiento a través de diferentes vistas de los datos.

3.3. Técnicas de Visualización

La literatura en el campo de visualización de información es copiosa; ha acumulado en las últimas décadas un extenso catálogo de técnicas y estrategias de visualización que resuelven tanto problemas genéricos como aspectos particulares del tipo de datos que se quiere representar. Una revisión pormenorizada de todas ellas excedería el espacio disponible en esta memoria, por lo que se remite al lector a alguno de los numerosos libros que las recogen (Card [67] [105] [27] [104] [43] [15] [13] [97] [14], Ware [134] [133] [135] [95] [49], Fry [30] [100] [101] [31], Spence [122], Tufte [128] [129] [130], etc.).

A continuación se exponen aquéllas técnicas o estrategias que son más relevantes para la presente tesis.

Vistas enlazadas: Las vistas enlazadas constituyen una técnica muy importante en los entornos visuales porque permiten a los usuarios comparar rápida, dinámicamente e interactivamente los datos visualizados en ventanas o vistas diferentes y, por tanto, analizar grandes cantidades de información mediante diferentes tipos de representaciones de los mismos datos [118]. Por ejemplo, supongamos que se conoce una proteína que es esencial para un virus y no se encuentra en el humano; sin embargo, no se ha reportado su estructura tridimensional, por lo cual un químico se da a la tarea de modelar una a partir de diferentes moldes (otras proteínas). Al resultado se le conoce como modelo. Para tener un mayor éxito en la construcción del modelo se generan varios para posteriormente seleccionar el mejor. Por lo que se podría tener una ventana consistente en una tabla con los resultados que miden la calidad de construcción del modelo y otras dos, que reflejen los mismos datos pero representados en una gráfica de dispersión y una en donde se muestren las estructuras tridimensionales de los modelos. Es habitual para realizar una mejor exploración, que dicha información esté representada en dos o más visualizaciones enlazadas para estudiar y representar esos datos [89]. Las vistas múltiples permiten al usuario utilizar diferentes representaciones para aspectos diferentes de los datos, dando así a los usuarios perspectivas adicionales. El concepto de enlace se refiere al hecho de que una interacción con cualquiera de las vistas supondrá un cambio en la representación -atendiendo a un determinado criterio- que se propagará al resto de las vistas; regresando al ejemplo anterior, mediante la gráfica de dispersión se puede agrupar por colores los modelos en que sus aminoácidos presenten una energía diferente a la del molde -lo que significaría que son modelos de mala calidad-. En contraparte el químico observaría los modelos en la representación tridimensional e incluso contrastaría los modelos con algún molde que empleó para su construcción, y, de esa manera, podría confirmar la calidad del modelo. Más aun, el enlace entre las vistas las mantiene sincronizadas du-

rante la interacción, permitiendo al usuario poner en relación la información entre las vistas, y por consiguiente ayudando a la navegación de la información [87]. North *et al.* [88], usaron esta técnica para visualizar múltiples tablas de varias bases de datos relacionales al mismo tiempo. El enlace es una técnica muy poderosa de interacción; por ejemplo, al seleccionar un grupo de datos en una vista que está enlazada con otra vista, provoca la selección de los mismos datos en la segunda vista. Así, si se seleccionan unas filas o una columna de la tabla de resultados, esos datos tendrían que ser resaltados automáticamente en el diagrama de dispersión y en la representación tridimensional. De esta manera, es posible que resulten evidentes nuevas relaciones, como la distribución, agrupación o subordinación entre los datos, que de otra manera permanecerían ocultas [9].

Reducción de Amontonamiento: Al visualizar un conjunto de datos de gran tamaño, por la cantidad de los propios datos (escala) y/o por la cantidad de atributos de los mismos (dimensionalidad), es habitual obtener una representación que produce una imagen amontonada -imagínese por ejemplo un grafo que represente las relaciones de amistad entre los alumnos de una universidad, en el que los arcos de relación se cruzan de tal forma que produce una gran maraña de líneas incomprensible, imposibilitando distinguir las propias relaciones-. Esto dificulta al usuario la tarea de distinguir patrones o relaciones posiblemente formadas por la información; es más, dadas las limitaciones de espacio en pantalla, al aumentar el número de elementos individuales a representar, prácticamente ninguna técnica de representación puede mostrar la información detalladamente. Esto ocasiona que muchos elementos gráficos se solapen y se pierdan partes de información que son útiles. Por tanto es necesario emplear técnicas que reduzcan el amontonamiento para así realizar un mejor análisis de los datos, entre las técnicas que pueden ayudar a reducir el amontonamiento están el foco+contexto y el barrido.

Foco+Contexto: Uno de los principios en los que se basan las distintas técnicas de visualización de información intenta evitar que el usuario pierda el contexto del problema cuando está examinando algún detalle en concreto del mismo. Una forma de resolver esto es proporcionar en todo momento una vista general (representación de todos los datos disponibles), junto a la vista del detalle (vista general+detalle). Una alternativa, conocida como foco+contexto, proporciona en una única representación tanto la vista general como la vista detallada, de forma que se debe recurrir a algún mecanismo complejo de transformación, que permita integrar ambas vistas. Al hablar de sistemas foco+contexto, es intrínsecamente necesario tener una noción de qué partes de los datos se consideran estar en foco y cuáles no. Una forma de visualización de foco+contexto se refiere a una distribución desigual del espacio de visualización tal que el espacio dedicado a cierto subconjunto de datos (datos en foco) es mayor. Al mismo tiempo, el resto de la visualización se comprime (distorsiona) para seguir mostrando el resto de los datos como contexto para que sirva de orientación al usuario [42] [70] [25]. Otra forma consiste en usar el color y la transparencia en los elementos visuales para mantener el contexto en segundo plano de la imagen representada. En otras palabras, centrándonos en la vista con el diagrama de dispersión y la representación tridimensional, en el primero le sería fácil ver aquellos modelos que sus aminoácidos presentan una energía diferente del molde, sin embargo, al pasar a la vista tridimensional resultaría complicado poder distinguir los átomos que conforman esos aminoácidos, ya que solo podrá ver una maraña de esferas y líneas interconectadas; la gran cantidad de esferas y líneas acabarían ocupando todo el espacio y por lo tanto el químico no vería nada. Para resolver esto, aplicaríamos el foco+contexto de la siguiente manera: supongamos que deseamos saber los átomos de esos aminoácidos para evaluarlos dependiendo de su posición en el espacio. Pintaríamos en color azul en el diagrama los modelos que cumplan esa condición, y, en el caso del resto de los datos, cambiaran todos a un color distinto y translúcido; como es también una vista enlazada, en la representación tridimensional se resaltarían los átomos cambiando de tamaño

y color, mientras el resto de átomos cambiarían a un color translúcido, así tendríamos en "foco" los de color azul que resaltarían los modelos que tienen aminoácidos con energía diferente del molde, mientras que los que están en otro color y translúcidos constituirían el "contexto".

Barrido (*Brushing*): El barrido es una técnica interactiva en la que el usuario puede seleccionar (arrastrando el ratón sobre un área de la gráfica para seleccionarla) un subconjunto de elementos sobre una subregión de la representación de datos, lo que desencadena la ejecución de una operación específica en ese subconjunto de datos [7]. La operación puede ser seleccionar, deseleccionar, ocultar, suprimir, resaltar, etc. Esta función sirve de criterio de contención para los puntos de referencia que son seleccionados por el barrido. Los principios del barrido fueron introducidos por Becker y Cleveland en [7], donde se aplicaban diagramas de dispersión de grandes dimensiones. En este sistema se mostraba un conjunto de datos multidimensional a través de diversas proyecciones de los datos en diagramas de dispersión en 2D; el usuario especificaba una zona rectangular en una de las proyecciones y, dependiendo del modo de operación activo, en las otras vistas los puntos que estaban dentro del barrido eran resaltados, borrados o etiquetados. Aplicando el barrido al ejemplo del foco+contexto, pero centrándonos en un mapa, podría servir para etiquetar aquellos que están en el mismo sitio geográfico lo cual sería difícil de saber en el diagrama de dispersión, sin embargo, también podría utilizarse a la inversa (seleccionar en el diagrama y ver los resultados en el mapa).

Estas técnicas de visualización se aplican cada día a un mayor número de áreas, tales como las páginas Web [44], análisis de redes [71], bases de datos [63], minería de datos [61], bioinformática [110] [17] [34] [93] [107] [108] [72] [33], entre otras muchas más. La exploración y el análisis de los datos son pasos cruciales en la investigación científica; médicos, físicos, matemáticos, y otros científicos examinan, exploran, y analizan datos para conseguir un mayor conocimiento de los problemas. Es razonable, por tanto, que la meta del software de visualización deberá ser facilitar este proceso de una manera intuitiva [81]. La visualización de datos ha capturado un interés muy alto entre los científicos y muchos sistemas de visualización tanto comerciales como públicos han aparecido en años recientes [62].

La sección siguiente, se da una breve noción sobre el diseño de herramientas de visualización.

3.4. Diseño de Herramientas de Visualización

Una de las funciones de los investigadores en visualización de información es la de proporcionar pruebas convincentes de la utilidad de las herramientas, lo cual es difícil para cualquier nueva tecnología; sin embargo, representa nuevos retos para la visualización de información. Teniendo en cuenta las técnicas antes mencionadas (entre otras), los diseñadores procuran crear herramientas que sean altamente interactivas y que ayuden al proceso de adquisición de conocimiento. En otras palabras, que el usuario de estas herramientas forme ideas concretas en cada interacción –recordemos la metáfora de las búsquedas- con la visualización y le proporcione respuestas para tomar una decisión. Más aun, es importante que los diseñadores provean a los usuarios la personalización de la visualización, esto le permitirá encontrar patrones del comportamiento de los datos y por tanto del problema en general. Por lo tanto, la visualización de información se puede describir como una manera de responder a las preguntas que no se sabía que se tenían. Dicho de otro modo, es importante que la visualización ayude a resolver el problema de representar una gran cantidad de información y de que mediante interacción con los datos responda a las preguntas que emergen durante el proceso de visualización.

Según sea tipo de datos que se tenga (datos unidimensionales, bidimensionales, tridimensionales o multidimensionales), estos pueden ser abordados de diferentes maneras. Esta característica de los datos será determinante a la hora de diseñar una herramienta de visualización. En el caso de los datos tridimensionales (3D), por lo general representan objetos del mundo real, como moléculas, el cuerpo humano, edificios, etc. Se trata, por tanto de objetos que tienen volumen y algún tipo de relación compleja con otros objetos. En el caso del diseño de fármacos cae dentro de los datos tridimensionales, dado que las macromoléculas y los compuestos que interactúan con ellas son objetos del mundo real, sin embargo, la herramienta que se pretende diseñar bien podría emplear otro tipo de visualización de datos no necesariamente tridimensional, sin dejar a un lado la interacción entre los dos tipos de visualización (3D y 2D).

Shneiderman [114] plantea un análisis de las tareas que se realizan en función de los tipos de datos. Por lo que para diseñar nuestra propuesta tendremos en cuenta los siguientes puntos:

- *Vista global:* Mostrar una vista general de toda la colección de datos. La información general es una estrategia que incluye una vista reducida (gráfico compactado) de cada tipo de dato para poder mostrar todo el conjunto de datos, más una vista adjunta donde se mostrará el detalle de los tipos de dato. Esto puede resolverse de otro modo a través de un enfoque foco+contexto. Si retomamos el ejemplo de las redes de interacción proteína-proteína, será imprescindible para el químico tener una visión global de todos los nodos de la red de interacción. Esto le proporcionará una manera rápida a dónde dirigir los esfuerzos de análisis sin necesidad de inspeccionar cada nodo por separado, suponiendo que varios nodos de la red tienen una gran cantidad de conexiones.

- *Zoom:* Acercamiento visual sobre zonas de interés. Cuando un área es de interés, se utilizan herramientas que controlen el foco del *zoom* y un factor de escala. Es deseable considerar una implementación del *zoom* suave, que ayude a no perder la sensación de posición y contexto de todo el conjunto de datos. El *zoom* podría realizarse moviendo los controles de una barra de *zoom* o ajustando el tamaño de un cuadro donde muestre el foco. La visualización de la red, deberá permitir al químico centrar su atención sobre un grupo determinado proporcionándole amplificación visual de ese grupo seleccionado, esto es, el químico tendrá en una ventana o en un apartado, solamente el grupo de manera ampliada. En otras palabras, dicho grupo se redimensionará, así podrá realizar un análisis sobre el foco, por otro lado, en la vista global, el químico podrá ver marcado en otro color el grupo seleccionado.

- *Filtro:* Filtro de elementos de interés/no interés. Al permitir a los usuarios controlar el contenido que se muestra en la representación, los usuarios pueden centrarse en lo que es de su interés mediante la eliminación de elementos no deseados. Deslizadores, botones u otros componentes de control, deben desencadenar una rápida actualización de pantalla (menos de 100 milisegundos), incluso cuando hay decenas de miles de elementos que se tengan que mostrar. Un claro ejemplo de esto es la técnica del barrido. Supongamos que la red cuenta con una etiqueta que marca las proteínas que son ortólogas a las de los humanos, y otra con las que no lo son. Por lo tanto, el químico se centraría en aquellas que no están presentes en humanos. Por otro lado, esto traducido en la visualización equivaldría a ocultar (o difuminar) aquellos nodos que contengan la etiqueta de ortólogo-humano.

- *Detalles bajo demanda:* Seleccionar un elemento o grupo y obtener información cuando se necesite. Por ejemplo mostrar etiquetas en los elementos con información relacionada al elemento o al conjunto que pertenece.

- *Relacionar:* Vista de relaciones entre los elementos. Hace hincapié en la exploración de relaciones entre atributos y en encontrar correlaciones entre los pares de atributos numéricos.
- *Historial:* Mantener un historial de acciones para poder deshacer, repetir, y rehacer mediante un detallado progresivo. Es poco común que un usuario produzca el resultado deseado en una sola acción. Por lo tanto, la exploración de información debe ser un proceso con muchos pasos, por lo que se debe mantener un historial de las acciones y, además, es importante permitir a los usuarios volver sobre sus pasos. Suele suceder en la mayoría de los casos que los usuarios realicen filtrados sobre algo que ha sido filtrado muchas veces y se pierdan en la cantidad de cambios que han ocurrido, por eso las herramientas deben ser capaces de regresar al punto de inicio siguiendo el camino en dirección contraria al último filtrado, para posteriormente, si se desea, seguir refinando el análisis.
- *Extracción:* Permitir la extracción de subconjuntos de datos y de parámetros de consulta. Una vez que los usuarios han obtenido el elemento o conjunto de elementos deseados, es útil que estos se puedan extraer de esos conjuntos y guardarlos en un fichero con un formato que facilite a otros su uso (por ejemplo, obtener sub-redes, extraer un conjunto de atributos de la red, exportar la red o sub-redes a otros formatos, o que puedan ser analizados por algún paquete estadístico externo a la herramienta).

Finalmente, no es suficiente con desarrollar herramientas que incorporen las más novedosas técnicas visuales, si no que es necesario realizar algún proceso de validación de las herramientas visuales que se desarrollen. Con el objetivo de encontrar fuerzas y debilidades en lo que se pretende transmitir a través de la visualización. En otras palabras, la visualización debe cumplir el objetivo de crear modelos mentales de todo el contexto del problema que se requiere visualizar. Sin olvidar la interacción, que juega uno de los papeles principales para generar estos modelos mentales.

Una revisión de los métodos utilizados actualmente en la evaluación de herramientas y técnicas de visualización, identifica las siguientes categorías: Comparación de rendimiento en tareas similares con diferentes visualizaciones, evaluaciones por parte de los usuarios, y casos de estudio [116]. Por otra parte Plaisant [94], hace una revisión de los tipos de experimentos de evaluación:

1. Experimentos controlados en la comparación de elementos de un diseño. Se centra en la comparación de componentes gráficos específicos, por ejemplo, Ahlberg y Shneiderman [2], implementaron una técnica a la que llamaron deslizadores alfa, estos deslizadores seleccionan elementos de una lista enorme sin la utilización de un teclado, comparado contra la selección de un área por medio del ratón, entonces se procede a evaluar, el resultado de la selección; otro ejemplo sería la comparación de la distribución de la información por medio de gráficos en la herramienta.
2. Evaluación de usabilidad de una herramienta. Proporciona información de los problemas que encontró el usuario al utilizar la herramienta y muestra a los diseñadores cómo refinar el diseño. En otras palabras, se toma en cuenta la satisfacción del uso de la herramienta por parte del usuario, por ejemplo, si la herramienta es lo suficientemente interactiva, sencilla de manejar, y si ofrece la automatización de ciertos procesos para la extracción o procesamiento de sus datos.
3. Experimentos controlados en la comparación de dos o más herramientas. Es el caso más común de un estudio. Por lo general, estos estudios se centran en comparar una técnica nueva con el estado del arte.

4. Casos de estudio reales. Es de los menos comunes, ya que realizarlos consume demasiado tiempo, y los resultados no siempre son reproducibles o generalizables. Sin embargo, la ventaja que poseen es que proporcionan un informe sobre los usuarios realizando tareas reales, lo que demostraría la viabilidad y utilidad en su contexto.

En este capítulo se ha definido el área de la visualización de información, sus técnicas y cómo cada una de ellas sirve para poder comprender y obtener conocimiento, que por medios tradicionales es imposible de abordar. Sin embargo, en un sistema que explote técnicas de visualización de información, no necesariamente se aborda una tarea analítica o se implementan algoritmos avanzados para el análisis de datos. Por otro lado, para nuestro problema, encontramos que las técnicas de visualización como el foco+contexto, vistas enlazadas, barrido (para reducir el amontonamiento), así como la interacción entre las vistas y el químico, ofrece una solución para abordar la exploración de los resultados de experimentos de *docking* en el diseño de fármacos. En el capítulo siguiente, se examina otra área estrechamente ligada a la visualización de información en la que podremos encontrar respuestas en nuestros datos al aplicar técnicas visuales de análisis de datos.

Capítulo 4

Analítica Visual

Hoy en día constantemente nos enfrentamos a una gran cantidad de información generada por Internet y otros medios. Aun cuando hemos superado el problema de almacenamiento, somos incapaces de procesarla y asimilarla en su totalidad para obtener el máximo beneficio de ella. Esta situación la podemos encontrar en redes sociales, empresas, gobiernos, medicina, economía, etc. En el caso del diseño de fármacos, por la gran cantidad de datos (resultados de *docking*) que se generan, aplicar los métodos tradicionales de análisis resulta insuficiente. Incluso el realizar una inspección visual de todos los resultados resulta extenuante para el químico, incluso si varios químicos se dieran a la tarea de inspeccionar un lote de cientos de compuesto, cada uno tendría un punto de vista diferente , por tanto, la calidad de los resultados variaría, al no seguir un criterio objetivo. Esto representaría un problema cuando se necesita identificar un candidato a fármaco para atacar una enfermedad específica -como la crisis de la gripe H1N1 del 2009- de manera rápida. Por otro lado, el tratar de consensuar todos los criterios, tomaría posiblemente un tiempo valioso que bien podría invertirse en otros estudios. Ante esta situación, los químicos están en una posición en que es difícil abordar esta vasta cantidad de información y, por consiguiente, encontrar respuestas fácilmente. Sin embargo, aplicando los avances en analítica visual es posible combinar las capacidades de exploración visual que posee el ser humano uniéndolo con el poder de procesamiento de los ordenadores para crear un entorno más útil para la extracción del mayor conocimiento posible. De esta forma si se le presenta al químico una alternativa visual para analizar e inferir respuesta en un ambiente altamente interactivo con los datos (compuesto y diana), esto es, que pueda manipular libremente lo que está viendo, y que, cada interacción que realice, signifique una reestructuración interna de sus ideas, así, podrá ir adquiriendo un mayor conocimiento del problema. Esto es a lo que se refieren Miksch *et al.* [139] cuando dicen: “El usuario deja de ser un mero espectador pasivo que sólo interpreta datos, en lugar de eso se convierte en el primer actor de todo el proceso de análisis”.

Se puede decir que el campo de investigación denominado Analítica Visual, (forma compacta de su denominación más precisa, Ciencia Analítica Visual) tiene su documento fundacional en el libro “Illuminating the Path: Research and Development Agenda for Visual Analytics” [127]. En este libro, se define de la siguiente manera: “La analítica visual es la ciencia del razonamiento analítico facilitado por interfaces visuales altamente interactivas” y que posteriormente complementará Daniel Keim “La analítica visual busca representar la información visualmente, permitiendo al usuario interactuar con los datos para comprender mejor, sacar conclusiones y finalmente tomar mejores

decisiones basadas en conjuntos de datos extensos y complejos" [64] [60].

Otros autores aportan otros matices a las dos definiciones previas, como en el caso de Miksch *et al.* [139] que tienen en cuenta a la psicología cognitiva e incluso se mencionan algunos campos de aplicación: "La analítica visual es un campo intrínsecamente multidisciplinar que va desde la psicología cognitiva a la investigación en bases de datos y cuyas áreas potenciales de aplicación son la medicina y biotecnología, negocios, seguridad y gestión de riesgos, el clima y el medio ambiente, entre otras". Por otro lado, Chen [19] hace énfasis en el aumento de la capacidad de análisis: "A través de la analítica visual se aumentan las capacidades de análisis y toma de decisiones para poder comprender situaciones complejas y llegar a decisiones de manera informada". También es interesante mencionar el punto de vista de Cook *et al.* [23] que hace referencia a las metáforas visuales abstractas para descubrir lo inesperado: "La analítica visual es la formación de metáforas visuales abstractas en combinación con un discurso de información humano (usualmente alguna forma de interacción) que permite la detección de lo que se estaba esperando y, además, la posibilidad de descubrir lo inesperado dentro de espacios de información que son masivos y cambiantes". Y, recientemente, Keim *et al.* [24] dan una definición más específica en el libro "Mastering the Information Age Solving Problems with Visual Analytics", la cual caracterizan de la siguiente forma: "La analítica visual combina técnicas de análisis automatizado con visualizaciones interactivas para la comprensión eficaz, razonamiento y toma de decisiones en base a datos muy grandes y complejos".

Por tanto, se puede afirmar que, la analítica visual ofrece una perspectiva guiada de exploración, gracias a las herramientas interactivas y de componentes analíticos subyacentes; los usuarios pueden explorar datos de grandes dimensiones que son previamente procesados por algoritmos de ordenador que se apoyan en diversas disciplinas como la estadística, minería de tatos, recuperación de información, inteligencia artificial, etc. El resultado es la combinación de los puntos fuertes de cada enfoque: el enfoque analítico de modelos estadísticos, y del derivado de la exploración visual [137]. Keim *et al.* [60] [65] [66], hacen un estudio de posibles áreas de interés en donde se aplica la Analítica Visual: biología y medicina, ingeniería, análisis financiero [39], socio-economía, Seguridad pública, seguridad y seguridad geográfica, física y astronomía [5] [4].

4.1. Elementos de la Analítica Visual

Actualmente, para una persona que se dedica a tomar decisiones, como en el caso de gerentes, bioinformáticos, químicos, médicos, etc., es crucial el disponer de la mayor cantidad de datos para extraer conclusiones precisas cuando se necesite. En otras palabras, la información en bruto (tablas, ficheros de texto, resultados de consultas a bases de datos, imágenes, etc.) aportan muy poco o nada, de forma individual y aislada, a la solución del problema. Con el objetivo de hacer posible la extracción de tales conclusiones ante apabullantes volúmenes de datos, la analítica visual se centra en el manejo de grandes volúmenes de información heterogéneos y dinámicos a través de la integración del juicio humano por medio de representaciones visuales y técnicas de interacción en el proceso de análisis [65].

Uno de los principales problemas que se provocan cuando no se manejan adecuadamente grandes volúmenes de información es que el observador (usuario) se pierde entre toda esa gran cantidad de datos. Esto ocurre cuando:

a) Al tener que reducir la cantidad de datos que se pueden representar, no se muestren los que

en ese momento sean relevantes para responder una cuestión en particular. Por ejemplo, volvamos al caso de las redes de interacción proteína-proteína; supongamos que la visualización muestra las interacciones de la red pero en vez de mostrar las proteínas que son esenciales al virus muestra todas, tanto las esenciales como las que no, sin embargo, lo que se deseaba analizar eran las que cumplieran la condición de ser esenciales para poder filtrar las que fueran ortólogos de humano y, por consiguiente, ser consideradas como blancos farmacológicos.

b) Se procesa de manera inadecuada. Supongamos ahora, que la información presentada nos muestra lo que se desea analizar, sin embargo, por la forma en que fue procesada, en la visualización se consideran dos variables como si fueran del mismo tipo. En nuestro ejemplo, se procesaron igual las proteínas que siempre están presentes (tienen una alta interacción) que las que sólo en ocasiones especificas interactúan (tienen baja interacción).

c) Se representa de manera inadecuada. Esto hace referencia al tipo de visualización para representar los datos analizados. Por ejemplo, si hay muchas variables a considerar y se desea analizarlas todas en una sola representación, los métodos tradicionales en los que solamente se pueden representar hasta tres dimensiones quedan automáticamente descartados. Supongamos que se ha analizado la red mediante algún método, y nos arroja los siguientes resultados: coeficiente de agrupamiento, grado de identidad con otras proteínas, sitios de interacción, ortólogos, no ortólogos, proteínas siempre presentes, proteínas ocasionalmente presentes, etc. Es claro que representar esto en la red directamente no es posible, es necesaria una alternativa.

Cualquier proceso analítico es complejo y consume mucho tiempo, por lo que se debe investigar en el desarrollo de software para afrontarlo y de esa manera el usuario podrá responder a cuestiones más complejas [60]. Actualmente se está popularizando el uso de herramientas de analítica visual y técnicas que sintetizan la información para generar conocimiento a partir de datos masivos, dinámicos, ambiguos y por lo general conflictivos; por ejemplo, las redes de interacción proteína-proteína suelen analizarse mediante grafo dirigido por fuerzas, sin embargo, estos presentan dos puntos débiles, el primero es que es difícil volver a reproducir la misma imagen entre ejecuciones, el segundo es que no necesariamente se representa información biológica complementaria, aparte de nombres de las proteínas y sus identificadores, pero no así la ontología. Fung *et al.* [32], propusieron una visualización complementaria al grafo dirigido por fuerzas, esta visualización se basa en una representación circular de grupos a partir del grafo de fuerza, los grupos se forman en base a la ontología, por lo que obtienen la distribución de las proteínas en diferentes tipos de componentes subcelulares.

Así, este tipo de herramientas ayudan a los químicos a ver lo que en un momento dado supusieron como resultado mientras recolectaban o generaban su información. En el caso del diseño de fármacos, este importante momento es cuando configuran los parámetros para realizar experimentos de *docking*, por ejemplo, el espacio en donde rotará, expandirá o se contraerá el compuesto, así como el lugar donde tenderá a moverse (efecto de la interacción con los aminoácidos). De la misma manera, el químico podrá encontrar nuevas preguntas que hasta ese momento no habían surgido.

En las definiciones previas ha quedado patente que la analítica visual es un campo multidisciplinar que incluye las siguientes áreas interrelacionadas, como se puede ver en la figura 4.1:

- Técnicas de razonamiento analítico, que permitirán a los usuarios obtener una visión profunda que ayude directamente a realizar evaluaciones, planteamientos y toma de decisiones.

- Representaciones visuales y técnicas de interacción, que explotan la potencia de percepción del ojo humano como vía para llegar a la mente y permitir que el usuario vea, explore y comprenda grandes cantidades de información simultánea.
- Representación y transformación de datos, que convertirán los tipos de datos conflictivos y dinámicos para que puedan ser visualizados y analizados correctamente.
- Técnicas de producción, presentación y difusión de los resultados de un análisis, para comunicar la información en un contexto apropiado para diferentes tipos de audiencias.

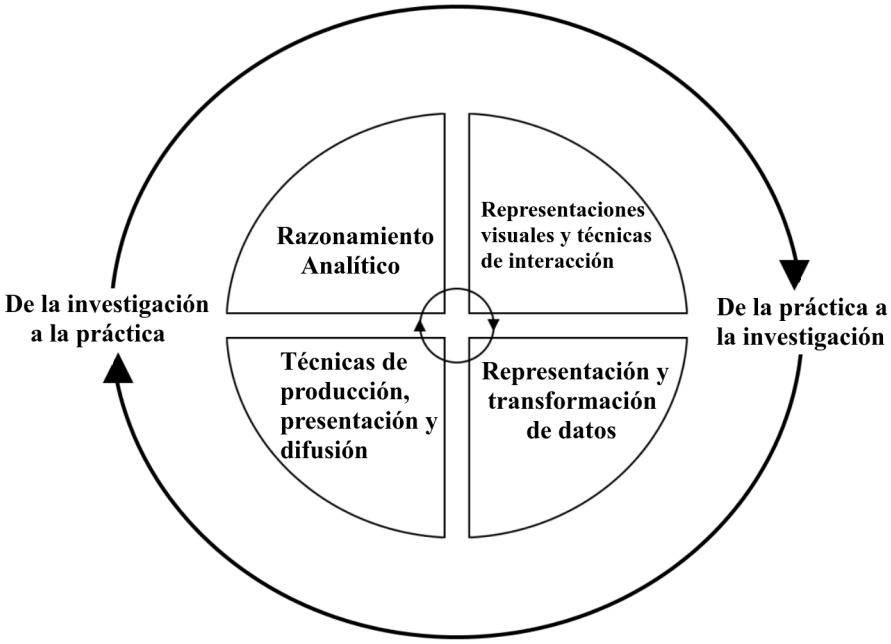


Figura 4.1: La analítica visual es un campo multidisciplinar que incluye áreas interrelacionadas: el razonamiento analítico como fruto de las representaciones visuales e interacciones con ellas, las cuales provienen de la transformación de los datos en representaciones, y por medio de técnicas de producción, presentación y difusión, transmiten a diferente tipo de usuarios la información necesaria del contexto del problema (fuente: Illuminating the path [127]).

4.1.1. Razonamiento Analítico

El gran desafío de la investigación en analítica visual es desarrollar una o varias visualizaciones para realizar el análisis de información, así como facilitar un razonamiento estructurado. Dado que el objetivo es facilitar el proceso de razonamiento analítico, esto se aborda por medio de la creación de software que maximice la capacidad de percibir, comprender, y razonar incluso en situaciones con datos dinámicos y complejos, obteniendo así un juicio de calidad con un esfuerzo mínimo en tiempo de análisis [127]; todo deberá concluir en un solo momento, como un asentimiento de cabeza, “allí está” o “mira esto”, etc.

Es sabido que la percepción juega un papel importantísimo en el desarrollo de conocimiento, ya que es el medio por el cual las personas interpretan lo que sucede a su alrededor y por tanto extraen conocimiento a partir de las visualizaciones presentadas [24]. Por eso hacemos hincapié en la percepción como clave fundamental del razonamiento analítico; nuestra propuesta debe ser capaz de producir en los químicos ese asentamiento de cabeza, en el que confirmará o descartará los resultados obtenidos de los experimentos de *docking*, el cual, sólo podrá conseguirse gracias a las diversas técnicas computacionales, estadísticas y de visualización, sumando la interacción con los datos visualizados. Sin embargo, otros autores no consideran a la percepción como conocimiento, uno de estos autores es Chang *et al.* [18]; ellos consideran que limita la capacidad de las visualizaciones para estructurar conocimiento y mostrar información. En otras palabras, podríamos decir que consideran a la analítica visual como una representación de datos y realmente no hay una diferencia entre los dos mundos; más aun, no lo ven como disciplinas complementarias. En suma, nuestra propuesta debe dar soporte al proceso de razonamiento analítico, y facilitar al químico que pueda enfocarse en lo que realmente es importante [127].

A través de herramientas y técnicas se debe proporcionar al usuario la capacidad de análisis del o de los problemas en múltiples niveles de abstracción y facilitar el razonamiento acerca de situaciones o eventos cambiantes con el tiempo, incluyendo a los que cambian extremadamente rápido, como el *framework* propuesto por Yedendra y Jarke [117], que está completamente enfocado al proceso de razonamiento analítico. Dicho *framework* consta de tres vistas enlazadas: la vista de datos que consta de varias herramientas interactivas de visualización, la vista de conocimiento, en particular proveen de un editor de gráficos en el cual se plasman los modelos mentales del análisis y, por último, la vista de navegación que ofrece un resumen de todo el proceso de exploración al hacer capturas de los estados de las visualizaciones al realizar alguna interacción con ellas. De esta manera el usuario podrá ir revisando, validando y re-evaluando lo que haya descubierto durante el proceso de análisis en cada una de las visualizaciones. En definitiva, este tipo de herramientas ayudaran al usuario a organizar su información, tener una idea general para explorar y obtener así información potencialmente utilizable [64].

Es crucial identificar las diversas tareas analíticas que deben apoyar las herramientas visuales, en el caso del análisis de resultados de *docking* del diseño de fármacos, las podemos resumir en:

- Comprender la tendencia de agrupación de las poses en el espacio conformacional de la diana; en otras palabras, debe explicarnos de forma clara y precisa, porqué la mayoría de los resultados de *docking* posicionan a los ligandos en una zona determinada.

- Identificar los aminoácidos que cubren esa zona y determinar su importancia en la interacción con los ligandos o los sustratos. Además, en el caso de sitios poco conocidos -en caso de que el hueco corresponda al de una encima-, extraer aquellos aminoácidos que no habían sido considerados hasta ese momento y buscar en la literatura por si han sido reportados.

- Apoyar a la toma de decisiones, si se habrá que cambiar o redefinir (aumentar o reducir) el espacio conformacional que se ha usado hasta ese momento.

Como derivado del proceso de razonamiento analítico, surge un diálogo entre el químico y la información. Al dar paso a este discurso, se forma el corazón de la misión de la analítica visual, que se conoce como el discurso analítico.

4.1.2. Discurso Analítico

Definición: "Es la tecnología para mediar un diálogo entre el usuario y su información para emitir una evaluación o juicio acerca de algún problema" [127].

Este discurso es un proceso evolutivo e iterativo en el cual se va construyendo un camino que va desde la definición del problema, el encadenamiento de evidencias y la creación de hipótesis hasta poder emitir un juicio o una sentencia. Por ejemplo, supongamos que se tiene el gráfico de un *microarray* en una ventana divida en dos, en la parte superior de la ventana se representa al *microarray* en su estado inicial, sin ningún tratamiento (datos en bruto, *microarray* original) –entiéndase por esto que sólo se ha representado sin aplicar ningún tipo de análisis-, el segundo muestra al *microarray* después de aplicarse técnicas de agrupamiento y de visualización (*microarray* ordenado, parte inferior). Siguiendo con el mismo ejemplo, supongamos que se aplicó un algoritmo de clasificación jerárquica y se está explorando un nivel de la jerarquía. A su vez, el analista, selecciona un grupo de ese nivel. Al realizar esto, automáticamente se marcaría en el *microarray* ordenado (parte inferior) un recuadro que abarcaría todas las casillas de ese grupo, pero, al mismo tiempo, las columnas del ordenado se unirán a las columnas del original por medio de líneas, haciendo así una conexión con los datos no agrupados y proporcionando respuestas a cómo se han agrupado los datos. Finalmente, el usuario podrá realizar un informe que incluiría:

- El problema a tratar.
- La información que el usuario ha reunido con respecto al problema, que puede o no incluir evidencias relevantes.
- La evolución del conocimiento del usuario sobre el problema, incluyendo suposiciones, hipótesis, escenarios, modelos o argumentos.

Más aun, Keim *et al.* [60] definen el objetivo de la analítica visual como una forma transparente para procesar la información y datos para el discurso analítico. La visualización de estos procesos, proporcionará los medios de comunicación acerca de ellos, en vez de quedarse solamente con los resultados. En un discurso analítico, se aprovecha el potencial tanto de los sistemas por ordenador como de los humanos para mejorar el proceso de análisis. Por un lado los ordenadores se encargan de encontrar patrones en la información para después organizarla y presentarla al usuario de forma que ésta sea reveladora. Por su parte, el usuario provee su conocimiento de forma que el ordenador refine y organice la información más apropiadamente.

4.1.3. Hacer que Todo Tenga Sentido (*Sense-Making*)

Una vez que se recopilan y organizan los datos en formas que facilitarán futuras cuestiones, el usuario debe realizar una serie de actividades para que esto tenga un sentido. Desgraciadamente, no siempre se puede obtener una visión de lo que está ocurriendo, porque simplemente es difícil en la práctica que la evidencia y el descubrimiento encajen tan perfectamente y nos revelen todo [23]. Por eso, el usuario debe hacer sus conexiones de piezas de datos dispersos para construir escenarios plausibles del gran todo. Así como el concepto del discurso analítico representa a una perspectiva de investigación aplicada, la investigación sobre el *sense-making* o "que tenga sentido" provee una

base teórica para comprender muchas de las tareas del razonamiento analítico que el usuario debe realizar (figura 4.2).

Muchas de las tareas del razonamiento analítico siguen un proceso de:

- Recolección de información.
- Re-representación de la información en formas que ayuden al análisis.
- Obtención de un descubrimiento a través de la manipulación de estas representaciones.
- Producción de resultados a partir del conocimiento o una acción directa basada en el conocimiento del descubrimiento.

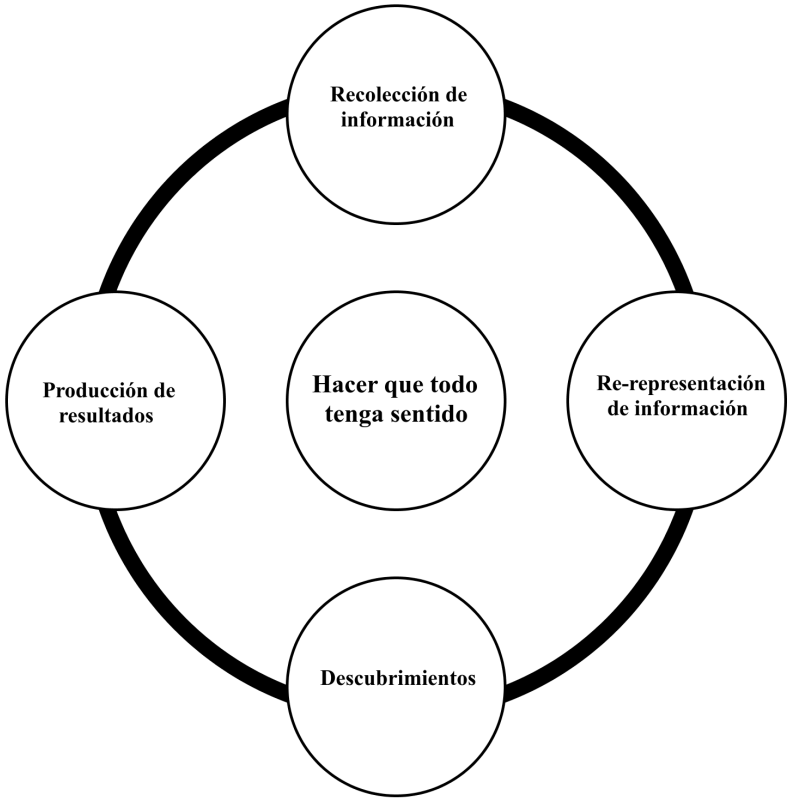


Figura 4.2: Proceso del razonamiento analítico (fuente: Illuminating the path [127]).

Para lograr el flujo del discurso analítico es necesario comprender la interacción entre la percepción y la cognición y cómo se ven afectados cuando se trabaja con ayuda externa. En otras palabras, son el proceso de percepción y cognición y el resultado de nuestras interacciones los que actualizan nuestro entendimiento. El proceso de percepción nos liga a nuestro ambiente y es crucial para asimilar el mundo que nos rodea. Esto es, a través de representaciones visuales el usuario explora sus

datos. En esta exploración se requiere que exista una interacción con los datos para comprender anomalías, separarlas y reorganizar la información apropiadamente, y después participar en el proceso del razonamiento analítico. Y gracias a estas interacciones el analista consigue comprender mejor. Por otro lado, la cognición permite comprender lo que se está visualizando (información visual) para hacer deducciones basadas principalmente en formación previa.

En el caso de estudio de Fung *et al.* [32], aplicaron su propuesta para analizar una red de interacción proteína-proteína en la replicación de ADN humano. La visualización consistía en un grafo dirigido por fuerzas, ellos mencionan que este tipo de visualización debe resaltar proteínas con un gran número de conexiones, y distinguir dos tipos de nodos, unos que se conocen como "nodos siempre presentes" (*party hubs*) y "nodos ocasionales" (*date hubs*). Esta clasificación la hacen en base al número de conexiones que posee el nodo. Asimismo, debe mostrar nodos conocidos como "cuellos de botella" -estos nodos sirven como puente entre nodos-. Por otra parte, la organización de la red se dividirá en dos módulos de interacción: funcionales y físicos. Sin embargo, ellos aseguran que con ese tipo de visualización (grafo dirigido por fuerzas), sólo consiguieron: la interacción entre nodos y la interacción entre subredes y cada una con una topología diferente. En contraparte, mediante la implementación de su visualización (agrupamiento en base a la ontología), consiguieron: localizar proteínas en complejos de proteína que interactúan, probables proteínas que son cuellos de botella y la localización de proteínas y complejos en múltiples orgánulos. En su caso de estudio, ellos encuentran al aplicar su propuesta que una proteína marcada como un nodo ocasional en el grafo dirigido por fuerzas, bien podría ser un cuello de botella al unir dos grupos de ontología diferente, lo que la catalogaría en el módulo funcional. En su experiencia, ellos intuían que no era un nodo ocasional. Retomando los párrafos anteriores, es claro que se generó un discurso analítico entre lo que ellos estaban viendo, su conocimiento previo y los datos en cuestión. En otras palabras, después de aplicar su propuesta, algunas de las suposiciones que tenían han tomado sentido.

4.1.4. Representaciones Visuales y Técnicas de Interacción

El mayor desafío al elegir una representación visual está en encontrar la que sea adecuada para la tarea que se está haciendo en ese momento y, lo más importante, es que no vale cualquiera. El uso de representaciones visuales y técnicas de interacción para obtener una comprensión de datos complejos es lo que distingue al software de analítica visual de otras herramientas analíticas. Las representaciones visuales traducen los datos en formas visibles para destacar características importantes, incluyendo tanto los aspectos comunes como las anomalías.

4.1.4.1. Representaciones Visuales

Al seleccionar una visualización principal -la primera que se muestra al usuario-, esta debe cumplir los requisitos antes mencionados, porque a partir de ese momento, dará comienzo todo el proceso analítico, sin embargo, es difícil cubrir todo los aspectos en una sola representación. Las herramientas de analítica visual saben de la importancia de dicha limitación, para resolverlo, emplean múltiples vistas y la suma de cada una, incrementa el conocimiento del usuario. Por ejemplo, supongamos que un químico desea alinear varias secuencias, porque necesita saber si un trozo de las secuencias -bien podría ser por donde interactúan con otras proteínas- se ha mantenido a lo largo de la evolución. La visualización elegida consistiría en mostrar las secuencias alineadas y resaltando los trozos que comparten, así como proporcionarle información biológica adicional. Tal vez, esto podría ser "suficiente"

y cumplir con lo que el químico deseaba saber en ese momento. No obstante, si el químico pudiera ver la estructura tridimensional de cada secuencia y, además, se mostrarán esos aminoácidos en otro color y forma, el químico tendría un panorama enriquecido en comparación con el anterior. Ya que una parte de la información se traduciría en conocer si los aminoácidos están en la superficie o en el interior -cada una de ellas tiene una implicación biológica diferente-, conocería el plegamiento de la estructura -tal vez compartan algún sitio de interacción y puedan ser consideradas como dianas-, entre otra información relevante.

Con esto, queremos dejar en claro la importancia de seleccionar una visualización apropiada para cada situación, no es que una sea mejor que otra, sino que son complementarias. Por eso, las vistas múltiples permiten una manipulación más intuitiva. Nosotros -los seres humanos-, no interactuamos con información de una sola dimensión, porque somos capaces de procesar información en varios niveles, tales como: percepción, emoción y cognición. Comúnmente, los usuarios se conforman con una respuesta del tipo "con esto es suficiente para mí", y esto ocasiona que se detenga el proceso analítico antes de identificar información critica que podría llevarlo a una conclusión totalmente diferente [47], como en el caso del ejemplo del párrafo anterior. Por eso cuando la información es sometida a la exploración, se convierte en semánticamente rica y puede ser visualizada a través de varios niveles de categorías. Las visualizaciones deben tener como prioridad categorizar la información para ayudar al usuario a dirigir su atención a una vista en particular, por ejemplo la implicación que los aminoácidos estén expuestos (en la superficie) y no ocultos. Es más, al utilizar múltiples vistas de organización de la misma información, el resultado es una poderosa ayuda -porque, lo que en una es difícil de mostrar en otra se resalta fácilmente-. Recordemos que, mientras el analista interactúa con la información en cualquiera de las vistas, las demás vistas también se actualizan (por estar relacionadas), y es por medio de estas interacciones que se hace posible que el humano y el ordenador proporcionen, tomen y generen conocimiento [39].

4.1.4.2. Transformación de Datos

Para poder visualizar, analizar y hacer un informe, la información debe ser transformada desde su forma original en bruto a una representación que sea susceptible de ser manipulada. En otras palabras, se debe procesar y transformar esta cantidad de datos en dimensiones manejables, sin que se pierdan sus características. Esto es lo que hace que las representaciones y transformaciones sean el fundamento sobre el cual la analítica visual está construida. A través de las representaciones se resaltan las características de los datos en lugar de mostrar cada detalle de la información, por lo cual deben existir procesos de abstracción de datos que son importantes. El nivel al que puede llegar una herramienta (*software*) en el escalado de datos está directamente influido por el tipo de representación visual que seleccionó el desarrollador. En otras palabras, se trata de comprimir al máximo la información que describe a los datos de tal forma que pueda ser representada en un objeto visual, por ejemplo, en [33], aplicaron un algoritmo de clasificación a un *microarray* con 17 pruebas, posteriormente, los grupos se visualizaron en un diagrama de dispersión en 3D. Para poder realizar la proyección y posterior análisis de los grupos, emplearon una reducción dimensional de los datos aplicando análisis de componentes principales. Principalmente emplearon el diagrama de dispersión en 3D para validar su algoritmo de clasificacion. Otro ejemplo lo encontramos en lo que propuso Yang *et al.* [142], ellos proponen una herramienta de visualización que reduzca las dimensiones de varios conjuntos de datos que a su vez son multidimensionales. Para representar los conjuntos de datos emplean un escalado multidimensional para después transformarlos en glifos -un glifo es una representación gráfica de un significado-. Principalmente tratan de representar la relación entre

dimensiones, detectar grupos y valores atípicos.

Es por tanto claro, que, a través de la creación de representaciones apropiadas, se pueden producir representaciones visuales significativas. El método utilizado para representar los datos debe facilitar los métodos de razonamiento analítico para capturar los resultados intermedios y finales del proceso de razonamiento.

4.1.4.3. Técnicas de Interacción

En muchas ocasiones en el proceso de desarrollo de herramientas de analítica visual, los investigadores tienden a enfocarse solamente en las representaciones visuales de los datos dejando a un lado la interacción [127], por tanto, hay que ser consientes de no dejar a un lado la interacción. El mantra de búsqueda de información, de Shneiderman [114], que dice "partir de una vista general, hacer un *zoom*, filtrar, detalles bajo demanda" (*Overview first, zoom and filter, details on demand*, en inglés), enfatiza el papel de la visualización en el proceso de descubrimiento o generación de conocimiento. Sin embargo, Keim [60] ajusta este mantra y lo enfatiza con respecto a la analítica visual: "Analizar primero, mostrar lo importante, hacer un *zoom*, filtrar y analizar más en detalle, detalles bajo demanda". Lo importante de este nuevo mantra es que nos llama a combinar enfoques analíticos junto con técnicas avanzadas de visualización e interacción.

Los nuevos desarrollos en las tecnologías de interacción persona-ordenador proporcionan mejores formas de manipulación de datos porque son específicas y, sin embargo, no están atadas a una sola herramienta. Una técnica de interacción es la forma en la que se utiliza un dispositivo de entrada/salida para realizar una tarea específica en un diálogo de persona-ordenador [53]. Este diálogo se aplica a las visualizaciones y una de las muchas propiedades del diálogo es el foco, con el que se logra llegar a una parte específica de toda la tarea que se esté realizando en ese momento e invertir los esfuerzos necesarios para la toma de decisiones. Una de las maneras de interactuar es llevar un historial del foco actual del usuario, con el que posteriormente se llegarán a nuevas conclusiones o hipótesis. Por ejemplo, en el caso del análisis de redes, el usuario podría haber filtrado varias veces hasta llegar al punto deseado, sin embargo, supongamos que dejó por un tiempo el análisis y no recuerda cómo ha llegado a ese punto, entonces es necesario para él retroceder en sus pasos para poder retomar su análisis; esto es muy común que suceda, por eso es importante llevar un historial de lo que ha ocurrido, para poder avanzar en las dos direcciones. Nuevos esfuerzos en las técnicas de interacción persona-ordenador están emergiendo. Las máquinas están siendo diseñadas para sentir o inferir los atributos de los usuarios y utilizar un gran número de modalidades disponibles para interactuar con el usuario, como en el caso de invidentes o discapacitados [50], por eso debemos aprovechar estas nuevas tecnologías y explotarlas en las herramientas visuales de análisis.

Ware *et al.* [133], por su parte, explican que las interacciones en la visualización consisten en ciclos de retroalimentación que caen en tres clases: manipulación de datos, exploración y navegación, en el que los usuarios van encontrando el camino hacia sus respuestas. El más sencillo -y a la vez el más común- de los tres tipos de interacción es la manipulación de datos en 2D y 3D, a través de objetos gráficos -iconos o figuras determinadas- en la visualización y se realiza a través de dispositivos como: ratón, *joystick*, teclado o dispositivos inalámbricos y, recientemente con movimientos del cuerpo (*Kinect*). Esta manipulación en un ambiente 2D y 3D comúnmente se realiza al arrastrar el ratón sobre un área determinada. Por ejemplo, la selección de átomos que conforman un aminoácido. Otro punto importante en la selección y en la interacción en general, es el tiempo de espera, ya que varios segundos pueden ser cruciales para el éxito como herramienta de análisis

visual -desde el punto de vista del usuario, porque la espera causa enfado en el uso de este tipo de herramientas-. En la exploración y navegación abarca a las herramientas de visualización en 3D en el sentido de percepción de los datos. En el caso del diseño de fármacos correspondería a la representación tridimensional de las proteínas. La interacción en este tipo de datos debe ser fácil de explorar y navegar, en otras palabras, que no ocurra la sensación de perder el contexto cuando estemos enfocados en una zona especifica de la representación visual -esto suele ocurrir al navegar entre los objetos tridimensionales-. Por ejemplo, al pensar que estamos seleccionando un punto que aparentemente esta cerca de un grupo, cuando en realidad se encuentra a distancia mayor; otro caso común es colocar un objeto en donde aparentemente se encuentra otro objeto cercano. Una técnica que es sencilla de implementar y útil, es mostrar al usuario la orientación en todo momento, de esta forma no perderá la noción de navegación.

Tanto las representaciones visuales como las tecnologías de interacción, proporcionan los mecanismos que permiten a los usuarios ver y comprender grandes volúmenes de información de forma instantánea. Por medio de los principios científicos que rigen la representación de información, se deben proveer las bases para las representaciones visuales, y estos principios son necesarios para los nuevos enfoques en la interacción que darán soporte a las técnicas analíticas y, en conjunto, estos fundamentos proporcionaran las bases para los nuevos paradigmas visuales que soportaran el razonamiento analítico en diversas situaciones.

4.1.5. Analítica Visual versus Visualización de Información

El trabajo realizado hasta ahora en la visualización de información está estrechamente relacionado con la analítica visual, sin embargo, no necesariamente aborda una tarea analítica o implementa algoritmos avanzados para el análisis de datos. Por su parte, la analítica visual es más que visualizar. En otras palabras se puede entender como la integración, en la toma de decisiones, de visualización, factores humanos y análisis de datos. La analítica visual se distingue por automatizar en mayor parte el proceso de análisis, teniendo en cuenta las limitaciones de los ordenadores y, técnicas y algoritmos, para ofrecer soluciones adecuadas por medio de visualizaciones interactivas [60].

Las visualizaciones no sólo deben soportar representaciones de características críticas de los datos, sino también proveer suficientes pistas contextuales para ayudar al usuario a rápidamente interpretar lo que está viendo. Una persona puede ver la información desplegada en tiempo real o explorar un espacio de información utilizando técnicas de interacción. Sin embargo, lo que el cerebro puede recibir es limitado en términos de la información que debe procesar para poder hacer un juicio acerca de eso [23].

4.2. Producción, Presentación y Difusión

En el capítulo 5 de [127], se discute la importancia de comunicar de manera efectiva el proceso de razonamiento analítico a las diferentes tipos de usuario, a través de la producción, presentación y difusión, enfocados a la seguridad nacional después del ataque del 9/11. Sin embargo, esto es aplicable en el contexto de nuestra propuesta. La producción y presentación se reflejan en la creación de representaciones visuales que hacen más sencillo la interpretación del problema (el diseño de fármacos), después de la aplicación de métodos de computacionales, adquisición de datos, de técnicas

computacionales y visuales, resumiendo todo en términos que son significativos para los químicos. Por otra parte, la difusión abarca no sólo a los usuarios destinados, sino también a otros que no necesariamente son expertos en el problema, para los que también se logra transmitir el contexto del análisis.

4.3. Metodologías de Evaluación de la Analítica Visual

En el capítulo 6 del libro [127], mencionan las ventajas de incorporar la evaluación, tales como la comprobación de hipótesis, comparación de técnicas, determinar si una herramienta cumple con sus objetivos, etc. Ellos consideran que hay tres niveles de evaluación: componentes, sistema y de trabajo. El nivel de componentes abarca las representaciones visuales, técnicas de interacción y algoritmos de análisis. Cuando los algoritmos de análisis (nivel de componentes) no son propiamente de interacción con el usuario, suelen medirse en cuanto a la rapidez, precisión, o mediante sus límites. Por otro lado, cuando los componentes son de interacción, se miden en términos de efectividad, eficiencia y satisfacción del usuario. El nivel de sistema, la evaluación se realiza comparando los sistemas con los programas que actualmente están utilizando los usuarios. Las mediciones son realizadas en base a la capacidad de aprendizaje y utilidad, así como la satisfacción del usuario. Finalmente, el nivel de trabajo, las mediciones de evaluación son referentes a la aceptación y confianza de la nueva tecnología, así como la productividad.

Por otro lado, en el capítulo 8 del libro "Mastering the Information Age - Solving problems with Visual Analytics" [24], hacen un resumen del estado del arte en la analítica visual considerando los aspectos del párrafo anterior. Para ellos la evaluación concierne a la ***calidad*** de ***artefactos*** relacionados con la analítica visual. Los artefactos incluyen varios elementos que van desde el software, técnicas, métodos, modelos y teorías. Para la calidad consideran tres aspectos: efectividad, eficiencia y satisfacción de uso. Los aspectos de la evaluación que ellos proponen, se muestran en la figura 4.3, y estos son: los artefactos, usuarios, tareas, y datos.

La relación de esos aspectos conlleva varios niveles de evaluación, en los artefactos se puede evaluar de diversas formas, ya que, el ámbito que abarcan es amplio, incluye desde visualizaciones, automatización de análisis, procesamiento y tratamiento de datos, entre otras. Por ejemplo, en un nivel básico, se evaluaría la efectividad de la representación o técnica de visualización empleada (lado izquierdo de la figura 4.3), sin embargo, en otros niveles se evaluaría el material que se proporciona con la herramienta (por ejemplo, manual, tutorial, vídeo, etc.), también si es adaptable a diferentes ambientes e incluso el costo de desarrollo. Por otro lado, una parte importante de la evaluación son los usuarios, y estos se dividen en dos grupos, expertos y usuarios comunes; tomando en consideración a los expertos, es difícil hacer una evaluación ya que se debería tomar en cuenta las necesidades y expectativas del experto. De igual forma las tareas que realiza con esos datos son evaluadas en relación a la satisfacción del uso de los artefactos, por ejemplo, los datos podrían provenir de diferentes fuentes y se estaría evaluando la eficiencia de la visualización así como la efectividad de la interacción con la herramienta.

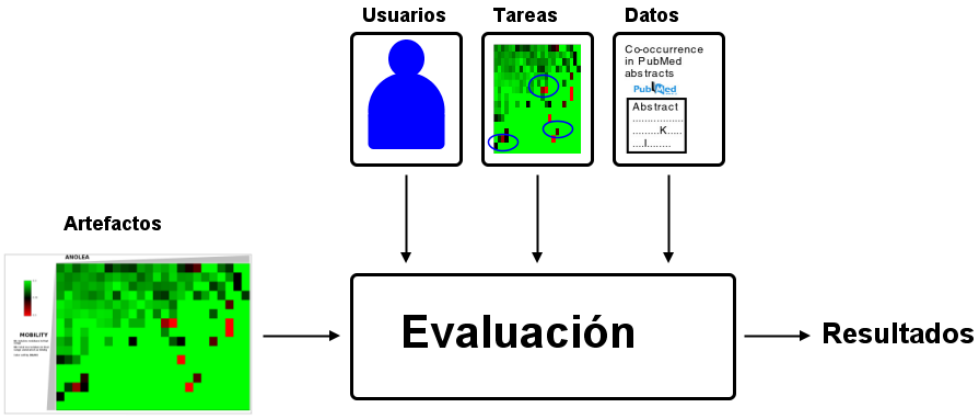


Figura 4.3: Los artefactos abarcan diversas formas como las visualizaciones, técnicas de interacción, proceso de minería de datos, etc., por otro lado, están las tareas, que realiza el usuario, como la selección de algún grupo en los datos, filtrado, resaltar un variable; los datos suelen venir de diferentes fuentes, las cuales pueden contener distinto tipo de información. Posteriormente los resultados se miden en relación a la efectividad, eficiencia y satisfacción de uso (fuente: Mastering the Information Age - Solving problems with Visual Analytics [24]).

Parte III

Estado del Arte

Capítulo 5

Clustering en el Diseño de Fármacos

El principal objetivo del diseño de fármacos basado en la estructura es encontrar compuestos activos frente a una diana biológica de entre la enorme cantidad de compuestos que existen o que pueden existir. Por lo general, los experimentos de cribado virtual se realizan con subconjuntos de compuestos de características definidas por el investigador en función del problema planteado. En una primera aproximación, el proceso es más útil cuando se parte de compuestos que presenten características fisicoquímicas apropiadas (lo que habitualmente se conoce como *drug-like*) ya que la distancia posterior a recorrer hasta un fármaco potencial es menor, pero en ocasiones se prefieren moléculas más pequeñas (fragmentos) que puedan posteriormente ensamblarse y, en otras ocasiones, uno o varios subconjuntos específicos de moléculas. Sea cual sea la situación, los experimentos virtuales generan una enorme cantidad de datos (proporcional al número de compuestos empleados) en los que se incluye de forma más o menos explícita una evaluación virtual de la interacción de cada compuesto con la diana. Una parte importante del proceso de análisis de los resultados implica la agrupación de los compuestos en función de estas interacciones con la diana, que son diferentes de las agrupaciones de los ligandos en función de su estructura (metodología empleada en el diseño de fármacos **NO** basado en la estructura de la diana). Una herramienta útil en este proceso es la minería de datos, ya que a través de algoritmos de clasificación (*clustering*) se extrae la información que es útil. En este capítulo se definirán los métodos y algoritmos de *clustering* que comúnmente se emplean en la minería de datos, posteriormente se hará una revisión aquellos aplicados hasta el momento al diseño de fármacos.

5.1. *Clustering*

La organización de objetos en grupos de afinidad es una de las formas de obtener conocimiento, así como un factor clave en el aprendizaje automatizado. La clasificación de objetos de acuerdo a su grado de similitud es uno de los principales procesos en Minería de datos. El análisis de *cluster* es el estudio formal de algoritmos y métodos de agrupamiento o clasificación de objetos. Los objetos se describen mediante un conjunto de medidas o relaciones entre ellos. En el análisis de *cluster* los objetos no están etiquetados con anterioridad ni se posee información alguna sobre la clasificación de los objetos. El objetivo del análisis de *cluster* es encontrar una organización de los datos conveniente

y válida. El sentido no es establecer reglas para separar los datos en categorías en el futuro sino que los algoritmos de *clustering* están enfocados a encontrar estructuras en los datos. Un *cluster* consta de una colección de un determinado número de objetos similares coleccionados o agrupados en un conjunto. Las definiciones más aceptadas de *cluster* [54] son:

1. Un *cluster* es una concentración de puntos en el espacio, tales que la distancia entre dos puntos dentro del *cluster* es menor que la distancia existente entre un punto del *cluster* y otro punto fuera de éste.
2. Los *clusters* pueden ser interpretados como regiones conectadas de un espacio $n-dimensional$, con una densidad de puntos relativamente alta, separadas de otras regiones, por una región con una densidad de puntos relativamente baja.

Estas definiciones asumen que los objetos que se están agrupando están representados como puntos en el espacio. Aunque no es difícil dar una definición funcional de *cluster*, en cambio es complejo dar una definición operacional, ya que los objetos pueden ser agrupados en *clusters* según el propósito planteado. Por otra parte, la pertenencia de un objeto a un *cluster* puede verse afectada en el tiempo y el número de objetos en los *clusters*. Por lo tanto, el problema crucial en la identificación de *clusters* en los datos es especificar cuál es la proximidad que se empleará y cómo medirla. Como es de esperar la noción de proximidad es dependiente del problema.

Los algoritmos de *clustering* están muy ligados al tipo de dato, por lo cual, si no se comprenden factores como escala, normalización y tipo de medición de proximidad pueden cometerse errores en la interpretación del resultado de un algoritmo. Dado que el análisis de *cluster* es una herramienta para explorar datos, desde sus orígenes se ha usado asociado a algún tipo más o menos complejo de técnica de visualización de datos. La visualización más directa se tiene en el plano donde se muestran los objetos, como puntos proyectados, lo que permite verificar los resultados de los algoritmos de *clustering*, sin embargo, existen otras visualizaciones más desarrolladas que permiten un mejor análisis.

Los algoritmos de *clustering* agrupan objetos (o individuos) en base a índices de proximidad entre pares de objetos. Los objetos (datos en bruto) que serán analizados por un algoritmo de *clustering*, pueden describirse en dos formatos:

1. Matriz de objetos.
2. Matriz de proximidad.

La matriz de objetos está formada por los $n-objetos$ que serán agrupados (filas) y consta de *d-características* (medidas, atributos, puntuaciones), por lo que la matriz tendrá $d-columnas$; es decir, cada objeto es un vector de d componentes y la matriz de objetos es de orden $n \times d$ características. Las *d-características* se representan usualmente como un conjunto ortogonal y los $n-objetos$ son embebidos en un espacio $d-dimensional$.

Un *cluster* es una colección de objetos, los cuales están próximos unos a otros o satisfacen alguna relación especial. La tarea del algoritmo de *clustering* es identificar los *clusters* existentes. Por otra parte, una matriz de proximidad $D = [d(i, j)]$ acumula índices de proximidad entre pares de objetos, donde cada fila y columna representa a un objeto. Se asume que la matriz de proximidad es simétrica

dado que todo par de objetos tienen el mismo índice de proximidad independientemente de su orden. La diagonal principal de esta matriz es ignorada dado que cada objeto tiene el mismo índice de proximidad consigo mismo. Los índices de proximidad se definen como índices de similaridad o índices de disimilaridad (por ejemplo, coeficiente de correlación o distancia euclídea, respectivamente).

Normalización: preparar los datos para un análisis de *cluster* requiere algún tipo de normalización que tenga en cuenta las medidas de proximidad, un ejemplo de ello es la distancia Euclídea. Es una métrica que asigna mayor peso a las características con valores grandes que a las características con valores pequeños, es decir, las características con mayor magnitud dominarán a las características con menor magnitud. Existen muchos esquemas de normalización (como los propuestos por Jain y Dubes [54], Berrar *et al.* [8], Geoffrey *et al.* [38], etc.).

5.2. Métodos y Algoritmos de *Clustering*

El análisis de *cluster* es el proceso de clasificación de objetos en subconjuntos, de tal forma que los agrupamientos hechos (subconjuntos) tengan sentido en el contexto de un problema en particular [54] [55]. Un *clustering* es un tipo de clasificación sobre un conjunto finito de datos, siendo la matriz de proximidad la única entrada al algoritmo de *clustering*. Los dos tipos de métodos más empleados son los jerárquicos y los particionales.

5.2.1. *Clustering* Jerárquico

Un método de *clustering* jerárquico es un procedimiento de transformación de la matriz de proximidad en una sucesión de particiones anidadas. Un algoritmo de *clustering* jerárquico es la especificación de los pasos para llevar a cabo un *clustering* en jerarquías. Como *clustering* jerárquicos se pueden mencionar el método aglomerativo y el método divisivo, se diferencian en que el primero inicia la clasificación tomando a cada objeto como un *cluster* y gradualmente comienza a unirlos (lo que se conoce como *clusters* atómicos) para formar nuevos *clusters*, se repite este proceso hasta que todos los objetos estén en solo un *cluster*. El método divisible es lo opuesto al método aglomerativo, este comienza con un único *cluster* que contiene a todos los objetos y gradualmente comienza a subdividirlo en piezas más pequeñas hasta que cada objeto forme un *cluster* atómico.

Para visualizar la información de un *clustering* jerárquico se utiliza una estructura especial de árbol que proporciona una conveniente representación. A esta forma de visualizar la jerarquía se le llama dendrograma o árbol, el cual consiste en capas de nodos, donde cada nodo representa un *cluster* (tal como muestra la figura 5.1). Las líneas conectan a los nodos representando *clusters* anidados unos con otros. Cortando el dendrograma horizontalmente obtenemos un *clustering*. El nivel o el valor de proximidad donde los *clusters* se forman pueden ser mostrados también en el dendrograma.

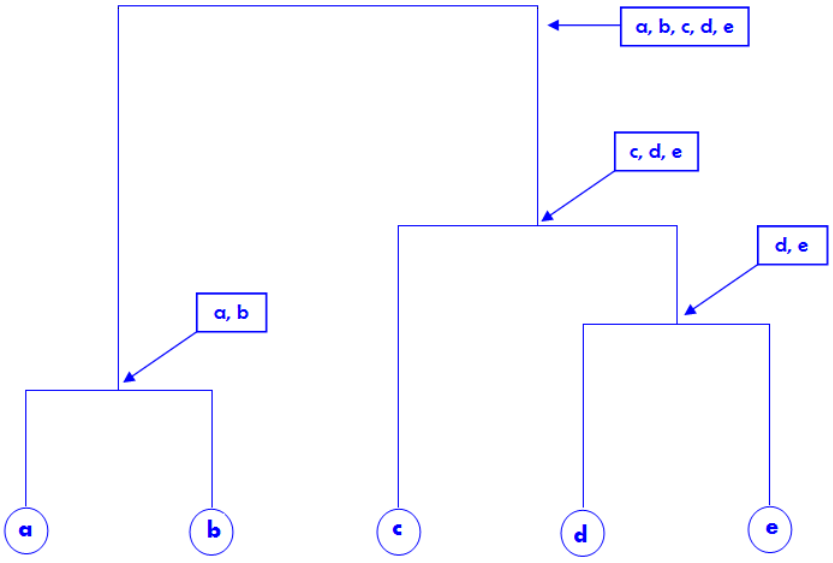


Figura 5.1: Estructura jerárquica en forma de dendrograma.

5.2.2. Algoritmo de Distancia Mínima y Algoritmo de Distancia Máxima

Existen muchos algoritmos aglomerativos, pero según el criterio utilizado para la distancia entre dos *clusters* es lo que los diferenciará. Los criterios más usados son: distancia mínima (*single-linkage*), distancia promedio (*average-linkage*) y distancia máxima (*complete-linkage*) entre dos *clusters*. El algoritmo *single-linkage* se refiere a la menor distancia existente entre los objetos de dos *clusters*, por otro lado, *complete linkage* se refiere a la mayor distancia entre dos objetos de dos *clusters*. Ambos algoritmos asumen que no existen distancias repetidas en la matriz de disimilaridad y producen una sucesión anidada que devuelve una jerarquía única de *clusterings* que posteriormente puede ser mostrada a través de un dendrograma.

5.2.3. *Clustering* Particional

El método particional construye k *clusters*. Esto es, clasifica un conjunto de datos n en k grupos que en conjunto satisfacen los requisitos de una partición: cada grupo debe contener al menos un objeto, cada objeto debe pertenecer exactamente a un grupo. Estas condiciones implican que hay por lo menos tantos grupos como objetos: $k \neq n$.

Es importante aclarar que el factor k lo debe proporcionar el usuario. Por lo tanto, el algoritmo construirá una partición con el número de *clusters* deseados. Cabe mencionar, que no todos los valores de k tienden a formar grupos naturales, por lo que es necesario ejecutar el algoritmo varias veces con diferentes valores de k y después seleccionar el valor de k con el que se hayan obtenido los mejores resultados de acuerdo a sus características o que gráficamente se vea mejor o con el que se puedan obtener mejores conclusiones. También es factible que esta decisión se tome automáticamente, esto es, que el ordenador pruebe todos (o la mayoría) los posibles valores de k y seleccione

el que mejor se adapte a un criterio numérico.

En términos generales el algoritmo tratará de encontrar una buena partición en el sentido de que los objetos del mismo *cluster* deben estar cerca o relacionados unos con otros, además de que los objetos de *clusters* diferentes deben estar lejos o ser muy diferentes. El objetivo es desvelar una estructura que esté presente en los datos; sin embargo, el algoritmo se utiliza para imponer una nueva estructura. El algoritmo particional más conocido es *k-means*.

5.2.4. *K-Means*

El algoritmo *k-means* es el más empleado y simple. Se comienza con una partición inicial aleatoria y continúa reasignando los objetos a agrupar, basándose en la similaridad entre el objeto y el centro del *cluster* hasta que se alcance un criterio de convergencia. La popularidad del algoritmo *k-means* se debe a que es sencillo de implementar, además de que su complejidad es $O(n)$, donde n es el número de objetos. Una desventaja importante de este algoritmo es que es sensible a la selección inicial de partición y puede converger a un mínimo local del criterio de convergencia si la partición inicial no se elige correctamente [55].

5.3. *Clustering* y el Diseño de Fármacos

El gran volumen de compuestos que potencialmente pueden convertirse en fármacos y el énfasis en automatizar su clasificación, ha supuesto la creación de algoritmos para el descubrimiento de líderes (gracias a la capacidad de estos algoritmos para comparar estructuras químicas), de forma que diversos algoritmos de *clustering* se han venido proponiendo a un buen ritmo. Estas herramientas son útiles y se emplean en la organización de conjuntos de estructuras, en la visualización del contenido de bases de datos o para seleccionar compuestos líderes [123]. Sin embargo, aún existe incertidumbre en cuanto cuáles de las posibles combinaciones de algoritmos de *clustering*, mediciones, tipos de representación de datos, y reglas para determinar el número de clusters apropiados (por ejemplo, la selección del nivel en el caso de los algoritmos jerárquicos) son mejores o más robustos con respecto a los datos [76].

Por otro lado, existe el problema de cómo expresar un objeto irregular (como es el caso de una estructura química) en una forma regular que permita comparar y contrastar con otras estructuras [74]. Para esto, se emplean diferentes descriptores moleculares, como las propiedades moleculares: hidrofobicidad, polaridad, flexibilidad, forma geométrica, volumen, propiedades de enlaces de hidrógeno, entre otras.

En la literatura se pueden encontrar ejemplos de uso de algoritmos de clasificación en el diseño de nuevos fármacos tanto jerárquicos [80] [92] [3] [112] [35] [76] [74] como particionales [73] [120] [48]. Por otro lado, también se puede encontrar algunos tipos de algoritmos menos extendidos en este campo como: híbrido [123], estocástico [102], algoritmos genéticos [12], etc. Estos últimos se mencionan como métodos alternativos, aunque están fuera del ámbito de este trabajo de tesis. En el caso de los algoritmos jerárquicos podemos afirmar que son los más utilizados y que, en general, muestran mejores resultados en moléculas pequeñas (como es el caso en este trabajo de tesis).

Por ejemplo, en los algoritmos jerárquicos podemos mencionar a: Paris *et al.* [92], que emplean

un algoritmo jerárquico para encontrar inhibidores de VIH-1, en el cual utilizan dos métodos (*single-linkage* y *complete linkage*). Aplicando esos métodos de agrupación a 99 estructuras del sitio de unión de la proteasa del VIH-1 pudieron obtener información acerca de las características de las conformaciones asociadas a la unión de inhibidores no nucleosídicos, permitiendo una exploración más completa de la energía libre de unión. Por un lado, la tendencia del método *single-linkage* forma *clusters* bien conectados (distancia mínima) mientras el método *complete-linkage* forma *clusters* más compactos (distancia máxima). Esto, los llevó a concluir que las agrupaciones son intrínsecas a los datos y no al método. Por otro lado, del análisis que hicieron de los *clusters*, los condujo a suponer: (i) los *clusters* con mayor número de elementos sugieren que las conformaciones de la proteína son localmente flexibles, pero debido a que el conjunto de proteínas estudiado no es un conjunto aleatorio, sino que los inhibidores han sido diseñados en base a inhibidores anteriores o de estructuras de proteínas previas a las actuales, su número no representa la energía libre relativa de cada *cluster*, y, (ii) los *clusters* pequeños, son una oportunidad para la exploración o el desarrollo de inhibidores potentes que se unen a conformaciones alternativas del sitio de unión.

Por otro lado, Amaro *et al.* [3], exploran la flexibilidad del receptor al analizar la interacción entre moléculas ligando y el receptor (diana). Parten de una gran cantidad de estructuras (proteínas) generadas por distintos métodos, y los ligandos los obtienen de bases de datos públicas y comerciales. El problema al que se enfrentan es la cantidad de estructuras, para solucionarlo, dentro de todo su proceso aplican un algoritmo jerárquico para filtrar, lo que repercutirá en una reducción tiempo del análisis general. En términos generales, lo que realizan, es un alineamiento de las moléculas y posteriormente encuentran la similitud entre ellas aplicando un algoritmo jerárquico. De esta manera, centran todo su esfuerzo en analizar aquellos que sean de mayor interés del químico. Además de la mejora del tiempo, el clasificarlas les proporcionó información valiosa de la flexibilidad del receptor tanto local como globalmente; por ejemplo, el número de *clusters* de estructuras (diana) contra el número de ligandos que caen en esos *clusters* -en nuestro caso, nos interesa saber el número de ligandos que cubren ciertas zonas de la diana-.

Por su parte, Meslamani *et al.* [80], aplican un algoritmo jerárquico para abarcar todo el espacio conformacional. Posteriormente, llevan a cabo un filtrado de conformaciones redundantes seleccionando un nivel de la jerarquía y extrayendo un representante de cada *cluster*. Como se ha expuesto anteriormente, es complicado saber el nivel óptimo que contendrá los mejores *clusters*, por lo que ellos se basan en el criterio de corte llamado *clustering gain* (1):

$$G = \sum_{j=1}^{j=K} (n_j - 1)\, d^2\,(g_j, g) \quad (1)$$

Donde, n_j, es el tamaño del *cluster* j, y $d(g_j, g)$, es la distancia del punto medio (*cluster*) al punto medio de todo el conjunto de n conformaciones. Este tipo de aproximación de corte nos es interesante, ya que en nuestro problema nos enfrentaremos a información redundante que es imprescindible eliminar, por lo que consideraremos esta propuesto como otras.

Shao *et al.* [112] aplican diferentes tipos de algoritmos de *clustering* para clasificar trayectorias de dinámica molecular -la dinámica molecular estudia el movimiento de átomos y moléculas, se utiliza para hacer simulaciones tomando en cuenta la flexibilidad de las estructuras [1]- de varios sistemas biomoleculares, entre sus resultados el método aglomerativo mostró mejor rendimiento junto con los mapas auto-organizados (*SOM*). Por otra parte, ellos recomiendas el empleo de algoritmos jerárquicos cuando no se conozca a priori el número de grupos existentes en los datos -como en nuestro caso-. Llegan a la conclusión que el método de *single-linkage* (distancia mínima) no es ca-

paz de producir grupos con significado relevante a la clasificación de las trayectorias, en términos generales, argumentan que realiza una delimitación pobre, cuando los elementos están muy cerca, sin embargo, el método de *complete-linkage* y *average-linkage* si son capaces de mostrar grupos con mayor signigicado al tener *clusters* de tamaños diferentes; por lo cual, ellos lo emplean con frecuencia para sus análisis de trayectorias.

Por otro lado, están los particionales, Li [73] propone el algoritmo *clustering* particional cd-hit-fp -se basa en el algoritmo *single-pass*, a diferencia de *k-means*, encuentra los *clusters* en una sola iteración- para eliminar la redundancia en las bibliotecas virtuales de compuestos y así seleccionar los mejores compuestos para realizar experimentos de HTS. Las ventajas de aplicar este tipo de algoritmo de *clustering*, radica en que puede ser útil como paso previo para determinar el número de particiones, en otras palabras sirve como dato de entrada del *k-means*. Por otro lado, una de las principales desventajas que presenta, es que los resultados son muy dependientes del valor que se determina para detener el algoritmo.

En general, podemos nombrar las ventajas de los dos métodos, una ventaja de los algoritmos jerárquico aglomerativo es que son rápidos para calcular e implementar, encuentran todos los posibles grupos existentes, por otra, su desventaja más sobresaliente es la rigidez de la estructura de árbol: no se puede deshacer lo que ya agrupo. Los métodos particionales tienen la ventaja de ser eficientes para encontrar grupos, pueden manejar grandes cantidades de datos eficientemente, pero sus desventajas son muy remarcadas: necesitan un valor inicial del número de grupos a encontrar -probablemente la peor desventaja de todas-, débil en datos con ruido y la necesidad de iterar varias veces hasta encontrar un número optimo de particiones.

A modo de conclusión de este capítulo, se desea constatar que el estado del arte en algoritmos de clasificación aplicados al diseño de fármacos refleja una preponderancia de los algoritmos jerárquicos. Por otra parte, al aplicar este tipo de algoritmos nos proporcionarán soluciones afines al contexto de nuestro problema en: al reducir los datos a grupos que representen las diferentes zonas del espacio conformacional, conocer ese número de grupos [80] -seleccionando el mejor nivel de la jerarquía-, porque se desconoce, y, de los grupos formados analizar las zonas que cubren del espacio conformacional [3].

Capítulo 6

Visualización de Información en el Diseño de Fármacos

Como se mencionó en el capítulo 3, la visualización de información es una herramienta eficaz en el análisis de datos, sobre todo cuando se necesita explorar datos masivos. En este capítulo se procede a hacer una revisión de las técnicas de visualización en relación al diseño de fármacos. La primera sección se centra en los tipos de visualización de datos jerárquicos -como se expuso en el capítulo anterior, nuestro análisis se basa en la agrupación en base a jerarquías de moléculas ligando-. Posteriormente, se menciona cómo la visualización de información es una herramienta útil en el ámbito más amplio de la bioinformática. Se mostrará la valiosa aportación que hace cada una de las técnicas de visualización de información que se han implementado en distintas herramientas, sin dejar a un lado la experiencia del experto, que se incorpora a medida que va interactuando con las visualizaciones en conjunto o por separado. Sin embargo, existe la necesidad de incorporar una visualización complementaria, que se coordine e interaccione con las demás vistas, la cual ha de mostrar lo que sucede a una escala molecular, para ellos en la última sección se revisan diferentes herramientas de visualización molecular para comprender este fenómeno.

6.1. Visualización de Datos Jerárquicos

Una vez que hemos agrupado nuestros datos -de manera jerárquica-, nos enfrentamos a un problema de exploración de datos, para ellos es necesario contar con una representación que nos muestre la toda la jerarquía que se construyó. La visualización de esta jerarquía es importante en el análisis de *clusters*, ya que permite la toma de decisiones durante el proceso de análisis de los datos. En otras palabras, es importante porque desvela todos los grupos intrínseco de los datos, y, la navegación de la jerarquía permite al usuario seleccionar el nivel que satisfaga su criterio. Esta estructura jerárquica de *clusters* por lo general se visualiza a través de un dendrograma [143]. El principal objetivo de este tipo de visualización es encontrar subconjuntos de datos que estén bien definidos y que sean diferentes del resto de grupos [109]. Uno de los muchos ejemplos de herramientas que emplean dendrogramas es el popular explorador de clusters jerárquicos (*Hierarchical Clustering Explorer* HCE) [110] cuyo fuerte reside en visualizar árboles enormes de datos biológicos (*microarrays*) -Está completamente

orientado a analizar datos de bioinformática-; para abordar este reto, a medida que se navega (sube o baja de nivel) mediante una barra que señala el nivel actual en el dendrograma, las ramas se contraen o expanden, esto es un filtrado de datos aplicando la técnica de foco+contexto, de esta manera evitan el amontonamiento. Por otro lado, emplean vista enlazadas interactivas como: diagramas de dispersión, comparación de dendrogramas por diferentes algoritmos de *clustering* jerárquico. En los diagramas de dispersión es posible seleccionar un elemento el cual se reflejará directamente al mostrar pequeños triángulos en la parte baja del *microarray*. Una parte interesante de HCE, sin duda es, la comparación de los resultados de distintos algoritmos jerárquicos, al seleccionar un *cluster* con el ratón muestra donde están localizados esos mismos datos en el otro dendrograma.

Otra herramienta de visualización de datos biológicos es GAP [140]; la visualización principal está orientada al análisis de matrices y análisis de *clusters* para *microarrays*; emplean diversas técnicas de agrupamiento jerárquico y dendrogramas como visualización de la jerarquía, tanto de las columnas como de las filas de la matriz. Para explorar los *clusters* de cada nivel de manera interactiva, utilizan una barra deslizadora al igual que el HCE, a cada *cluster* se le asigna un color para resaltarlos. Dos características de interacción hacen resaltar a esta herramienta: la selección de ramas del dendrograma empleando el ratón y el intercambio de ramas al reordenar la matriz sin tener que construir de nuevo el dendrograma.

Otra herramienta que incorpora el uso de dendrogramas es gCLUTO [98]. La visualización es interactiva al aplicar una técnica de zoom semántico -el zoom semántico consiste en ir cambiando el contenido de la información que se muestra mientras se muestran más detalles a medida que se acerca a un área en particular de la gráfica [45]- contrayendo o expandiendo las ramas del dendrogram. En otras palabras, al contraer la rama, muestra de manera resumida en una línea toda la información de la rama, este tipo de técnicas se emplean cuando hay una gran cantidad de datos. Otra técnica que implementaron es el foco+contexto, al seleccionar ciertas zonas, muestran de manera ampliada esas ramas, mientras las otras se contraen por la técnica del zoom semántico. Lo que permite centrarse en *clusters* de interés.

Hasta este punto, hemos descrito cómo las técnicas de visualización de información ayudan a comprender datos inherentemente jerárquicos en matrices de datos biológicos (*microarrays*). Por otro lado, se sabe que la visualización de información ha sido aplicada anteriormente en el tratamiento de enfermedades [22], tales como el procesamiento de imágenes, la presentación de informes médicos de pacientes, interpretación de diagnósticos, etc.

En nuestra situación, necesitamos analizar una gran cantidad de compuestos químicos, para poder encontrar aquellos con gran potencial para convertirse en fármacos y su posterior uso en el tratamiento de alguna enfermedad. Una vez que se han encontrado aquellos compuestos que puedan ser líderes a fármaco, habrá que refinarlos, por lo que la optimización de esos compuestos líderes determinará el éxito de un fármaco. Durante esta optimización, se mejoran aspectos como la actividad farmacológica, la farmacocinética y la seguridad de la utilización del fármaco (que no sea tóxico). Dada la gran diversidad de compuestos que se derivan de un compuesto líder (lo cual no supone necesariamente una ventaja durante el diseño de fármacos) es necesaria la selección del mejor compuesto que presente la mayor calidad en los aspectos mencionados anteriormente, los cuales determinarán el éxito en el descubrimiento y posterior desarrollo del fármaco [141].

El llevar a cabo la extracción y descubrimiento de estos compuestos líderes directamente de las fuentes de información en bruto es complicado, por eso, la información debe pasar de una información multidimensional a una de pocas dimensiones -la aplicación de algoritmos de *clustering*

jerárquico, ayudan a esta transición- para que esos compuestos resalten sobre los demás. Un caso en particular es el de Maniyar *et al.* [77], que ven en la visualización información el método para proyectar la información multidimensional en un espacio de pocas dimensiones pero manteniendo la mayor cantidad de propiedades (en este caso la información de los compuestos), y, sobretodo no perder la sencillez de interpretación de los datos, para extraer la mayor cantidad de información útil (compuestos líderes). Proponen una herramienta de visualización de compuestos químicos llamada *Hierarchical GTM* (HGTM). Esta visualización aplica diferentes técnicas de reducción de dimensionalidad, tales como el análisis de componentes principales, redes neuronales, mapas auto-organizados y generación topográfica de mapas. A partir de esta visualización, el químico puede obtener información de los compuestos que presenten propiedades similares, a su vez, estos *clusters* pueden extraerse y formar sub-ramas formando una jerarquía -es una especie de barrido de los datos-. Como soporte del análisis de *clusters*, las propiedades de interés son representadas por medio de coordenadas paralelas. En nuestro caso, también partiremos de una gran cantidad de compuestos que necesitamos agrupar según ciertos criterios, aun así, seguiremos teniendo una gran cantidad de datos que al representarlos, probablemente cubran gran parte de la visualización, por lo que es necesario emplear una extracción de los *clusters* que sean de interés, sin perder el resto del contexto de los datos.

Podemos concluir esta sección afirmando que hoy en día la visualización de información es una herramienta imprescindible para explorar datos masivos a los cuales se les ha aplicado un método de *clustering*. Por otra parte, estos métodos son aplicados al diseño de fármacos en la optimización de compuestos que en un futuro puedan convertirse en un fármaco. Por otro lado, aunque no es el objetivo de esta tesis, la visualización de información ha encontrado en la bioinformática un campo fértil para explotar sus logros, al ayudar en el estudio de la estructura y funcionamiento de componentes celulares, tales como moléculas grandes (proteínas). Estos problemas de bioinformática están emparentados muy cercanamente con nuestro objetivo, y comparten retos en lo tocante a la visualización de información, como es el ayudar a comprender la estructura tridimensional y el comportamiento dinámico de moléculas. Hasta este momento, hemos planteado una estrategia de visualización de reducción de dimensionalidad en base a los grupos que se forman de los datos, sin embargo, el proceso en el diseño de fármacos necesita analizar la interacción que se da entre la diana y el ligando, en otras palabras, no basta con conocer los ligandos, el químico necesita ver en donde se han situado en la diana esos compuestos, de igual manera la forma geométrica de cada compuestos que pertenece a un grupo. Por estos motivos, se hace necesario enlazar la visualización abstracta (dendrogramas) con una representación real de la diana y el ligando. Existen muchas herramientas especializadas en la visualización molecular que apoyan la investigación de las estructuras moleculares y el diseño de fármacos [11]. En la sección 6.3 se discuten algunos de estos enfoques.

6.2. Visualización de Información en Bioinformática

El propósito de utilizar la visualización de información en bioinformática recae en los objetivos de las técnicas de visualización de información: visualizar cantidades grandes de información y facilitar el análisis de información por medio del reconocimiento de patrones generados por la minería de datos. La bioinformática es conocida por generar mucha información biológica.

Los sistemas de visualización de información en 3D, construyen una imagen en 3D que los expertos (y los no expertos también) pueden comprender fácilmente [126]. La estructura de una

molécula se genera mediante coordenadas en el espacio (cada átomo tiene una coordenada en tres dimensiones (x,y,z)). Tratar de comprender esta estructura espacial únicamente a través de los números carece de sentido, pues hacerse una imagen mental es demasiado complicado incluso para un experto (considérese el caso de una proteína); por eso, al traducir tales coordenadas, a través de una herramienta informática a una imagen, es más sencillo comprenderla (otras ayudas visuales se pueden utilizar también, como asignar un color único a cada tipo de átomo). Además, a través de interacción se puede habilitar la exploración de ciertas zonas especificas, acercándose o alejándose según se requiera.

Un ejemplo de visualización de información en la bioinformática es el análisis de los resultados de *docking*. Se parte del alineamiento de diferentes compuestos, por lo general es un proceso automatizado, pero suele generar mucho ruido en los datos y no siempre es preciso. Por lo tanto, el usuario debe realizar una inspección de los resultados antes de tomar una decisión, lo cual hace necesario implementar algún tipo de visualización, ya que el objetivo es obtener el mayor conocimiento posible de la información generada, y más si se toma en cuenta que los resultados son datos numéricos y con muchas posibles interpretaciones. Por consiguiente, en nuestro problema, se aplicarán técnicas de minería de datos, principalmente el *clustering* jerárquico, en conjunto con técnicas de visualización como: dendrogramas, barrido y foco+contexto [131].

6.3. Herramientas de Visualización Molecular

Las herramientas de visualización de moléculas que hay disponibles son muchas, por lo que continuación se explicarán algunas de la tabla 6.1; no obstante si se desea profundizar más en dichas herramientas, Gu y Bourne [40] exponen una lista más completa.

PyMol: es una herramienta desarrollada en Python, que permite la visualización de moléculas tanto pequeñas (ligandos) como macromoléculas (proteínas). Crea la superficie de la proteína, hace cálculos de puentes de hidrogeno, además de mostrar la secuencia de la proteína, incluso permite descargar ficheros de macromoléculas de la páginas como PDB. Una función que hace resaltar esta herramienta es la posibilidad de analizar los resultados de *docking* en especial los de AutoDock. Sin embargo, no es tan eficiente, en ocasiones al tratar de manipular los resultados, se pierde la interacción entre la ventana que muestra los resultados y la visualización molecular, esto puede ser debido a que no es parte nativa de la herramienta por ser un complemento (plug-in).

Chimera: al igual que PyMol, está desarrollada en Python, lo que resalta de esta herramienta, es que puede ser empleada para preparar tanto la proteína como el ligando para realizar el proceso de *docking*. Esto consiste en añadir hidrógenos y cargas a las moléculas. Además de comparar macromoléculas mediante alineamiento, permite buscar aminoácidos de manera más sencilla y rápida que PyMol.

MOLMOL: es un programa de representación molecular 3D para el análisis y manipulación de macromoléculas biológicas, enfocándose en el estudio de estructuras de ADN, proteínas y ácidos nucleicos de datos provenientes de NMR (Resonancia Magnética Nuclear). Las características principales de esta herramienta son: la superposición de 20 conformaciones distintas, útil para la determinación por NMR, identificación de puentes de hidrógeno, identificación de distancias entre átomos de hidrógeno, entre otras.

VMD: es un visualizador molecular que antiguamente sólo podía correr en estaciones de trabajo,

sin embargo, ha sido portado a la mayoría de las plataformas actuales. Es un visualizador potente, pero con el detrimento de que es poco amigable para el usuario. Por otro lado, una vez que se ha dominado la curva de aprendizaje -por lo general es muy pronunciada- es posible visualizar simulaciones de dinámica molecular.

AutoDockTools: o ADT, es una herramienta de los mismo desarrolladores de AutoDock, con la cual se pretende analizar y extraer conocimiento de los resultados de *docking*, además de preparar todo lo necesario de manera visual para realizar *docking*. En cuanto a la visualización de los resultados de *docking* ofrecen una visualización de grupos muy rudimentaria y casi carente de interacción con la visualización molecular.

KiNG: es un visualizador molecular 3D que permite la manipulación y análisis de moléculas, está basado en Java al igual que Jmol, y se especializa en el formato de imágenes que genera, sobresaliendo por su alta calidad. Sin embargo, hace uso de una librería especial llamada JOGL que es un puente entre Java y OpenGL -librería de gráficos 3D-.

Si bien, son muchas las herramientas de visualización molecular, Jmol es una de las que destaca. Este visualizador consta de tres partes, que a su vez representan ventajas sobre otros visualizadores; estas partes consisten en: Jmol-Applet, un *applet* que se integra en un navegador web; como programa de escritorio; y JmolViewer, un *toolkit* que se integra en otros programas escritos en Java [138]. Por otra parte, como está escrito completamente en Java, hace que sea multiplataforma (puede ejecutarse en diferentes sistemas operativos gracias a su máquina virtual). Jmol puede visualizar la mayoría de los formatos de estructuras químicas, incluso puede leer ficheros comprimidos cuando se usa en modo web, aparte de que tiene compatibilidad con comandos de RasMol y Chime -son de los primeros visualizadores moleculares en 3D, en especial Chime era empleado en páginas web, sin embargo, no se han producido nuevas actualizaciones desde su lanzamiento, por lo que Jmol se concibió como el reemplazo de ambos-, además de que son herramientas conocidas por los miembros del laboratorio de química farmacéutica de la Universidad de Salamanca.

La ventaja más evidente de Jmol está en su motor de gráficos. El motor está completamente escrito en Java (no utiliza Java3D ni OpenGL o ningún otro tipo de acelerador de gráficos), su rendimiento es significativo. Es sumamente eficiente para visualizar ficheros que contienen una gran cantidad de moléculas o macromoléculas, porque está dedicado a la representación molecular (esferas y cilindros), por lo que no emplea ninguna maya triangular para hacer el renderizado de moléculas. El empleo de Jmol en la comunidad científica y educativa es grande; ejemplos del uso de Jmol se pueden encontrar en [136] [41] [75] [96] [79] [29] [138] [28] [46] [16] [124].

Programa	Sistema Operativo
Jmol[1]	**Mac, Win y Linux**
Chime[2]	**Mac y Win**
PyMol[3]	**Mac, Win, Linux y Unix**
MOLMOL[4]	**Win y Unix**
RasMoL[5]	**Mac, Win, Linux y Unix**
VMD[6]	**Linux y Unix**
AutoDockTools[7]	**Mac, Win, Linux y Unix**
Chimera[8]	**Win, Linux y Unix**
KiNG[9] **[21]**	**Mac, Win y Linux**

Tabla: 6.1: Programas de visualización molecular.

Jmol Visualizador Molecular

Hanson explica en [41] el empleo de Jmol como principal visualizador de cristalografía y generador de imágenes de calidad de esas estructuras [79], destacando los beneficios de la funcionalidad del visualizador para realizar diferentes tareas de manipulación de estructuras. Por otro lado, menciona que es empleado en la educación al ser mencionado en simposio de American Chemical Society. Un ejemplo del uso de Jmol en la educación e industria es el del Chemistry Development Kit (CDK), destacándose como visualizador 3D de estructuras Steinbeck [124], así como otros los podemos encontrar en [46], [16], entre otros.

Una propuesta interesante la encontramos en la tesis de master de Logtenberg [75], implementa Jmol dentro de un ambiente multitouch, resaltando que la tendencia de la educación se orienta a la interacción entre el ordenador-persona y el multitouch da una sensación de mayor interacción que el empleo de teclados y ratones. Por otra parte, Forlines *et al.* [29], extienden el ambiente multitouch a uno de pantallas múltiples, en donde el usuario interactúa con la visualización en varias pantallas, de esta manera, tiene varias perspectivas de la misma información, y, aunado al ambiente multitouch puede rotar en varias diferentes direcciones cada perspectiva de la visualización. Por otro lado, es utilizado ampliamente por páginas web como PDB[10], The Protein Model Portal[11], ZINC[12], entre otras [28]. La ventaja de incluir en una página web a Jmol es que no necesita de alguna librería en especial para generar sus gráficos en 3D [138], como sería el caso de OpenGL o DirectX.

No obstante, Jmol tiene sus detractores, como: White ***et al.*** [136] y Nicolay ***et al.*** [96] . En el caso del primero, compararon la visualización de Jmol contra una herramienta de simulación. Por

[1] http://jmol.sourceforge.net/index.en.html
[2] http://www.umass.edu/microbio/chime/getchime.htm
[3] http://www.pymol.org/
[4] http://www.cecalc.ula.ve/BIOINFO/servicios/herr3/MOLMOL/molmol.php
[5] http://www.umass.edu/microbio/rasmol/
[6] http://www.ks.uiuc.edu/Research/vmd/
[7] http://autodock.scripps.edu/resources/adt
[8] http://www.cgl.ucsf.edu/chimera/
[9] http://kinemage.biochem.duke.edu/
[10] http://www.rcsb.org/
[11] http://www.proteinmodelportal.org/
[12] http://www.zinc.docking.org/

una parte, la simulación muestro que los usuarios tenían una comprensión de las estructuras de las proteínas, sin embargo, el estudio mostró que mediante el empleo de la visualización molecular -con Jmol- los usuarios identificaban más átomos que utilizando la simulación. Por su parte, Nicolay ***et al.*** argumentan que Jmol no es eficiente para la representación de superficie de las proteínas, como resaltar los sitios de actividad, sin embargo, esto puede solucionarse mediante el empleo de scripts para solucionar este tipo de deficiencia.

De este capítulo podemos resaltar que la unión de estas dos herramientas, minería y visualización de datos, dan como resultado un mejor análisis del problema en cuestión. Esto da paso a realizar un discurso analítico del diseño de fármacos a través de la analítica visual, el cual será ampliado en el siguiente capítulo.

Capítulo 7

Analítica Visual en el Diseño de Fármacos

Como se mencionó en el capítulo 4, el objetivo de la analítica visual a través de un discurso analítico es aprovechar la potencia de los ordenadores y los del usuario para mejorar el proceso de análisis de un problema (Keim *et al.* [60]). El uso de algoritmos de clasificación permite encontrar patrones que se encuentran en los datos, mientras que el empleo de visualizaciones adecuadas ayuda a organizar y presentar estos patrones de tal manera que el usuario tenga una visión general del problema en cuestión. A través de la experiencia del usuario en el problema a analizar, el ordenador refinará y organizará la información más apropiadamente. A continuación se revisan los pocos ejemplos en que el enfoque de analítica visual ha sido aplicado al problema del diseño de fármacos.

Saffer *et al.* [106] aplican la analítica visual como método de descubrimiento de posibles dianas biológicas y sus respectivos ligandos. Este es un claro ejemplo de cómo aplican la analítica visual para comprender datos de diferentes ámbitos, en este caso particular la biología y la química. En una primera aproximación realizan el alineamiento de varias secuencias para determinar el grado de identidad. Posteriormente, realizan cálculos de similaridad de las proteínas que son visualizadas mediante un programa llamado OmniViz[1]. Una de las visualizaciones consiste en proyectar a través de un mapa la proximidad de cada proteína, de esta manera, de acuerdo a la proximidad de cada punto -proteína- es posible mostrar las moléculas que son parecidas, lo que significaría que tendrán dianas parecidas, esta visualización la llamaron OmniViz Galaxy. Por otra parte, también es posible determinar la similaridad de compuestos -ligandos- en base a su actividad biológica, cada compuesto es representado en la visualización (OmniViz Comet) como una celda asignándole un color (rojo presencia alta, azul poca presencia), al final se forma una matriz de color (*heat map*), y para encontrar los grupos dentro de la matriz, aplican un algoritmo jerárquico y su respectivo dendrograma para explorarlos. Al combinar las visualizaciones y enlazándolas para que tengan interacción, el químico podrá encontrar nuevas dianas o ligando; bastara con seleccionar grupos en cualquiera de las dos visualizaciones.

Ray [99] emplea mediciones de correlación a través de una visualización de coordenadas paralelas [52] modificadas, para mostrar la correlación entre secuencias, cuando no se tiene la información

[1] http://www.biowisdom.com/tag/omniviz/

estructural de la secuencia. Las coordenadas paralelas son un tipo de visualización multidimensional, en el que cada objeto es representado por una polilínea, utilizando $n-copias$ del $eje-y$ que representan $n-dimensiones$ en el espacio. Al emplear las coordenadas paralelas para visualizar la correlación de requeriría que por cada posición en la secuencia existiera un eje que lo representara, lo cual generaría un amontonamiento de los datos y por tanto sería imposible extraer información útil del análisis. Por lo que aplican una reducción de dimensiones para tener las que son más relevantes. La modificación que realizan sobre las coordenadas paralelas es que pasan de una representación vertical a una circular, de esta manera, cada línea se conecta con uno o varios puntos, de esta manera resaltan a los aminoácidos que tienen mayor número de interacciones; por lo que asegura que es más representativo que emplear las coordenadas paralelas tradicionales. En una etapa del desarrollo de nuestra investigación, se diseño e implemento a las coordenadas paralelas como visualización principal, sin embargo, debido a la cantidad de datos nos creaba el problema del amontonamiento, por lo que a primera instancia era imposible deducir un camino hacia donde explorar en profundidad.

Por otra parte, MassVis [68] a través de una visualización de diagrama de dispersión, paneles de control enlazados, como el zoom, selección de elementos individuales con el ratón, y, técnicas de minería de datos, como la correlación y la clasificación (*clustering* aglomerativo) realizan análisis de interacción proteína-proteína. Existen otros métodos que se apoyan en el análisis visual, por ejemplo, Stone *et al.* [125] y ODonoghue *et al.* [90], lo emplean para extraer conocimiento las representaciones tridimensionales de las estructuras. En el primer caso, emplean el análisis para el modelado molecular en base a realidad virtual en conjunto con el visualizador VMD [51], argumentan que la interacción entre visualizaciones en 2D y 3D ofrecen al químico un mejor entendimiento al centrarse en zonas especificas. En el segundo caso, mediante diversas técnicas tanto interactivas como estadísticas, logran obtener información relevante de las visualizaciones, tales como: la comparación de estructuras al realizar una superposición, de esta manera, se puede determinar si comparte el mismo sitio activo, construcción de superficies de las estructuras -mediante la visualización de la superficie es posible determinar si la zona activa es superficial o interna-, análisis de interacción entre el ligando y la diana este tipo de interacción ocurre en zonas internas de la proteína, por lo que debe ser posible extraer el sitio activo completo junto con el ligando; de esta manera será más sencillo el extraer información de la interacción, por ejemplo, si se forman puentes de hidrogeno, lo que supondría una mayor afinidad del ligando con la diana.

Una herramienta especializada en la búsqueda de compuestos líderes y filtrado de compuestos tóxicos es HAD [113], aparte de realizar estas búsquedas también hace predicciones de estructuras similares, las cuales son visualizadas a través del programa de análisis Spotfire[2]. Otro aporte importante es la agrupación de estructuras similares en forma, además de agruparlas por su actividad biológica. Konecni *et al.* [69] proponen un modelo de analítica visual para el descubrimiento de compuestos líderes. El modelo incorpora varias visualizaciones, así como métodos de minería de datos. Entre las herramientas de visualización está el *UMass Lowell Universal Visualization Platform* (UVP) [37], coordenadas paralelas modificadas, *heat maps*, comparación y visualización de *clusters* con CComViz [144].

Finalmente, también es necesario analizar y extraer conocimiento de información generada por otros métodos, como por ejemplo los datos generados por experimentos de laboratorio. Una herramienta diseñada con la intención de evaluar este tipo de información es InfVis [91], que a través de representaciones de glifos (pequeñas representaciones gráficas a modo de símbolos) en 3D, exploran, analizan y aplican minería de datos sobre información relevante en los datos generados por

[2]http://spotfire.tibco.com/

estos experimentos.

A lo largo de los capítulos anteriores hemos expuesto las diversas técnicas que se aplican cuando se toma la determinación de resolver un problema multidimensional por medio de la analítica visual. En el capítulo 4, se explico las áreas que la conforman como: el razonamiento analítico, las representaciones visuales, la transformación de los datos y la manera adecuada de presentar y transmitir el conocimiento generado a partir del análisis. Es importante que por medio de nuestra propuesta se genere un dialogo entre los datos, las visualizaciones y el químico de manera interactiva. El problema es claro, el tratar de analizar los resultados tal como los arroja el programa de *docking* -en este caso Autodock-, requiere un esfuerzo enorme por parte del químico para filtrar las poses que cumplan su criterio de las que no. Para automatizar el proceso de reducción es necesario contar con un método que agrupe las poses que sean similares para reducir el tiempo de selección; por otra parte es evidente que este tipo de resultados -de *docking*- contiene intrínsecamente grupos independientemente del método que se emplee para agruparlos. Dado que desconocemos la cantidad de grupos existentes en los datos, la mejor opción es emplear un algoritmo jerárquico, sin embargo, una vez superada la parte de agrupar, persiste el problema de exploración de esos grupos, será sencillo cuando se tengan unos cuantos, pero se complicará al aumentar el número -que es lo habitual-, por eso, es necesario emplear un método para automatizar ese proceso también [80].

Por otro lado, no podemos dejar a un lado el proceso de razonamiento analítico, al analizar los resultados de la agrupación y selección de poses, debe surgir una pregunta: ¿tiene sentido lo que estoy viendo? Si es afirmativa, continua el proceso de análisis, si la respuesta es negativa, entonces nuestra propuesta debe ofrecer al químico la manipulación de los resultados del agrupamiento, por ejemplo, tal vez es necesario separar un grupo en dos, y dado que se decantó por una representación de dendrograma y dotada de interacción, será posible separar esos grupos con bajar uno o varios niveles sin tener que volver a recalcular la jerarquía -una clara ventaja sobre los algoritmo particionales-, si aun persistiera la disconformidad de los resultados, nuestra propuesta deberá ofrecer más de un método de agrupación. Para una mejor compresión de una representación abstracta de las agrupaciones (dendrograma), se pretende añadir una visualización realista del espacio conformacional con las zonas que cubren las poses, de esta manera, se entabla un discurso analítico entre la interacción de las visualizaciones y el químico. En este sentido, se optará por emplear un visualizador con el que el químico, esté familiarizado, de esa manera, la curva de aprendizaje de manipulación de nuestra propuesta se reduce, además, por ser conocida previamente, ofrece cierta comodidad para interactuar con ella.

Por otra parte, una vez que se tiene automatizado el proceso de selección de poses que son de interés, surge un nuevo obstáculo, la necesidad de comparar de alguna forma las poses seleccionadas, a su vez, se mantiene la dificultad de ser demasiadas y añadiendo que son distintas, lo que representaría un retardo en la decidir si se parecen o no -hay una necesidad de automatizar este proceso también-. Una vez que esté automatizado, es necesario proveer de algún tipo de exploración sencilla, en la que pueda ver el químico al mismo tiempo los resultados de las agrupaciones y una representación realista de cómo sería la interacción entre las poses y la diana. De primera instancia, esto es abrumador, por lo que se dotaría de una guía visual para resaltar grupos interesantes [106], resaltando esos grupos que son interesantes, dirigen al químico hacia donde comenzar la exploración. Los siguientes capítulos detallaran la implementación de todas estas ideas en una herramienta para el análisis de resultados de *docking*.

Parte IV

Análisis de Resultados de *Docking*

Capítulo 8

Proceso de Análisis de Resultados de *Docking*

El *docking* es una herramienta ampliamente utilizada en el diseño de nuevos fármacos. El objetivo es encontrar el mayor número de moléculas que presenten afinidad por una determinada diana terapéutica para convertirlas posteriormente en fármacos. Los resultados del *docking* deberán reproducir razonablemente resultados experimentales. Para que esto sea eficaz, deberán probarse una gran cantidad de moléculas, de forma que, con los resultados obtenidos se tenga que sintetizar, adquirir y/o ensayar un número menor de compuestos reales a diferencia de cuando no se empleen herramientas virtuales. Su función es reducir el número de posibles moléculas a ensayar en campañas de búsqueda de líderes o de moléculas activas. Por ello el éxito de los experimentos de *docking* dependerá de que puedan realizarse de forma rápida y reproducible, esto es, que tengan la capacidad de refinar con facilidad los resultados (para adaptarse a los datos experimentales a medida que estos se obtienen), y, que su análisis sea lo más rápido y eficiente posible (para agilizar el proceso). En este trabajo se propone la generación de una herramienta que facilite el análisis de los resultados de *docking* y su posterior refinado de los mismos; apoyándose en Autodock[1] (versiones 3 [83] y 4 [84]), una herramienta de *docking* muy utilizada para el diseño de fármacos desarrollada por el Instituto Scripps de Investigación. Además, con una opción para el refinado de los resultados de Autodock propicia la transformación de dicho programa en una herramienta de búsqueda por farmacóforos en ausencia de información estructural sobre la diana. Por otro lado, hay que resaltar que, al realizar esto, se aporta una solución al diseño de fármacos cuando se carece de una estructura de la proteína o cuando la información de la estructura está incompleta, como suele suceder en muchos de los casos.

La selección del programa Autodock se realizó por ser un programa ampliamente utilizado, de libre disposición bajo la licencia GNU, y por qué se encuentra en continuo desarrollo y es utilizada por investigadores del Departamento de Química Farmacéutica de la Facultad de Farmacia de la Universidad de Salamanca para el diseño de nuevas moléculas antitumorales. Junto al programa Autodock, los investigadores del instituto Scripps han desarrollado una interfaz gráfica basada en Python, encaminada a facilitar la preparación de los ficheros necesarios para ejecutar Autodock. Esta herramienta también permite analizar los resultados de Autodock, pero es bastante limitada en la

[1] http://autodock.scripps.edu/

versatilidad de sus visualizaciones y sobre todo en la interacción con el experto para realizar análisis más detallados. Por otro lado, hasta ahora Autodock no se ha utilizado para realizar búsquedas por farmacóforo en situaciones en las que no se dispone de la estructura tridimensional de la diana, por lo que es uno de los objetivos de este trabajo. A continuación, se explicará el proceso de análisis de los resultados del programa Autodock, esto es, análisis de *docking*.

Durante el proceso de *docking* se genera para cada molécula virtual ensayada un número de posibles disposiciones (orientación en el espacio) de la misma cuando se enfrentan a la diana. Estas disposiciones se evalúan con distintas funciones que estiman de diversas formas la calidad de la interacción con la diana. Desafortunadamente, las funciones utilizadas para evaluar la interacción entre la diana y las moléculas (en sus diferentes disposiciones) no permiten discriminar adecuadamente la calidad de la interacción (un problema agravado por el poco tiempo disponible para el cálculo) y se hace necesaria la inspección visual de los resultados para seleccionar las mejores opciones. Este proceso de inspección visual se facilita considerablemente con herramientas de visualización como PyMol, Jmol, Marvin[2], RasMol, etc. Sin embargo, estas herramientas están diseñadas para la visualización de moléculas y complejos ligando-diana, pero no para el estudio de las interacciones ni para el análisis y comparación de muchos complejos simultáneamente.

Para lograr una comprensión detallada del problema de investigación es necesario realizar una breve descripción del proceso de cribado virtual y del análisis de los resultados del mismo. Para realizar el cribado virtual se necesita la estructura de la diana y los compuestos que se desean ensayar. A partir de la estructura de la diana, Autodock genera una malla (*grid*) centrada en el sitio en el que se considera que se pueden unir los compuestos (ligandos). La posición del *grid* y su resolución (espaciado de sus puntos) es configurable por el usuario. El programa coloca en cada uno de los puntos del *grid* cada uno de los átomos que forman los ligandos a ensayar (típicamente C, N, O, H, S, F, Cl, Br, I, B y cualquier otro átomo adicional), y evalúa la energía de interacción de ese átomo con la diana para cada uno de los puntos del *grid*. Genera así un mapa (*map*) por cada tipo de átomo. El programa utiliza estos *maps* para evaluar la energía de interacción de los compuestos en cada una de sus disposiciones con la diana, asignándole a cada uno de sus átomos la energía calculada para su posición en el *map* correspondiente y sumando los valores de todos los átomos del ligando. Además se calculan *maps* que consideran la repulsión estérica y se introducen factores que permiten estimar la energía intrínseca de cada disposición evaluada. Estos *maps*, son representaciones de la calidad de la interacción de la diana con cada tipo de átomo colocado en ese entorno (el equivalente a un farmacóforo pero definido por la estructura de la diana y no por la comparación de la estructura de distintos ligandos activos, ver figura 8.1).

[2]http://www.chemaxon.com/products/marvin/

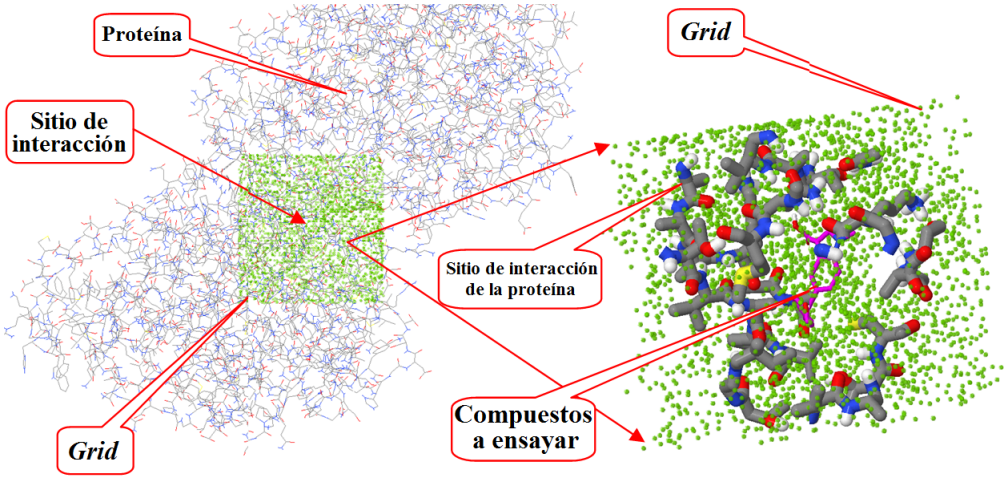


Figura 8.1: A la derecha, la proteína con su sitio de interacción y el *grid* de un tipo de átomo en color verde. A la izquierda, el sitio de interacción resaltado, el *grid* en verde y, en el centro del sitio, una molécula a ensayar.

Tras evaluar distintas disposiciones de los ligandos en el *grid*, se selecciona para cada compuesto un número prefijado de ellas -por lo general 100 disposiciones-. Frecuentemente, muchas de estas disposiciones son similares entre sí. Cada una de estas disposiciones recibe una puntuación que se relaciona con la afinidad de la asociación entre el ligando y la diana (*binding affinity*). Generalmente se observa que una disposición que se repite un número elevado de veces se asocia con una buena energía de interacción. En cualquier caso, aún cuando las repeticiones de una disposición de un compuesto son indicativas de interacciones favorables, para la inspección visual de los complejos sólo aportan información redundante, por lo que pueden resumirse en un ejemplo representativo (normalmente el de energía más favorable). El cálculo de las distancias entre distintas disposiciones de una misma molécula es trivial, y puede hacerse fácilmente mediante el RMSD (*root mean square deviation*), lo que facilita su agrupación. La dificultad en este caso consiste en decidir la distancia entre disposiciones (o lo que es lo mismo, el número de *clusters*) que es significativa en cada caso.

La complejidad del problema aumenta al intentar comparar ligandos diferentes. En este caso, la ocupación de los mismos sitios y el posicionamiento de átomos o agrupaciones de átomos de propiedades similares en posiciones cercanas aumenta la similitud entre disposiciones de compuestos diferentes. Alternativamente, la ocupación de posiciones diferentes o el posicionamiento de átomos o de agrupaciones de átomos de propiedades diferentes en posiciones cercanas aumenta la distancia entre ellas. Sin embargo, el peso asignado a cada una de estas evaluaciones varía frecuentemente con el criterio y experiencia del evaluador -son subjetivas- y dificultan la automatización del proceso. En la mayoría de los casos el proceso se realiza por visualización simultánea de las disposiciones a comparar en presencia y ausencia de la diana. Como el proceso es lento, con frecuencia se recurre a soluciones que conllevan pérdida de información, como es considerar solamente aquella disposición de cada compuesto que presenta menor energía y/o que ocurre con más frecuencia o que se superpone con un ligando de referencia, se reduce el número de compuestos a evaluar visualmente a un número reducido de los mismos -por ejemplo el 5 % mejor valorado por las funciones de evaluación-, etc. Para asistir en la toma de decisiones, con frecuencia se hace *docking* de una serie de ligandos de

referencia -normalmente activos- simultáneamente con el grupo a ensayar. Estas referencias permiten decidir qué estrategia tomar en cada caso particular y nos informan de la fiabilidad del protocolo empleado.

Una vez seleccionados los compuestos para los que se predice una mayor afinidad, estos se ensayan experimentalmente. Lo ideal sería que la información obtenida tras la evaluación pudiera incorporarse en el modelo, pero Autodock está diseñado para evaluar interacciones en muchos contextos, así que incorporar estos resultados no suele ser fácil, salvo que puedan asignarse a un problema en la parametrización del sistema. Si los resultados no son los esperados, con frecuencia se recurre a modificar la función de evaluación o a incrementar el tiempo de cálculo dedicado a cada compuesto.

Para demostrar la utilidad de la herramienta desarrollada en este trabajo se ha seleccionado la tubulina como diana terapéutica debido a que se conocen un número importante de ligandos de estructura muy diversa y a veces compleja que se unen a ella. Esta diversidad estructural de compuestos capaces de unirse en un mismo sitio hace difícil extraer las características estructurales que comparten estos ligandos por las cuales son reconocidas por la tubulina.

Proteína de Tubulina

La tubulina es una proteína que mediante procesos muy dinámicos de polimerización y despolimerización (reacción química que la que varias moléculas pequeñas se unen para formar una más grande y viceversa) da lugar a la formación y degradación de los microtúbulos, filamentos que constituyen el citoesqueleto en la célula. Los microtúbulos desempeñan funciones muy importantes; participan en el desarrollo y mantenimiento de la célula, en el transporte de vesículas, mitocondrias, en la mitosis [119] y segregación cromosómica [82] durante el proceso de la división celular. También juegan un papel importante en los procesos de señalización, migración de las células eucariotas [26] y otros componentes a través de las células. Los microtúbulos son heterodímeros formados por subunidades α y β de tubulina (Ver figura 8.2). Constituyen una diana muy importante para el desarrollo de fármacos contra el cáncer. También pueden ser una diana para el desarrollo de fármacos antifúngicos y antiparasitarios, e incluso como pesticidas [26].

Sustancias estructuralmente muy diversas, mayoritariamente de origen natural, son capaces de unirse a la tubulina y alterar el equilibrio o el dinamismo de la polimerización o despolimerización. Este hecho hace suponer que debido al importante papel que esta proteína juega en los procesos celulares, los seres vivos han sido capaces de generar sustancias para alterar su equilibrio durante el proceso evolutivo. Una prueba de ello es que seres vivos muy alejados desde el punto de vista filogenético (aparentemente muy separados en la rama evolutiva) como plantas vasculares, bacterias, moluscos, esponjas y otros animales marinos secretan metabolitos capaces de alterar la dinámica de los microtúbulos.

Debido a esta diversidad estructural mencionada, hay compuestos que se unen en distintos sitios de la tubulina y actúan mediante mecanismos distintos; así, hay sustancias que actúan inhibiendo su polimerización como por ejemplo, los alcaloides de la vinca o la podofilotoxina y otros lo hacen impidiendo su despolimerización como el taxol. En cualquiera de los casos, el efecto es el mismo puesto que ambos tipos de sustancias producen una parada en la división celular al alterar la dinámica de los microtúbulos que constituyen el huso acromático, elemento imprescindible para la migración de los cromosomas durante la mitosis.

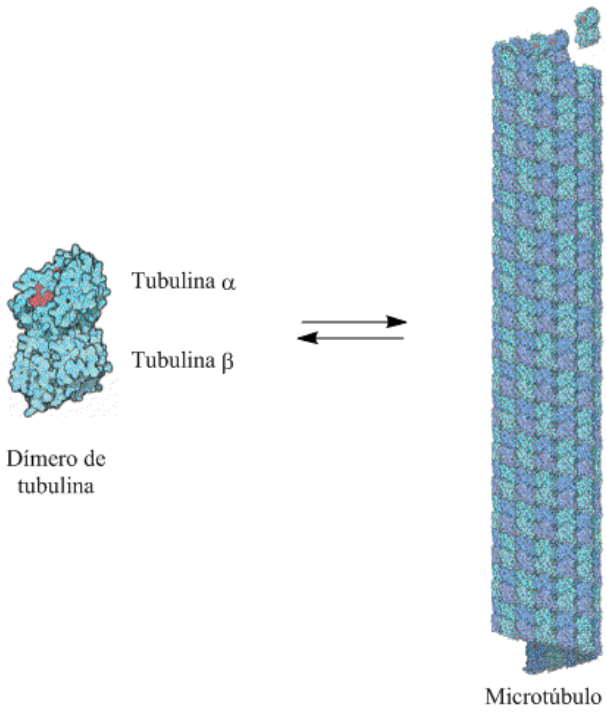


Figura 8.2: Formación de los microtúbulos a partir de α y β tubulina.

Las sustancias se unen principalmente en tres sitios distintos de la tubulina y reciben su nombre de tres sustancias de referencia. Así, se habla del sitio del taxol, donde además del taxol, importante fármaco también se unen otros taxanos, epotilonas, eleuterobinas entre otras; el sitio de los alcaloides de la vinca, donde además de importantes fármacos antitumorales como la vinblastina y vincristina se unen también dolastinas entre otros; por último el sitio de la colchicina, donde además de esta sustancia se unen la podofilotoxina, las combretastatinas, esteganacinas, curacinas y el $2-medroxiprogesterol$.

Estos agentes de unión de tubulina *Tubulin-Bindig Agents* (TBA) [59], además de incorporar una extrema diversidad y complejidad estructural, son menos sensibles a los mecanismos de resistencia, aumentando su selectividad (sólo se unirán a esa proteína en particular) y reduciendo los efectos secundarios, característica muy importante en los fármacos utilizados contra el cáncer. En el proceso de diseño de fármacos, una etapa clave después de la identificación de una diana terapéutica consiste en desarrollar ligandos potentes y selectivos capaces de unirse a ella. En el caso que aquí se aborda, se pretende diseñar ligandos capaces de alterar la dinámica de los microtúbulos mediante su unión a la tubulina. Conocida la estructura 3D de la diana, o mejor, como en este caso de complejo de interacción, el primer paso consistirá en realizar estudios de verificación virtual de un número de compuestos procedentes de fuentes muy diversas como por ejemplo la base de datos Zinc[3]. Mediante *docking* estos compuestos o fragmentos son colocados en sitios seleccionados de la diana (proteína),

[3] http://zinc.docking.org/

lo que se conoce como pose. Para cada uno de los compuestos o fragmentos se generarán *n* poses, a las que la aplicación informática le asignará una puntuación en función de la interacción más o menos favorable con la diana; además, clasificará a las poses y compuestos de acuerdo a la puntuación otorgada.

En la figura 8.3a, se muestra como ejemplo muy sencillo el resultado de procesar la información que se obtiene después de un estudio de *docking* realizado con un único compuesto, la colchicina, una de las sustancias de referencia que da nombre a uno de los sitios de unión de la tubulina. En la parte izquierda se representan cien poses distintas que este ligando ha ocupado en el sitio durante el estudio de *docking*. Una opción para el análisis de los resultados de este estudio consistiría en una inspección visual para seleccionar las mejores conformaciones/poses. El primer paso consistiría en agrupar las conformaciones más parecidas. En la figura 8.3a aparecen cuatro agrupaciones de conformaciones distintas que han sido agrupadas visualmente basándose en el carácter diferencial del color amarillo del átomo de azufre (encerrados en círculos de color verde en la figura 8.3). Se trata de un ejemplo relativamente sencillo, pues somos capaces de apreciar el carácter diferenciador del átomo de azufre. Sin embargo, la situación puede llegar a ser mucho más compleja como el ejemplo que se presenta en la figura 8.3b.

En este segundo ejemplo no es tan evidente ni sencillo saber cuántos grupos funcionales presenta cada compuesto, puesto que no hay un color por el cual guiarse ni otro elemento tan evidente como lo era para la colchicina. Su clasificación de forma visual sería un proceso muy tedioso, considerando que en un proyecto de investigación se pueden someter a estudios de *docking* miles de compuestos. Una vez obtenidas las agrupaciones de las poses de cada uno de los compuestos, habría que volver a filtrar esos grupos para obtener un representante de cada uno de ellos, para continuar con el proceso de diseño de un fármaco. En promedio, agrupar las conformaciones de un solo compuesto y obtener los representantes de cada grupo toma entre 10 a 15 minutos, considerando un promedio de 100 conformaciones para cada compuesto. Si multiplicamos este tiempo por el número de compuestos a estudiar en un proyecto, 1000000, nos daría un total de 15000000 minutos, lo que supondría más de 28 años. Puede ser que en los primeros cien compuestos el químico muestre optimismo, sin embargo, al cabo de un tiempo es prácticamente imposible realizar esta tarea en un tiempo considerablemente corto. Una manera de abordar este problema, sería hacer grupos de cien mil compuestos y analizar los primeros mil de cada grupo. Posteriormente al tener los representantes de cada grupo, se aplicaría un método de agrupación para filtrar y obtener aquellas conformaciones que compartan similitud de estructura; por último se continuaría con los pasos restantes del diseño de fármacos.

La cantidad de información que se genera es abrumadora y por lo tanto el proceso de diseño se puede ver comprometido al no fluir rápidamente para la obtención de resultados. Este cuello de botella en el diseño de fármacos es una clara oportunidad para aplicar técnicas de analítica visual, automatizar partes del proceso (en especial para agrupar moléculas que compartan similitudes estructurales), y lo más importante, permitir que un químico con experiencia en el campo, pueda tomar decisiones en todas las etapas del proceso.

En el siguiente capítulo se propone una solución para este problema. Por lo que a lo largo de las sucesivas secciones se expondrán al lector las tareas que se deberían seguir para abordar y solucionar el problema, tales como el manejo de la información, el filtrado y la visualización final para analizar los resultados.

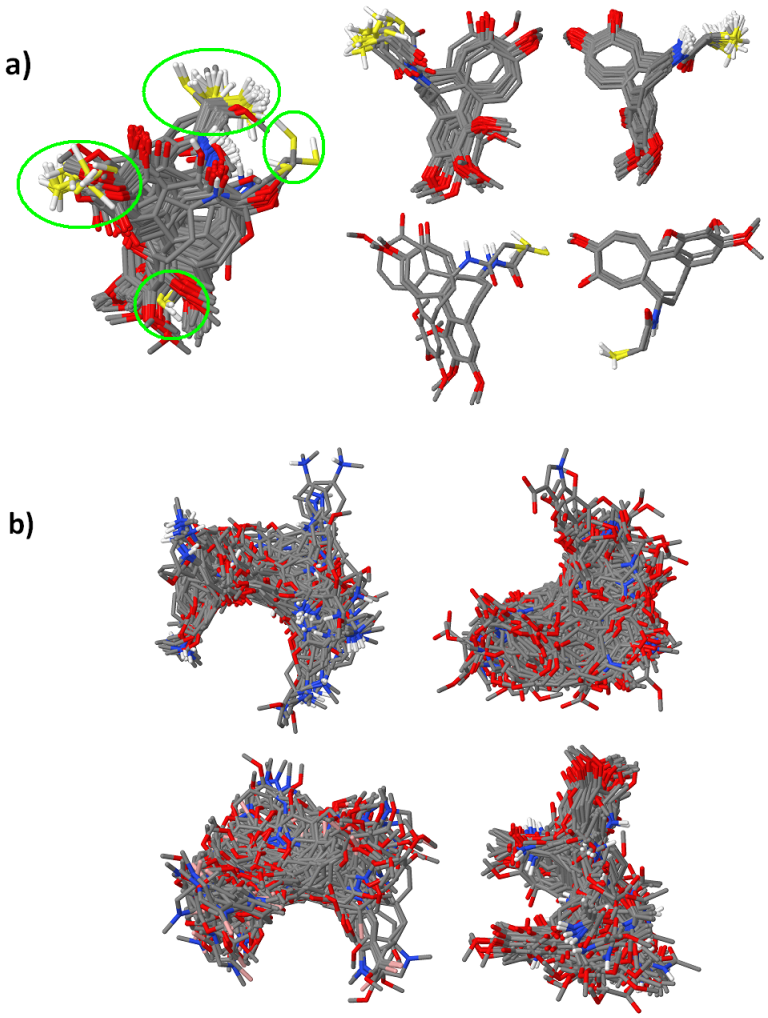


Figura 8.3: a) Resultado del *docking*: cien conformaciones de la colchicina. Agrupaciones manuales (visualmente) de los resultados del *docking*. b) Resultado del *docking*: diferente cantidad de grupos por compuesto.

Parte V

Solución Propuesta

Capítulo 9

Java based Autodock Preparing and Processing Tool (JADOPPT)

La solución que se propone en este trabajo de tesis al problema planteado en el capítulo anterior, está basada en consideración de los tipos de representación y convenciones (químicas) con las que los químicos farmacéuticos están familiarizados, lo que condiciona la forma de presentación de los datos. Esto representa un reto ya que, como se mencionó antes, la información, representación y el análisis están sujetos a la interpretación de datos químicos. Por otro lado, está la integración de métodos de extracción de información, procesamiento, representación (información relevante), interacción y toma de decisiones que emplea la analítica visual para proporcionar soluciones al combinarse e integrarse con herramientas más puramente químicas. Inicialmente se exploró la preferencia de los químicos por distintas herramientas de visualización molecular. De entre las opciones disponibles se seleccionó como herramienta de visualización de estructuras químicas Jmol, por ser código de abierto y permitir su modificación para integrarla en otras herramientas.

En esta elección también fue determinante la familiaridad de los químicos farmacéuticos participantes en el proyecto con la mencionada herramienta. En etapas iniciales del proyecto, para que existiera un dialogo entre farmacéuticos e informáticos era importante que los farmacéuticos se encontraran en un entorno familiar y cuando los informáticos estuvieran familiarizados con el problema, se introducirían nuevas representaciones y visualizaciones que apoyarían el análisis hecho por los farmacéuticos.

El trabajo de tesis doctoral se ha articulado en diferentes tareas:

1. Manipulación de la información química generada por el programa Autodock para generar datos en formatos que fueran reconocidos por los programas de visualización molecular (Jmol) y procesamiento de la información para que puedan ser procesados por los algoritmos de clasificación y técnicas de analítica visual. Esta manipulación exigió una familiarización con la terminología y los conceptos químicos y con los formatos reconocidos y/o generados en cada caso.
2. Agrupación de las soluciones de *docking* (poses de las moléculas ensayadas) en grupos que tuvieran significado químico y que reprodujeran, de la forma más automatizada posible, las

decisiones que los químicos tomaban en las clasificaciones manuales que ellos realizaban. Puesto que las agrupaciones de los químicos son una interpretación de los datos -y por tanto subjetivas- y debido a que los resultados de *docking* pueden presentar distintos grados de complejidad y dificultad dependiendo de la naturaleza de los conjuntos ensayados, se decidió desde un primer momento habilitar alguna forma de interacción visual con los resultados de los algoritmos de clasificación, permitiendo su refinamiento hasta conseguir los resultados esperados. También se consideró importante que la herramienta fuera capaz de seleccionar las moléculas de forma no supervisada en una primera aproximación, proporcionando una respuesta inicial que permitiera al usuario estimar los posibles puntos de mejora.

3. Visualización y manipulación de las soluciones de agrupación proporcionadas. Cuando se manejan grandes cantidades de datos (como suele ocurrir en experimentos de cribado virtual), el número total de disposiciones es muy elevado, lo que dificulta enormemente su análisis. Por ello se hizo necesario incorporar métodos de reducción de la complejidad de los datos, pero que permitieran al mismo tiempo el análisis visual de las estructuras que habitualmente se realiza con las herramientas de visualización molecular, por lo que se hizo necesario integrar las herramientas de visualización de agrupaciones (resultados de los algoritmos de clasificación) y de estructuras (Jmol), al tiempo que se ofrecía información relevante para el análisis en una sola herramienta visual.

4. Generación de una nueva herramienta que permita refinar tras el análisis los experimentos de *docking* y convertir Autodock en un programa de búsqueda por farmacóforos, cuando se carezca de información sobre la estructura de la diana. Para ello se propuso incorporar en la herramienta un método visual que modificara interactivamente los ficheros *maps* de Autodock. Estos *maps* modificados podrían utilizarse en etapas posteriores para experimentos de cribado masivo refinados -modificando *maps* generados a partir de una diana- o para realizar búsquedas por farmacóforos definidos por el usuario a partir de superposiciones de moléculas activas y/o información de relaciones estructura-actividad.

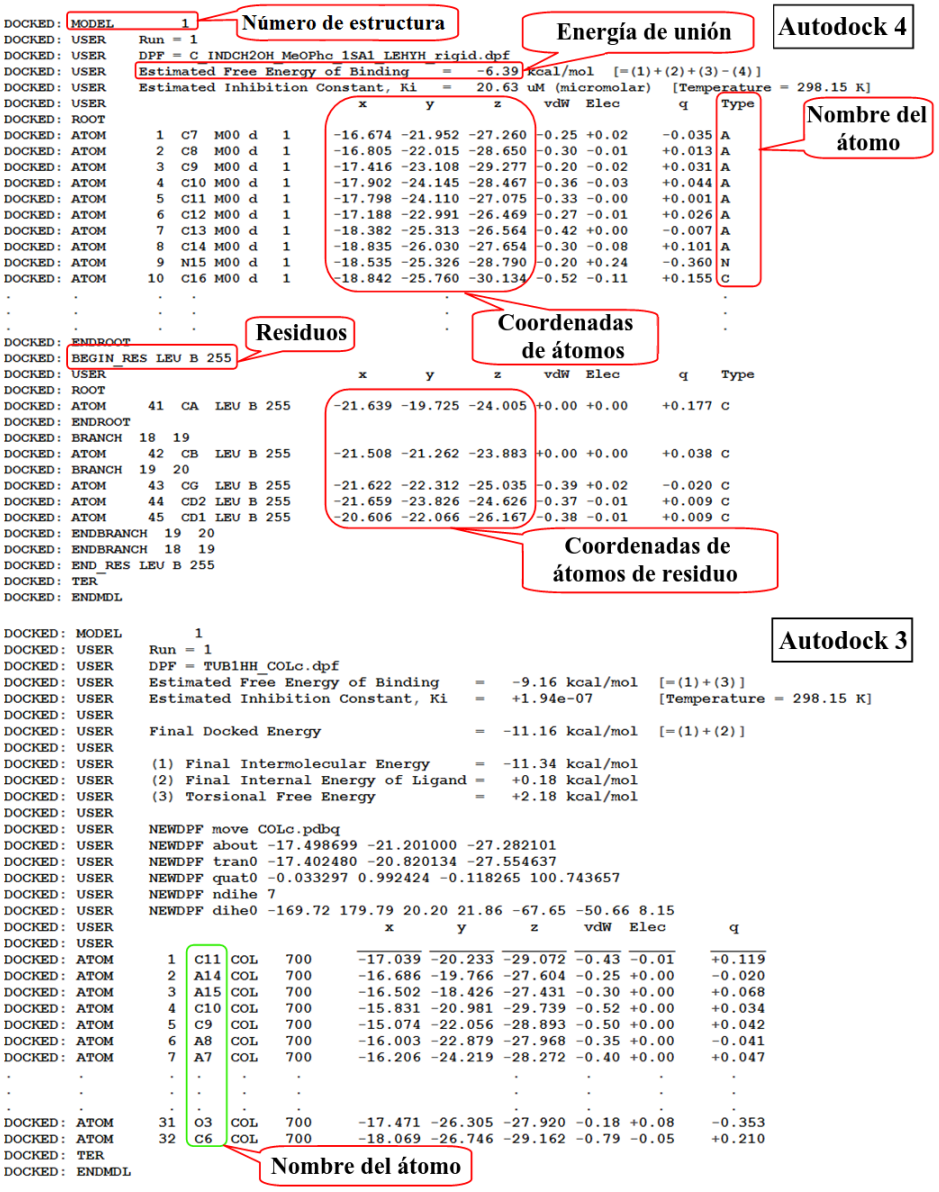
9.1. Manipulación de la Información Química Generada por Autodock

Los resultados del *docking* realizado por el programa Autodock (versiones 3 y 4) son almacenados en unos ficheros con extensión "dlg" (ficheros de registro *dock-log*), en los que se encuentran los detalles del *docking*, tales como: los parámetros utilizados en los experimentos -seleccionados en el fichero de *input*-, el número total de ejecuciones realizadas para cada molécula, el número de poses resultantes de cada ejecución, el tiempo que tardó para cada estructura, la energía de unión de cada pose con la diana (*Estimated Free Energy of Binding y Estimated inhibition constant*), sus coordenadas en el espacio (x,y,z) para cada átomo y su tipo, así como para los átomos de la diana que se consideran móviles (*docking* flexible).

No toda la información contenida en los ficheros dlg es procesada por la herramienta, sólo aquella con relevancia química es extraída, procesada y después visualizada para su análisis. El proceso de extracción de la información es la siguiente (ver figura 9.1): se extraen todas las líneas que están etiquetadas con la palabra *DOCKED*. Posteriormente se elimina la etiqueta *DOCKED* y se separa cada estructura por el número de ejecución, y se almacenan los valores de unión con la diana, el

tipo de átomo, los átomos de residuo, si es que los tiene, y las coordenadas tridimensionales de las estructuras. En la versión 3 de Autodock la estructura de la información era un poco diferente, inicialmente el tipo de átomo de carbono aromático estaba representado por la letra A y esta debía convertirse en la letra C para que pudiera ser reconocida por Jmol, esto ya no sucede en la versión 4 de Autodock; así, la herramienta lo mantiene por compatibilidad de ficheros procesados con la versión anterior (ver la figura 9.1); sin embargo, es la letra A la que es almacenada como tipo de átomo, para diferenciarla de los átomos de carbono. Finalmente, ese fichero será convertido en un fichero PDB (el cual es reconocido por el programa Jmol).

La estructura final -obtenida a partir de los datos de la figura 9.1- se almacena en un fichero PDB y su representación visual proporcionada por Jmol se puede observar en la figura 9.2. Se han eliminado todas las etiquetas que comienzan con *DOCKED* y se ha mantenido el resto de la línea. Las etiquetas que comienzan por *USER* son interpretadas como comentarios -no son procesadas por Jmol-, se muestra también que se incluye el nombre de la fuente original con la que fue creado ese PDB y la energía de unión. En la parte inferior de la figura se muestra cómo Jmol interpreta cada línea del PDB; por ejemplo, la flecha roja indica que la línea 9 corresponde a un átomo de oxígeno de la estructura, a diferencia de la flecha verde, que indica que la línea 43 corresponde a un átomo de hidrógeno, sin embargo, dicho átomo pertenece a los residuos.




```
DOCKED: MODEL        1
DOCKED: USER    Run = 1
DOCKED: USER    DPF = C_INDCH2OH_MeOPhc_1SA1_LEHYH_rigid.dpf
DOCKED: USER    Estimated Free Energy of Binding    =   -6.39 kcal/mol  [=(1)+(2)+(3)-(4)]
DOCKED: USER    Estimated Inhibition Constant, Ki   =   20.63 uM (micromolar)  [Temperature = 298.15 K]
DOCKED: USER                                  x       y       z     vdW  Elec       q    Type
DOCKED: ROOT
DOCKED: ATOM      1  C7  M00 d   1     -16.674 -21.952 -27.260 -0.25 +0.02    -0.035 A
DOCKED: ATOM      2  C8  M00 d   1     -16.805 -22.015 -28.650 -0.30 -0.01    +0.013 A
DOCKED: ATOM      3  C9  M00 d   1     -17.416 -23.108 -29.277 -0.20 -0.02    +0.031 A
DOCKED: ATOM      4  C10 M00 d   1     -17.902 -24.145 -28.467 -0.36 -0.03    +0.044 A
DOCKED: ATOM      5  C11 M00 d   1     -17.798 -24.110 -27.075 -0.33 -0.00    +0.001 A
DOCKED: ATOM      6  C12 M00 d   1     -17.188 -22.991 -26.469 -0.27 -0.01    +0.026 A
DOCKED: ATOM      7  C13 M00 d   1     -18.382 -25.313 -26.564 -0.42 +0.00    -0.007 A
DOCKED: ATOM      8  C14 M00 d   1     -18.835 -26.030 -27.654 -0.30 -0.08    +0.101 A
DOCKED: ATOM      9  N15 M00 d   1     -18.535 -25.326 -28.790 -0.20 +0.24    -0.360 N
DOCKED: ATOM     10  C16 M00 d   1     -18.842 -25.760 -30.134 -0.52 -0.11    +0.155 C
.       .         .   .                                                              .
.       .         .   .                                                              .
.       .         .   .                                                              .
DOCKED: ENDROOT
DOCKED: BEGIN_RES LEU B 255
DOCKED: USER                                  x       y       z     vdW  Elec       q    Type
DOCKED: ROOT
DOCKED: ATOM     41  CA  LEU B 255     -21.639 -19.725 -24.005 +0.00 +0.00    +0.177 C
DOCKED: ENDROOT
DOCKED: BRANCH  18  19
DOCKED: ATOM     42  CB  LEU B 255     -21.508 -21.262 -23.883 +0.00 +0.00    +0.038 C
DOCKED: BRANCH  19  20
DOCKED: ATOM     43  CG  LEU B 255     -21.622 -22.312 -25.035 -0.39 +0.02    -0.020 C
DOCKED: ATOM     44  CD2 LEU B 255     -21.659 -23.826 -24.626 -0.37 -0.01    +0.009 C
DOCKED: ATOM     45  CD1 LEU B 255     -20.606 -22.066 -26.167 -0.38 -0.01    +0.009 C
DOCKED: ENDBRANCH  19  20
DOCKED: ENDBRANCH  18  19
DOCKED: END_RES LEU B 255
DOCKED: TER
DOCKED: ENDMDL

DOCKED: MODEL        1
DOCKED: USER    Run = 1
DOCKED: USER    DPF = TUB1HH_COLc.dpf
DOCKED: USER    Estimated Free Energy of Binding    =   -9.16 kcal/mol  [=(1)+(3)]
DOCKED: USER    Estimated Inhibition Constant, Ki   =   +1.94e-07       [Temperature = 298.15 K]
DOCKED: USER
DOCKED: USER    Final Docked Energy                 =  -11.16 kcal/mol  [=(1)+(2)]
DOCKED: USER
DOCKED: USER    (1) Final Intermolecular Energy     =  -11.34 kcal/mol
DOCKED: USER    (2) Final Internal Energy of Ligand =   +0.18 kcal/mol
DOCKED: USER    (3) Torsional Free Energy           =   +2.18 kcal/mol
DOCKED: USER
DOCKED: USER    NEWDPF move COLc.pdbq
DOCKED: USER    NEWDPF about -17.498699 -21.201000 -27.282101
DOCKED: USER    NEWDPF tran0 -17.402480 -20.820134 -27.554637
DOCKED: USER    NEWDPF quat0 -0.033297 0.992424 -0.118265 100.743657
DOCKED: USER    NEWDPF ndihe 7
DOCKED: USER    NEWDPF dihe0 -169.72 179.79 20.20 21.86 -67.65 -50.66 8.15
DOCKED: USER                                  x       y       z     vdW  Elec       q
DOCKED: USER
DOCKED: ATOM      1  C11 COL   700     -17.039 -20.233 -29.072 -0.43 -0.01    +0.119
DOCKED: ATOM      2  A14 COL   700     -16.686 -19.766 -27.604 -0.25 +0.00    -0.020
DOCKED: ATOM      3  A15 COL   700     -16.502 -18.426 -27.431 -0.30 +0.00    +0.068
DOCKED: ATOM      4  C10 COL   700     -15.831 -20.981 -29.739 -0.52 +0.00    +0.034
DOCKED: ATOM      5  C9  COL   700     -15.074 -22.056 -28.893 -0.50 +0.00    +0.042
DOCKED: ATOM      6  A8  COL   700     -16.003 -22.879 -27.968 -0.35 +0.00    -0.041
DOCKED: ATOM      7  A7  COL   700     -16.206 -24.219 -28.272 -0.40 +0.00    +0.047
.       .         .   .    .                                                     .
.       .         .   .    .                                                     .
.       .         .   .    .                                                     .
DOCKED: ATOM     31  O3  COL   700     -17.471 -26.305 -27.920 -0.18 +0.08    -0.353
DOCKED: ATOM     32  C6  COL   700     -18.069 -26.746 -29.162 -0.79 -0.05    +0.210
DOCKED: TER
DOCKED: ENDMDL
```

Figura 9.1: Formato de salida de Autodock 4 y 3. En la parte superior (Autodock 4) se muestra la información que se extrae del fichero dlg. En la parte inferior, el formato antiguo (Autodock 3), que era muy similar, salvo que era necesaria la extracción del tipo de átomo y que en ésta versión no se incluían los residuos.

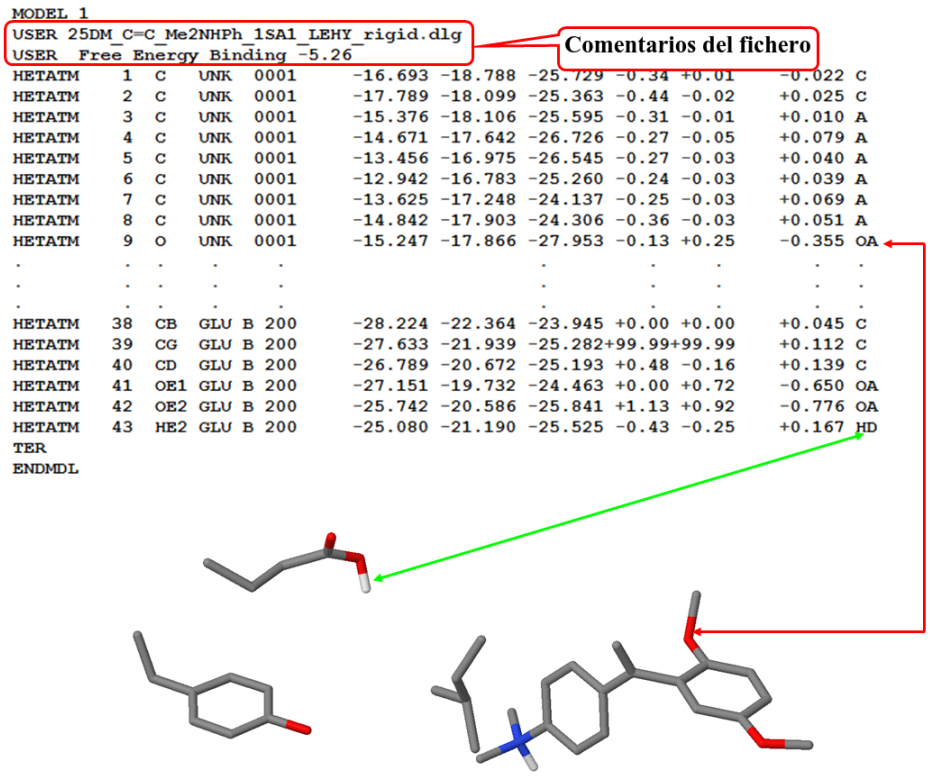


Figura 9.2: Estructura del PDB final después del procesamiento de extracción y conversión de un dlg. En la parte inferior la representación visual del mismo PDB por Jmol. La flecha roja corresponde a un átomo de oxígeno de la estructura y la flecha verde a un hidrógeno de los residuos.

Otro de los ficheros que genera Autodock es un fichero de tipo *map* -con la extensión *map*-. Para cada tipo de átomo que esté presente en las moléculas que han sido sometidas a *docking*, se crea un nuevo fichero *map*. Estos ficheros contienen los valores de energía que asigna Autodock al situar un átomo del tipo *map* en cada uno de los puntos del *grid*, definidos en los parámetros proporcionados al programa. Los valores están ordenados de forma que su número de orden se relaciona con la posición en el *grid* y, consecuentemente con su posición el espacio. Valores negativos implican situaciones favorables y valores positivos situaciones desfavorables en la interacción con la diana. Las primeras tres líneas del fichero contienen información referente a los ficheros de entrada utilizados por Autodock, mientras que las tres siguientes definen el *grid* al que corresponden los valores del *map*: separación entre los puntos del mismo (en la figura 9.3 0,375Å), número de puntos que el *grid* tiene en las tres direcciones del espacio alrededor del punto central (en este caso 60, que se traduciría en un *grid* de $61 \times 61 \times 61$ puntos, lo que resultaría en 226981 valores) y las coordenadas cartesianas del punto central del *grid*. La asignación de los valores en la malla de puntos se hace utilizando un bucle $z(y(x))$ (ver figura 9.3).

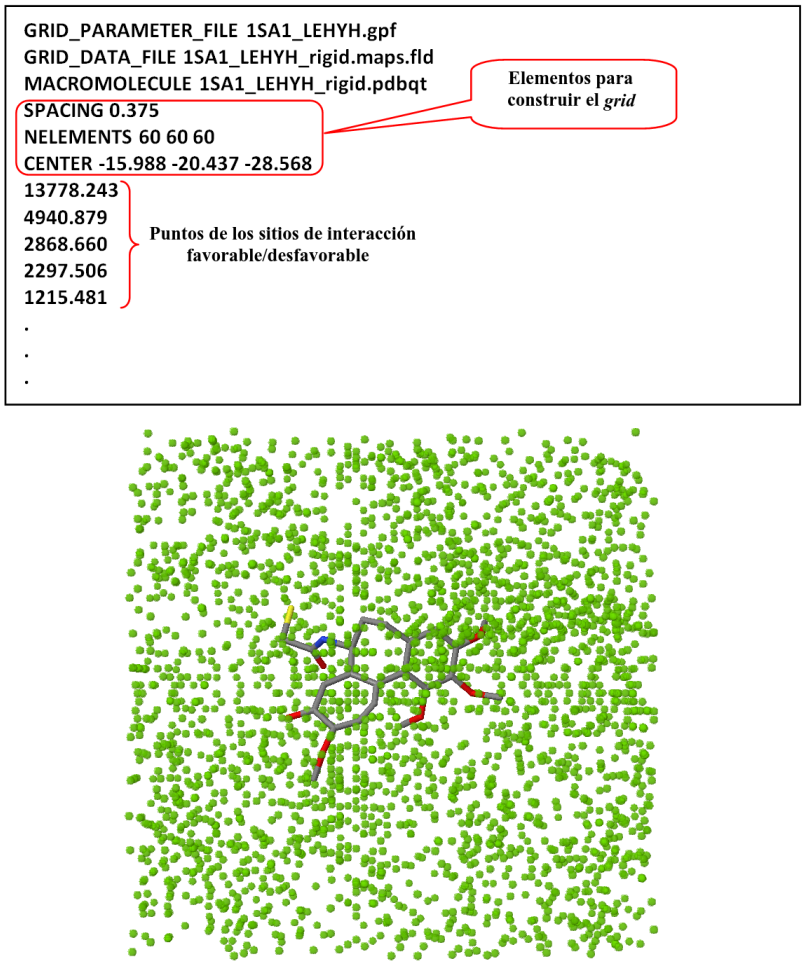


Figura 9.3: Estructura de un fichero tipo *map*. De la línea cuatro a la seis contiene la información de cómo están calculadas las coordenadas de los sitios favorables/desfavorables para un átomo. En la parte inferior se muestra cómo se verían los puntos en el espacio de los sitios favorables.

Del fichero tipo *map* se extraen todas las líneas a partir de la cuarta. Con la información de las líneas 4 a 6 se regenera la malla virtual. Posteriormente, a partir de la línea 7 se realiza una conversión inversa para calcular las coordenadas de cada valor de la malla virtual aplicando la misma función con la que fueron calculadas $z(y(x))$; una vez obtenidas estas coordenadas, se eliminan los valores mayores que cero -los valores positivos son desfavorables y se descartan para reducir el tamaño de los ficheros, ya que la información sobre interacciones desfavorables se puede extraer de los *maps* que no corresponden a un tipo de átomo-. El resultado de extraer esos valores del fichero se muestra en la parte inferior de la figura 9.3, cada punto en el espacio corresponde a un valor favorable del fichero *map*. Estos puntos serán utilizados posteriormente como puntos observadores de moléculas con el fin de comparar estructuras diferentes en cuanto a tamaño y tipo de átomos. Además, contribuyen al refinado del proceso de *docking*, lo que se explicará en detalle en la sección 9.2.3.

9.2. Agrupación de las Soluciones de *Docking*

AutodockTools[1] (ADT) es una herramienta desarrollada en paralelo a Autodock para poner a punto y analizar experimentos de *docking* y cribado virtual realizados con Autodock. ADT proporciona un método de visualización de los resultados del *docking* (ver figura 9.4), ya que realiza un *clustering* para analizar las poses de una misma molécula, representando visualmente los *clusters* mediante gráficas de barras. La interacción entre el visualizador de moléculas y el diagrama de barras permite cargar los *clusters* individualmente y visualizar uno a uno sus representantes mediante otra ventana de control. Una desventaja notable es que no es posible visualizar todo un conjunto de moléculas diferentes. El programa permite cargar varias moléculas diferentes y manipular cada una de ellas de forma individual, pero no permite agrupar moléculas diferentes ni manipularlas como un conjunto. Esto dificulta notablemente el análisis de conjuntos de moléculas diferentes.

Uno de los principales objetivos de la herramienta que se ha desarrollado se dirige a solucionar el problema antes mencionado mediante la automatización de esta parte del proceso. Para ello, se ha dividido el problema en partes: 1) agrupación de los resultados del *docking* de cada molécula por separado, 2) elección de un número de *clusters* que sea representativo del conjunto total de disposiciones para cada molécula, 3) selección de representantes de cada uno de dichos grupos para su utilización en la comparación con los de otras moléculas diferentes, 4) comparación (estimación de la similitud/diferencia) de los representantes de las distintas moléculas y agrupación de los representantes en *clusters*, 5) manipulación de los *clusters*, los representantes y sus visualizaciones.

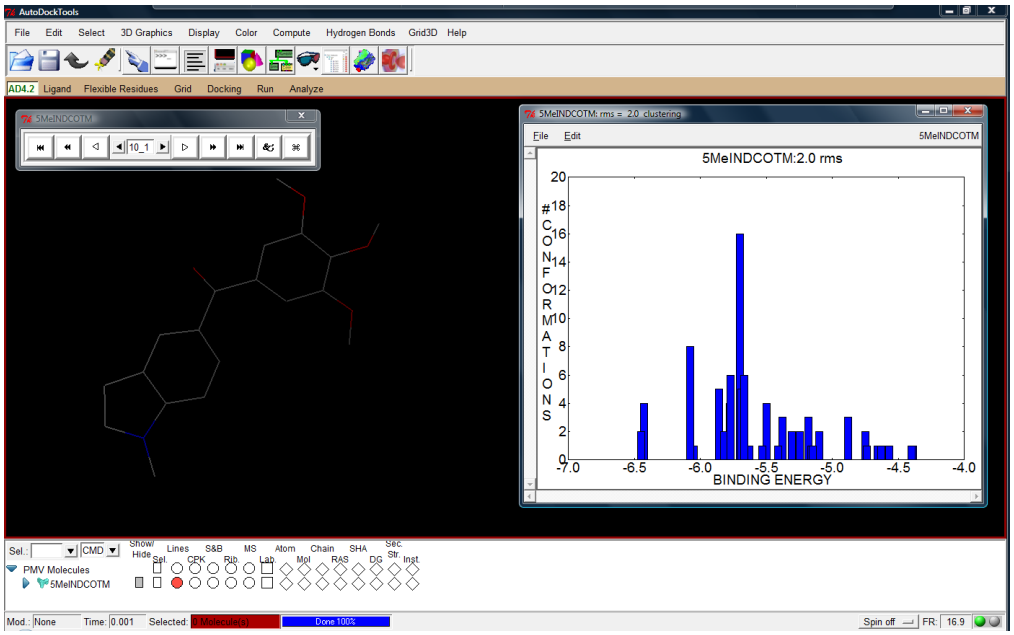


Figura 9.4: Vista general de AutodockTools. Análisis de un compuesto.

[1] http://autodock.scripps.edu/resources/adt

9.2.1. Agrupación de los Resultados del *Docking* de Cada Molécula por Separado

La herramienta cuenta con dos métodos para agrupar los resultados del *docking*: el primero es por RMSD (*Root Mean Square Deviation*). El cual incorpora el mismo método de ADT para generar *clusters*. El segundo método consiste en la generación de *clusters* mediante un algoritmo no supervisado. A continuación se detalla cada método.

9.2.1.1. Agrupación Mediante RMSD

La agrupación de elementos se basa en el cálculo de las distancias (diferencias) que existen entre los distintos representantes. En el caso de disposiciones de una misma molécula, el procedimiento más habitual es calcular el RMSD entre sus átomos. El cálculo de RMSD se puede realizar mediante la función (1)

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_{ix}-B_{ix})^2+(A_{iy}-B_{iy})^2+(A_{iz}-B_{iz})^2} \quad (1)$$

Donde A y B representan dos estructuras del mismo tipo, n el total de átomos presentes en la estructura y x,y,z a las coordenadas en el espacio de cada átomo. Sin embargo, este método considera como diferentes disposiciones equivalentes (o incluso idénticas) que surgen por la presencia de elementos de simetría, como se indica en la figura 9.5 para un anillo de trimetoxifenilo, ya que la identificación de los átomos en los modelos es inequívoca, aunque que en la realidad sean indistinguibles.

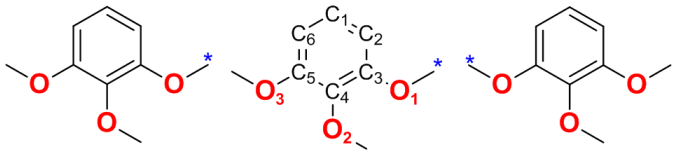


Figura 9.5: El problema de la simetría en el cálculo del RMSD. Los anillos de trimetoxifenilo a ambos extremos son idénticos. Sin embargo, al dar nombre a los carbonos y oxígenos (centro) se desimetriza (indicado por el asterisco azul), de modo que para el cálculo del RMSD son diferentes.

Una posibilidad es detectar estos elementos de simetría y permitir en el cálculo del RMSD el intercambio de las posibilidades relacionadas por el elemento de simetría. Habitualmente esto se realiza proporcionando a los programas de cálculo del RMSD los grupos de átomos intercambiables, pero como en este caso se pretende que el proceso sea automático habría que calcularlo a partir de la estructura. Esto plantea notables dificultades y se ha considerado que ralentizaría el proceso. Si se ignora este hecho en el cálculo, el resultado es que se considerarán más agrupaciones de las debidas (tantas como agrupaciones equivalentes genere el/los elemento/s de simetría). En el peor de los casos, el resultado sería que se seleccionarían varios representantes para las etapas subsiguientes, pero a priori esto no parece un problema grave, por lo que se decidió seguir con dicha estrategia (ver figura 9.6).

En cualquier caso, el cálculo del RMSD (aún en ausencia de elementos de simetría) no permite la sustitución o reemplazo de unos grupos por otros, al tratarse de un cálculo biunívoco (figura 9.7).

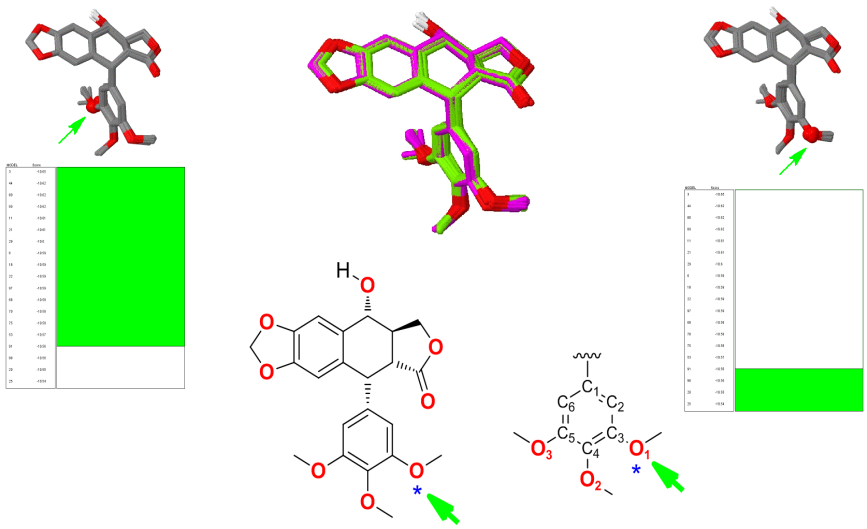


Figura 9.6: Agrupaciones idénticas que se diferencian en la posición de los oxígenos. La agrupación de las disposiciones de podofilotoxina (mostradas en el centro con los carbonos coloreados en rosa o verde, junto con la desimetrización causada al etiquetar los átomos) utilizando el RMSD genera dos agrupaciones (derecha e izquierda) que se diferencian sólo en la identidad asignada a los oxígenos (resaltados como *cpk* e indicados con flechas de color verde). Como se ve en el centro, a excepción de esto los representantes de las dos agrupaciones son idénticos. El *cluster* seleccionado se muestra en verde en el *treemap*.

Al igual que con el problema de los elementos de simetría, esto no tiene más inconveniente que el de seleccionar un mayor número de representantes para la etapa siguiente.

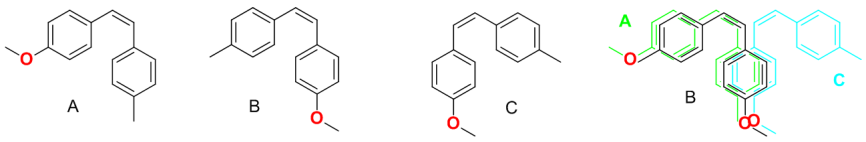


Figura 9.7: El cálculo del RMSD no permite reemplazos ni sustituciones. La comparación de las disposiciones A, B y C (representadas en la parte derecha con un ligero desplazamiento para permitir su visualización y dibujadas individualmente en la parte izquierda para facilitar su visualización) del mismo compuesto utilizando el RMSD como criterio daría una mayor distancia entre A y B que entre B y C.

La opción inicial para el cálculo del RMSD fija un umbral de distancia de 2.0Å, que es la misma distancia que emplea ADT. Sin embargo, para reducir el número de *clusters* finales, la herramienta repite el cálculo dos veces más aumentando la distancia primero en 0.50Å y finalmente en 0.25Å. En otras palabras, una vez que se han formado los *clusters* la herramienta hace una segunda vuelta tomando un elemento del *cluster* y lo comprará primeramente con *clusters* de un solo elemento, si cumplen el criterio de distancia es incorporado a ese *cluster*, posteriormente hace lo mismo con los

clusters restantes. De esta forma se reduce el número de *clusters* y se trata de subsanar el problema de simetría.

El resultado del cálculo de RMSD a diferencia de las barras de ADT (ver figura 9.4), es visualizado mediante un conjunto de contenedores anidados llamado *treemap* -el *treemap* es un tipo de visualización que se emplea para representaciones de jerarquías-, este tipo de visualización ayuda a ver claramente qué *clusters* tienen mayor número de elementos. Así, es sencillo para el químico enfocarse en aquellos que tienen mayor población, a diferencia de lo que ocurre en la visualización de ADT. Uno de los principales problemas que presenta la visualización de barras en ADT es la poca interacción con el usuario, además, las barras se sitúan en relación a número de elementos (*Conformations*) en el $eje-y$ y a la energía de unión (*Binding Energy*) en el $eje-x$; lo cual ocasiona superposición de las barras, debido a los valores de energía de cada una de las poses. Por otro lado, para cada barra utilizan un color, lo cual, dificulta saber el total de *clusters* que se han formado, si observamos en la figura 9.4, hay barras que están detrás de otras y comparten una energía de unión muy cercana, pero, el número de elementos es menor al de la barra de enfrente. Incluso, hay ocasiones, en las que tienen el mismo número de elementos al igual que la energía de unión, pero no necesariamente las poses están orientadas igual, esto es recurrente en los resultados de *docking*, por lo que, el tratar de analizar los *clusters* uno ocultaría al otro, haciendo imposible su análisis.

En contraparte, en la visualización del *treemap*, es sencillo de interpretar, así como encontrar fácilmente los *clusters* de interés. En la figura 9.6, el químico centraría su atención en el cuadro más grande al igual que lo haría en la visualización de ADT, sin embargo, en el caso de que hubiera dos *clusters* con el mismo número de elementos y sus valores de energía de unión fueran muy cercanos en ADT estaría superponiéndose mientras que en la visualización de *treemap* esto no sucedería, por lo tanto no habría perdida de información para el análisis.

9.2.1.2. Agrupación Mediante el Algoritmo Aglomerativo

En el caso de la agrupación por medio de un algoritmo aglomerativo, se extraen las coordenadas cartesianas de cada átomo en las distintas disposiciones, la cual, genera una matriz de distancias (obviamente se trata de una matriz simétrica respecto a la diagonal, ya que la distancia entre las poses 1 y 2 es la misma que entre las poses 2 y 1, cuyos valores en la diagonal son necesariamente ceros, ya que la distancia de una disposición consigo misma es cero), y el cálculo, se realiza fácilmente utilizando la expresión (2), para el caso de la distancia Euclidia.

$$d(P,Q) = \sqrt{\sum_{i=1}^{n} (P_i - Q_i)^2} \quad (2)$$

Y en el caso de la distancia Manhattan la expresión (3):

$$d(P,Q) = \sum_{i=1}^{n} |P_i - Q_i| \quad (3)$$

En donde P y Q representan a las disposiciones en forma de vector (figura 9.8). Ambas fórmulas se pueden emplear para calcular las matrices de distancia, siendo la Euclidia la más común y, por ende, la que se emplea por defecto en nuestra propuesta. Sin embargo, para compatibilidad con la mayoría de herramientas estadísticas, en nuestra propuesta también se incluye la distancia Manhattan.

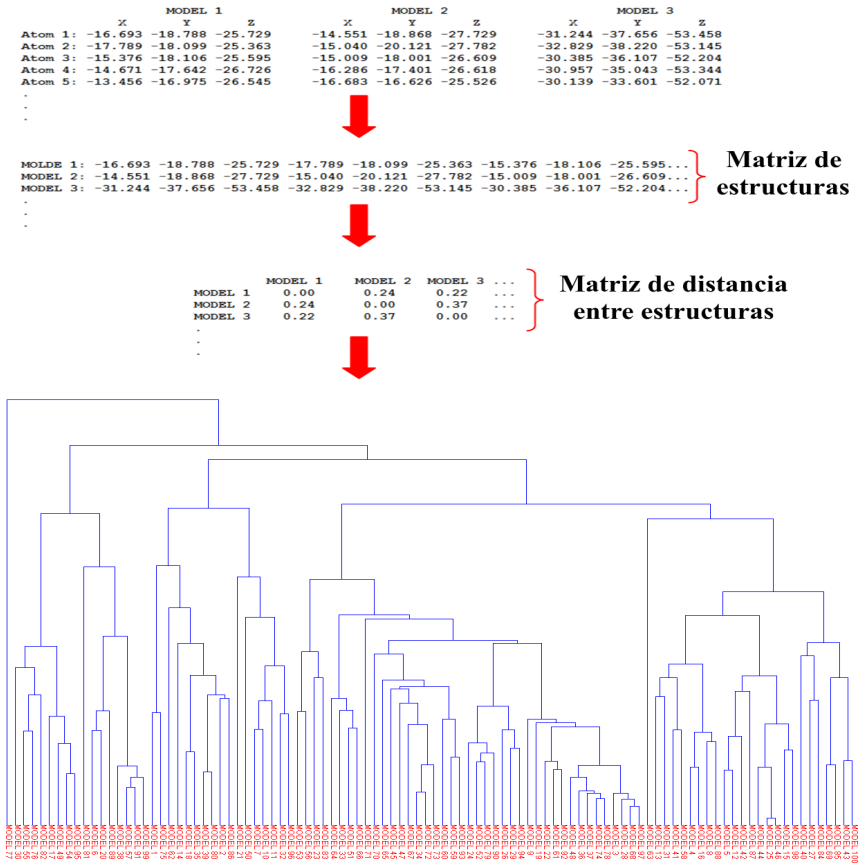


Figura 9.8: Flujo del algoritmo para agrupar poses de la misma molécula. Cada estructura (*MODEL*) es convertida en vector para formar una matriz de estructuras. Para cada estructura se calcula la distancia entre ellas y se construye una matriz de distancias entre estructuras. Se aplica un algoritmo que usa la distancia entre *clusters* para crear una jerarquía que es visualizada a través de un dendrograma.

Una vez que se dispone de la matriz de distancias se procede a agrupar los elementos de la misma (disposiciones de la molécula). En las etapas iniciales de este trabajo de tesis, se utilizó el algoritmo particional *k-means*, debido a que en la literatura se mencionaba como el algoritmo de clasificación más sencillo de implementar y por ser razonablemente rápido. No obstante, este método presenta una desventaja importante, pues es preciso conocer previamente el número de particiones que se desea analizar; sin embargo, el número de particiones que se desea en el caso del análisis de los resultados de *docking* este dato no se conoce de antemano, ya que varía considerablemente en función de cada compuesto (figura 9.9). Entre las distintas opciones, se consideró fijar a priori el número de *clusters* para cada compuesto y hacer esta opción configurable. Sin embargo, un análisis preliminar de resultados de *docking* realizados con anterioridad mostró rápidamente que esta aproximación era poco adecuada, ya que el número de *clusters* encontrado para moléculas de tamaño y propiedades similares en una misma diana y sitio variaba considerablemente.

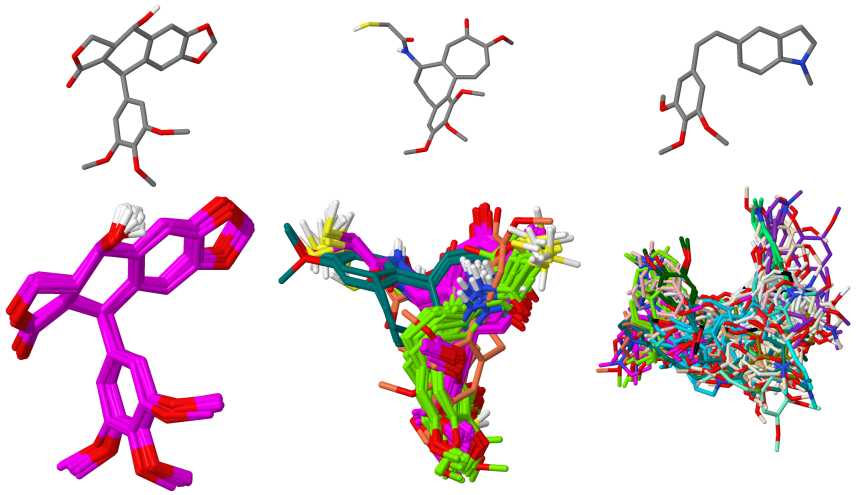


Figura 9.9: Variabilidad en el número de *clusters* para moléculas similares en un mismo experimento de *docking* [78]. En la línea superior se muestra la estructura tridimensional de tres ligandos diferentes del sitio de la colchicina en tubulina: podofilotoxina, colchicina y una indolcombretastatina. La primera molécula (podofilotoxina) contiene un sólo *cluster*; la del centro (colchicina) contiene cuatro *clusters* definidos por colores (verde oscuro, verde claro, naranja y magenta) y, por último, la combretastatina presentó un total de 22 *clusters*.

Por ello, surgió la necesidad de utilizar un algoritmo no supervisado, entre los que se optó por el uso de un algoritmo aglomerativo. Inicialmente se planteó llevar a cabo el proceso de modo que produjese un resultado similar al que produce por defecto ADT. Esto permitiría establecer comparaciones directas entre los dos métodos y establecer un valor por defecto. ADT emplea un algoritmo jerárquico de agrupación basado en el RMSD en el que se construyen *clusters* partiendo de la estructura de mínima energía e incluye en dicho *cluster* todas las disposiciones con un RMSD por debajo de un valor umbral[2], que por defecto se sitúa en 2Å. Cuando no encuentra disposiciones que cumplan ese criterio, genera un nuevo *cluster* con la siguiente disposición de menor energía, y así hasta agruparlas todas. En caso de que el punto de corte no se considere adecuado, el usuario de Autodock puede elegir un nuevo valor umbral y recalcular el *clustering*.

Finalmente, aplicando un algoritmo para calcular la distancia entre *clusters*[3] de la matriz de distancias, se construye una jerarquía de *clusters*; en otras palabras, el algoritmo a medida que va iterando va formando *clusters* y a cada iteración se le conoce como nivel. La herramienta cuenta con tres algoritmos para calcular la distancia entre *clusters*, tales como: la distancia promedio (*average-linkage*), es la distancia promedio entre todas las moléculas de *cluster* con respecto a otro, por lo que para formar el siguiente nivel calculará las distancias promedio de todas las moléculas de los *clusters* existentes (todos contra todos) y unirá solamente dos *clusters* en los que su distancia promedio sea menor. Esta aproximación mostró mejores resultados en moléculas pequeñas, con respecto a las otras dos; distancia mínima (*single-linkage*), también conocida como el vecino más

[2]http://mgltools.scripps.edu/api/AutoDockTools/AutoDockTools.cluster.Clusterer-class.html

[3]Hay que recordar que inicialmente el método de *clustering* aglomerativo considera a cada elemento (disposición/molécula) de la matriz de distancias como un *cluster*, ver sección 5.2.1.

cercano, unirá dos *clusters* siempre y cuando exista una distancia mínima entre dos moléculas de *clusters* diferentes. Esta aproximación muestra buenos resultados en moléculas pequeñas y también al aplicarse a proteínas (experimentos ajenos a este trabajo de tesis); distancia máxima (*complete-linkage*) es el mismo criterio que el de distancia mínima salvo que se considera la máxima distancia entre dos moléculas (ver figura 9.10). Esta aproximación mostro mejores resultados en moléculas grandes como los inhibidores de VIH-1 (ver sección 10.4 para más detalles) y para proteínas.

Esta jerarquía es visualizada a través de un dendrograma interactivo que permite visualmente analizar su estructura. Además, utilizando la herramienta de visualización vinculada al dendrograma (que se explicará más adelante) puede analizarse visualmente la composición de los *clusters*. Utilizando esta herramienta de visualización se intentó reproducir los resultados de *clustering* manual y los que producía ADT para una serie de ejemplos, haciendo posible estimar unos valores por defecto para el umbral de corte.

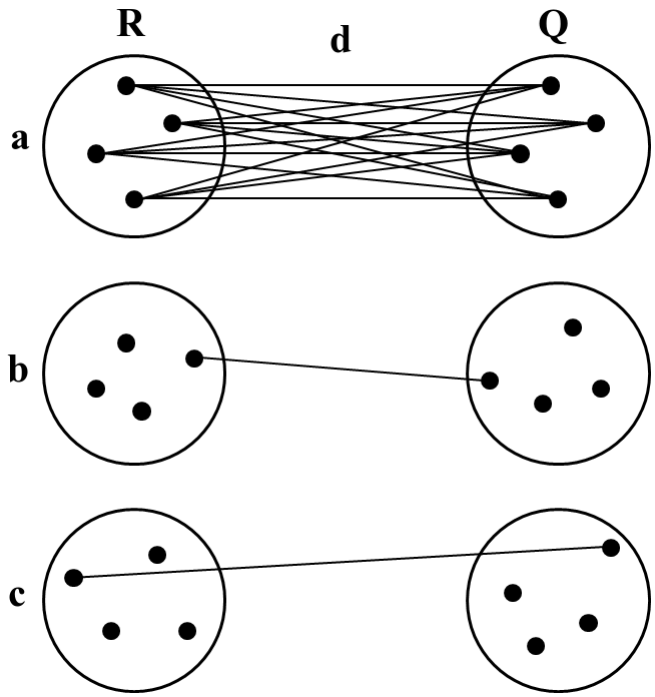


Figura 9.10: Representación de cálculo de distancia entre *clusters* de moléculas. Sean *R* y *Q* dos *clusters* y d la distancia entre cada molécula de *R* y *Q*. a) Distancia promedio entre todas las moléculas de *R* y *Q* (*average-linkage*). b) Distancia mínima entre una molécula del *cluster R* y una molécula del *cluster Q* (*single-linkage*). c) Distancia máxima entre una molécula del *cluster R* y una molécula del *cluster Q* (*complete-linkage*).

El problema en este punto consiste en saber qué nivel de la jerarquía contiene los mejores *clusters* y hacer un corte en ese nivel. En un principio se trató de emular el corte partiendo de la manera en que lo hacía ADT. Se realizaba una búsqueda por cada nivel y aquellos *clusters* que fueran más parecidos al valor de RMSD que utiliza ADT eran seleccionados, en otras palabras, se extraía el valor de distancia de unión de cada nivel y se comparaba con el valor de RMSD de ADT. Sin embargo, los *clusters* no siempre coincidían con los del ADT.

En la figura 9.11 se compara un ejemplo del resultado del *clustering* obtenido con ADT con la respuesta generada por el algoritmo de *clustering* cuando el número de *clusters* es el mismo. Sin embargo, los *clusters* son un poco diferentes los. Los *clusters* de la izquierda pertenecen al RMSD de ADT y los de la derecha fueron calculados con el método de estimación de corte. Claramente coinciden los tres *clusters* que contienen el mayor número de estructuras tanto en ADT como en nuestra propuesta. En cambio en la figura 9.12 nuestra propuesta propone dos *clusters* de todo el conjunto de estructuras. Sin embargo, al examinar el *cluster 2* en detalle, resalta que puede separarse para formar otros *clusters*. Interactivamente, al arrastrar la barra verticalmente, el *cluster 2* se separa en otros *clusters*. No es posible esta interacción entre ADT y su visor de moléculas; una vez calculados los *clusters* no hay forma de separarlos mediante alguna función visual interactiva, la única opción disponible es rehacer el *cluster* cambiando los parámetros y, por ende, cambiando los elementos y el número de *clusters*. Si bien esto es una opción válida, no siempre es funcional si lo que se desea es solamente separar un grupo en el que hay claramente dos o más *clusters* bien definidos, como en el caso de la figura 9.12. Esta no es la única ventaja que ofrece nuestra propuesta, también tiene la posibilidad de rehacer el *cluster* aplicando otro de los algoritmos antes mencionados. Un aporte adicional que ofrece la herramienta es unir *clusters* que estén en diferentes niveles para formar uno sólo.

Al finalizar el proceso de selección de *clusters* ya sea de forma automática o manual (explorando uno o todos los niveles y seleccionado uno o todos los *clusters* de ese o esos niveles) se procede a seleccionar una estructura de cada *cluster* que ha de representar a ese *cluster* en etapas posteriores. Para cada *cluster* se elegirá como representante aquel que posea la mejor unión con la diana, es decir, aquel cuya energía de unión sea la más negativa (*free binding energy*, ver figura 9.1).

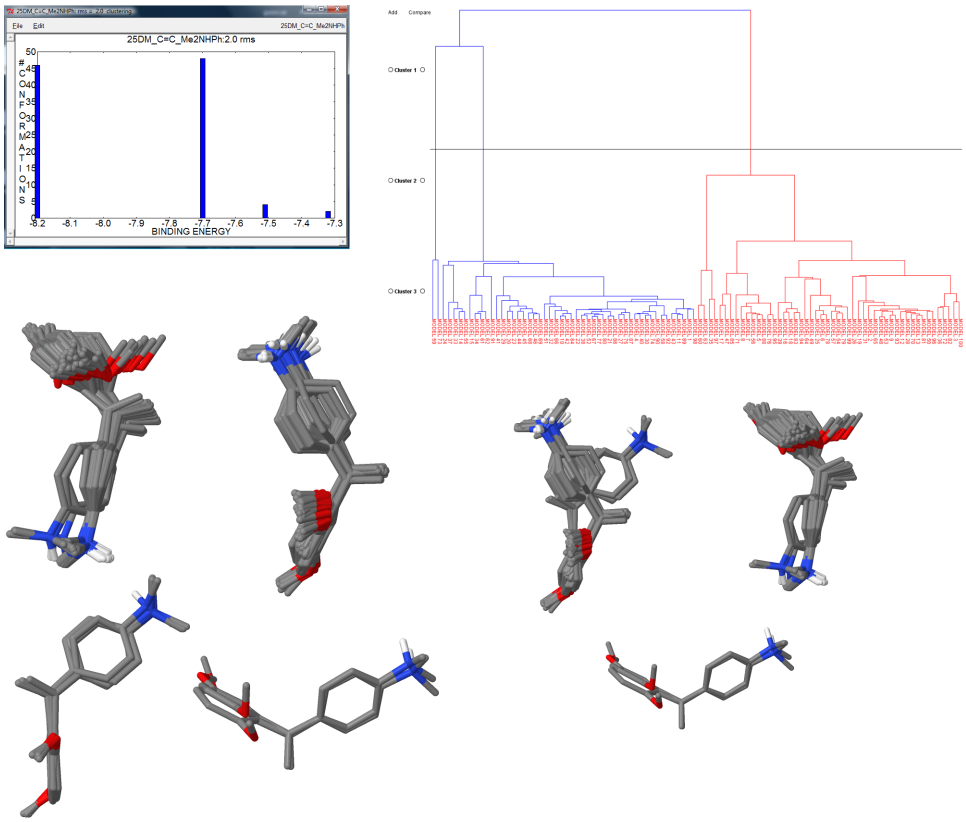


Figura 9.11: A la izquierda *clusters* realizados con el RMSD de ADT. A la derecha *clusters* realizados con nuestra propuesta. Las dos herramientas muestran su ventana de análisis de *clusters*. Sin embargo, en nuestra propuesta es posible cambiar interactivamente el número de *clusters* a explorar con la barra de corte.

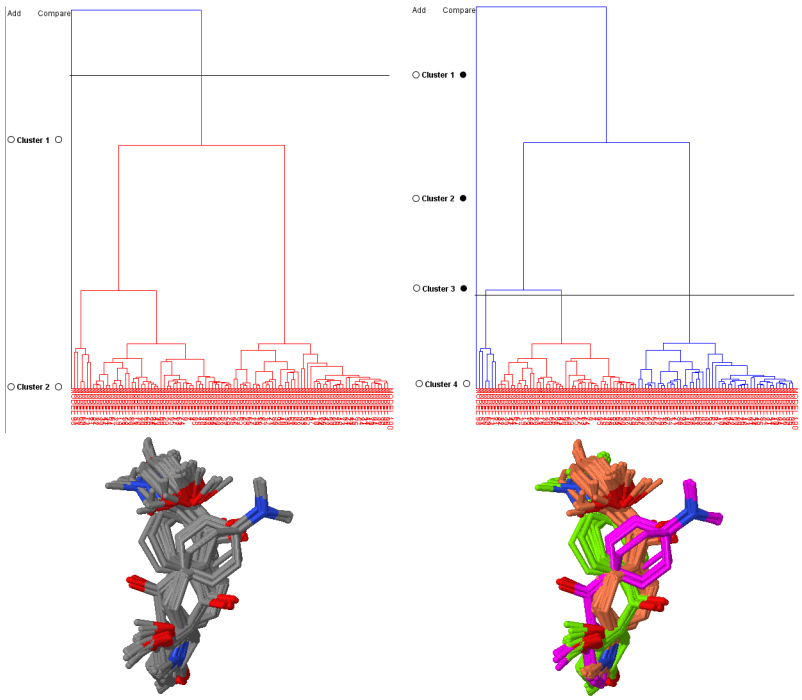


Figura 9.12: Selección manual de nivel a través de la manipulación interactiva de la barra horizontal. El número de *clusters* cambia al arrastrar la barra verticalmente.

9.2.2. Selección de Representantes de Cada Uno de los *Clusters*

Previo al proceso de selección de representantes, es necesario definir el número de *clusters* a tener en cuenta, en otras palabras, el químico definirá qué *clusters* pasarán a la segunda etapa del proceso, como ya se mencionó previamente, se pueden seleccionar de manera manual o automática. En el caso de que el químico haya optado por el método de RMSD y haya seleccionado la manera automática, la herramienta tomará de cada *cluster* aquél elemento que tenga una energía de unión más negativa. En el otro caso, si el químico optó por el método aglomerativo, se tiene que emplear un método diferente dado que no hay un número definido de *clusters*, sino que en la jerarquía existe diferente número de *clusters* formados. El proceso de selección de *clusters* se explica a continuación.

9.2.2.1. Elección de un Número de *Clusters* Adecuado Para el Método Aglomerativo

De acuerdo con la filosofía propuesta, una vez agrupadas las disposiciones en distintos niveles de la jerarquía era necesario elegir un número de *clusters* tal que, seleccionando un miembro de cada *cluster*, quedase suficientemente representado el conjunto total de disposiciones.

Para la solución del problema de selección de *clusters*, se emplearon dos métodos con el fin de estimar el nivel de corte óptimo, esto es, encontrar el nivel en la jerarquía donde estén mejor representados los grupos. El primer método consiste en calcular el grado de separación entre *clusters*

(4) que mencionan Daxin *et al.* [57] en su revisión de análisis de *cluster*. Donde S_1 es el grado de separación entre dos *clusters*, C_1 y C_2 son *clusters* que contienen moléculas, d es la distancia entre una molécula i que pertenece a C_1 y j una molécula que pertenece a C_2. El grado de separación se basa en que los elementos de un *cluster* son muy parecidos pero al mismo tiempo diferentes a los elementos de otro. En otras palabras, la distancia entre los elementos de un *cluster* es mínima pero de mayor distancia con referencia a otro *cluster* o, lo que es lo mismo, todas las moléculas de un *cluster* deben tener la misma orientación y posición en el espacio (ver figura 9.12). Después se calcula la separación promedio de *clusters* para cada nivel (5) definido por Shamir *et al.* [111]; donde S_{ave} es la separación promedio de todos los *clusters*. Finalmente se estima el corte con el valor máximo de separación promedio entre *clusters* (6) de cada nivel.

$$S_1(C_1, C_2) = \sum_{i \in C_1 \wedge j \in C_2} \frac{d(i,j)}{|C_1| \cdot |C_2|} \tag{4}$$

$$S_{ave}(k) = \frac{1}{\sum_{C_i \neq C_j} |C_i| \cdot |C_i|} \sum_{C_i \neq C_j} |C_i| \cdot |C_i| \cdot S_1(C_1, C_2) \tag{5}$$

$$nivel = \arg_k max\{S_{ave}(k)\} \tag{6}$$

Al finalizar el proceso de elección del número de *clusters*, ya sea de forma automática o manual, es necesario seleccionar una o varias estructuras representativas de cada *cluster*, para reducir el número de disposiciones que se manejan en las etapas posteriores de comparación con los de otras moléculas diferentes. Si los *clusters* contienen representantes con una RMSD pequeña, es de esperar que todos los integrantes sean pequeñas variaciones de una misma disposición. Una opción sería elegir la estructura promedio de cada *cluster* como representante. Sin embargo, como cada representante posee una afinidad predicha por Autodock (*free binding energy*, ver figura 9.1), que es la característica que define los representantes más interesantes, se decidió seleccionar en cada *cluster* la disposición con mayor afinidad. Esto permitirá además comparar los *clusters* y etiquetarlos en función de ella, además de obviar la necesidad de calcular una estructura promedio.

9.2.3. *Clustering* de Representantes

Con el programa AutoDock se pueden realizar de forma automatizada estudios de *docking* sobre una misma diana con múltiples ligandos. Estos ligandos pueden ser muy diferentes en cuanto a su estructura y tamaño (tipo y número de átomos). El problema que se plantea es agrupar estas moléculas diversas en función de cómo han sido orientadas en el hueco durante el proceso de *docking*. De esta manera, se podrán encontrar características comunes a todas ellas, facilitando el análisis de los resultados. Para agrupar moléculas con diferente tipo y número de átomos no puede utilizarse el RMSD como en la sección 9.2.1, ya que este cálculo se basa en la identidad de los átomos en las diferentes disposiciones. Como el problema reside en que el número de elementos (número de átomos de cada tipo) no es común en las diferentes moléculas, se propusieron dos estrategias para llevar a cabo el agrupamiento de moléculas diferentes: 1) reducir el número de elementos a un máximo que todas las moléculas poseyeran y calcular las diferencias intermoleculares teniendo en cuenta solamente estos y 2) proyectar la información molecular en un número fijo de elementos, de forma que todas las moléculas poseyeran el mismo número de descriptores y calcular la distancia

intermolecular a partir de estos. A continuación se describen los resultados encontrados para ambas aproximaciones.

Para reducir el número de elementos descriptores de cada molécula a un número que todas poseyeran, la primera consideración a tener en cuenta es el número de átomos que pueden tener los ligandos más pequeños. El problema reside en que, en los experimentos de *docking*, se emplean desde fragmentos pequeños (entre 50 y 250 Daltons de masa, o lo que es lo mismo, entre 5 y 25 átomos pesados como máximo) hasta compuestos con propiedades tipo fármaco (hasta 500 Daltons de masa, es decir, hasta 40 átomos pesados) o incluso mayores, pasando por lo que se conoce como compuestos con características de líder (intermedias entre los anteriores). Además, es habitual mezclar unos y otros en el mismo experimento. Por ello, se decidió establecer como parámetro configurable el tamaño mínimo a considerar en función de la biblioteca de compuestos utilizada. Inicialmente se decidió seleccionar los 20 átomos de cada molécula con interacciones más favorables con la proteína para calcular las distancias entre moléculas, bajo la asunción de que situaciones similares presentarían sus átomos más favorables en posiciones similares. Posteriormente es necesario calcular la distancia mínima entre cada pareja de disposiciones, para lo cual, se ordenaban los valores de acuerdo a su valor de interacción, a continuación se empleaba la fórmula de la distancia Euclidia para construir la matriz de distancias (matriz simétrica). Al comparar moléculas con número de átomos diferentes se observó que el método no funcionaba correctamente, por lo que se optó por considerar la segunda opción.

Para calcular la similitud/distancia entre dos conformaciones de moléculas distintas (diferente tipo y número de átomos) mediante un número fijo de elementos externos, se modificó el método propuesto por Jain [56], que se basa en el sumatorio de las distancias desde los átomos de las moléculas a puntos "observadores", fijados independientemente de la molécula o disposición considerada. En principio, se consideró que los puntos del *grid* reunían las características de los observadores. Sin embargo, su elevado número hace impracticable la opción de utilizarlos todos, por lo que se decidió su reducción a un conjunto formado por los que tuvieran un mayor significado desde el punto de vista químico. Así, se decidió seleccionar aquellos puntos en el espacio de los ficheros tipo *map* (ver figura 9.3) que correspondieran a un mínimo local (energía de interacción lo más favorable posible para ese tipo de átomo) de un tamaño suficiente (para no considerar aquellos puntos en los que, por su reducido volumen, es improbable que sean ocupados por átomos de las moléculas).

El proceso de selección es el siguiente: en primer lugar se seleccionan todos los valores negativos (representativos de una interacción favorable), como se explicó en la sección 9.1. Una vez que se obtienen estos valores, se forman nubes de puntos (zonas en el espacio con muchos puntos). El proceso de creación de estas nubes empieza seleccionando el valor más negativo de todos los puntos (será un punto en el espacio que se utilizará como centro de una esfera). Posteriormente se crea una esfera de un radio predeterminado (se propuso que fuera tres veces la separación entre puntos de la malla, ver figura 9.3); al hacer esto se evitan nubes muy pequeñas o que tengan una forma alargada, tal como se muestra en la figura 9.13; los puntos que estén dentro de la esfera son seleccionados, almacenados y finalmente eleminados, para que no sean considerados en las iteraciones siguientes. Este proceso continuará hasta terminar con todos los puntos en el espacio. Por otra parte, estas nubes son empleadas por la herramienta como ayuda visual para la creación de nuevas zonas favorables, las cuales se pueden emplear para la búsqueda por farmacóforos (se explicará en detalle en la sección 9.3).

Por lo tanto, un observador será aquel punto de la nube con el valor más negativo. Para cada fichero *map* se tendrán *n* observadores; por ejemplo si se forman cincuenta nubes entonces habrá

cincuenta observadores de ese fichero *map*, y de esta manera habrá observadores de todos los átomos presentes en las moléculas. Evidentemente, el número de observadores dependerá del número de *maps* (tipos de átomos) presentes en el conjunto de moléculas. Por tanto, se ha dejado en manos del usuario la selección o descarte de algunos de ellos.

Este proceso concluye con un número fijo de observadores (posicionados en el *grid* utilizado por Autodock) que corresponden a cada tipo de átomo, pero el total es fijo. Para el proceso de cálculo de las distancias se puede calcular la distancia mínima a cada observador de cada átomo de ese tipo de la molécula en consideración. Esta opción excluiría el reemplazo de átomos por otros químicamente similares, por lo que una segunda opción que se estudió fue el cálculo de la distancia mínima a cada observador de cualquiera de los átomos que se consideraron a priori similares a ese tipo de la molécula en concreto. Así, los carbonos aromáticos (*maps* A) y los no aromáticos (*maps* C), considerados de manera diferente por Autodock, podían ser considerados como químicamente asimilables. Lo mismo se hizo para oxígenos y nitrógenos y también se agrupó los halógenos en un grupo común (Cl, Br, I y F). Por tanto, los grupos formados como observadores en nuestra propuesta son:

- C, A, N
- OA, SA, NA
- F, Cl, Br, I
- HD
- e

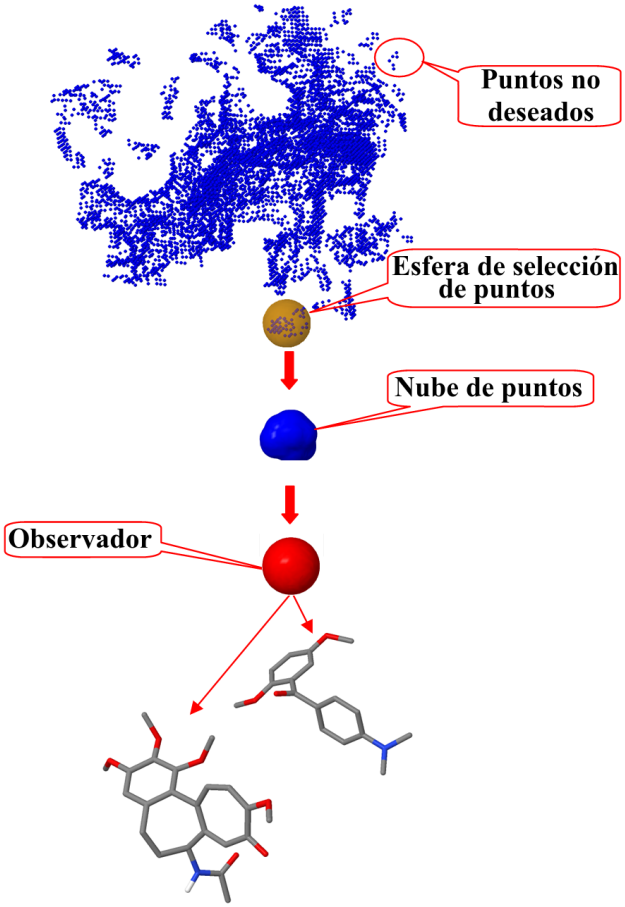


Figura 9.13: Creación de nubes de puntos. Cada nube se crea a partir del punto en el espacio que contenga el valor más negativo de todos. Posteriormente de cada nube se extraen los observadores.

Para el cálculo de la similitud entre un observador y las moléculas puede utilizarse una función gaussiana inversa centrada en el observador, de modo que a medida que el átomo se aleja más de una cierta distancia (configurable), del observador la similitud se hace mínima. Al considerar las distancias calculadas por este método, se observó que átomos situados muy cerca en el espacio podían recibir puntuaciones bastante diferentes, debido a la pendiente de la curva. Se pensó entonces en recurrir a un sistema de puntuación basado en escalones, que pese a incrementar de golpe las diferencias, mantenía constante la puntuación en tramos de distancia interatómica que se consideraban no significativos. El cálculo se llevó a cabo de acuerdo con la siguiente expresión (7).

$$sim = min_i^a\,(d_{j\in a}, a_i) \begin{cases} \text{si } d > 0 \;\&\; d \leq 0{,}2 & \text{entonces } d \\ \text{si } d > 0{,}2 \;\&\; d \leq 0{,}5 & \text{entonces } d \times 0{,}75 \\ \text{si } d > 0{,}5 \;\&\; d \leq 1 & \text{entonces } d \times 0{,}5 \\ \text{si } d > 1 \;\&\; d \leq 2 & \text{entonces } d \times 0{,}2 \\ \text{si } d > 2 & \text{entonces } d = 0 \end{cases} \quad (7)$$

Donde d es la distancia (Euclidia) mínima de un observador O de tipo *map* a cada átomo de una molécula (figura 9.14), O_i es un punto observador de tipo *map*, a_j es cada uno de los átomos de la molécula de la que se calcula la similaridad; *sim* es la distancia mínima calculada entre un observador y la molécula. Por ejemplo, si se seleccionan cuatro ficheros de tipo *map* como H (hidrógeno), S (azufre), A (carbonos aromáticos), y C (carbonos no aromáticos) y hay 10 observadores de tipo H, 25 de S, 8 de A y 14 de C, el número total de observadores sería 57. Ahora supongamos que empieza primero con H, cada observador calculará la distancia a todos los átomos de la molécula observada (figura 9.14) y sólo se almacenará la distancia que sea más cercana al observador; una vez obtenidas esas distancias, se aplicará la formula (7) para determinar el rango en el que están. Esto es, calculando todos los átomos de la molécula, sin embargo, en la sección 9.2.4 se da una descripción detallada cuando sólo se utiliza unos cuantos observadores y sus diferentes resultados.

Una vez establecido el método de cálculo de la distancia, es posible obtener una matriz de distancias de cada disposición a todos los observadores (matriz de observaciones). A partir de ella se calcula la distancia entre disposiciones de distintas moléculas. En otras palabras, a partir de la matriz de observaciones se obtiene la matriz simétrica para que sirva de entrada al algoritmo de *clustering*. Este método de cálculo resuelve algunos de los problemas encontrados con el cálculo de distancias mediante RMSD: permite sustituir átomos por sus equivalentes en sistemas con elementos de simetría y permite en cierta medida sustituir átomos o grupos de átomos por otros de propiedades similares si se selecciona la opción de agrupar distintos tipos de átomos con similares propiedades químicas.

Un aporte extra que se hace es añadir las coordenadas de los átomos de las referencias y considerarlos como observadores. Las referencias son moléculas para las que se conoce su modo de interacción con la diana, por lo que sirven como punto de partida para comparar los resultados del *docking* y del método de *clustering* entre moléculas diferentes.

Es preciso remarcar la flexibilidad que posee esta herramienta puesto que permite al usuario con experiencia decidir qué tipo de observadores van a ser considerados en el proceso. La figura 9.15, muestra la ventana de configuración para la agrupación de representantes. En la parte derecha aparecen distintas opciones de selección para los observadores. Por defecto aparece seleccionada la opción *all observers*, que utiliza todos los observadores sin importar el tipo de átomo a evaluar. Si observadores distintos (*map*) ocupan el mismo lugar en el espacio, evaluarán igual a un átomo de una molécula, lo que significa que esa molécula tendrá buena interacción si tiene un átomo del tipo de *map* que la observó. La segunda opción *Each atom* restringe los observadores considerados para que sólo evalúen un tipo de átomo en concreto, La tercera opción *Group Atoms* permite llevar a cabo la evaluación por grupos de átomos. Por otro lado, la herramienta permite visualizar de forma simultánea los observadores, las referencias y la proteína (ver figura 9.16), de manera que el usuario puede hacerse una idea general de todos los elementos que participarán, tanto para calcular la similitud como del posible resultado final previamente a la selección de una de las tres opciones.

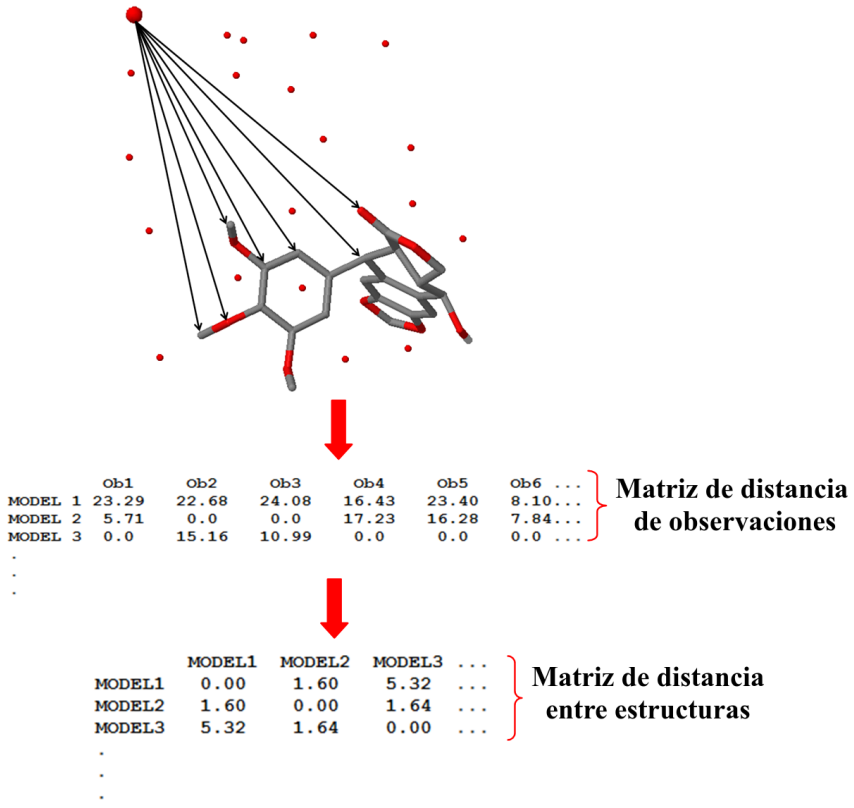


Figura 9.14: Cálculo de distancia de un observador a cada átomo de una molécula. Generación de la matriz de observaciones de estructuras y matriz de distancias de las observaciones de estructuras.

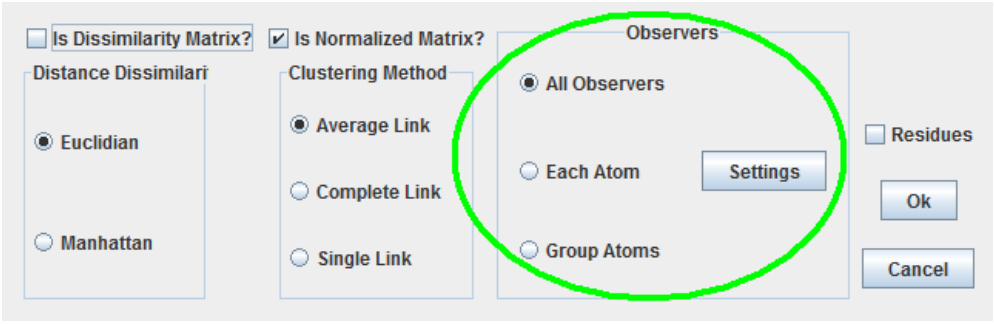


Figura 9.15: Opciones para los observadores.

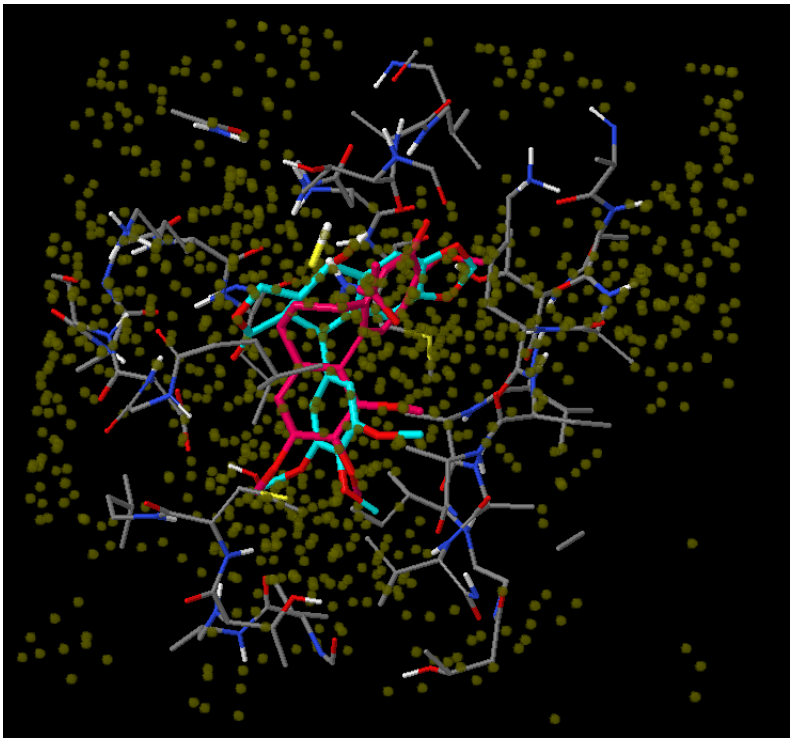


Figura 9.16: Visualización de los observadores, referencias y proteína.

9.2.4. Manipulación de los *Clusters*, los Representantes y sus Visualizaciones

Una vez elegida una de las tres opciones para la selección de los tipos de observadores, el resultado final es analizado a través de un sistema de visualización múltiple de vistas enlazadas para facilitar el análisis de los resultados. A través de estas vistas se pueden observar simultáneamente las estructuras de las moléculas evaluadas, las moléculas utilizadas como referencias y la proteína. En la figura 9.17 se observan tres ventanas diferentes, la de la izquierda es el visor de estructuras Jmol; a la derecha están las vistas enlazadas con los resultados del *clustering* en forma de dendrograma. Para formar un nivel que contiene los *clusters* (unión entre elementos, elemento con *cluster* o *cluster* con *cluster*) el valor de distancia más pequeño que no sea cero es tomado para hacer la unión y formar un *cluster*; este procedimiento continua para calcular una nueva matriz de distancias y repite estos dos pasos hasta formar todos los niveles de la jerarquía. En la parte superior derecha de la figura 9.17 se muestra el dendrograma general, mediante el cual se exploran por nivel el número de *clusters* formados; también en color café y verde se resaltan las referencias, y su correspondencia en Jmol a la izquierda. En la parte inferior derecha se muestra la ampliación de la extracción del *cluster* seleccionado en color rojo.

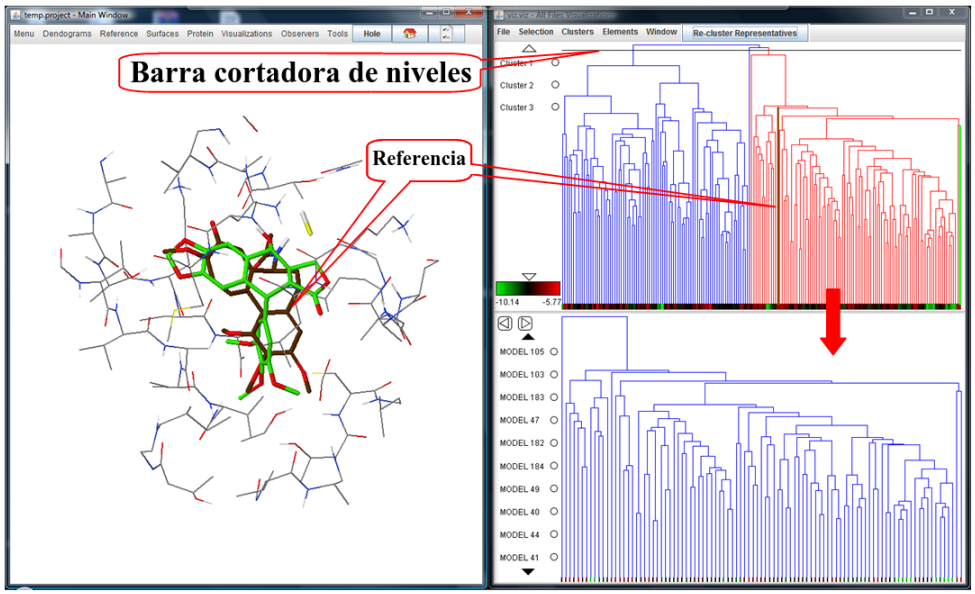


Figura 9.17: Vista general del resultado de la evaluación de los observadores a través del algoritmo de *clustering*.

La herramienta permite visualizar las estructuras de las moléculas de forma simultánea con el dendrograma. Dado que son vistas enlazadas permiten interacción entre ellas. En otras palabras, al ir cortando con la barra horizontal (figura 9.17) el número de *clusters* cambia (aumenta o disminuye) y son mostrados mediante etiquetas a la derecha del dendrograma, el usuario podrá seleccionar o comparar los *clusters* mediante estas etiquetas o seleccionando ramas individuales, lo cual facilita el análisis de los resultados. Por otra parte, la herramienta permite la introducción de las referencias al conjunto de moléculas bajo estudio; para diferenciarlas, se codifican siempre en un color específico como por ejemplo, el café, verde u otro, tanto en el dendrograma como en el visor. Así, el usuario puede determinar los grupos que están cerca o se superponen a las referencias. Otra de las facilidades que ofrece la herramienta, es una barra de colores que aparece debajo del dendrograma que indica la energía de unión de cada una de las moléculas (*free binding energy*) con la diana, de acuerdo al color de energía puede ser un factor de decisión para la selección (figura 9.17 y figura 9.18) de *clusters*; la barra de colores que aparece en la parte inferior de la izquierda del dendrograma marca el rango de valores de interacción de las moléculas con la diana; el color verde indica el valor más negativo y el rojo el menos negativo (figura 9.18). En otras palabras, el color verde siempre significará el mejor valor de interacción que existe en ese conjunto de moléculas, mientras el color rojo significará lo contrario. De esta manera el usuario dispone de una visión completa de la información. Esta visualización es de gran interés químico, ya que permite a vista de pájaro determinar qué ramas del dendrograma contienen disposiciones que Autodock analiza como favorables. En la figura 9.18 se puede ver la barra de la figura 9.17 ampliada, en la que se observan cinco zonas donde se agrupan los resultados más favorables. Es notable que las referencias (líneas marrones) se sitúen en una zona del dendrograma poco favorable (indicado por el color rojo). Esto puede sugerir la necesidad de modificar el sistema, en un intento de optimizar su comportamiento, sobre todo si las moléculas cercanas a la referencia en el dendrograma corresponden al *docking* con ellas mismas.

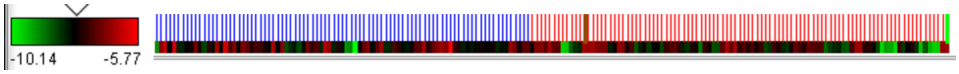


Figura 9.18: A la izquierda, barra de colores que marca los valores de interacción de un conjunto de moléculas y la proteína; el verde es el mejor valor de interacción mientras que el rojo es el de menor interacción. A la derecha una barra de colores que representa el valor de interacción de cada estructura.

Por otro lado, en la parte inferior aparece otra ventana (figura 9.17), que muestra una vista ampliada del grupo seleccionado en el dendrograma principal; además, se puede profundizar/filtrar y explorar otras ramas de ese subgrupo seleccionado con esa vista. Por ejemplo, si se ha seleccionado el *cluster 3* y después en la ventana de ampliación se selecciona una sub-rama, automáticamente las tres vistas se modifican e interactúan entre sí, esto es, en la ventana superior cambia el color de la sub-rama seleccionada en la parte de abajo (de rojo a amarillo) y consecuentemente se hace una ampliación de esa rama en la parte de abajo; a su vez se envía una notificación al visor y este muestra las moléculas correspondientes a esa sub-rama (ver sección 10.4). Una vez que el usuario ha terminado con el análisis de esa sub-rama y le interesa otra, la herramienta permite ir hacia la rama anterior o hasta la principal, modificándose interactivamente todas las vistas. Entre otras opciones, la herramienta permite seleccionar una de las moléculas del análisis y considerarla como una futura referencia en posteriores análisis. El proceso es el siguiente, supongamos que se ha analizado un *cluster* y se considera que por sus características tanto visuales como químicas, bien podrían valer como una referencia, entonces se procedería a crear la referencia a partir del *cluster*, en otras palabras, se seleccionaría del *cluster* aquella molécula que tenga mejor interacción con la diana y se procedería a crear un fichero PDB de esa molécula. Incluso, si el usuario lo desea, puede crear un fichero que contenga todo el *cluster*.

Si es preciso hacer una exploración en particular sobre un *cluster* (por ejemplo, uno que contenga las referencias), como en el caso de la figura 9.19a, se muestra la selección de la última rama (*cluster*) de la derecha (color rojo), esas estructuras han sido extraídas y visualizadas en una nueva ventana (figura 9.19b). Continuando con el análisis de esas estructuras, se han seleccionado los *clusters 4* y *cluster 6* (círculo verde) y se ha aplicado de nuevo el proceso de comparación de moléculas diferentes con los observadores; sin embargo, se optó por utilizar la opción que restringe a evaluar solamente a un tipo de átomo y empleando el algoritmo de distancia máxima (*complete linkage*). El resultado final de evaluar los átomos tipo A, C, HS, Cl y F se muestra en la figura 9.19c. Este proceso de filtrado y evaluación con los distintos métodos ha llevado a encontrar aquellas estructuras que están en la zona de las referencias (color verde y café). Y aun así, es posible seguir filtrando hasta encontrar nueva información.

Supongamos que de todos los ficheros dlg hay uno en el que sus poses se emplearían como control, esto es, que los representantes de ese dlg cumplen con ciertos requisitos o bien, por alguna razón se toman como punto de partida. Lo ideal sería tener todos los resultados de los *dockings* esparcidos en los mismos sitios en los que el control tiene situados los suyos. Para lograr esto, la herramienta a través de una ventana muestra los dlgs que han participado, y, al seleccionar uno de ellos, estos se resaltan en colores en el dendrograma, mientras el resto cambia a un color gris tenue, facilitando su localización en el dendrograma (se ha adoptado aquí una estrategia de foco+contexto, para evitar que al analizar el detalle del problema se pierda la visión global del mismo). Por otra parte, las referencias se mantienen resaltadas, por lo que, si el químico selecciona el dlg de control, confirmaría, si los elementos (representantes) quedaron dentro del *cluster* donde se encuentran las

o, por lo menos, una referencia. Continuando con el ejemplo anterior, ahora estaría en el interés del químico localizar aquellos representantes que cumplan las dos condiciones: que estén cerca de las referencias e identificar qué representantes están en el mismo *cluster* que el de control. En la figura 9.20, se muestran dos dlgs seleccionados en los que claramente se aprecia que no están cerca de las referencias (color café y verde).

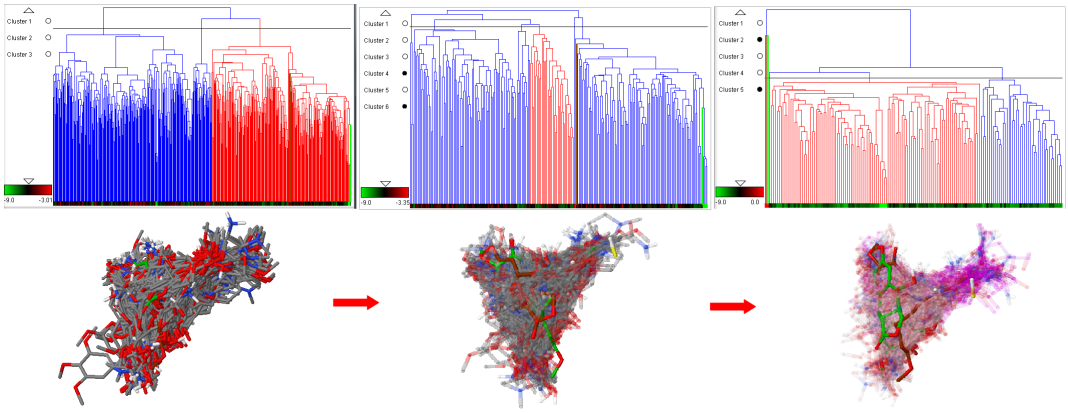


Figura 9.19: a) Selección y filtrado de moléculas cercanas a las moléculas de referencia. b) se han seleccionado los *clusters* 4 y 6 (círculos en negro) para un análisis en más detalle. c) el resultado final al emplear la opción de evaluar un tipo de átomos por los observadores.

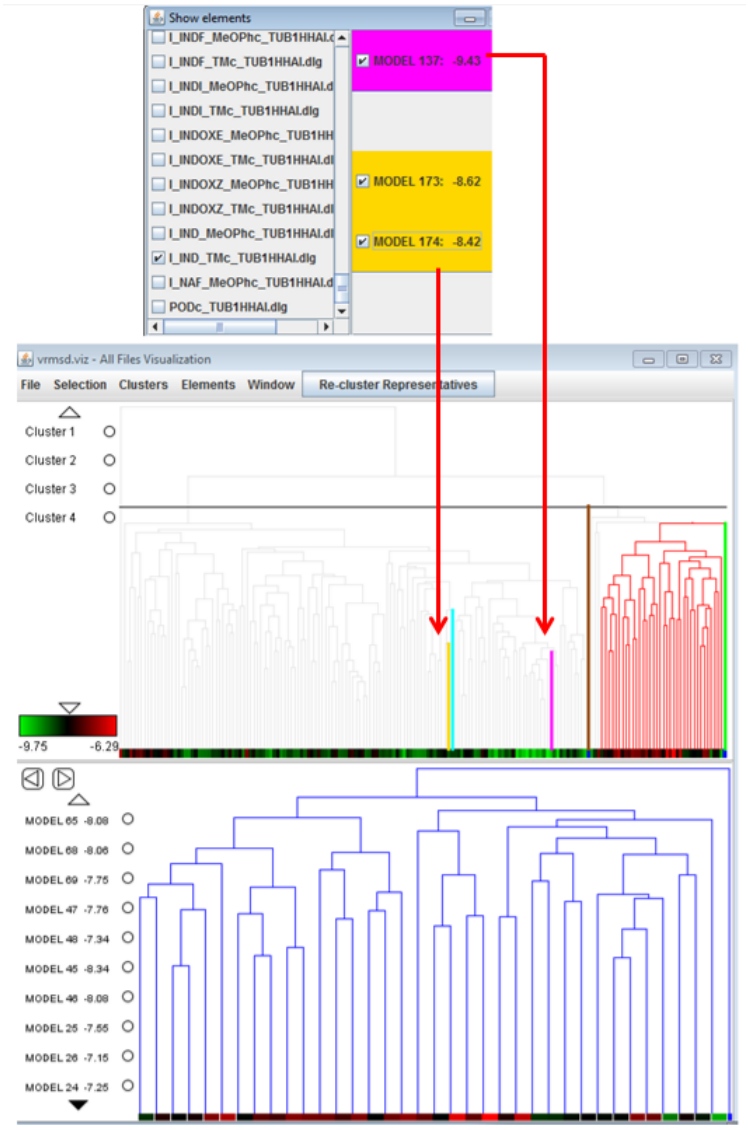


Figura 9.20: Elementos seleccionados de algunos *clusters*. El fondo cambia a un color gris claro para resaltar los elementos seleccionados (foco+contexto), de ésta forma el químico tiene un panorama general de cómo se agruparon los representantes de cada dlg.

Todas las opciones que proporciona la herramienta ayudan al químico a hacerse una idea general del contexto de los resultados de los *dockings*. Así, da paso al refinamiento del análisis, esto es, a seleccionar un número manejable visualmente, porque es mucha la información presentada y, de otra forma, sería imposible hacer un análisis a mayor profundidad. De esta manera, es posible realizar un filtrado reduciendo el número de representantes por dlg; en otras palabras, el químico podrá decidir el número de elementos en el dendrograma. Para realizar esto, la herramienta cuenta con cuatro

tipos de filtrado para reducir el número de representantes (ver figura 9.21). Similarmente, en la parte inferior de la ventana de control se muestra el número de representantes antes y después de realizar el filtrado. Las opciones de filtrado son las siguientes:

1. **RMSD.** El químico podrá reducir el número aplicando una distancia de RMSD que él elija. Este proceso es similar a la opción del primer paso de agrupar por el RMSD de ADT.

2. **Energía de unión (*Score*).** Mediante una barra deslizadora el químico selecciona representantes de acuerdo al puntaje de unión de Autodock. El rango para seleccionar los representantes cae entre el representante con la energía de unión más negativa y la menos negativa. Hay que remarcar que sólo aquellos representantes que estén en el rango serán seleccionados, por lo que es posible que algunos dlgs queden fuera y no estén presentes en el resultado del filtrado.

3. **Primeros con mejor energía de unión (*High Score, First*).** Con esta opción, el químico selecciona un número determinado, del cual, la herramienta tomará de cada dlg sólo los primeros tantos de acuerdo al número introducido. Por ejemplo, podría ser el caso que se deseen los 10 primeros representantes con mejor energía de unión -en otras palabras, sólo le interesan los 10 primeros con energía de unión más negativa-.

4. **Número de *clusters* (*Number of clusters*).** Finalmente, la aplicación permite al químico seleccionar por número de *clusters*. Por ejemplo, el químico podría pedir a la herramienta que seleccione 5 *clusters* de cada dlg. El método de selección se basará en los *clusters* con mejor energía de unión, en otras palabras, seleccionará aquellos *clusters* con energía de unión más negativa. Si llegará a darse el caso en el que el número de *clusters* por dlg es menor al introducido, todos los *clusters* estarán presentes en el resultado del filtrado.

Finalmente, si el químico requiriere realizar un segundo filtrado, podría hacerlo sobre el filtrado actual o regresar al número de representantes original, ya que la herramienta cuenta con esta opción.

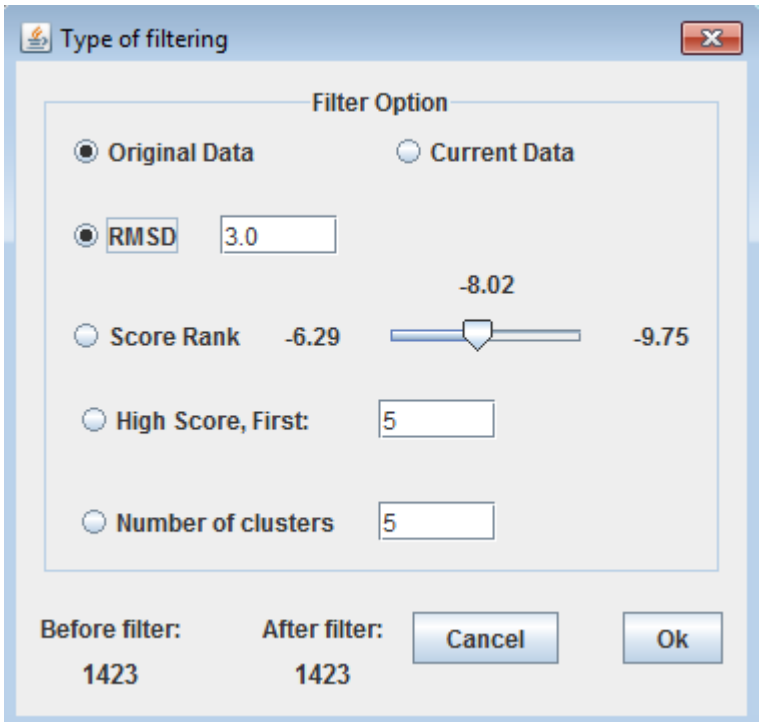


Figura 9.21: Ventana de control para filtrado de representantes.

9.3. Refinado, Modificación y/o Creación de Nuevos *Maps*

Puesto que la aplicación está desarrollada para manipular los *maps* de Autodock, y estos son la base a partir de la cual el programa calcula las disposiciones y su energía de interacción, se consideró que una opción muy favorable podría ser proporcionar al usuario la capacidad de manipular la información de los *maps* para incorporar información adicional que no puede deducirse de la imagen estática de una disposición única de la proteína. Así, la aplicación puede permitir incorporar al cálculo de Autodock la información de modelos adicionales (*docking* múltiple) en un único experimento, incluir la información de relaciones-estructura actividad, o incluso, a partir de resultados de *docking* de las moléculas de referencia (como se ha indicado anteriormente, si el programa no predice adecuadamente los resultados experimentales y es posible atribuirle una explicación estructural). En último caso, cuando no se disponga de información tridimensional sobre la diana, la aplicación proporciona un medio para poder definir farmacóforos en forma de zonas en los *maps* atómicos configurados a voluntad por el usuario. Para ello, la visualización simultánea de la proteína, las referencias y las zonas de la proteína que han sido seleccionadas como favorables (información extraída de los ficheros *map*) proporciona una ayuda inestimable. Con esta información presente se diseñarán nuevas zonas favorables que a su vez servirán de retroalimentación al proceso de *docking*.

Supongamos que se tiene un fichero de tipo *map* al que se desea modificar sus zonas de interacción con la proteína. Para realizar esto, sería útil poder visualizar qué zonas de interacción son

favorables (figura 9.22 parte superior), sin embargo, no bastaría con sólo visualizarlas, habría que resaltar la diana de la proteína (figura 9.22 parte inferior izquierda). Por otra parte, debido al tamaño de la proteína, es preferible visualizar solamente la diana así como las moléculas de referencia (figura 9.22 parte inferior derecha). La herramienta que se propone, ofrece la facilidad de modificar de manera visual e interactiva las zonas de interacción con la proteína, pero además de visualizar y manipular, también puede salvar en ficheros esas nuevas zonas para posteriormente ser cargadas en el mismo proyecto o en uno nuevo.

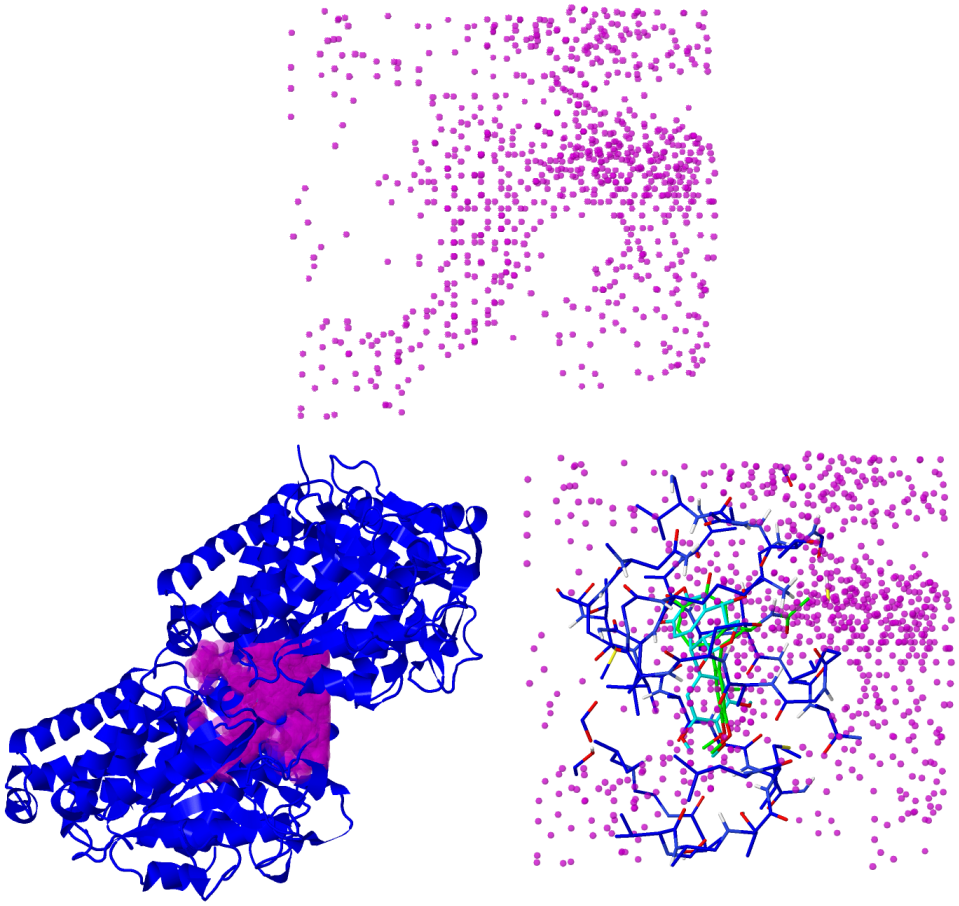


Figura 9.22: En la parte superior las zonas de interacción con la diana favorables. En la parte inferior izquierda, la proteína en *cartoon*, las zonas favorables en forma de nubes en la diana. A la derecha la diana en forma de hilos con las zonas de interacción favorables y las referencias (colchicina en verde y la podofilotoxina en cyan).

9.3.1. Visualización y Manipulación de Zonas Favorables de los *Maps*

La herramienta cuenta con dos ventanas que interaccionan con el visualizador molecular y permiten manipular los *maps*. Una ventana le permite al químico manipular las zonas favorables; en otras palabras, la manipulación consiste en mostrar u ocultar las zonas según se requiera (ver figura 9.23). Hay dos formas de manipular la visualización:

1. A través de un botón de seleccionar/deseleccionar; esta opción hará que todas las zonas de un tipo de *map* se oculten o se muestren dejando visibles al resto. La utilidad de esta opción radica en que el químico podrá hacer comparaciones de zonas y así decidir una estrategia de diseño para las nuevas zonas.

2. ADT permite mostrar las zonas favorables y mediante una barra deslizadora el químico puede ir filtrando para mostrar en un rango los puntos que caen dentro. Nuestra propuesta incorporó esa idea y se diseñó una opción similar. A diferencia de ADT, nuestra herramienta muestra puntos en lugar de figuras triangulares, pero su funcionamiento es básicamente el mismo. Esta opción permitirá al químico definir qué zonas del *map* son sitios de interés para modificarlos. Por ejemplo, para un *map* se selecciona un rango y esto ocasiona que todos los puntos que no están dentro de él se oculten. Y dado que en la visualización se tiene a las referencias y a la diana, es posible definir con exactitud nuevas zonas, tanto favorables como no favorables, pues la suma de todos los elementos visuales (proteína, referencias y zonas favorables) ayudan a diseñar sitios estratégicos, para posteriormente emplearlos en búsquedas por farmacóforo.

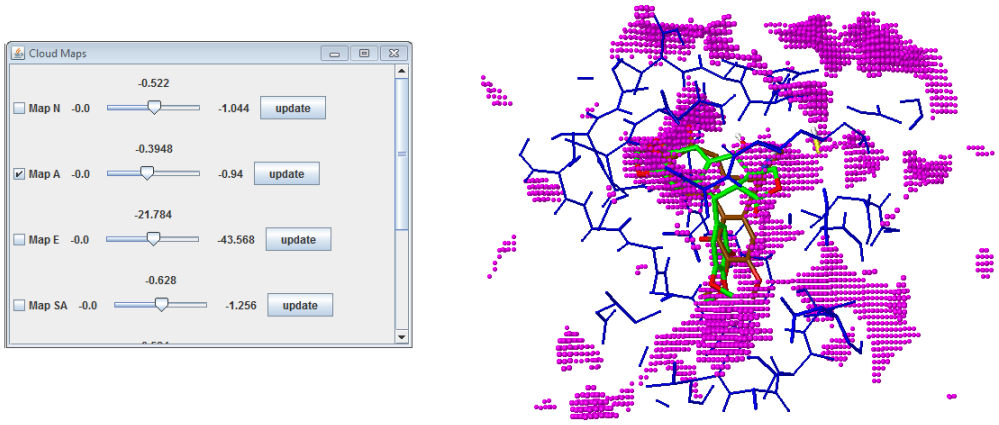


Figura 9.23: Ventana de control para visualizar zonas favorables.

9.3.2. Diseño de Farmacóforos

La segunda ventana le permitirá al químico diseñar nuevas zonas a través del control de configuraciones de diseño. Al poder observar las zonas de interacción más favorables, el químico podrá agregar de manera interactiva nuevos elementos al farmacóforo. La idea principal es poder crear zonas distintas (favorables o desfavorables) a las del programa de *docking*. Estas nuevas zonas se

definirán visualmente mediante esferas que se corresponderán con un tipo de átomo, y, una vez definidas, modificarán o generarán un fichero *map* con esta nueva información. La presencia y observación del sitio de la diana le ayuda a determinar los límites en el espacio disponible. De esta manera no pierde la información de las referencias por lo que puede ir diseñando farmacóforos por tipo de átomo, incluso superponer esferas de diferente tipo de átomo. Esta modificación interactiva puede realizarse fácilmente a través de la ventana de control, puesto que permite seleccionar, agregar, eliminar, cambiar de lugar las esferas e incluso salvar las esferas en un fichero para posteriormente volver a cargarlas en el mismo diseño o en uno nuevo. La aplicación se ha diseñado para que el usuario pueda cambiar de lugar las esferas con la ayuda del puntero del ratón y, posteriormente, pueda utilizar el control para realizar movimientos más finos. Por otro lado, también cuenta con rotaciones de las vistas para facilitar la colocación de las esferas en el sitio deseado. Más aun, para no perder la orientación, el químico dispone de ejes de coordenadas para guiarse y así saber en qué sentido está moviendo las esferas.

En la figura 9.24 se muestran tres esferas de distinto tamaño, así como las referencias y diana. Cada esfera representa un conjunto de puntos que le indicarán a Autodock que existen unos sitios en un rango de valores, partiendo desde el centro de la esfera con el valor más negativo y degradándose hasta llegar a la superficie de la esfera con el valor menos negativo. De esta forma, el químico define los valores de ciertas zonas de su interés. El químico podrá definir el rango de valores proporcionando el tamaño del radio de la esfera, así como el valor máximo negativo y minino negativo. Además, se le proporciona al químico la información necesaria para definir el rango de valores; en otras palabras, la herramienta muestra el valor máximo negativo y positivo de cada *map*, facilitando así el diseño del farmacóforo. En la figura 9.24 se han definido tres esferas -cada color representa un fichero *map*-, lo cual quiere decir que solamente sobre esos *maps* se modificarán o crearán nuevos ficheros tipo *map*.

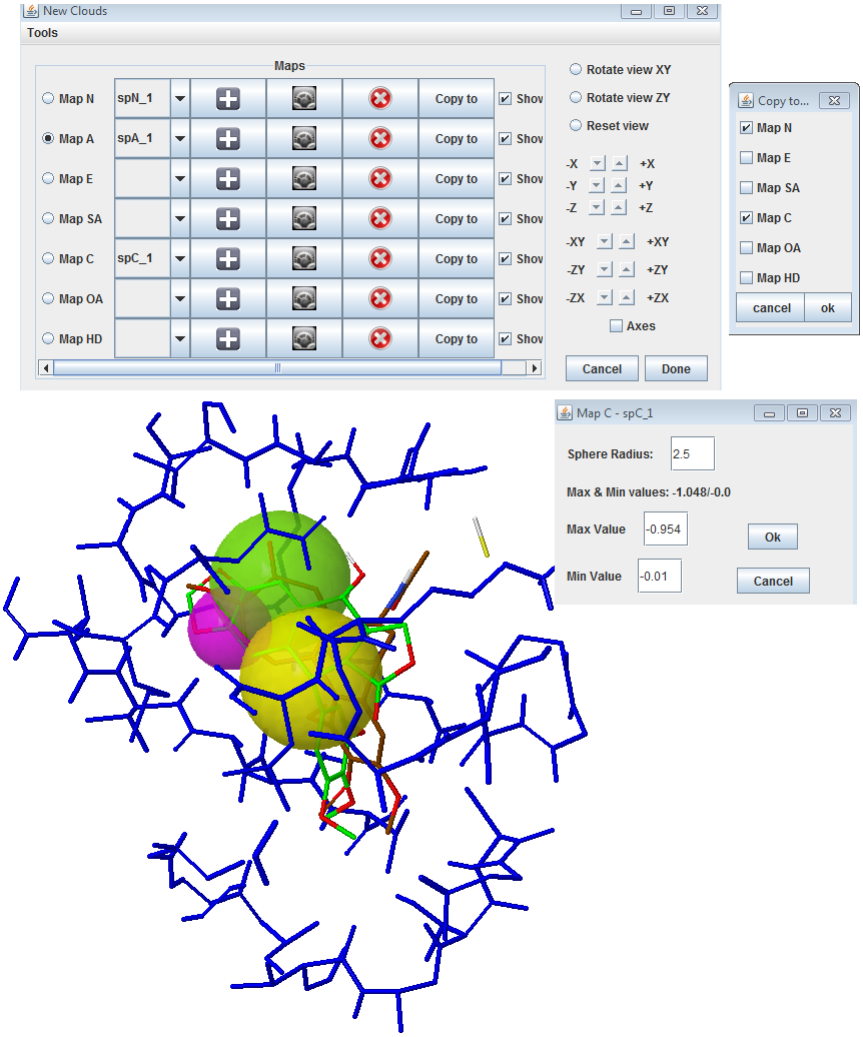


Figura 9.24: En la parte superior la ventana de control, al lado derecho la ventana de copiado a otros tipos de átomos; en la parte inferior la ventana de configuración de las esferas (define el tamaño, valores máximos y mínimos de interacción con la proteína). En la parte inferior izquierda las nuevas zonas favorables diseñadas en forma de esfera; hay tres zonas nuevas que modificarán los ficheros tipo *map* (fichero A, N y C). Las esferas pueden ser reposicionadas mediante el uso del ratón o mediante la ventana de control.

9.3.3. Diseño de Farmacóforos por Grupo de Esferas

Dentro de las opciones del diseño de farmacóforos, cuenta con la creación de grupos de esferas en un solo paso. Supongamos que el químico desea crear un conjunto de esferas con forma hexagonal, para realizar esto manualmente tomaría un tiempo considerable si se realizara empleando la ventana de control descrita en la sección anterior. La aplicación cuenta con una forma sencilla de generar estos grupos a través de una ventana de control para grupos. Esta ventana de control (figura 9.25), cuenta con las siguientes opciones:

1. Cinco formas geométricas predefinidas: triangulo, cuadrado, pentágono, hexágono y octágono, además de permitir al usuario definir el número de esferas. Estos grupos son creados alrededor de una esfera central, la cual puede eliminarse según el diseño del químico sobre ese grupo de farmacóforos.

2. Rotación del grupo de esferas: una vez puestas en un lugar, el químico puede rotar las esferas definiendo los grados a rotar; las rotaciones pueden ser en los tres ejes (x, y, z).

3. Traslación del grupo de esferas: al igual que la rotación del grupo, el químico puede mover en los tres ejes (x, y, z) al grupo de esferas, además de poder definir la distancia del traslado.

4. Eliminación de esferas del grupo (figura 9.26): si el químico deseara replicar la figura exacta de un ligando de referencia, la forma más sencilla es empleando el diseño de grupos con el control anterior, para esto sería necesario poder eliminar algunas esferas para no superponer unas con otras. De esa forma el químico podría obtener la forma exacta del ligando.

Finalmente, es posible dar un orden de escritura a cada esfera creada, esto no aplica a los grupos de esferas, solo al método anteriormente descrito. Esta opción se planteó porque al escribir las esferas, si estas se superponen modificarían los valores de la anterior esfera, lo cual generaría un resultado desfavorable al ser introducido en un ensayo de *docking*. Una vez terminado el proceso de diseño de los farmacóforos se continuaría con la generación de los nuevos *maps*.

En este punto, se le ofrecen tres opciones al químico, la primera consiste en modificar el fichero *map* para sobrescribirlo incorporando la información de la esfera. En otras palabras, la herramienta buscará cada punto dentro de la esfera con su valor y reemplazará en el fichero el viejo valor por el nuevo, creando así un fichero con la información nueva. Se debe tener cuidado, ya que el viejo fichero *map* será reemplazado perdiendo la información original -por lo que sería recomendable crear una copia antes de realizar los cambios-. La segunda opción consiste en generar un fichero *map* solamente con los valores de la esfera y asignándoles un valor de cero al resto del *map*. La ventaja de generar un fichero *map* de esta naturaleza es poder convertir al programa de Autodock en una herramienta de búsqueda por farmacóforos cuando se carezca de la estructura de la diana. La última opción, permite al químico realizar las dos opciones anteriores, esto es, modificar el fichero *map* y generar uno con ceros y valores de las esferas. Para agilizar el diseño de las esferas, la herramienta permite realizar copias de las esferas que están actualmente en el visor de moléculas. Esto representa una ventaja ya que si es de interés ese sitio se copia el tamaño a los demás *maps* con solo seleccionarlos a través de una ventana en la que aparecen los nombres de los *maps* cargados en ese proyecto. Finalmente, en un nuevo proceso de *docking* estos valores -de las esferas- influirán en las puntuaciones que se les dé a los compuestos a ensayar y por tanto en disposiciones diferentes.

Una vez establecida nuestra propuesta, en el siguiente capítulo se corroborará la efectividad del uso de la herramienta mediante dos casos de estudio. El primero será sobre un conjunto de compuestos derivados del proceso de *docking* tomando como diana la Tubulina. Y en el segundo se validarán nuestros resultados mediante la comparación con otra herramienta ampliamente utilizada.

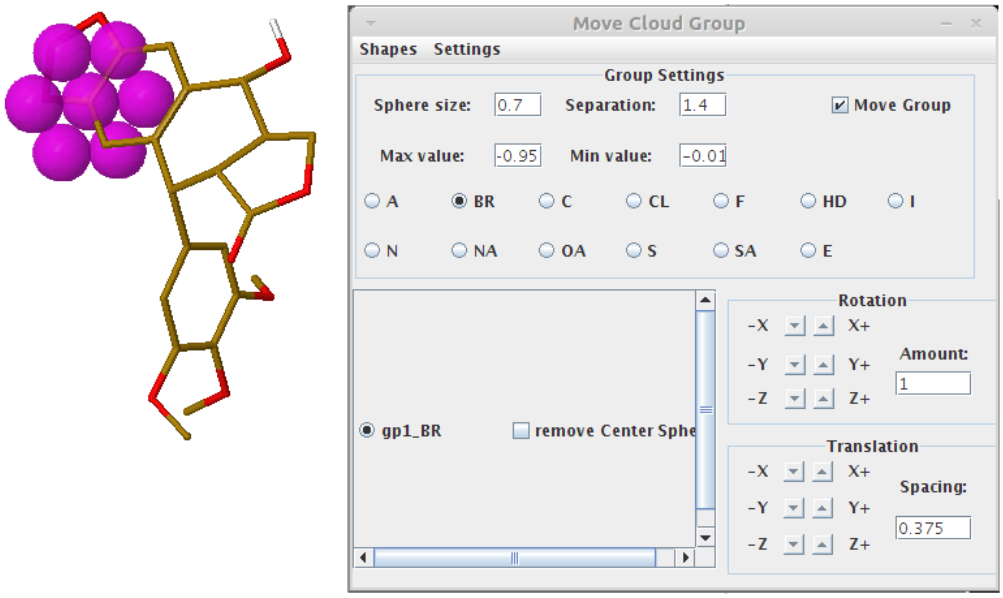


Figura 9.25: Ventana de configuración de grupo de esferas. A la izquierda, un grupo con forma de hexágono. A través de la ventana el químico define el radio de las esferas, sus valores mínimos y máximos, así como el tipo de fichero *map* al que modificará. La ventana también controla la interacción con el visualizador, puesto que permite rotar y trasladar al gupo.

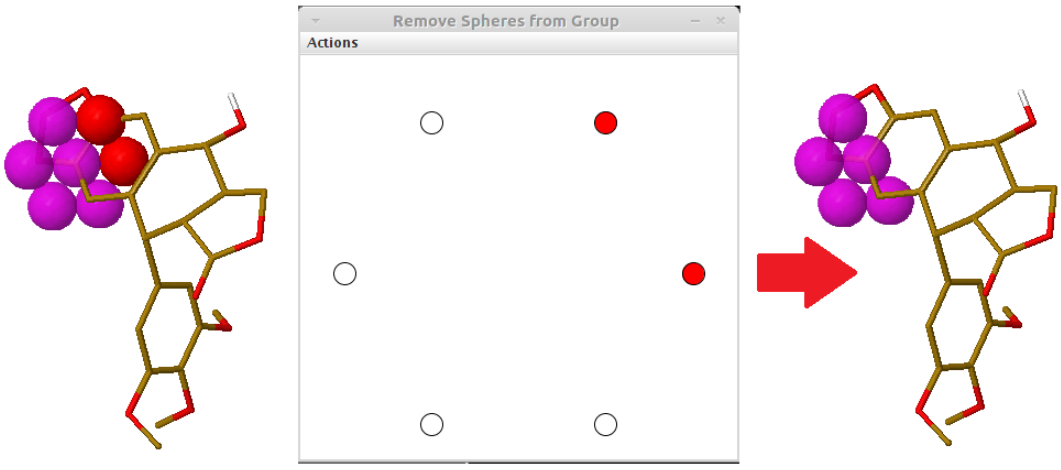


Figura 9.26: Ventana de control para remover esferas de los grupos. En la parte de la izquierda se muestran seleccionadas dos esferas, en el centro una representación en 2D muestra la figura del hexágono con dos esferas seleccionadas, las cuales se corresponden con las seleccionadas en la parte izquierda. Por otro lado, en la parte derecha de la figura, se muestra el resultado de remover las esferas selccionadas.

Parte VI

Resultados

Capítulo 10

Resultados

A continuación se presentan los resultados de JADOPPT aplicado a diferentes resultados de *docking*. Las pruebas realizadas se efectuaron sobre un conjunto de compuestos derivados del proceso de *docking* tomando como diana la tubulina. El conjunto tenía un total de 168 ficheros tipo dlg (168 estructuras con 100 conformaciones por estructura), la estructura de la tubulina, varias moléculas de referencia (compuestos cuya estructura cuando se unen a tubulina se ha determinado experimentalmente: colchicina, DAMA-Colchicina, podofilotoxina, soblidotina, fomopsina, GTP, GDP, ABT-751, NSC-613862, NSC-613863, T138067, TN16), 14 ficheros tipo map que correspondían a los átomos presentes en las moléculas virtuales estudiadas (A, Br, Cl, C, F, HD, HS, I, NA, N, OA, SA, S), un fichero map que representa a los electrones y otro para la solvatación. El programa está escrito en Java y se ha utilizado Jmol como el visualizador molecular incrustado dentro del programa. El equipo usado para realizar las pruebas cuenta con 12 procesadores Intel Core i7 (X980) a 3.33GHz, 12 GB RAM y el sistema operativo LinuxMint (Debian) de 64bits. Para validar el método de *clustering* propuesto en este trabajo se compararon los resultados con los resultados de otro método, en este caso se empleó el programa AuPosSOM [10], utilizando inhibidores de la proteasa VIH-1. El programa está disponible en `http://www.aupossom.com`. Este programa emplea inicialmente mapas auto-organizados (SOM) de Kohonen en una red neuronal y posteriormente se aplica un algoritmo de *clustering* para explorar los resultados.

10.1. Resultados del Análisis Visual en Tubulina

El análisis parte de un total de 16800 conformaciones. El primer paso es tratar de reducir el número de conformaciones a uno manejable y el segundo es agrupar aquellas conformaciones que sean parecidas. Cada método implementado en el primer paso lleva a una reducción diferente, cada una con sus ventajas y desventajas. Tomando como ejemplos los resultados del *docking* flexible (se permite el movimiento de algunos aminoácidos de la tubulina) de podofilotoxina en el sitio de la propia podofilotoxina (proceso conocido como *autodocking*) en tubulina y de IR3b-LEYHr. Para la podofilotoxina, se obtuvieron ocho *clusters* al aplicar el método de RMSD, por medio del método jerárquico: seis *clusters* con distancia promedio, cinco *clusters* con distancia máxima y seis *clusters* con distancia mínima. Por otra parte para IR3b-LEYHr, se obtuvieron treinta y dos *clusters* con

el método de RMSD, y con el jerárquico: diecinueve *clusters* con distancia promedio, catorce con distancia máxima y veinticinco *clusters* con distancia mínima. La ventaja de utilizar el jerárquico está en que las agrupaciones son intrínsecas a los datos, sin embargo, la desventaja principal es explorar visualmente el resultado y seleccionar un nivel de la jerarquía para continuar con el filtrado del resto de resultados de *docking*.

Aun cuando JADOPPT cuenta con un método de selección de nivel de corte, no deja de ser aproximado. Por otro lado, el emplear el RMSD con un corte de 2.0 Å, se obtiene prácticamente el total de grupos sin tener que realizar un paso extra para encontrarlos. No obstante, como se explicó anteriormente en la sección 9.2.1.1, basta con tener un giro en una de las conformaciones para que éste lo considere diferente, por lo que obtendremos un mayor número de *clusters,* llegando incluso a tener las 100 conformaciones del dlg. En otras palabras, el método de RMSD evalúa una conformación contra el resto para formar grupos, a diferencia del jerárquico donde se evalúa a todos contra todos para formar los grupos.

10.1.1. Análisis de Podofilotoxina y IR3b-LEYHr

En el análisis de *clusters* de podofilotoxina por RMSD se observa que hay 4 *clusters* mayoritarios (figura 10.1) que además corresponden a las mejores opciones de energía. La visualización cuenta con una escala de colores en la parte superior para representar la energía de unión, de esta forma se indica la evaluación de los elementos del *cluster* por Autodock, en otras palabras, se guiará al químico en la selección de aquella conformación que presente la mejor energía de unión con la diana. Al comparar los c*lusters 1* y *4*, podemos observar que se trata de la misma pose en diferentes *clusters*, posiblemente se deba a la simetría del anillo.

En el *cluster 1* hay 36 conformaciones en el rango de -6.34 a -8.5 de energía de unión, y, en el *cluster 4* hay 16 conformaciones en el rango de -6.37 a -7.31 la diferencia de energía es de -1.1, lo que indica un error del programa de *docking*, ya que son dos valoraciones de la misma situación. El número de moléculas en el *cluster 1* y *4* es relativamente similar, lo que sugiere que el programa encuentra ambas respuestas con igual facilidad. Por otro lado, El *cluster 2* es la respuesta correcta (coincidente con lo observado experimentalmente -molécula marrón-). El rango de energía de unión es de -7.6 a -8.93 y es casi tan frecuente (20 conformaciones) como los *clusters 1* y *4*. El *cluster 3* es la última opción mayoritaria de energía más favorable (el rango va de -6.2 a -7.63) y contiene 18 conformaciones. Finalmente los *clusters 5* a *8*, son los de peor energía de unión (color rojo) ya que el rango de todos esos *clusters* es de -5.03 a -6.49.

En ausencia de la información proporcionada por el ligando de referencia (molécula marrón), probablemente el químico habría seleccionado la opción 1+4, pues agrupa a un mayor número de representantes y la energía está dentro de la más favorable según el programa. Sin embargo, el disponer de la información proporcionada por la referencia podría permitir manipular los mapas de Autodock para penalizar estas disposiciones respecto a la “correcta” (ver sección 10.5).

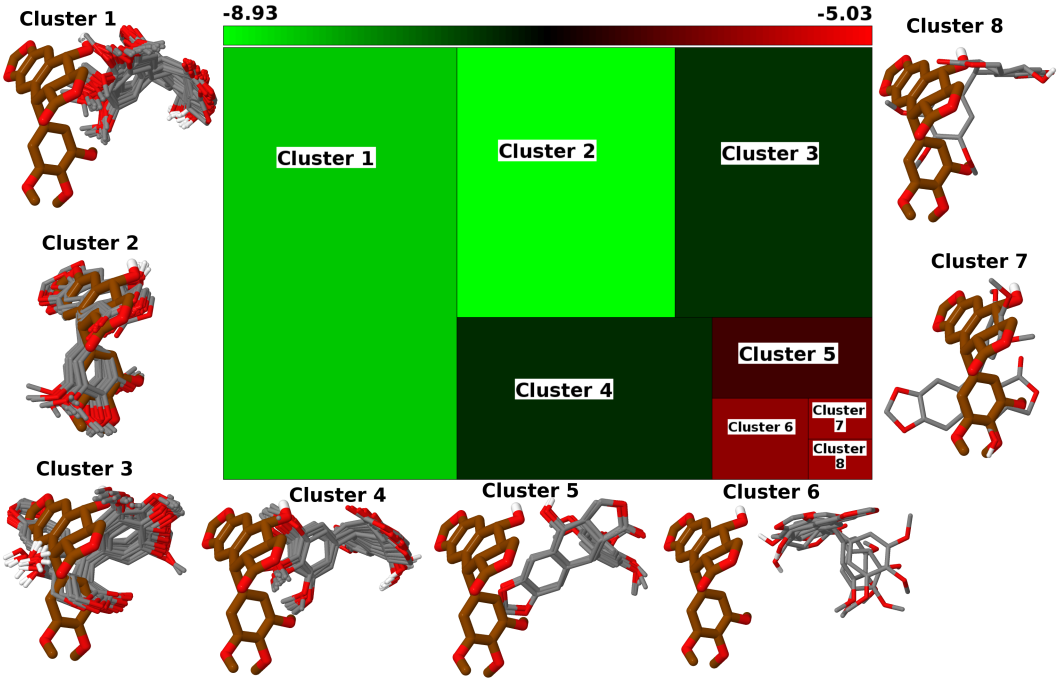


Figura 10.1: *Clustering* mediante el cálculo de RMSD para los resultados de *docking* de PODc-LEHYHr. Resaltan a la vista 4 *clusters* por su tamaño, además de que sus elementos están en la zona favorable de unión de acuerdo a la escala de colores en la parte superior (-5.03 a -8.93). Con este método no es obligatorio hacer un refinamiento de los resultados, salvo que sea necesario. Claramente, el *cluster 2* contiene las conformaciones similares a la referencia de podofilotoxina (en color marrón). De cada *cluster* se seleccionará un representante, el cual pasará a un segundo paso para encontrar aquellos que sean similares a él.

Por otra parte, en la figura 10.2, se muestran los seis *clusters* pertenecientes al método jerárquico con distancia promedio. Es claro que el resultado es muy aproximado al que se consigue mediante el cálculo de RMSD. Se obtienen 6 *clusters*, de los cuales el *cluster 4* es el mayoritario seguido del *cluster 3* y *cluster 1*. En el *cluster 4* hay 53 conformaciones, el rango de energía de -6.34 a -8.5, que se corresponderían con la unión del *cluster 1* y *4* del RMSD. Por otra parte, podemos observar en el dendrograma que dicho *cluster* está formado a su vez por dos *clusters* mayoritarios similarmente representados. El *cluster 3* es la respuesta correcta, que contiene 20 conformaciones en el rango de -7.6 a -8.93, y se corresponde con el *cluster 2* del RMSD. Para el resto de los *clusters* del dendrograma con distancia promedio las correspondencias con los *clusters* del RMSD son: el *cluster 1* es igual al *cluster 3*, *cluster 2* al *cluster 7*, el *cluster 5* es el mismo para los dos, al igual que el *cluster 6*; por último, para el único representante del *cluster 8* del RMSD el representante está dentro del *cluster 4* en el dendrograma.

Los resultados para los métodos de distancia máxima y mínima fueron parecidos a los de la distancia promedio, para la respuesta correcta en la distancia mínima es el *cluster 4* y en la distancia máxima es el *cluster 2*; el *cluster* mayoritario en ambas distancias es el *cluster 3* que se corresponde con el *cluster 4* en distancia promedio y con los *clusters 1* y *4* del RMSD. El *cluster 1* de la distancia

máxima se corresponde a la unión de los *clusters 1* y *2* de la distancia mínima y de distancia máxima.

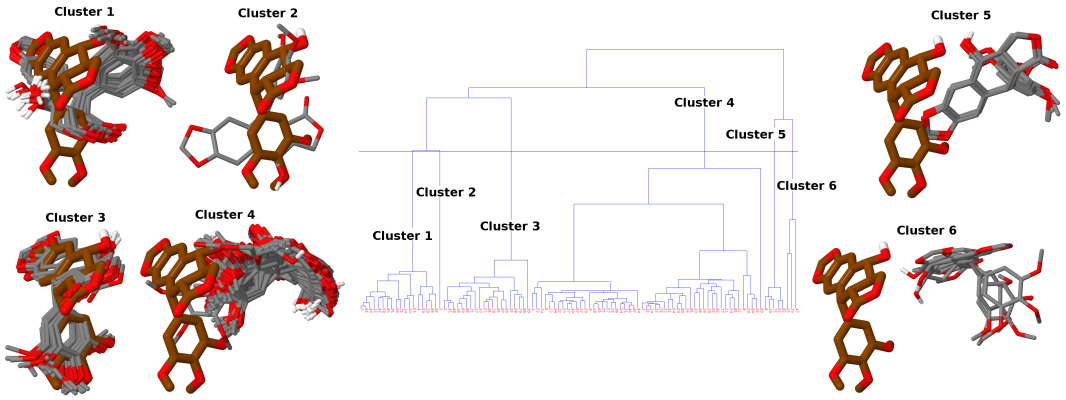


Figura 10.2: *Clustering* mediante el algoritmo jerárquico con distancia promedio para los resultados de *docking* de PODc-LEHYHr. La aproximación del nivel de corte, sugiere una exploración visual de los resultados del *clustering* para confirmar o, si es necesario, refinar los *clusters* seleccionados. Sin embargo, es evidente que el nivel de corte es satisfactorio, el *cluster 3* claramente es la respuesta correcta, además de corresponderse con el *cluster 2* del RMSD.

Es evidente que el método jerárquico con distancia promedio al igual que el de distancia mínima reducen satisfactoriamente el número de representantes. Por otro lado, el RMSD no lo hace tan mal, sin embargo, cuando se tengan resultados en que las conformaciones estén en aparente caos, esto es, no se apreciarán grupos definidos a primera vista; más un, si aplicamos el RMSD para agrupar, debido a que las conformaciones están muy dispersas en el espacio, no será capaz de encontrar elementos similares, lo cual originará un mayor número de representantes; es claro que el método jerárquico encontrará los grupos existentes, y, así se podrá reducir el número de clusters a uno manejable.

IR3b-LEYHr es un claro ejemplo de lo que se comentó en el párrafo anterior. La figura 10.3 muestra a la proteína y el resultado del *docking*. Como podemos observar, el programa de *docking* ha colocado en un aparente desorden cada conformación en el espacio: a simple vista es imposible poder decir qué conformación es la mayoritaria, a diferencia de lo que ocurría en el ejemplo anterior con podofilotoxina. Al aplicar el método de RMSD (figura 10.4), sobresalen ocho *clusters* por su tamaño, por otra parte, en este ejemplo no contamos con una molécula de referencia como en el caso anterior, lo cual indica que la selección de representantes por parte del químico se basaría en seleccionar los *clusters* de mayor tamaño y de mejor energía. También, podemos observar un recuadro en color azul encerrando a otros clusters de menor tamaño, estos *clusters* no se muestran en la figura 10.4 debido a que el número máximo de elementos es de tres, además de que sólo cinco *clusters* están dentro del rango de favorables, pero con una sola conformación, por lo que no merece la pena mostrarlos dado que no se cuenta con una molécula de referencia. Lo mismo se puede decir para el método jerárquico. Los *clusters 1*, 3 y *5* podrían formar un mismo grupo ya que cubren en forma similar el mismo espacio, lo mismo ocurre con los *clusters 4* y *6*. Finalmente, el *clusters 2* y *7* presentan la misma situación que los *clusters* anteriores. Por su parte, el *cluster 8* podría unirse con cinco *clusters* únicos del recuadro azul.

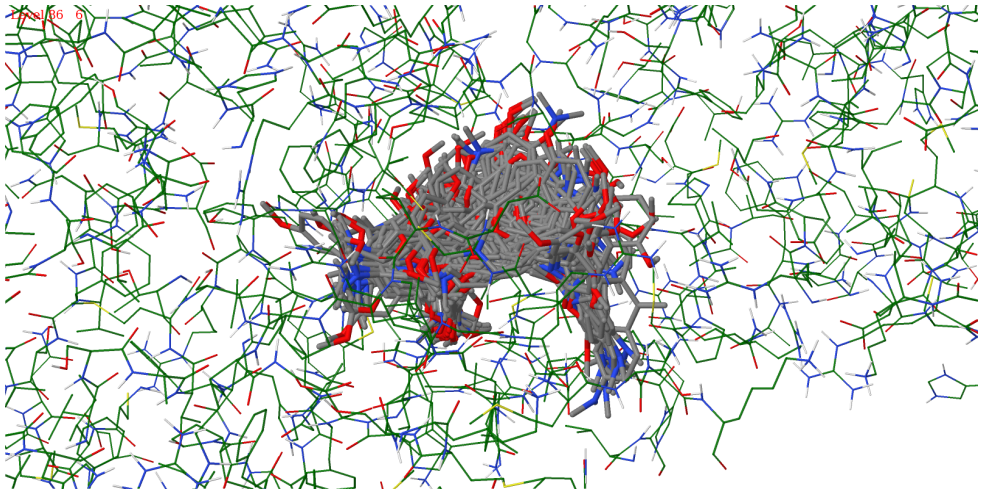


Figura 10.3: Los resultados del *docking* IR3b-LEYHr son difíciles de interpretar, ya que no se sigue un patrón claro, lo que se interpreta como un aparente caos.

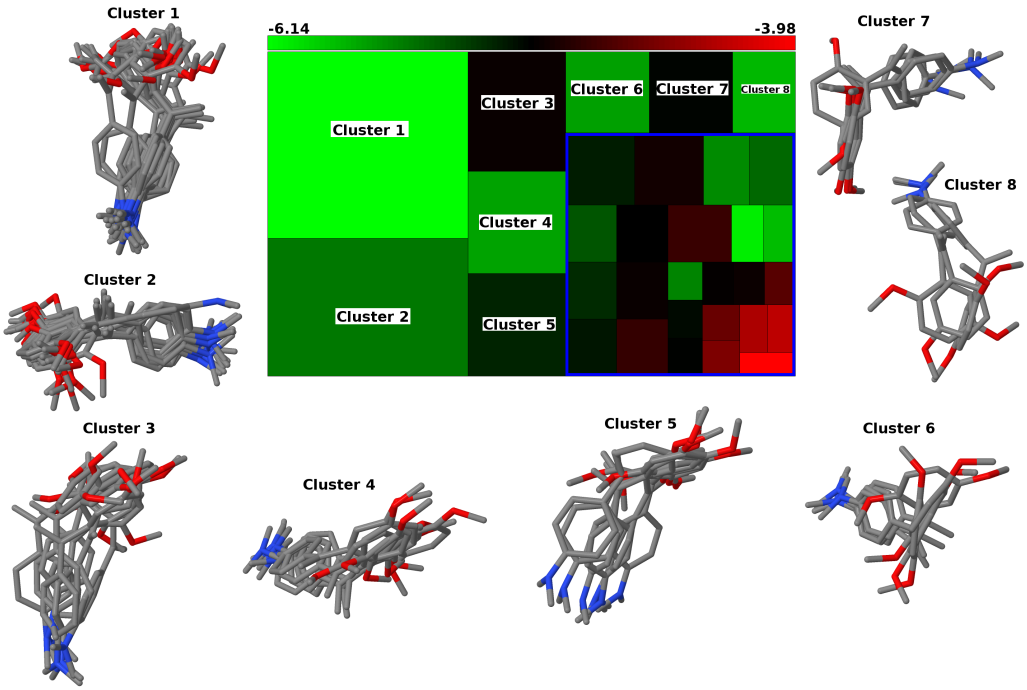


Figura 10.4: La visualización del *treemap* muestra 32 *clusters* generados por el método de RMSD. Claramente algunos de los *clusters* de mayor tamaño podrían unirse para formar uno solo. El recuadro azul encierra a *clusters* de tamaño mínimo.

Al igual que RMSD, el método jerárquico con las distintas distancias generó una gran cantidad de *clusters*, sin embargo, bastaría explorar el dendrograma con la barra horizontal para dividir o unir *clusters* y, de esa manera, reducir/expandir el número de representantes para la segunda etapa –la línea sólida en las figuras representa el nivel del corte que genera el método automatizado de selección de representantes-. Para el *clustering* de distancia mínima (figura 10.5), podemos observar que los *clusters 18* y *19* son similares y pueden unirse para formar uno solo (línea punteada en la figura 10.5). Por otro lado, vemos que los *clusters 1* y *8* son similares, sin embargo, tendríamos que subir la barra hasta tener solamente dos *clusters* para unirlos a diferencia de la unión de los anteriores.

Es claro que la distancia mínima entre los representantes no es una buena medida para agruparlos, ya que obtenemos casi la misma cantidad de *clusters* que el método de RMSD. En contraparte, en el *clustering* de distancia promedio, al subir la barra nos permite unir los *clusters 1* y *2*, además de los *clusters 8* y *9,* para de esa forma reducir los representantes (línea punteada figura, 10.6). Incluso podríamos subir más la barra hasta unir al *cluster 10* con el *12*; sin embargo, los *clusters* únicos no son tan parecidos a los *clusters* mayoritarios y, por lo tanto, tendríamos elementos muy dispares. Esto es bueno, ya que estaríamos filtrando aquellas conformaciones que inflarían el número de representantes para el segundo paso.

Finalmente el método que mejor resultados proporciona en el *clustering*, es el de máxima distancia (figura 10.7). Al subir la barra uniríamos cuatro *clusters* (*1-4*) que claramente pertenecen a la misma zona del espacio. Finalmente, después de seleccionar ese *cluster*, observamos que se han unido los *clusters 13* y *14*, aunque parece claro que son diferentes; bastaría con desplazar la barra hacia abajo para separarlos y seleccionarlos por separado (línea punteada en rojo).

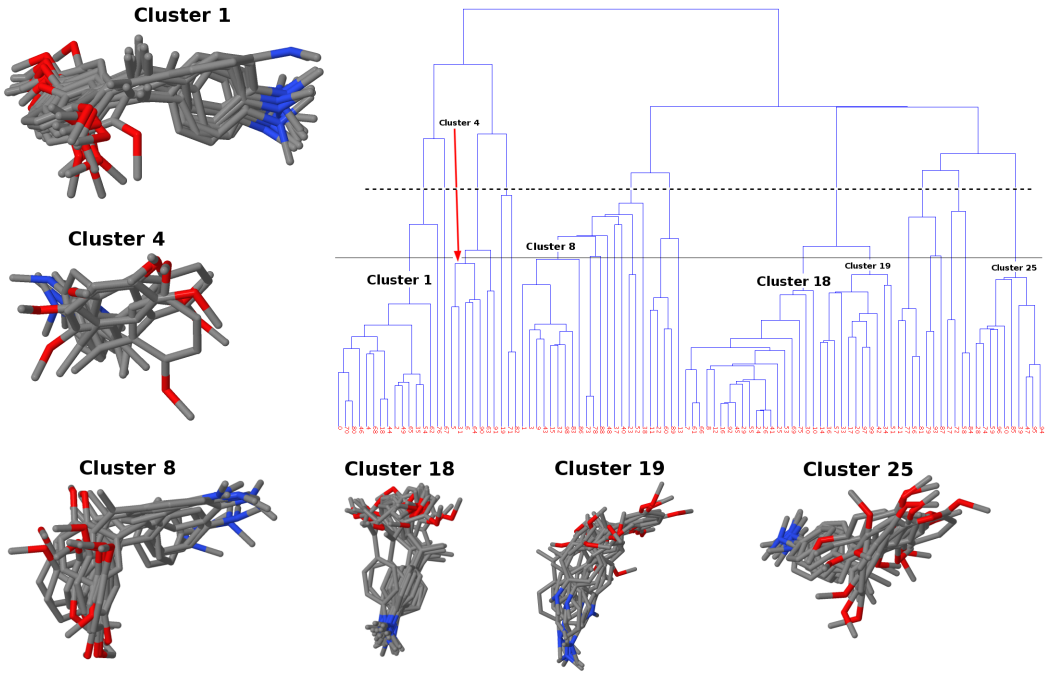


Figura 10.5: La línea punteada es el corte que realizaría el químico para seleccionar los representantes para el segundo paso. El número de *clusters* seleccionados de manera automatizada (línea solida) genera casi la misma cantidad que el método RMSD. El método busca los elementos más cercanos de cada grupo formado, y, separa grupos que podrían estar unidos -por ejemplo, *clusters 1* y *8*-, lo que indica que no es viable el *clustering* de distancia mínima en este ejemplo.

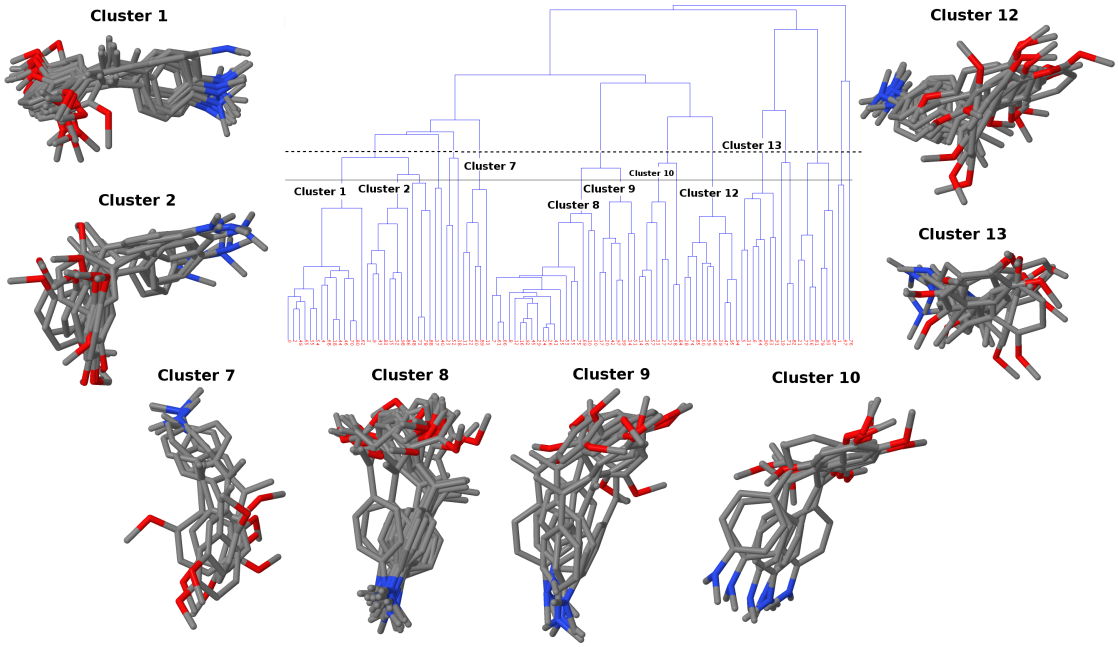


Figura 10.6: La línea punteada representa el corte que realizaría el químico para seleccionar el número de representantes para el segundo paso. La distancia promedio ayuda a reducir significativamente los representantes que pasarían al segundo paso.

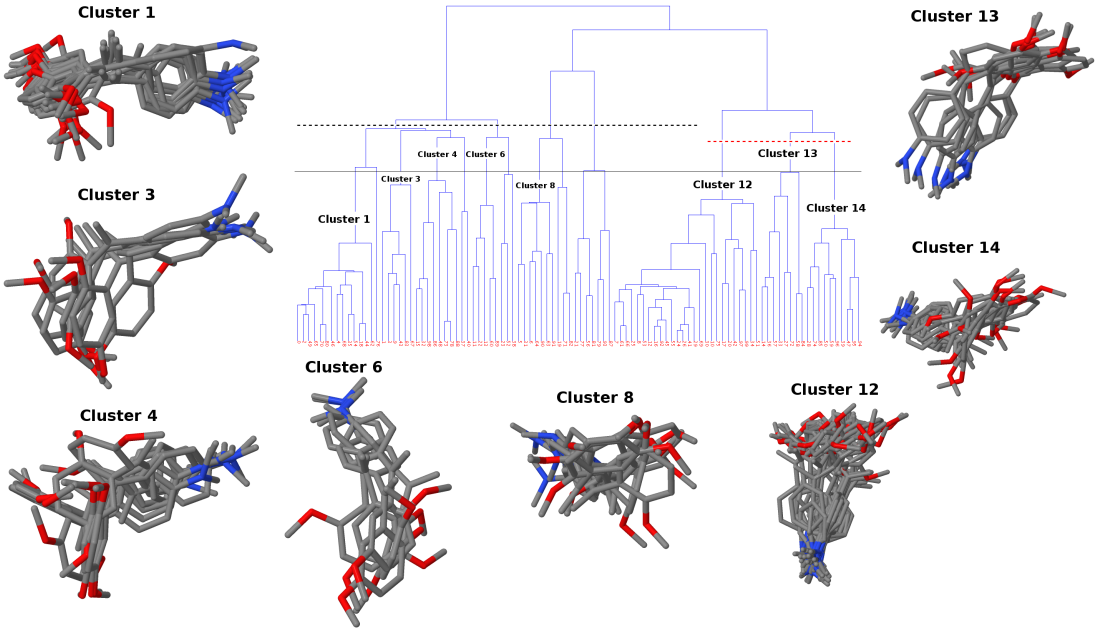


Figura 10.7: La línea punteada representa el corte que realizaría el químico para seleccionar el número de representantes para el segundo paso. La distancia máxima ha encontrado la menor cantidad de grupos existentes en los datos. Sin embargo, al subir la barra se pueden unir tres *clusters* mayoritarios (*1*, *3* y *4*). La línea punteada en negro representa la primera unión de *clusters* y la de color rojo representa un segundo filtrado para reducir a un número óptimo a los representantes para el segundo paso.

Es claro que el método jerárquico encontró los grupos existentes a diferencia que el método por RMSD. Por otro lado, se redujo el número de *clusters* a uno manejable. Sin embargo, la desventaja es latente, ya que es necesario hacer una exploración visual para afinarlo a un número satisfactorio de *clusters*, como se mostró. La barra de exploración de la visualización del dendrograma, al igual que la ayuda visual de la escala de colores del RMSD, ayuda al químico a desentrañar el caos presentado en la figura 10.3. Más aún, al emplear la ayuda visual en forma de tooltip[1], se facilita la extracción de información al posicionarse sobre las etiquetas marcadas como *clusters*; dicha información se da en relación a la interacción de las conformaciones con la diana (éstas se ordenan de forma ascendente mostrando la información de las cinco primeras conformaciones de cada *cluster*). Retomando lo anterior, aunque con la desventaja de realizar una inspección visual, es posible automatizar este proceso, pero con la reserva que sugiere una exploración para corroborar los resultados, como ya se demostró.

Es evidente que la automatización de agrupación de conformaciones por cualquiera de los dos métodos (RMSD o jerárquico) es fiable (en la figura 10.8 se presenta el mismo *cluster*). Más aun, podemos observar claramente que el *cluster* está sobre la referencia de podofilotoxina (color marrón); en otras palabras, la automatización ha sido capaz de reproducir el resultado que un químico reali-

[1]Recuadros que aparecen en pantalla al posicionarse sobre algún elemento, por ejemplo un botón o icono y proporcionan información específica sobre ese elemento.

zaría -ha seleccionado para todos los casos el *cluster* con la respuesta correcta, pero en un tiempo mucho menor-.

Cabe añadir que este logro fue determinante para de igual forma automatizar la selección de representantes, lo cual es uno de los objetivos de este trabajo de tesis.

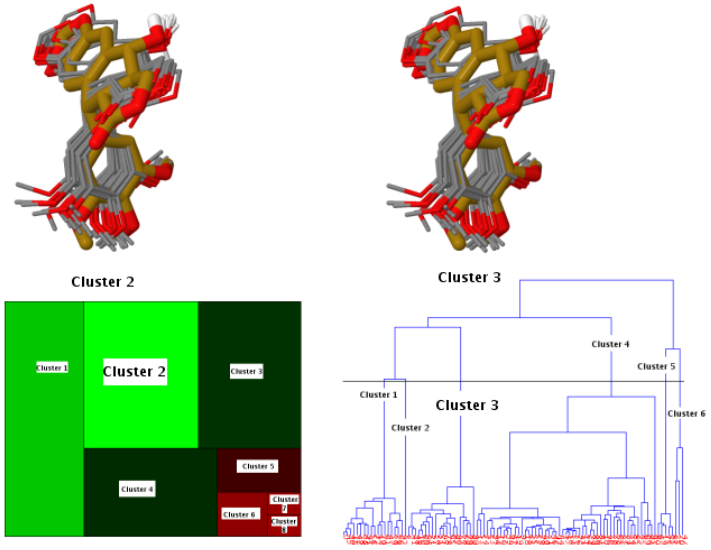


Figura 10.8: La automatización de agrupamientos y selección de *clusters* son fiables, pues con los dos métodos se ha llegado al mismo resultado.

Continuando con el análisis de resultados del primer paso, realizado mediante la selección automatizada para reducir las 16800 conformaciones; por el método de RMSD se filtraron 6926 conformaciones en total, y para el jerárquico: 1090 con distancia promedio, 1472 con distancia mínima y 946 con distancia máxima. Es claro que el número se reduce considerablemente con el método jerárquico por cualquiera de las opciones ofrecidas. Para ambos métodos, RMSD y jerárquico, el proceso de selección de representantes es automatizado, salvo que en el método jerárquico se basa en un punto de corte -con un significado físico-, lo que tiene el inconveniente de seleccionar más representantes de los necesarios o por el contrario unir *clusters* diferentes y seleccionar un único representante. Por esta razón se sugiere la inspección visual para realizar el refinamiento de los resultados. Por otro lado, los resultados de la selección de *clusters* son aceptables, además de ser más flexibles en cuanto a discernir la similitud entre conformaciones. En el caso del RMSD el punto de corte se da como entrada para realizar las agrupaciones, por lo tanto es un procedimiento totalmente automatizado; sin embargo, como se ha demostrado también tiene una desventaja al considerar como desiguales aquellas conformaciones que estén alejadas del umbral de distancia por muy poco, algo que en el método jerárquico no sucede.

Las repercusiones de elegir un método de reducción para el segundo paso son las siguientes –tomando como ejemplos a los dos casos, el de podofilotixina y IR3b-LEYHr:

1. RMSD: En el caso de la podofilotixina (figura 10.1), habría dos representantes de la misma conformación (*clusters 1* y *4*), el representante del *cluster 1* tendría la mejor energía de unión

con respecto al representante del *cluster 4*; en un principio esto no afecta más que al número de representantes de ese resultado de *docking*, puesto que están dentro del rango de unión satisfactorio. En cambio, en IR3b-LEYHr (figura 10.4) hay 7 representantes de más (*clusters 1*, *3* y *5*, *clusters 2* y *7*, *clusters 4* y *6*). El representante del *cluster 3* tiene una energía de unión poco favorable, además de tener un número de elementos menores con respecto al *cluster 1* –con quien tiene mayor similitud-, lo cual implica tener una conformación de más, de la cual se podría prescindir. Igualmente sucedería para el resto de los grupos.

2. Jerárquico: De los tres métodos para podofilotixina el resultado es muy similar. En cambio, para la otra molécula, es evidente que las divisiones existentes en el de RMSD se ven compensadas en el jerárquico, en otras palabras, un *cluster* del jerárquico –con los tres métodos- equivale a dos en el RMSD. En cambio para la otra molécula, la situación cambia, mientras que para el RMSD se generan demasiados *clusters* con elementos únicos, para el jerárquico se reduce el número de representantes; en otras palabras, se desvelan los grupos que hay intrínsecamente en los datos.

La figura 10.9a muestra los 6926 representantes elegidos como resultado del filtrado mediante el RMSD a partir de las 16800 conformaciones iniciales. Tan sólo se ha producido una reducción al 41 %. Esto implica que, como promedio, se han conservado 41 representantes de cada molécula inicial, un número muy superior al observado para la podofilotoxina (*autodocking*), que debe ser una excepción y no la regla (o bien el conjunto presenta dos grupos: unos con muchas respuestas y otros con pocas, de modo que el promedio es intermedio). Esto indica que el programa de *docking* encuentra un elevado número de respuestas posibles para cada molécula virtual. El significado de esta observación es claro: la podofilotoxina es rígida y encuentra pocas opciones favorables, mientras que otras opciones más flexibles encuentran numerosas respuestas. Por el contrario, en el caso del *clustering* jerárquico con máxima distancia (figura 10.9b) se ha producido una reducción de los datos al 5,1 %, conservando 5 representantes por cada molécula virtual, un número similar al de la podofilotoxina. Sin embargo, de la comparación de los resultados anteriores se deduce que esta reducción posiblemente se hace a costa de eliminar los representantes de *clusters* minoritarios (y generalmente de peor energía), pero a costa de una pérdida de información. Dependiendo de la situación, una u otra opción ofrecen alternativas útiles. Para un análisis inicial de los resultados la reducción de la dimensionalidad facilita un análisis global de los resultados, mientras que para el análisis detallado de los resultados individuales de cada molécula es deseable preservar el mayor número de representantes significativos. Cuanto más flexible y más alejada la estructura del ligando virtual de la que generó el sitio en la proteína, más probable es que el programa de *docking* tenga problemas en encontrar la respuesta "correcta", y más deseable es la opción que conserva más representantes.

De los tres métodos jerárquicos, el de distancia máxima reduce considerablemente el número de representantes, aparte de mostrar que es el mejor método para agrupar conformaciones en donde sus elementos están muy dispersos. Por otra parte, en la mayoría de los casos, los resultados de *docking* que se analizan en este trabajo de tesis son similares al ejemplo presentado anteriormente. Por lo tanto, para el segundo paso tomaremos el filtrado de RMSD y, del jerárquico, el de máxima distancia.

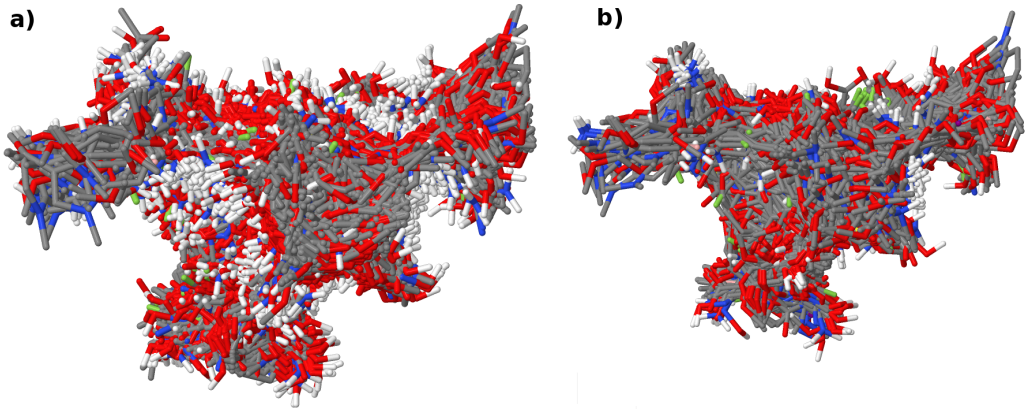


Figura 10.9: Resultado del agrupamiento de representantes. a) Resultado del filtrado por RMSD, 6926 representantes han pasado a la segunda etapa de análisis. b) 946 representantes como resultado del filtrado por el algoritmo jerárquico con distancia máxima.

10.2. Análisis General del Agrupamiento de Representantes

Hemos realizado el análisis de los dos tipos de filtrado RMSD y jerárquico (distancia máxima). Esto es, hemos realizado la comparación de moléculas diferentes a través del algoritmo jerárquico aplicando los tres métodos de distancia. El procedimiento para preparar a los representantes fue el siguiente:

1. Para cada representante se calculó una matriz de distancia a los observadores, empleando las tres opciones de que dispone JADOPPT (todos los observadores, selección de tipo de átomo como observador, y grupo de átomos como observadores).

2. Posteriormente a la matriz de distancias de observadores se calculó otra matriz, pero en esta ocasión de disimilaridad, una vez calculada, esta sirve como entrada al algoritmo jerárquico.

3. Se aplicó cada método con las diferentes opciones de los observadores.

4. Finalmente se generó un dendrograma para analizar los resultados de la agrupación, y mediante el visualizador molecular Jmol se exploró cada resultado.

10.2.1. Análisis del *Clustering* de Representantes por Filtrado Jerárquico de Distancia Máxima

La inspección visual del dendrograma nos arroja excesiva información sobre los *clusters*, pues al haber tantos representantes las líneas están muy juntas y sólo es posible apreciar la estructura en las primeras agrupaciones. Por tal motivo, se exploraron zonas en donde predominaba el color verde, ya que se indica el valor de unión con la diana de acuerdo al código de colores a la izquierda del dendrograma.

Distancia Promedio con Todos los Observadores (*Average-all*)

Las referencias se sitúan en cinco zonas a lo largo del dendrograma; diecisiete *clusters* se encuentran al bajar la barra para separar las ramas que contienen demasiados *clusters* poblados (figura 10.10), sin embargo, el color verde se concentra en la tercera parte del dendrograma hacia la izquierda –encerrado en un cuadro en la figura 10.10-, en el centro se encuentran demasiado mezclados los colores sin reportar un patrón definido, salvo una porción muy pequeña que sobresale por el color verde, y hacia la derecha predominan los colores oscuro a rojo, las referencias se dividen entre el extremo izquierdo y derecho. Dado que el color verde se encuentra en la izquierda y los colores rojos a oscuro indican que tienen una pobre energía de unión a la diana, nos enfocamos a analizar los *clusters* de esas zonas. Tres *clusters* sobresalen en esa zona y los analizamos de izquierda a derecha. El primer *cluster* no agrupa referencias, lo que indica que éstas no son similares o cubren otra zona del espacio, esto lo comprobamos al visualizar tanto el *cluster* como las referencias (figura 10.11). La energía de unión de los representantes de esa zona de color verde van de -5.24 a -7.69. Por otro lado, también observamos, que los representantes no son parecidos a las referencias, sin embargo, entre ellos sí son similares -donde podemos apreciar mejor esta similitud es en *cluster 1* y *2* de la figura 10.11-. De igual forma, los representantes al no tener una referencia y dado que están relativamente cerca de la zona que cubre las referencias visibles, podríamos asumir que esos representantes están en la zona correcta de interacción con la diana.

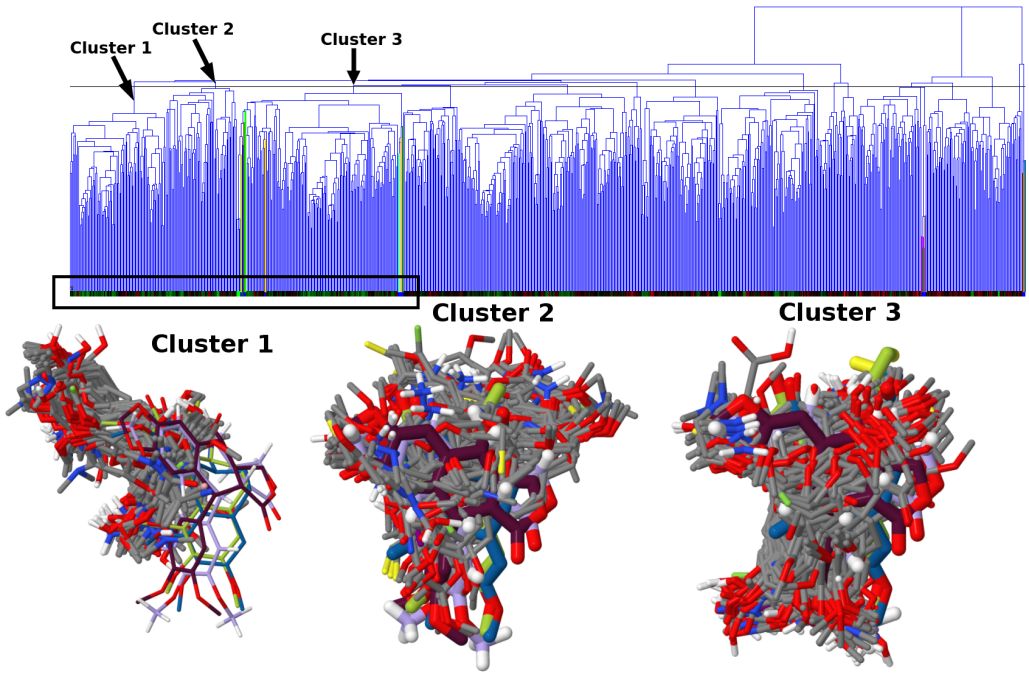


Figura 10.10: El cuadro negro resalta la concentración de representantes con mejor afinidad a la diana. Tres *clusters* sobresalen de esa zona, las referencias se agruparon en el tercer *cluster*, sin embargo, se muestran en los *clusters 1* y *2* para orientar la vista y así apreciar la zona que cubren.

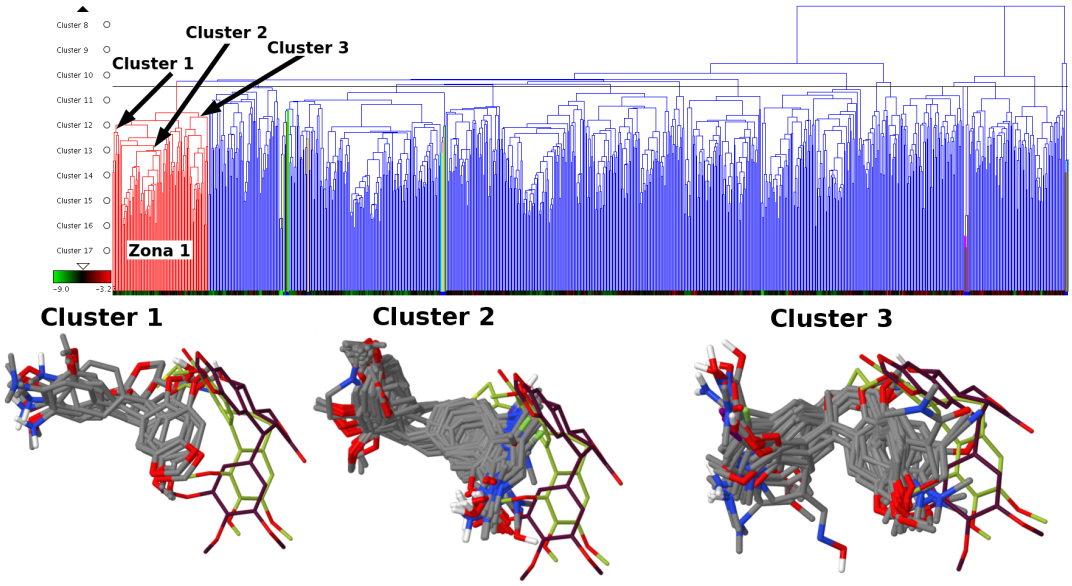


Figura 10.11: Es evidente que lo representantes no son parecidos a las referencias, sin embargo, al carecer de ellas y dado que están relativamente cerca de las referencias de podofilotoxina y colchicina podemos asumir que están en la zona correcta de interacción con la diana.

En el segundo *cluster* el número de representantes con color verde disminuye, esto lo podemos observar en la figura 10.12 en el extremo derecho, ya que se muestra una ampliación del *cluster*. La fragmentación del *cluster* mostró que se encuentran dos representantes para los que se cuenta con referencias (podofilotixina y colchicina). Sin embargo, este *cluster* tampoco agrupó ninguna referencia, pero al superponerlos observamos que los representantes para la podofilotixina se encuentran en una orientación invertida en relación a sus referencias y lo mismo sucede para la colchicina. Si nos guiaramos por el átomo en color amarillo de la referencia, observamos que estos átomos apuntan hacia la derecha, mientras los átomos de los representantes lo hacen hacia la izquierda (figura 10.12). En un caso especial, uno de los representante de colchicina tiene una energía de unión de -8.03 y esto lo convierten en el más favorable de todo el *cluster*, sin embargo, al superponer las referencias de colchicina observamos que el átomo de azufre (color amarillo) está orientado de forma diferente, por lo tanto podemos afirmar que ésa es la causa de que no forme parte del *cluster* donde se encuentran las referencias, aun cuando tenga una energía de unión muy favorable; probablemente la orientación se deba a los valores del *grid* que generó Autodock.

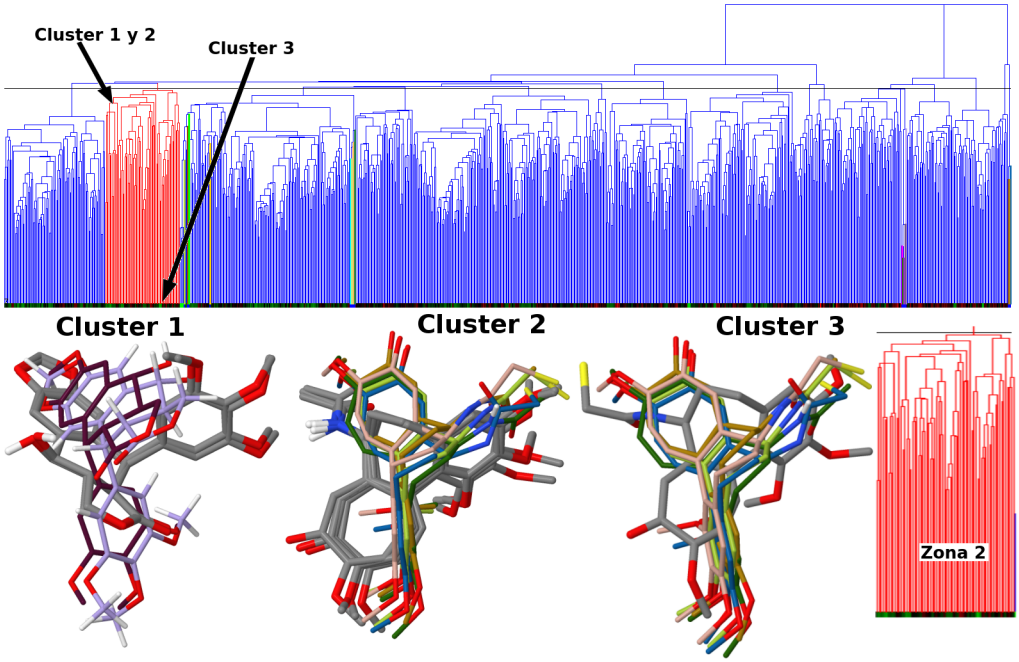


Figura 10.12: Los representantes del *cluster* seleccionado son similares a sus referencias. Los representantes de los *clusters* son similares entre ellos pero orientados de forma diferente a sus referencias. El *cluster 3* muestra al mejor representante de todo el *cluster* en rojo (-8.03 de energía de unión), pero al superponer a su referencia observamos la orientación diferente.

En el tercer *cluster* podemos observar que contiene referencias (figura 10.13), de igual forma, resalta una pequeña zona en color verde claro, lo que indica que son las respuestas correctas en relación a la podofilotoxina y colchicina. Esto lo corroboramos al superponer las referencias y los representantes –*cluster 1* y *2* en la figura 10.13-. Sin embargo, sólo dos representantes de colchicina son similares, al resto le sucede lo mismo que a la colchicina del *cluster* de la figura 10.12. Los representantes del *cluster 3* de la figura 10.13 también cubren la zona de las referencias y el rango de energía de unión es de -5.45 a -7.75. Finalmente, en el *cluster 4*, los representantes son similares y cubren la zona de las referencias de podofilotoxina, el rango de valores de energía de unión del *cluster 4* va desde -4.71 a -7.79.

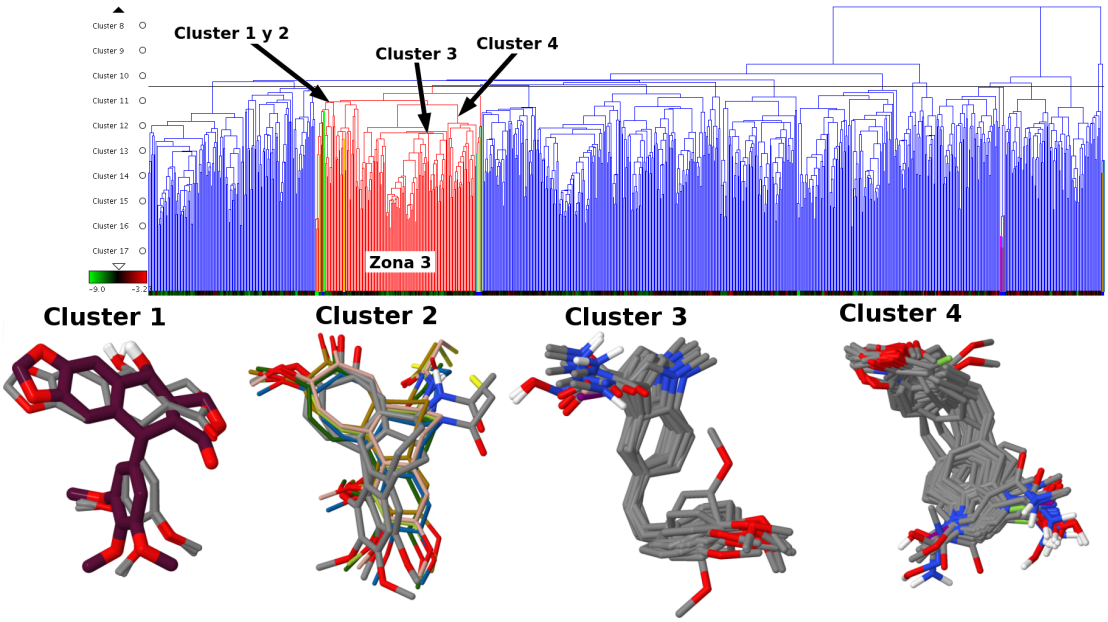


Figura 10.13: En este *cluster* tanto las referencias como los representantes del *cluster 1* y *2* son similares en todos los aspectos a las referencias. Los *cluster 3* y *4* cubren la misma zona que las referencias de podofilotoxina y colchicina, además de ser similares entre sus elementos.

Distancia Promedio Seleccionando los Observadores Br, Cl, HD, F, I, C (*Average-each*)

Se seleccionaron los átomos al azar y dado que no todos los representantes contienen los mismos tipos de átomos, se esperaría que agrupara de manera mezclada a los representantes. En total se pueden apreciar diez *clusters* de tamaño similar, además de distribuir a las referencias en cuatro zonas distintas. En cuanto a los representantes, se observan zonas muy pequeñas en la parte izquierda del dendrograma donde se puede apreciar el color verde (figura 10.14). Por tal motivo, se analizaron tres *clusters* de la izquierda del dendrograma. La figura 10.14 muestra una zona en amarillo sobre el rojo, esto indica que del *cluster* marcado en rojo se seleccionó una rama. Los representantes del cluster en amarillo son similares entre ellos mas no a las referencias. Finalmente, también es visible que no están sobre las referencias y, en términos generales, el *cluster* en amarillo es parecido al *cluster 2* de la figura 10.11 y figura 10.19, incluso sus rangos de energía de unión son similares, el *cluster* amarillo va de -5.12 a -7.69 y tanto el de la figura 10.11 como el de la figura 10.19 es de -5.24 a -7.69. Esto indica que la agrupación se debió a que la mayoría de los representantes contenían alguno de los átomos seleccionados.

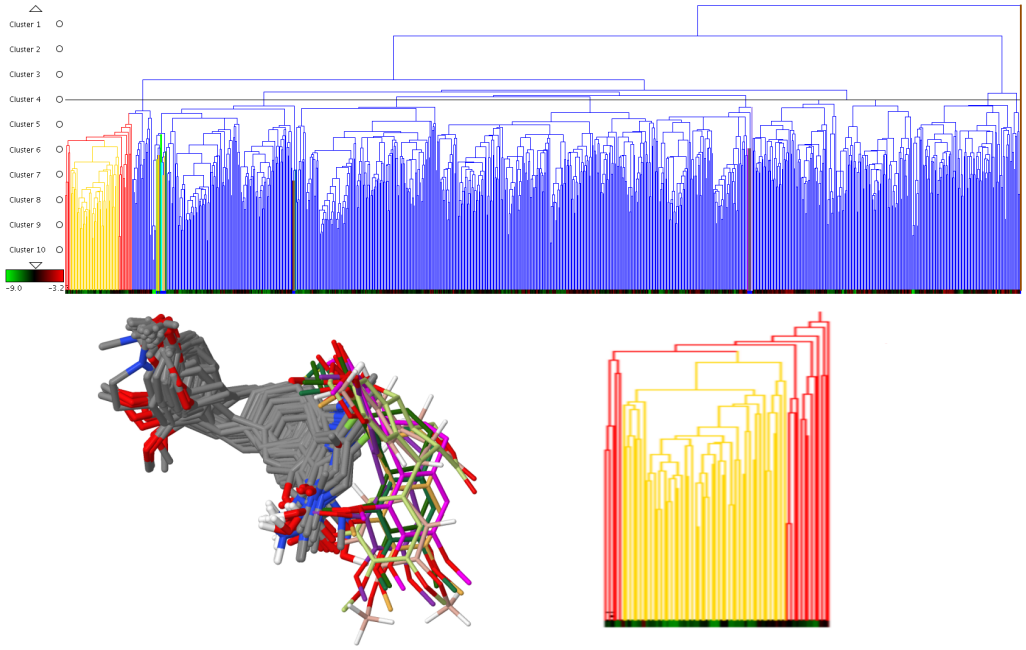


Figura 10.14: *Cluster* seleccionado en rojo muestra en su centro una mayor concentración de representantes con buena afinidad a la diana. Al superponer las referencias podemos observar que los representantes están hacia un lado de las referencias.

La segunda zona contiene menos representantes con color verde que el anterior, no obstante, en esta zona están la mayoría de las referencias. Una rama pequeña de representantes sobresale de las demás en el tono del color verde por ser más claro, además, en esa rama están las referencias, por tanto, nos centraremos directamente en analizar dicha rama. En la figura 10.15, en la parte inferior derecha se muestra una ampliación del *cluster* en color rojo; por otro lado, también podemos observar la rama donde se concentra la mayoría de las referencias del dendrograma. Gracias a la visualización de las referencias en el dendrograma es posible centrar el análisis sobre esa rama. En este caso en particular donde la cantidad de representantes con energía de unión favorable es muy reducida, las referencias resaltadas en el dendrograma anima a explorar dicho *cluster*. Así, al realizar la exploración de la rama que contiene las referencias pudimos observar que la podofilotoxina y colchicina están presentes y que son similares a sus referencias.

Finalmente la tercera zona contiene menos representantes con energía de unión favorable (figura 10.16). Solamente dos ramas contienen suficiente cantidad de representantes de color verde como para explorarlos. El rango de energía de unión va de -4.77 a -7.72. El cluster 1 es la rama de la derecha en la figura 10.16, el color verde es más claro y los representantes son de la colchicina y podofilotoxina, sin embargo, al superponer las referencias observamos que no son similares en la orientación. En el *cluster 2* los representantes son similares entre ellos y no a las referencias, pero aun así, están dentro de la zona de las referencias.

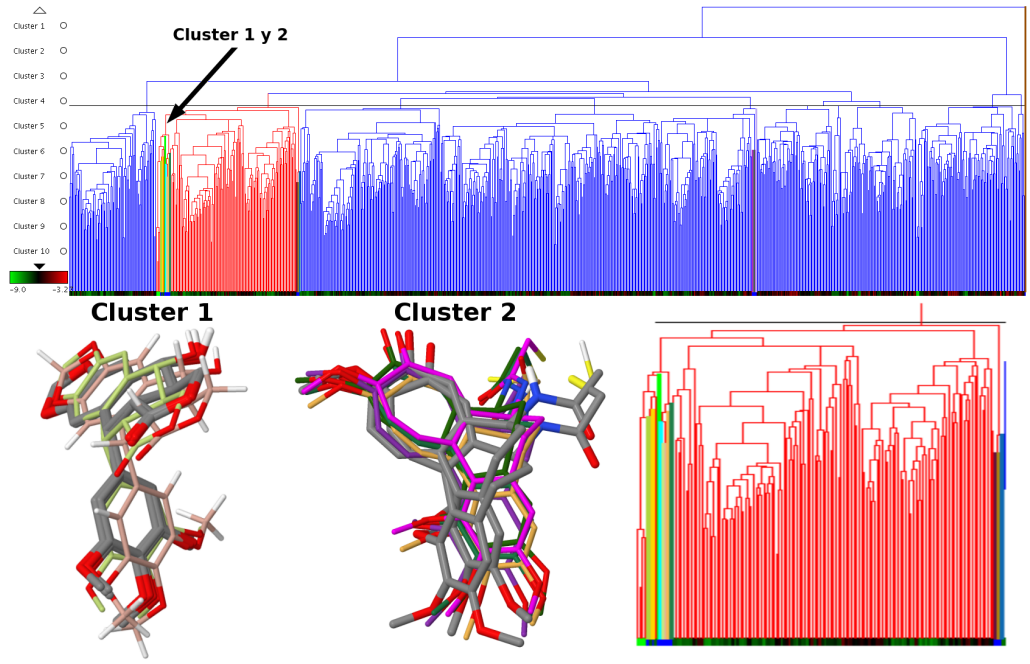


Figura 10.15: Resaltar a las referencias en el dendrograma ayuda a localizar aquellos representantes que son similares o que sugieren similaridad con ellas. El *cluster 1* es similar a sus referencias al igual que el *cluster 2*.

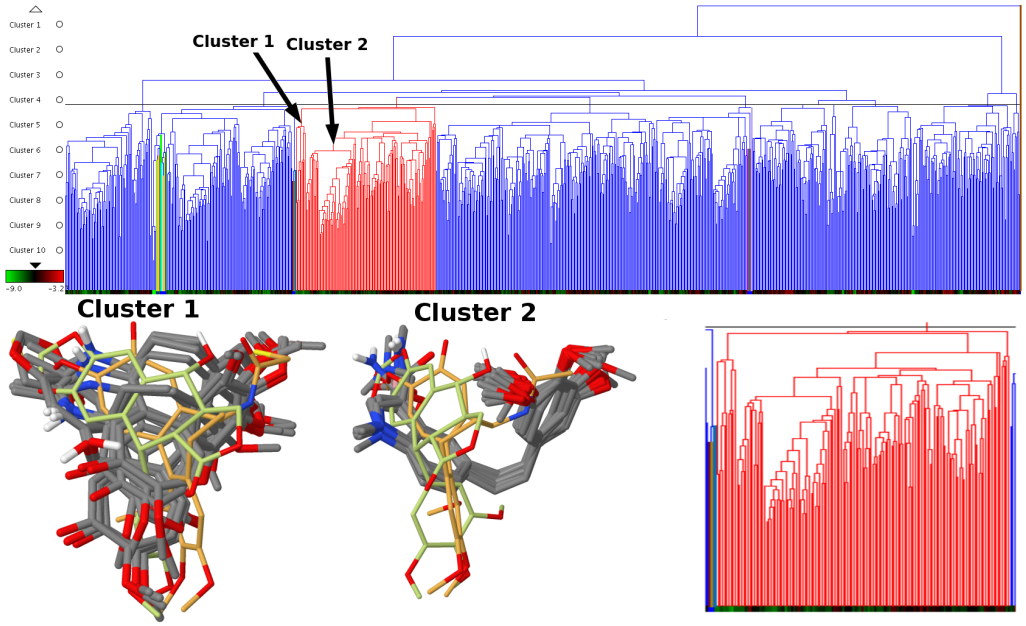


Figura 10.16: Los representantes son escasos en la zona. El *cluster 1* es similar a las referencias en la forma geométrica pero están orientadas de forma diferente. En el *cluster 2* sus representantes son similares, pero no a las referencias; no obstante, cubren la zona de las referencias.

Distancia Promedio con Grupo de Observadores (*Average-groups*)

De partida se puede observar que el método ha agrupado en un *cluster* a todas las referencias y los representantes que rodean a las referencias indican una energía de unión pobre en el rango de -4.83 a -7.02 (figura 10.17). Por lo tanto, nos enfocamos en tres zonas donde sobresale el color verde. El análisis lo realizamos de izquierda a derecha. En la figura 10.17, la consideramos la primer zona por ser la de mayor tamaño, en donde sobresalen cuatro *clusters*, y, dado que las referencias no forman parte de ese *cluster*, al superponerlas observamos que cubren la zona perfectamente; por otro lado, al visualizar individualmente cada *cluster*, observamos que la agrupación de cada uno está bien agrupada, en otras palabras, hay similaridad en los representantes. El rango de valores de la zona es de -5.45 a -7.79.

La segunda zona se divide en tres *clusters* (figura 10.18). El primer *cluster* no cubre bien el sitio de las referencias, lo que también se observa para el segundo y tercer *cluster*. En términos generales, los representantes están en el rango medio bajo de energía favorable, los valores van de -4.77 a -7.46. Finalmente, en la última zona (figura 10.19), todos sus representantes son similares, sin embargo, al superponer las referencias, los representantes están en otro sitio diferente. A pesar de ello, los representantes están en un rango de nivel medio y su rango es de -5.24 a -7.69. Hacia el extremo izquierdo, se encuentra un pequeño grupo de color verde, es evidente que contiene a los representantes con mayor similitud a las referencias, lo que comprobamos al superponerlas: están la colchicina, podofilotoxina y otra que no cuenta con referencia, aunque, está dentro del sitio de las referencias, lo que indica que posiblemente sea una respuesta correcta también.

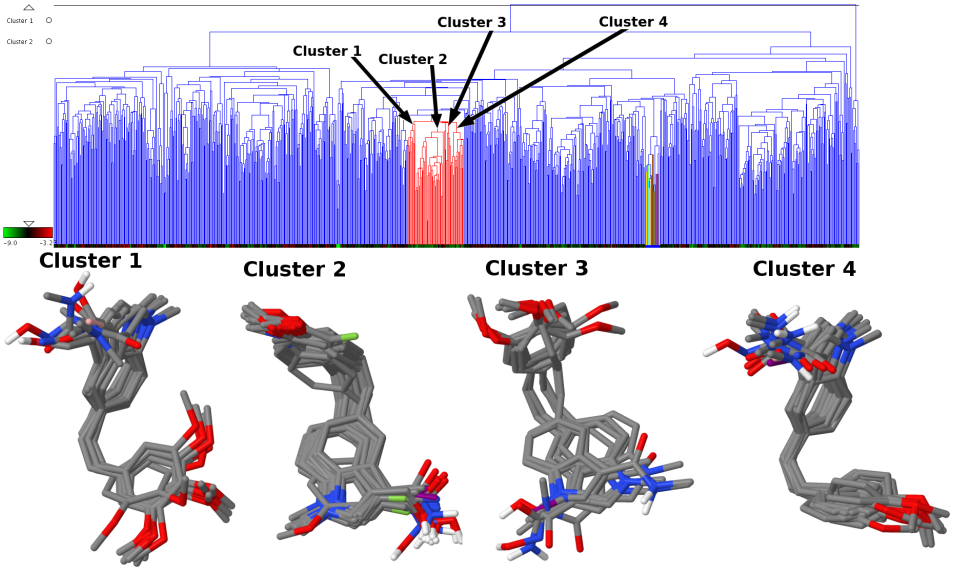


Figura 10.17: Las referencias han sido agrupadas en una rama y los representantes en ambos lados son pobres en afinidad a la diana. Esta es la mayor zona de representantes agrupados con color verde. Cuatro *clusters* son visibles en la zona. En cada *cluster* sus elementos son similares y la zona presenta una afinidad razonablemente buena con la diana.

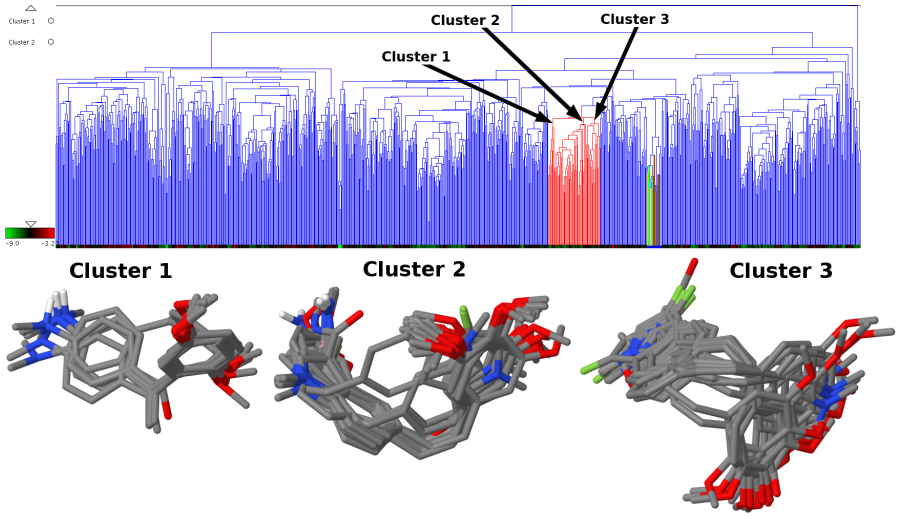


Figura 10.18: Tres *clusters* contiene la segunda zona con mayor cantidad de representantes con color verde. Todos los representantes cubren la zona de las referencias.

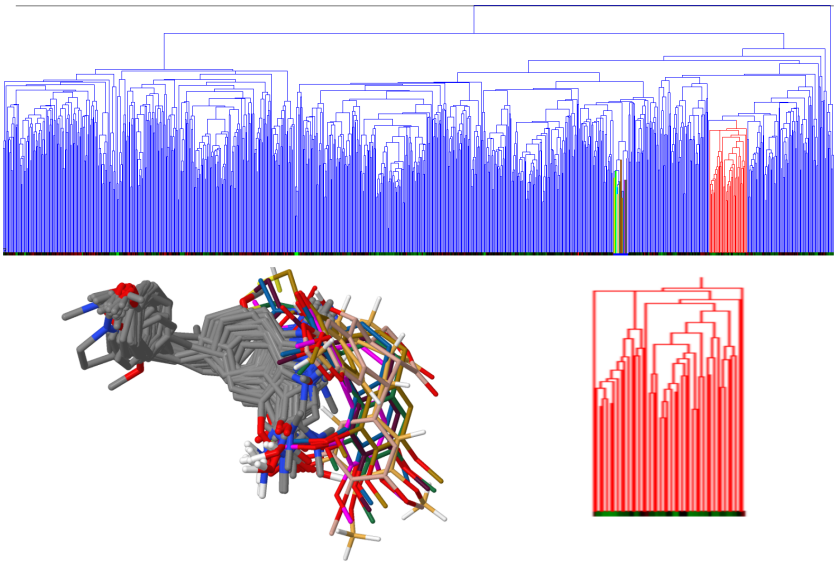


Figura 10.19: La última zona es pequeña en comparación con las otras zonas. Los representantes son similares entre ellos, pero cubren otra zona de las referencias. Este *cluster* es muy similar al de las figuras 10.11 (*cluster 2*) y figura 10.14.

Distancia Máxima con Todos los Observadores (*Complete-all*)

Aplicando la distancia máxima se observan 5 *clusters*, de los cuales solamente dos zonas contienen una cantidad suficiente de representantes con color verde. La primera zona se divide en tres *clusters* (figura 10.20) y el rango de energía de unión de la zona es de -4.98 a -8.93. Por otro lado, se trata de la zona que contiene la mayor cantidad de referencias, así como una pequeña zona de representantes para los que el color verde es claro; esto indica que contendrá una de las respuestas correctas, en otras palabras, que los representantes sean similares a sus referencias. En la figura 10.20, para el *cluster 1* los representantes son similares a sus referencias, lo que explicaría el rango maximo, además, los representantes están en la misma rama. En el *cluster 2* observamos que los representantes no son similares a las referencias pero si entre ellos, y lo mismo sucede con el *cluster 3*. La segunda zona (figura 10.21) es pequeña y no contiene referencias: al superponer las referencias observamos que los representantes cubren una zona diferente; sin embargo, hay similitud entre los elementos del *cluster*, así como con los *clusters* de la figura 10.19. El rango de energía de unión es de -5.12 a -7.69.

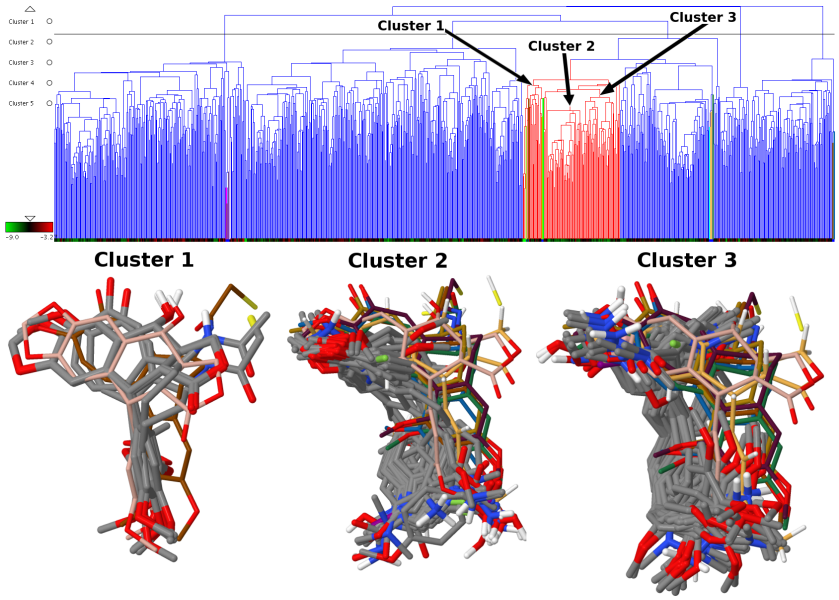


Figura 10.20: Por la escasa zona de color verde, el análisis se dividió en dos zonas. La zona uno se presenta aquí. Los elementos de una de las ramas sugieren similitud con las referencias, lo cual se corrobora con el *cluster 1*.

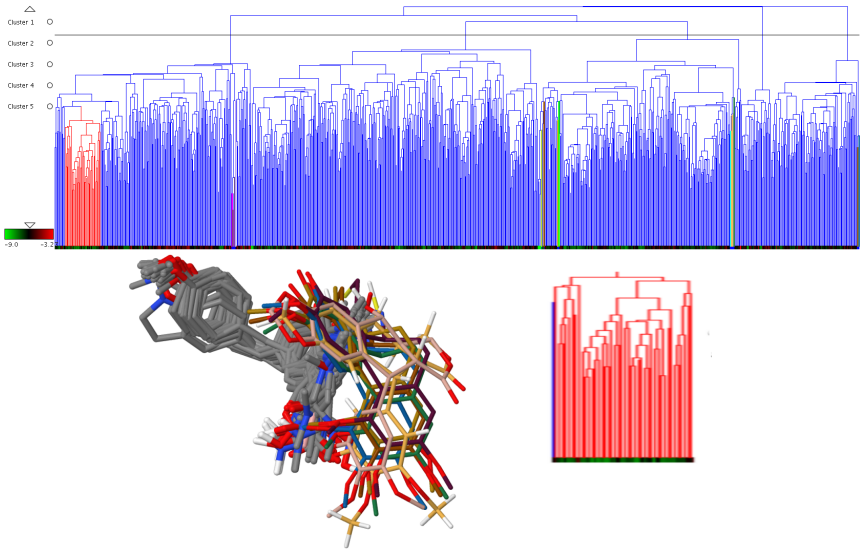


Figura 10.21: Esta pequeña zona de representantes con color verde está a un lado de las referencias, sin embargo, su rango de energía de unión es medio alto en afinidad a la diana, además de ser parecido al *cluster* de la figura 10.19.

Distancia Máxima Seleccionando los Observadores Br, Cl, HD, F, I, C (*Complete-each*)

Cuatro zonas resaltan debido a las referencias, al bajar la barra para cortar estas zonas, se obtienen cinco *clusters* –de izquierda a derecha-, de entre los cuales el tercero contiene la mayor cantidad de representantes con energía favorable (figura 10.22), y en el primero se observa una pequeña parte en color verde (figura 10.23), mientras los demás están muy mezclados en color. Aquí se presenta la misma situación que en el caso de distancia promedio seleccionando los mismos observadores (*average-each*), por lo que el análisis se basó en las zonas de color verde del *cluster 3*. Comenzando con la primera zona de verde (*cluster 3*). El rango de los representantes es de -4.71 a -7.79. Los *clusters* en su mayoría están sobre las referencias (figura 10.22), al centramos en el *cluster 1*, comprobamos que es muy similar a las referencias, aunque los representantes están orientados de manera diferente. En la segunda zona (figura 10.23) observamos que hay similitud entre los representantes, sus rangos son de -5.4 a -7.69; sin embargo, al superponer las referencias en el *cluster 1*, observamos que el sitio donde se encuentran es diferente al de las referencias. Por otro lado, el *cluster 2* es el que contiene a los representantes que son similares a las referencias, pero se esperaría que los valores del rango fueran los más cercanos al máximo valor de la escala, lo cual indica que probablemente son una variación de las referencias.

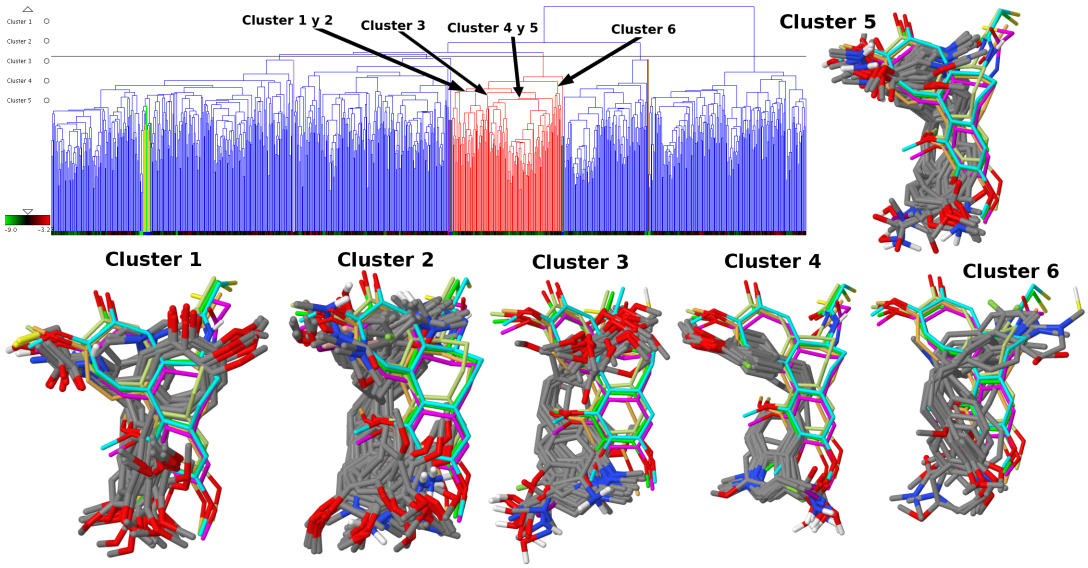


Figura 10.22: Seis *clusters* se encontraron en la primera zona (*cluster 3*), la cual concentra el mayor número de representantes con color verde. Se esperaría que el *cluster 1* fuera similar a las referencias, sin embargo, la orientación es diferente. Los demás *clusters* son similares entre ellos y cubren bien la zona de las referencias.

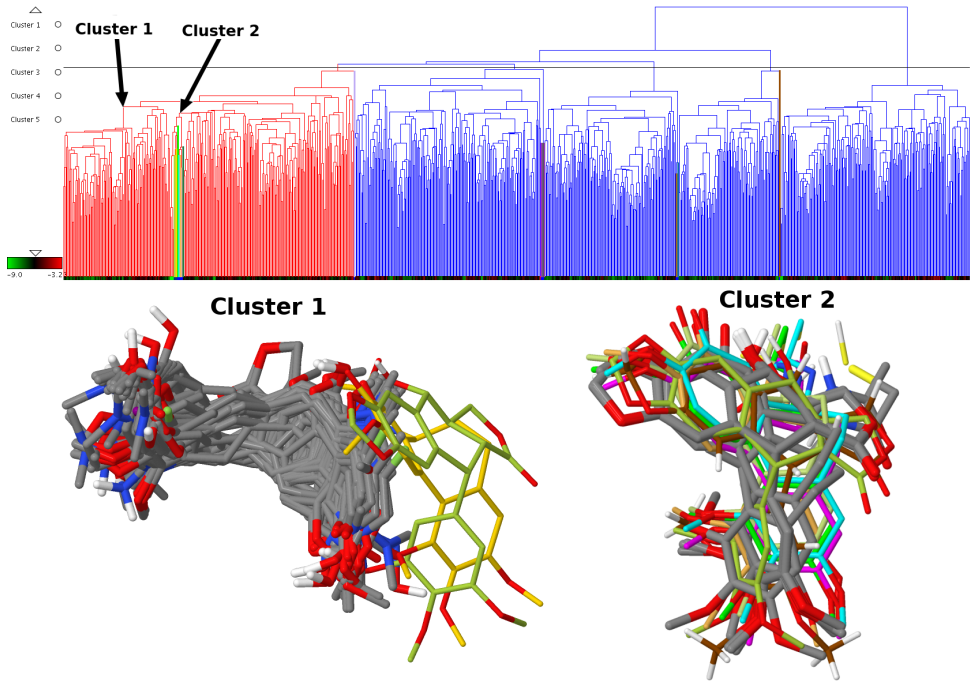


Figura 10.23: La segunda zona (*cluster 1*) muestra solamente dos *clusters*. El *cluster 2* contiene a la mayoría de las referencias y al superponerlas se observa que son similares, pero, según el código de colores de la izquierda, su energía de unión no es el valor máximo, lo que sugiere que son variaciones de las referencias.

Distancia Máxima con Grupo de Observadores (*Complete-groups*)

Se generó una gran cantidad de *clusters*, por lo que fue imposible definir un corte, pues al final había demasiados *clusters* unitarios, y, al igual que en el método de distancia promedio,el resultado fue una concentración de todas las referencias en un *cluster* rodeado de representantes de energía media. Sobresalen dos zonas con mayor número de representantes con color verde. El rango de la primera zona va desde -4.71 a -7.79. En la primera zona se forman tres *clusters* y estos cubren dos porciones diferentes, en otras palabras, las orientaciones están en sentidos opuestos (figura 10.24). El primer y segundo *clusters* son muy similares en orientación y el tercero está orientado hacia otro lado. En la otra zona (figura 10.25), los representantes son similares entre ellos y no hay necesidad de fragmentar la zona, el rango es de -5.24 a -7.69. Finalmente un *cluster* pequeño (figura 10.26) sobre sale por el color verde, en este *clusters* están los representantes similares a las referencias.

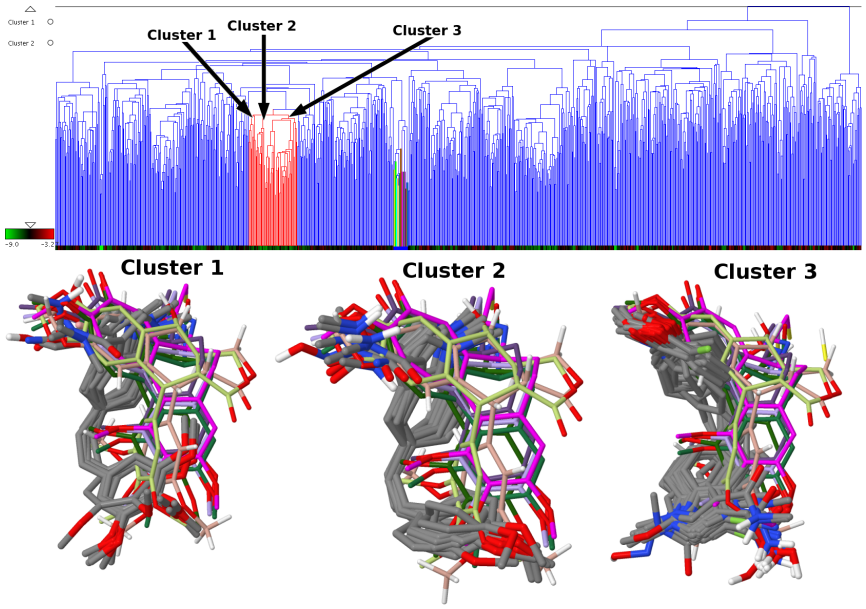


Figura 10.24: El método genera muchos *clusters*. Las zonas con representantes de color verde están muy fragmentadas. Esta zona contiene tres *clusters* dos de ellos son muy parecidos y, debido al algoritmo, la distancia de separación de las ramas del *cluster 1* y *2* es tan pequeña, que prácticamente parecen unidos. Por otra parte, el *cluster 3* contiene otros representantes que cubren la zona de las referencias.

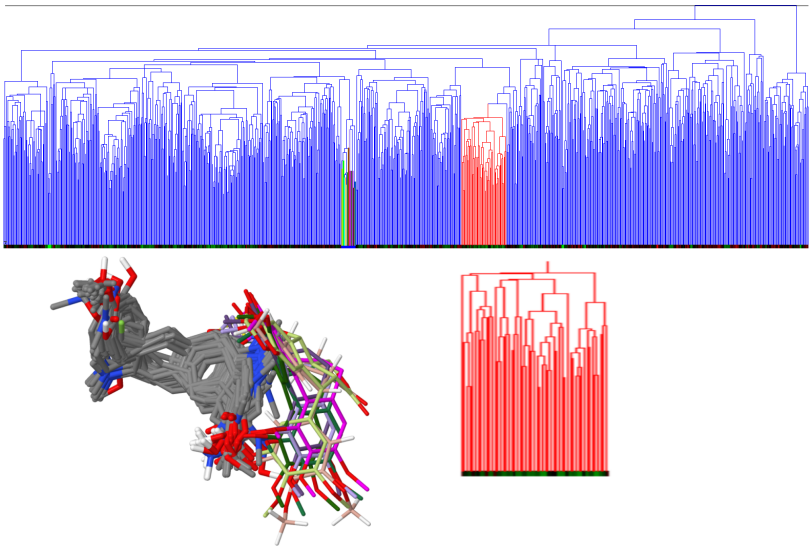


Figura 10.25: La segunda zona es similar a la de la figura 10.23.

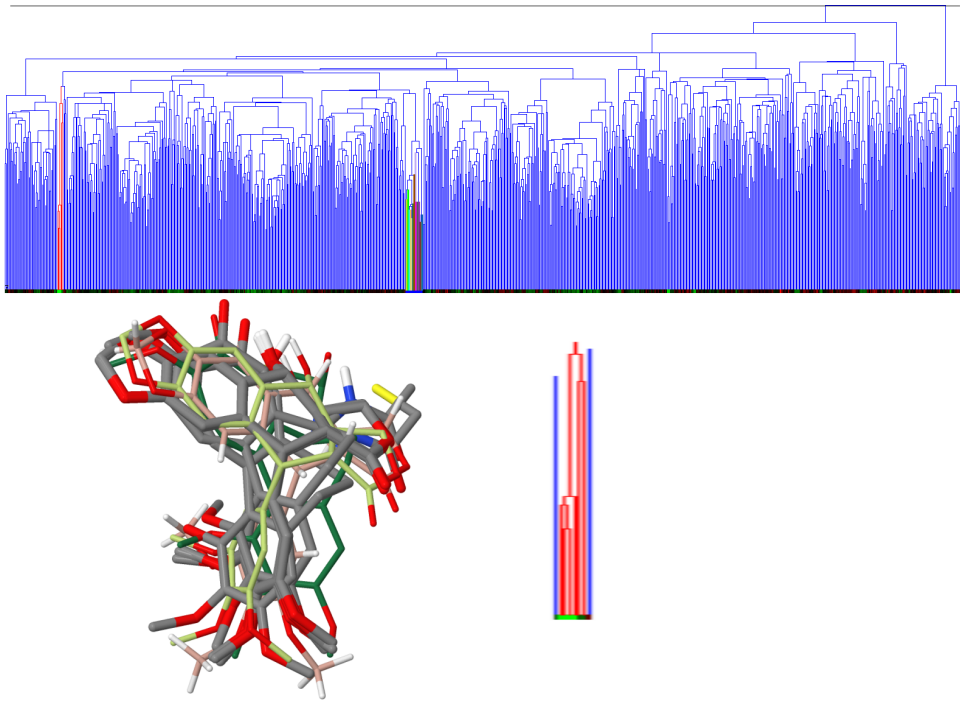


Figura 10.26: Esta pequeña zona que resalta por su color verde claro son las respuestas correctas, en otras palabras, son similares a sus referencias.

Distancia Mínima con Todos los Observadores (*Single-all*)

Se observan ocho *clusters* (figura 10.27), dos que concentran una gran cantidad de representantes y un *cluster* que contiene solamente referencias. El sexto *cluster* de izquierda a derecha contiene el mayor número de representantes con color verde, e incluye algunas referencias. Una pequeña zona, que en la figura 10.27 es el *cluster 1*, contiene a los representantes del máximo valor de energía de unión de acuerdo con el código de colores de la izquierda. Hasta el momento es el peor método para agrupar a los representantes, ya que la mayoría de los representantes de energía favorable están mezclados y no hay zonas en donde centrar el análisis, salvo ésta.

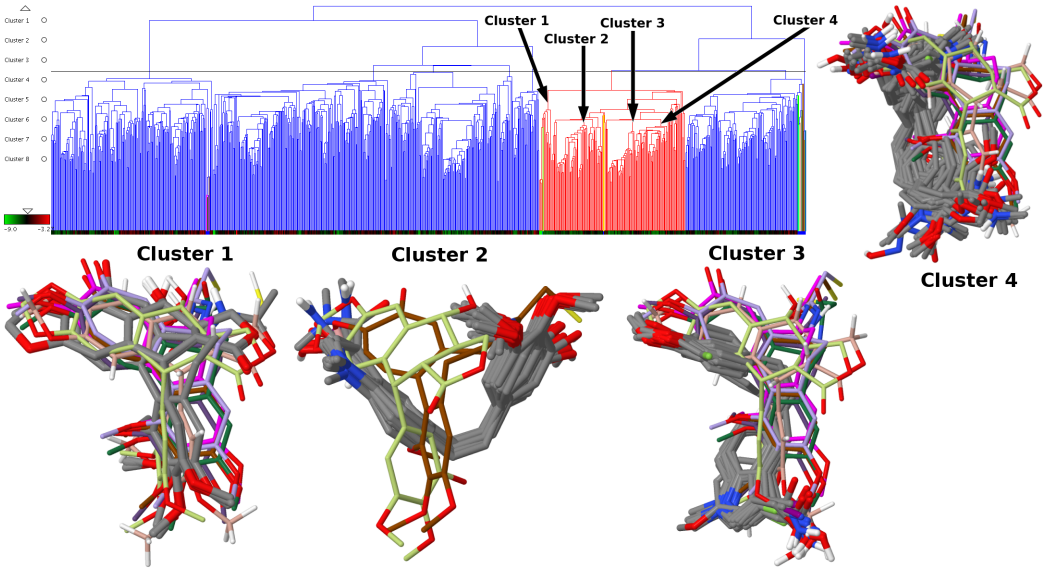


Figura 10.27: El método agrupó en diversas zonas muy pequeñas a los representantes de energía favorable. La zona muestra la mayor concentración de representantes con color verde. En la zona se encuentran los representantes que son similares a las referencias y están en el *cluster 1*. El resto de *clusters* están bien agrupados y están en la zona de las referencias.

Distancia Mínima Seleccionando los Observadores Br, Cl, HD, F, I, C (*Single-each*)

La concentración de representantes con color verde se distribuye en tres zonas. La figura 10.28 muestra los dos *clusters* en que se divide la zona, los rangos. Tanto en el *cluster 1* como en el *2*, los representantes están a un lado de las referencias, las cuales están agrupadas en su mayoría en el extremo izquierdo; sin embargo, en ambos lados los representantes en su vecindad son de energía media. Por tanto, al superponerlas con los representantes de los *clusters 1* y *2*, se ve claramente que no forman parte de la zona de las referencias, aunque, eso sí, los elementos del *cluster* son similares. En el caso de la segunda zona (figura 10.29), la cual es pequeña en comparación con las otras dos zonas, podemos decir que debido a que el método tiende a agrupar aquellos elementos que estén más próximos en términos de distancia, se ha agrupado a esa cantidad de representantes. Finalmente, la última zona (figura 10.30) es la que contiene a las respuestas correctas, donde se agrupó a otros representantes que están sobre la zona de las referencias. Este método encontró una mejor agrupación que el anterior, o, en otras palabras, resolvió de mejor manera el agrupamiento de representantes por tipo de átomo.

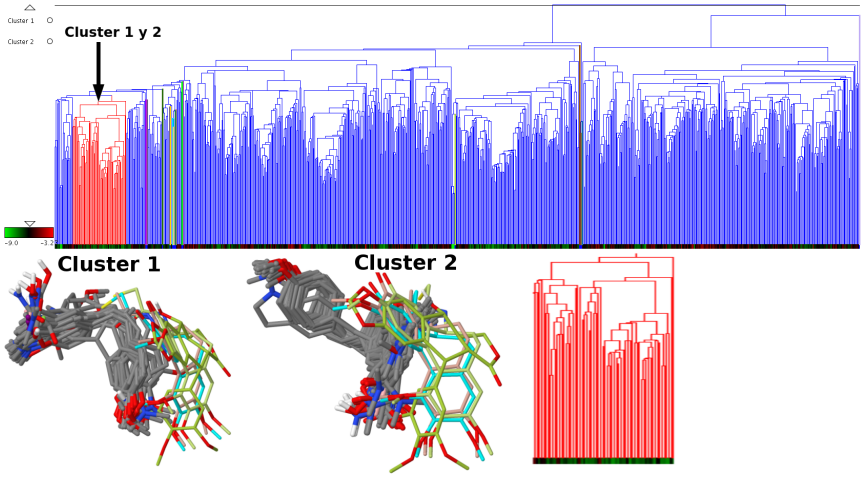


Figura 10.28: La zona está situada cerca de las referencias, pero al superponerlas los representantes no forman parte de la zona de las referencias, aunque entre los elementos sí exista similaridad.

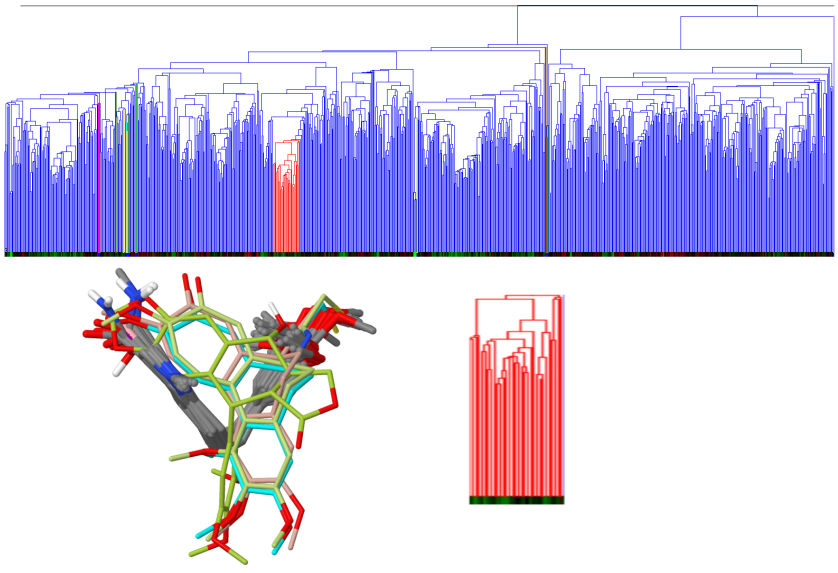


Figura 10.29: En otra de las zonas, la más pequeña, encontramos que sus elementos son similares, pero no forman parte de la zona de las referencias.

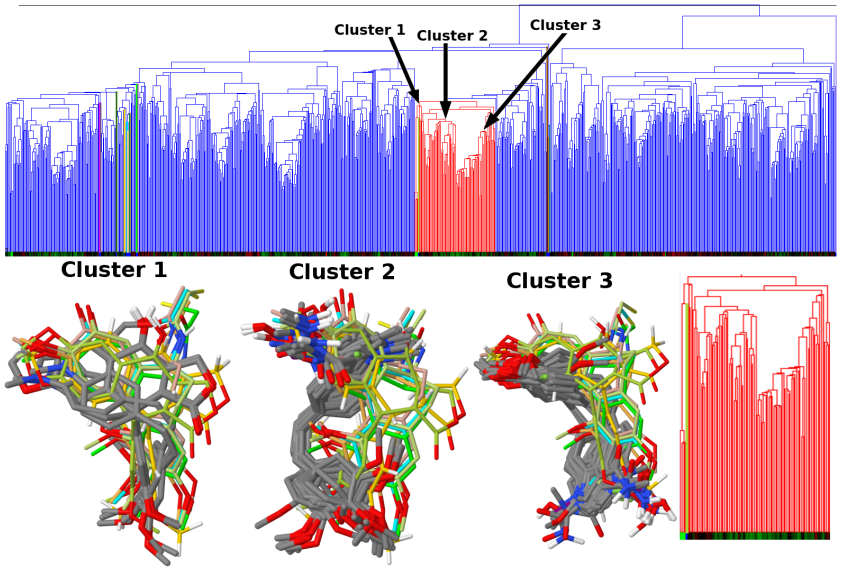


Figura 10.30: Esta zona es relativamente grande. En ella se encuentran los representantes que son similares a sus referencias, como es el caso del *cluster 1*. Por otro lado, los representantes de los otros *clusters* están en la zona de las referencias, aunque sin ser similares a ellas.

Distancia Mínima con Grupos de Observadores (*Single-groups*)

El análisis de todos los casos en que se aplicó el uso de observadores por grupo de átomos nos dice dos cosas: a) que ninguna de las referencias contiene átomos que pertenezcan a alguno de los grupos utilizados como observadores, lo cual es improbable, ya que debe contener por lo menos un átomo de alguno de los grupos para ser viable como fármaco; b) que el número de observadores de tipo de átomo es muy pequeño y/o está muy disperso en el espacio, lo que no logra aportar un peso adecuado en la comparación de representantes. En otras palabras, supongamos que hay un total de 20 observadores del tipo de átomo de flúor (F) y 500 observadores del tipo de átomo de carbono (C). Resulta evidente que los observadores de carbono tendrían más peso que los de flúor en la comparación de representantes. Así, el algoritmo realizaría realmente la agrupación en base a átomos de carbono. Supongamos ahora que hay el mismo número de átomos para los dos tipos, pero el flúor se encuentra disperso en el espacio, y para el caso del carbono hay zonas con suficientes puntos, pero que, al igual que los puntos del flúor están dispersos en el espacio. Pues bien, en esta situación sucedería lo mismo que en el ejemplo anterior, tendrían más peso los observadores de carbono, porque al estar muy dispersos los observadores del flúor se obtendrían solamente algunas observaciones significativas en comparación con las de los grupos de observadores del carbono. Por lo tanto, podemos afirmar que la agrupación de las referencias en un sólo *cluster* tiene como explicación que los observadores del grupo donde se encuentra el carbono y el grupo del electrón tuvieron más peso que el resto de grupos de observadores.

En el dendrograma observamos tres zonas donde se concentra la mayor cantidad de color verde. En la figura 10.31 podemos observar que el resultado es similar al *cluster* de la figura 10.25, el rango de valores de los representantes es de -5.55 a -7.69 -incluso los rangos de valores son parecidos-. Los representantes son similares, no están sobre la zona de las referencias, sino a un lado de ellas. La figura 10.32 muestra la segunda zona, la cual es similar a la de la figura 10.29, el rango de valores de energía de unión es de -4.77 a -7.46, al lado de esta zona hay otra pequeña, en la cual están los representantes que son similares a sus referencias (colchicina y podofilotoxina). Finalmente, en la última zona (figura 10.33), encontramos dos *clusters*, los cuales están sobre la zona de las referencias; sin embargo, están mezcladas con representantes de energía media.

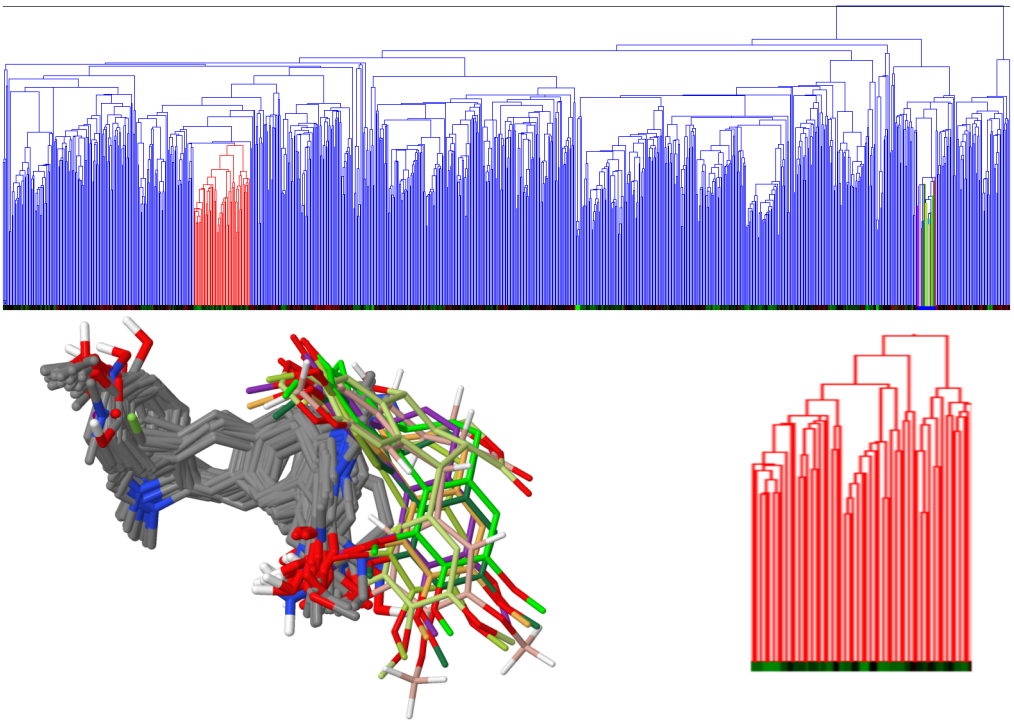


Figura 10.31: Todas las referencias han sido agrupadas en un sólo *cluster*. Los representantes son similares entre ellos mas no a la referencia, por otro lado, la figura es parecida a la de la figura 10.25.

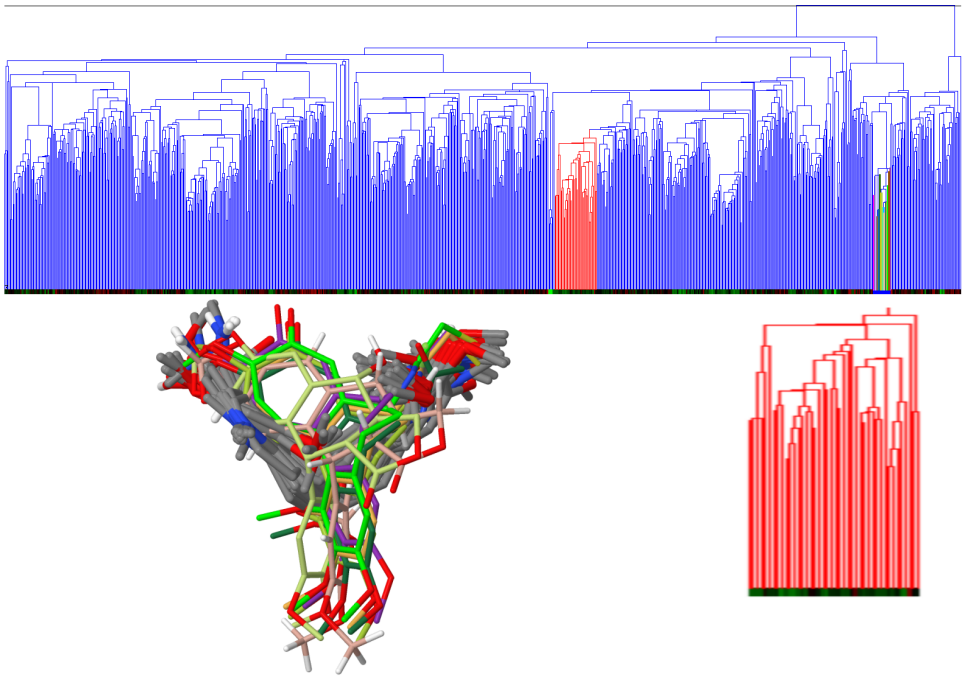


Figura 10.32: El *cluster* es parecido al de la figura 10.29.

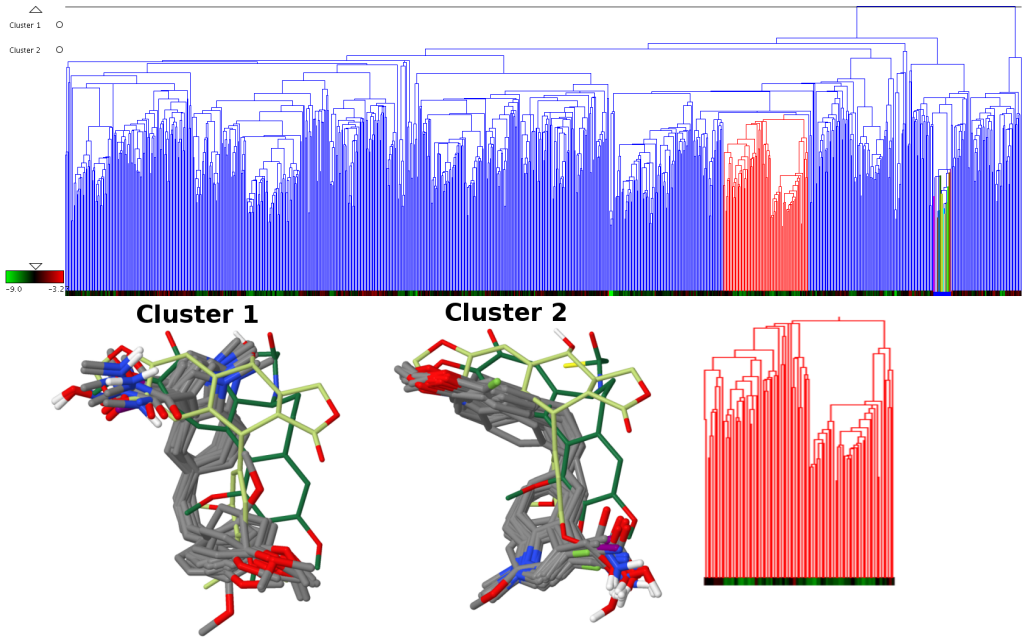


Figura 10.33: Se puede apreciar que hay una mezcla de representantes con energía de unión media.

10.2.2. Análisis del *Clustering* de Representantes por Filtrado de RMSD

Probamos las mismas opciones que en el caso anterior, pero debido a la gran cantidad de representantes, el color verde se mezcla aún más, por lo tanto aplicamos la misma forma de análisis que en los casos anteriores; nos centramos en visualizar las zonas donde se concentra la mayor cantidad de color verde. Llegados a este punto, hay que hacer una aclaración, en este caso han entrado una gran cantidad de representantes con energía de unión poco favorable, por lo que los rangos mínimos serán menores que en el caso anterior. Esto a priori podría parecer que tendría gran repercusión, sin embargo, dado que la mayoría de representantes no tiene una referencia asociada, necesitaríamos la mayor cantidad de información posible para determinar si el resultado del *docking* fue satisfactorio o no. En otras palabras, en términos de facilitar el análisis de *docking*, el químico preferiría analizar una menor cantidad de representantes si tuviera un punto de comparación como las referencias, pero en el caso de no tenerlas, se dificulta determinar si el *docking* fue satisfactorio o no. Por tal motivo, es razonable pensar que el químico preferiría tener la mayor cantidad de información posible, sin que llegue a ser abrumadora, o lo que es lo mismo, sería deseable disponer de un mayor número de conformaciones entre las que seleccionar una de acuerdo a su experiencia.

Distancia Promedio con Todos los Observadores

Al aplicar la distancia promedio utilizando todos los observadores se obtuvieron dieciocho clusters, siete de ellos muy grandes. Las referencias se distribuyeron en 7 sitios a lo largo del dendrograma, predominando el color rojo oscuro, con muy pocas zonas de color verde, debido a la gran cantidad de representantes. Se lograró detectar tres zonas. La zona 1, contiene cuatro *clusters* y una referencia, por lo que el análisis se centró en los representantes que estuvieran cerca de la referencia. El *cluster 1* es similar a la referencia pero orientado hacia otro lado. Los otros *cluters* son similares entre ellos y, aunque están sobre la zona de la referencia, no son similares a ella. La zona 2 (figura 10.34) es similar a la de la figura 10.28.

La tercera zona está en el noveno *cluster* del dendrograma (figura 10.35), empezando por la izquierda, que es, además, el *cluster* que contiene más referencias (figura 10.35). El *cluster 1* de la tercera zona es el que contiene a los representantes que son similares a su referencia, que en este caso es la colchicina. Las otras dos opciones de observadores (selección de átomos y grupo de átomos) presentaron los siguientes resultados: al seleccionar los mismos átomos que en el caso anterior, encontramos que los representantes fueron agrupados de acuerdo a su similitud, sin embargo, las zonas de color verde son demasiado pequeñas y la mayoría de los representantes con energía media y baja se mezclaron con los de buena energía. En otras palabras, hay un exceso de representantes que tienen una energía poco favorable y que son similares a otros de su mismo experimento de *docking*, pero probablemente tenían un pequeño giro o estaban alejados un poco más del umbral y el RMSD no pudo distinguir que forman parte del mismo grupo. El hecho de que no todos tienen los mismos átomos influyó al realizar la comparación. Finalmente, al aplicar el método de grupos de átomos, nos topamos con la misma situación que en los otros casos: las referencias estaban situadas en un mismo *cluster* y las zonas de color verde eran muy ambiguas, lo que dificultaba el análisis. Por tanto, podemos afirmar que el emplear a todos los observadores aplicando la distancia promedio logra agrupar satisfactoriamente a los representantes de mejor energía de unión con la diana.

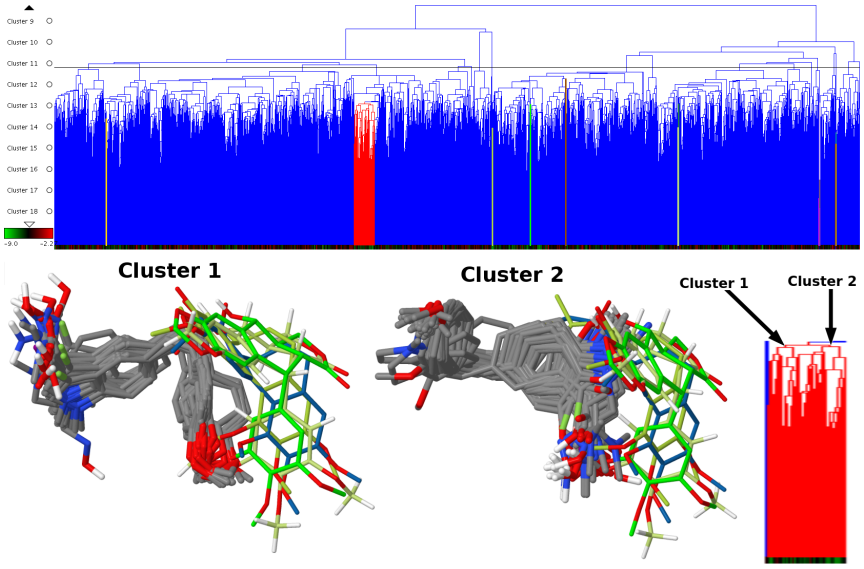


Figura 10.34: La visualización muestra dieciocho *clusters* -etiquetas a la izquierda del dendrograma- al bajar la barra en el dendrograma al intentar obtener *clusters* de similar tamaño. Las zonas con color verde son muy escasas, lo cual dificulta el análisis. Los representantes agrupados son parecidos al *cluster* de la figura 10.28.

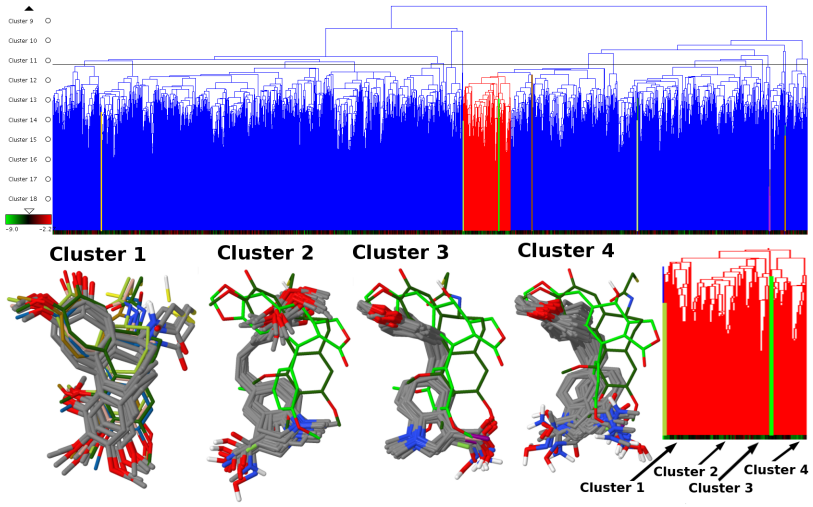


Figura 10.35: Empezando por la izquierda, el noveno *cluster* es el que contiene más referencias agrupadas, sin embargo, la zona de color verde es escasa. Por otro lado, el *cluster 1* es similar a su referencia, los otros *clusters* son similares solamente entre ellos.

Distancia Máxima con Todos los Observadores

Se esperaba que el empleo de este método diera mejores resultados que el anterior, ya que los representantes son demasiados y cubren diversas zonas de la diana. Sin embargo, el resultado arrojó pocas zonas con representantes de color verde empleando las tres opciones. Aun así, las tres opciones sí agruparon bien a los representantes. La opción de emplear a todos los observadores distribuyó a las referencias en los dos extremos (figura 10.36) y en ese sitio agrupó a unos representantes con energía favorable, sin embargo, al superponer sus referencias, se desveló que estaban orientadas en sentido contrario. La opción de selección de átomos (figura 10.37) agrupó a las referencias hacia la derecha. Una pequeña zona a la derecha del dendrograma -específicamente el *cluster* seleccionado en rojo en la figura- contiene dos *clusters* con energía de unión favorable. Al superponer tanto los representantes como las referencias, notamos que hay similaridad entre los representantes y que estos ocupan la zona de las referencias; no obstante, no son similares a ellas. Por último, quien peor agrupó a los representantes fue la opción de grupos de átomos, situación recurrente a lo largo de todo el análisis.

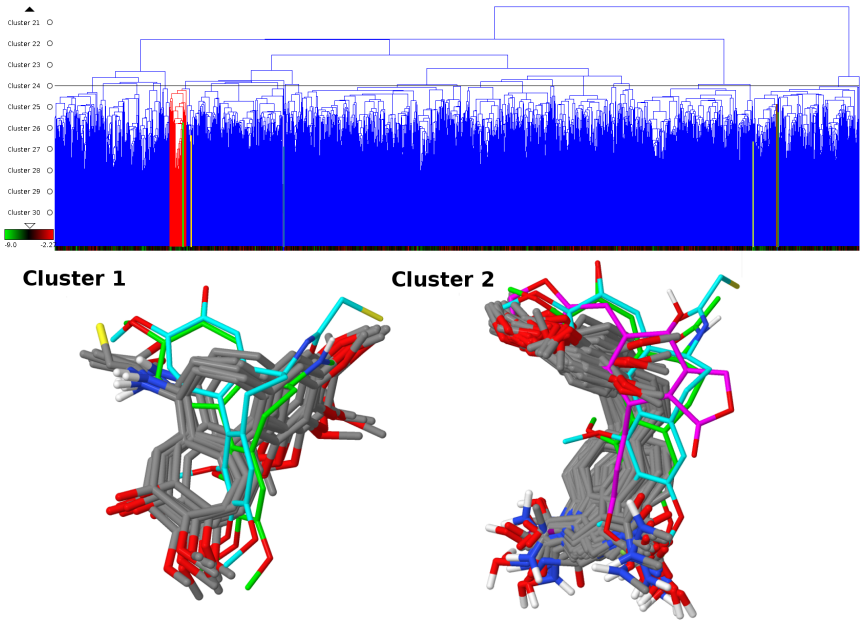


Figura 10.36: Se dividió el dendrograma en treinta *clusters* con la opción de todos los observadores. Una pequeña zona que contiene referencias resalta sobre las demás. Los *clusters 1* y *2* de dicha zona están sobre las referencias. En particular, el *cluster 1* es similar a su referencia, aunque presenta una orientación de los representantes diferente.

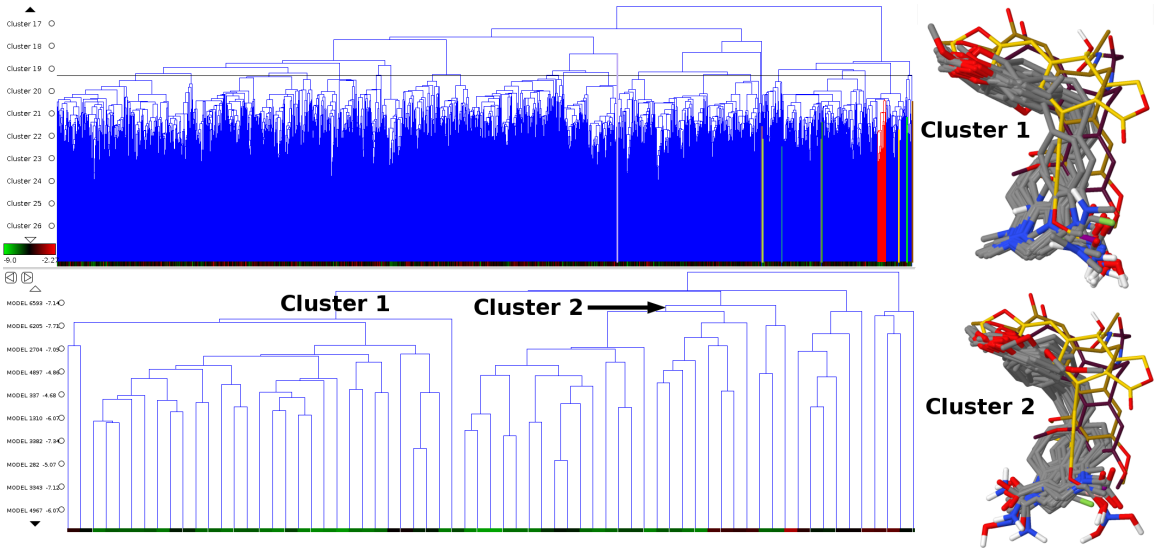


Figura 10.37: Resultado de la opción de selección de átomos. Se muestran veintiséis *clusters* de tamaño similar. El color verde está demasiado mezclado, Una pequeña zona ampliada en la parte de abajo muestra el resultado de la agrupación por el método. Los representantes son similares pero no hay referencias en el cluster. Al superponer las referencias, se muestran a los representantes en la misma zona que éstas.

Distancia Mínima con Todos los Observadores y Selección de Observadores

La opción del método de distancia mínima con las opciones de todos los observadores y la selección de observadores, dio resultados similares, hay varias concentraciones de representantes con color verde en ambos resultados. Aplicando todos los observadores, resaltan dos zonas donde hay representantes. Al visualizarlos encontramos que los representantes son similares entre ellos, pero las referencias son distintas (figuras 10.38 y 10.39), al superponer las referencias notamos que los representantes cubren el espacio de las referencias en ambas zonas de representantes. En el caso de la selección de observadores (figura 10.40), una zona resalta sobre las demás, pues el color verde es más claro que en el resto del dendrograma, además de estar entre dos ramas que tienen referencias. Al superponer los representantes y las referencias, observamos que los representantes son similares entre ellos pero no a las referencias. Una vez más, este resultado es prácticamente lo mismo que sucede en los casos anteriores en que se aplica la misma opción.

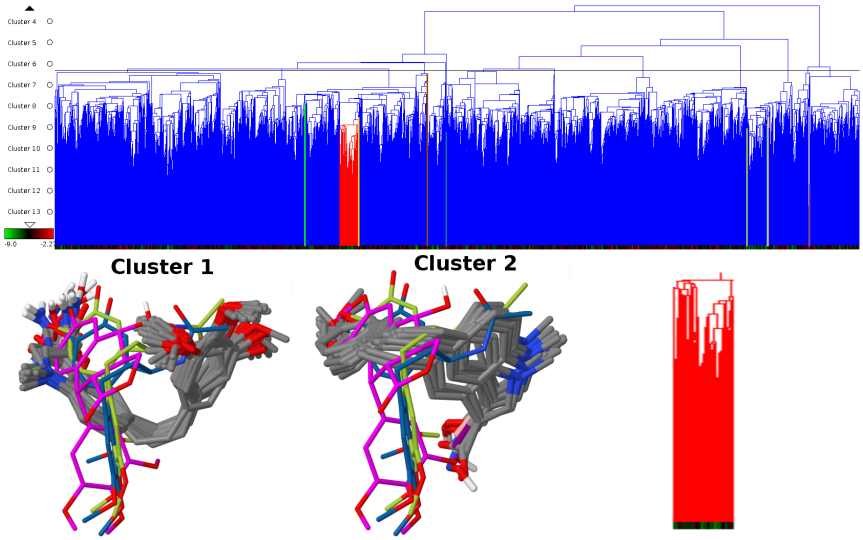


Figura 10.38: La zona seleccionada contiene una referencia. Se observa en la superposición que los representantes son similares entre ellos, y, aún cubriendo la zona de las referencias, no son similares a ellas.

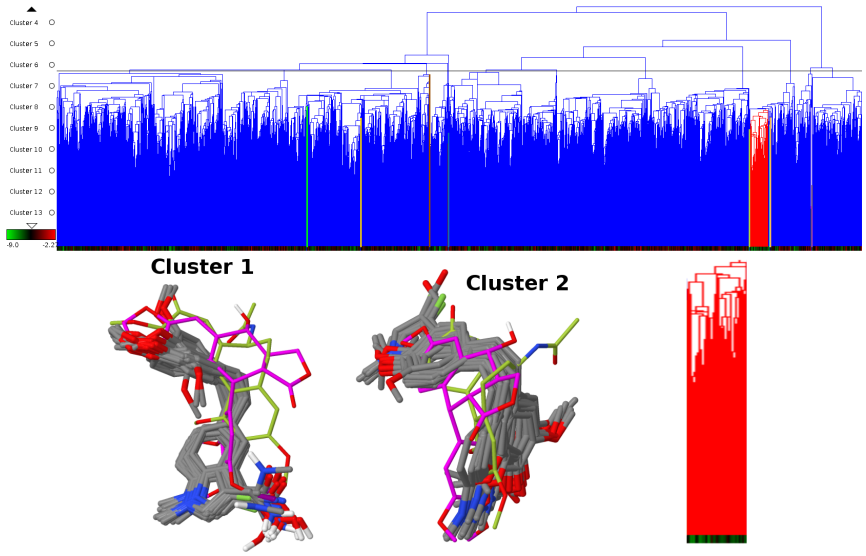


Figura 10.39: Los representantes están entre dos referencias, la zona de representantes con energía de unión favorable es pequeña en comparación con el resto de representantes donde la mayoría está en un rango medio bajo.

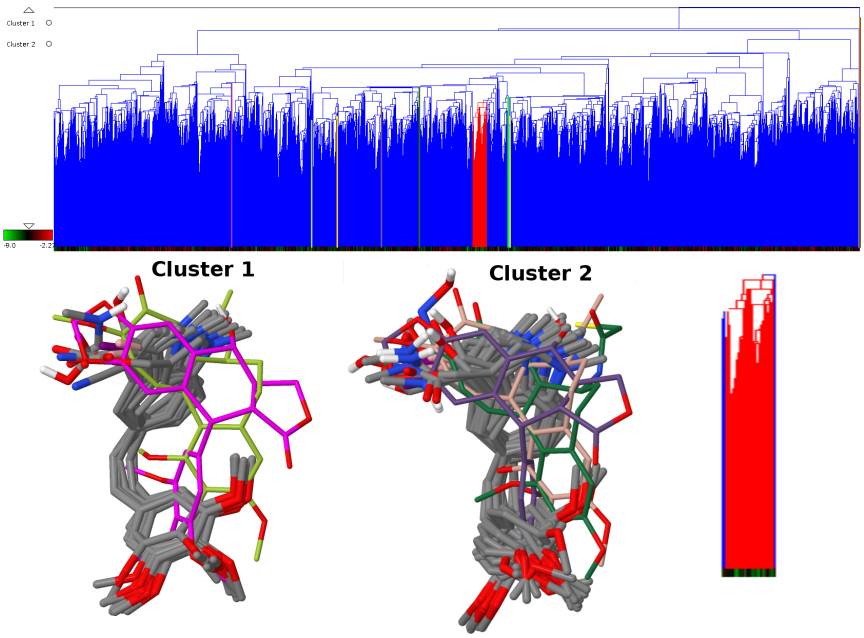


Figura 10.40: La zona resalta por el color verde que indica que hay representantes con energía de unión favorable. Los *clusters* muestran que están en el sitio de las referencias y son similares entre ellos.

En conclusión al análisis de resultados del filtrado de representantes por el método jerárquico de distancia máxima: podemos afirmar que aplicando el método de distancia promedio la opción de todos los observadores agrupó mejor a los representantes con energía de unión más favorable y a las referencias, otra opción teniendo en cuenta que las comparación las realizaría en relación a la existencia de los átomos en los representantes esta la selección de átomos. En todo caso aplicar este tipo de selección se adecuaría para cuando no se tengan referencias de los representantes y se conozcan los tipos de átomos que más relevancia tienen para la afinidad con la diana. Por otra parte, el método de distancia máxima obtuvo mejores resultados al aplicar la opción de grupos en comparación a emplear todos los observadores. Las zonas de color verde son continuas con esta opción que en la otra, donde es clara la fragmentación de los representantes, aun cuando las referencias han sido agrupadas en un sólo *cluster*, esto quiere decir que la mayoría de los representantes no son del tipo de las referencias y que tuvieron mayor peso los grupos de átomos al momento de la comparación y sobre todo a que los representantes están muy dispersos en el espacio. Finalmente el método de distancia mínima, se debaten la mejor agrupación la opción de todos los observadores y los de tipo de átomo. Sin embargo, la opción de todos los observadores es quien fragmenta menos al color verde. Y de todos los métodos el que mejor agrupó a los representantes con mejor energía es el método de distancia máxima empleando los grupos de átomos (figura 10.41), seguido de la distancia promedio empleado a todos los observadores. Y para el caso de filtrado por RMSD es el de distancia promedio empleando a todos los observadores.

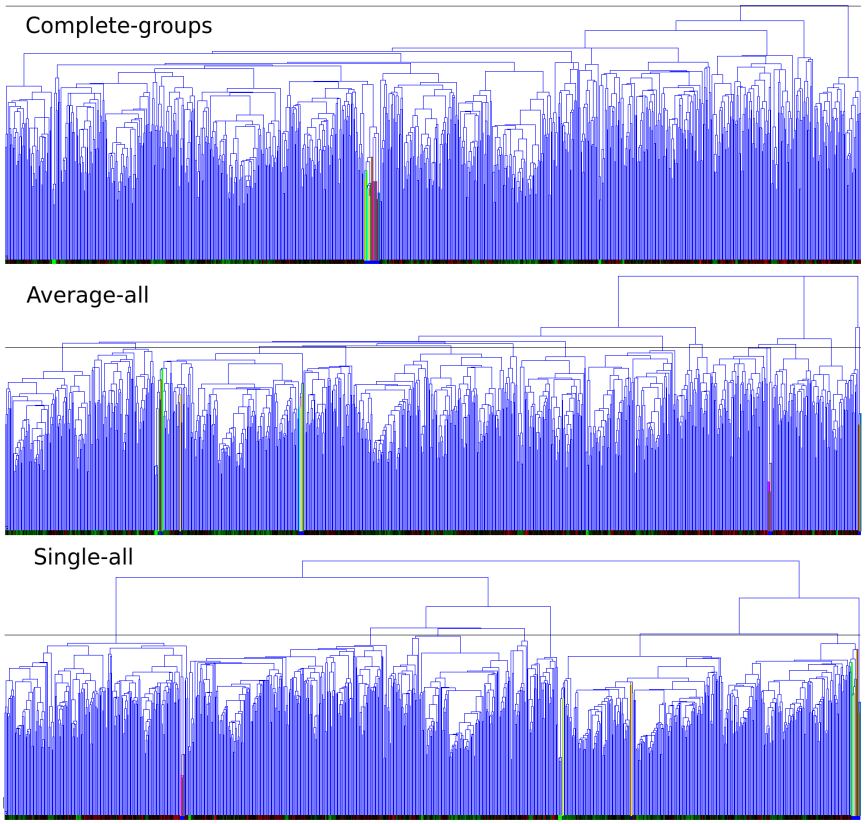


Figura 10.41: Comparación de dendrogramas. El primer dendrograma emplea el método de distancia máxima con grupo de átomos; el segundo, el método de distancia promedio con todos los observadores, al igual que el dendrograma de abajo, pero con distancia mínima.

No obstante, el número de representantes es demasiado en ambos casos, por lo cual para facilitar el trabajo de análisis se aplicaron tres filtros diferentes para reducir el número de representantes: RMSD, energía de unión, y, número de *clusters*. A modo de ilustración del funcionamiento de los filtrados, empleamos el análisis con mayor cantidad de representantes, esto es, el filtrado por RMSD y, específicamente el método de distancia promedio con todos los observadores, ya que es el segundo mejor y porque tiene una mejor distribución de las referencias en el dendrograma. Para cada tipo de representantes se calculó el RMSD a 4.0Å, en otras palabras, se reagruparon los representantes que estuvieran a una distancia menor o igual a 4.0Åy, posteriormente, se extraería un representante de nuevo; al aplicar este tipo de filtrado se reduce de 6926 representantes iniciales a 453. Para este filtrado se toman en cuenta todos los tipos de representantes. En cambio, el filtrado por energía de unión está basado en un rango; una vez establecido el rango, solamente aquellos representantes que estén dentro del rango permanecen y el resto es ignorado, esto significa que algunos representantes no estarán en el análisis. En este caso, hemos definido el rango entre -6.57 a -9.0, reduciendo de esta manera a 629 representantes. Por último, hemos filtrado los primeros dos mejores representantes de cada tipo, obteniendo así un total de 353.

10.3. Filtrado de Representantes por Distancia de 4.0 por RMSD

La escala de energía de unión ha cambiado ligeramente después del filtrado, primeramente iba de -2.27 a -9.0 y ahora es de -2.92 a -9.0. La distribución de las referencias también ha cambiado, los representantes con color rojo han sido filtrados en su mayoría y ahora resaltan más representantes en verde. La figura 10.42, muestra las zonas que se seleccionaron por contener una mayor cantidad de representantes en color verde. Los *clusters 4* y *5* de la figura 10.42 muestran a los representantes en sitios diferentes al de las referencias, al igual que en la figura 10.34, el *cluster 5*, está en la zona verde del dendrograma indicando que la orientación es afín a la diana, sin embargo, la superposición de las referencias indica otra cosa. En el caso del *cluster 4* los representantes no son muy similares y tampoco están cerca o sobre las referencias, sin embargo, son representantes con color verde; esto puede tener dos significados: 1) Los ficheros map consideran esa zona con más afinidad que el sitio de la colchicina y podofilotoxina o 2) Es un error en el cálculo de interacción y esto demanda un refinado de los ficheros map, con el fin de corregir dicho fallo. Para los otros casos, podemos decir lo siguiente: los *clusters 2* y *3* están cerca de las referencias en el dendrograma y predomina el color verde, lo cual indica que deben ser similares o que están en el mismo sitio que ellas. La superposición de las referencias mostró que el *cluster 3* está en el sitio de las referencias, y, además, es similar a los *clusters 2*, *3* y *4* de la figura 10.35. El *cluster 2* tiene mayor similitud con las referencias en su orientación a diferencia del *cluster 1*, que pasa por detrás de las referencias. En conclusión: el filtrado nos ayudó a encontrar de forma más sencilla a los representantes con mejor energía de unión a la diana; por otro lado, pudimos detectar posibles zonas de los *map* a mejorar o explotar en beneficio del ensayo de *docking*.

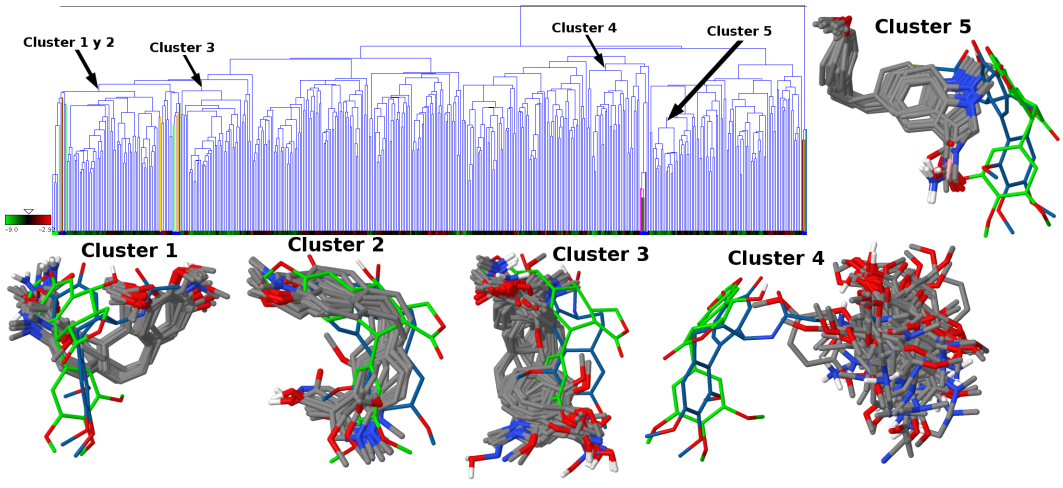


Figura 10.42: El filtrado mejora el análisis, pues muestra en mayor detalle las zonas donde se encuentran los representantes con mayor afinidad a la diana. Incluso, también se logró detectar conformaciones que posiblemente puedan modificarse, aumentando o eliminando las zonas por medio de JADOPPT.

10.3.1. Filtrado de Representantes por Rango de Energía de Unión.

El filtrado arrojó prácticamente todo el dendrograma con representantes en el rango inferior, pero consideremos que el rango inicial era de -2.27 a -9.0, esto indicaba que podíamos prescindir de aquellas conformaciones que no aportaran valor alguno al análisis de los resultados de *docking*. Por tal motivo, se elevó el rango hasta -6.57 y eso redujo los 6926 representantes de partida a 629 con energía de unión mayor a -6.5, en otras palabras, sólo el 9.08 % de los representantes son significativos en el ensayo del *docking*. Finalmente, la figura 10.43 muestra sólo dos pequeñas zonas (*cluster 1* y *4*) con representantes mayores a -6.5 de energía de unión. El *cluster 1* es similar a su referencia, al igual que el *cluster 2*, sin embargo, el *cluster 4* no es similar en orientación a su referencia. Por otro lado, los rangos de los *clusters 1* y *4* son muy cercanos, con una diferencia mínima: *cluster 1* -8.93 a -9.0 y *cluster 4*, -8.5. Es evidente que el ensayo de *docking* contiene un error en sus ficheros *map*, lo que supone una oportunidad para JADOPPT para refinar los ficheros. Los *clusters 2*, *3* y *5* se tomaron en cuenta por pertenecer a la misma rama que las referencias y al superponerlas se evidenció que, salvo el *cluster 2*, que es similar en todo a su referencia, los otros están sobre el sitio correcto y son similares entre ellos. En conclusión, esperábamos una mayor cantidad de color verde, pero el filtrado mostró que la mayoría de los representantes podría estar en el rango de -6.57 a -7.7. Esto quiere decir que posiblemente los representantes similares a las referencias son más afines a la diana que los otros. Por otro lado, también podrían modificarse los ficheros *maps* para mejorar los resultados de *docking*.

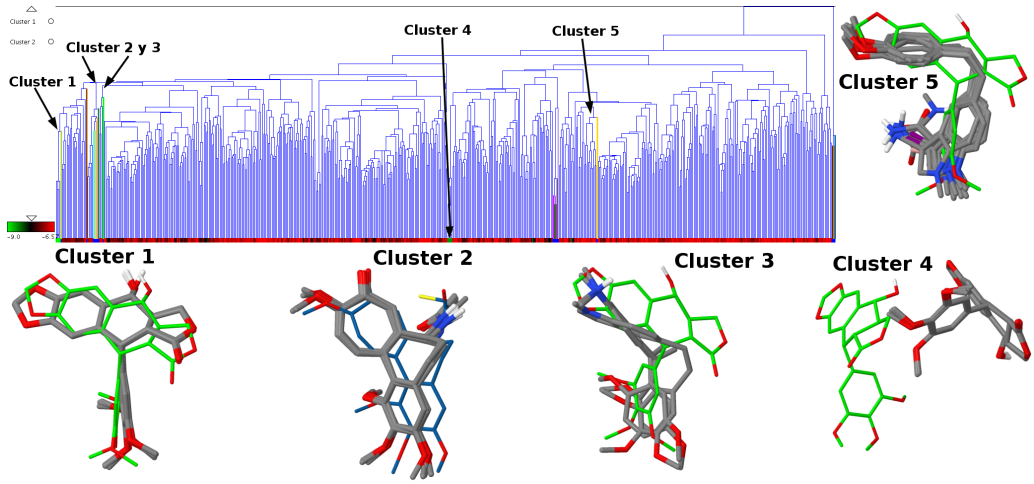


Figura 10.43: El filtrado sugiere que el rango de energía de unión de los representantes es de -6.57 a -7.7.

10.3.2. Filtrado de Representantes por los Primeros Dos Mejores de Cada Tipo.

Cuando se tiene una gran cantidad de resultados de *docking*, la estrategia a seguir es considerar solamente aquella conformación que obtenga la mejor energía de unión y a partir de allí, reducir el número de conformaciones a uno manejable. Sin embargo, esto no garantiza que las conformaciones

sean las esperadas, en otras palabras, que la conformación de mejor energía de unión sea similar a la pose de la referencia como hemos visto en los filtrados anteriores. En cambio, si partiésemos del análisis previo de todos los resultados de *docking*, podemos filtrar con garantías de seleccionar a los mejores, puesto que ya conocemos, a modo grueso, dónde están los representantes que tienen mayor similitud con las referencias –nos guiamos por los colores de las referencias en el dendrograma-. Por otro lado, la barra de colores nos guía para saber el punto de corte de los representantes, pues a mayor cercanía del color verde, mejor energía de unión existe. Esta última es la razón por la que, en este caso, hemos seleccionado las dos mejores puntuaciones. No obstante, podría ocurrir que no se encontrasen todas las conformaciones similares a las referencias. La figura 10.44 tiene pocos representantes en color verde, solamente los *clusters 1* y *3* contienen representantes con buena energía de unión. Podemos comparar los resultados de este último filtrado con los otros filtrados, por ejemplo, el *cluster 2* se corresponde con el *cluster 5* de la figura 10.43, así como el *cluster 3* con el *cluster 4*, y el *cluster 4* también con el *cluster 4* de la figura 10.42. Resaltan las referencias que se concentraron en el extremo izquierdo, indicando que posiblemente son los representantes similares a ellas, y lo corroboramos en la figura 10.44. Por otro lado, la figura 10.44 nos muestra que la mayoría de los representantes están en un posible rango de -7.7 a -9.0, de acuerdo a la escala de colores de la izquierda. No hay duda en el *cluster 1*, que son los representantes que exhiben el máximo valor de la escal. Sin embargo, el *cluster 3*, que también indica un valor máximo, cuando se visualiza, nos encontramos que los representantes no son similares en orientación a su referencia. Finalmente, en el caso del *cluster 2*, a pesar de que tiene una energía de unión medio baja, los representantes sí están en el sitio de las referencias, lo cual indica que posiblemente para sus referencias sea el sitio correcto.

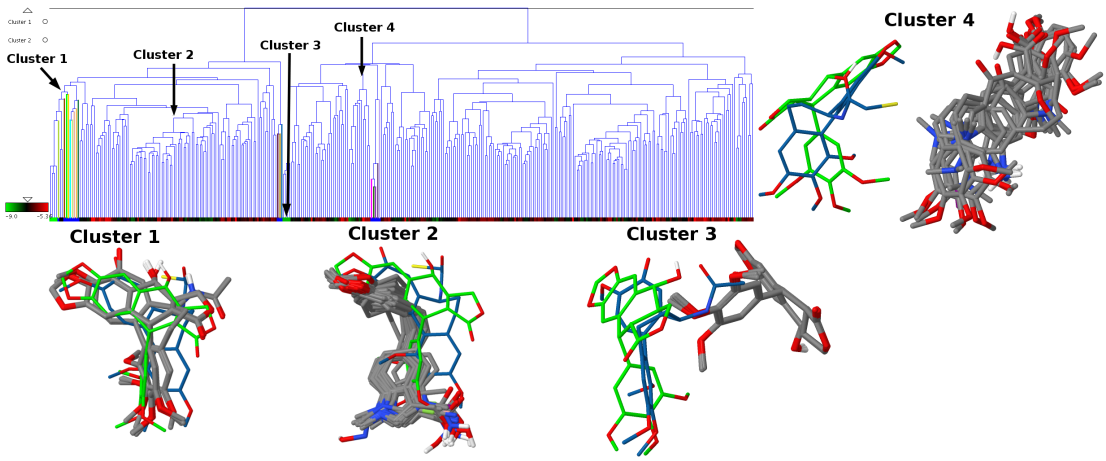


Figura 10.44: Filtrado por los primeros dos mejores representantes en relación a su energía de unión.

En conclusión, de los tres filtrados podemos decir que el filtrado por distancia RMSD podría servir para reducir el número de representantes para continuar explorando y analizando los resultados del *docking*, esto es, aunque nos permitiría ver detalles concretos lo hace de forma muy general. Por el contrario, el filtrado por energía determina qué representantes son más afines a la diana de acuerdo a la escala de colores. El inconveniente es que se pierden conformaciones que probablemente son similares a sus referencias –si las hay- pero que han sido evaluadas como poco favorables por el

programa de *docking*. Por último, el filtrado que selecciona sólo aquellos mejores resultados del *docking* podría ser empleado para un filtrado cuando se tengan enormes cantidades de representantes, pero tomando en cuenta la pérdida de información que podría o no ser útil en un futuro.

10.4. Resultados del Análisis de Visualización para la Proteasa VIH-1

Para contrastar la efectividad del método de agrupación de moléculas distintas se probó con otro conjunto de moléculas, mapas, referencias y proteína. En concreto, se trata de la proteasa del VIH-1, con un total de 1200 conformaciones (60 ficheros dlg con 20 poses por molécula), nueve ficheros tipo *map* y tres moléculas de referencia: A-98881, DMP323 y Atazanavir. El método de agrupación empleado para la selección de representantes de los ficheros dlg fue el jerárquico a fin de mostrar los dos métodos con los que cuenta JADOPPT.

La figura 10.45a muestra que se encontraron tres clusters empleando el método de distancia máxima (*complete-linkage*), sin embargo, la representación no favorece un análisis claro, debido al tamaño de las moléculas. Se puede apreciar un *cluster* en color verde, otro en color rosa y uno más en color marrón claro; los *clusters* verde y rosa aparecen claramente separados, como se muestra en la figura 10.45b. En consecuencia, todo el análisis se centrará en el *cluster 3*.

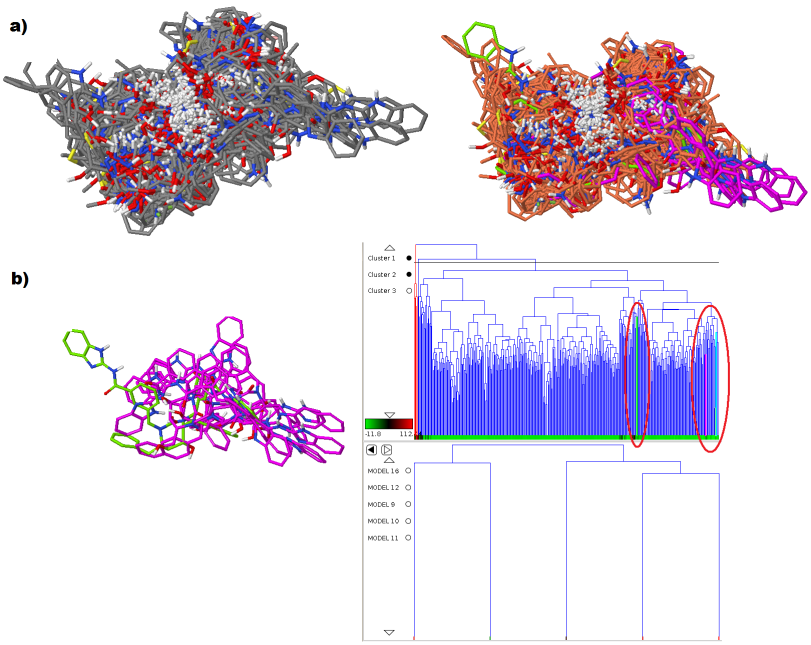


Figura 10.45: Conjunto de representantes. a) Debido al tamaño grande de las moléculas no son tan evidentes los *clusters* formados, sin embargo, el conjunto de datos está dividido en tres *clusters*. b) El *Cluster 1* y el *Cluster 2*, pese a su gran tamaño, aparecen separados.

Debido a la naturaleza de los datos, los representantes se encuentran en una misma zona -zona de interacción con la diana- muy reducida, sin embargo, hay diferencia en el tamaño y forma de algunos representantes. La figura 10.46 muestra el *cluster 3* -en la parte superior izquierda-, una sub-rama del *cluster* seleccionada en amarillo y su correspondiente representación individualizada abajo a la izquierda. Se puede apreciar que muchos de los representantes cubren la mayor parte de las referencias, por lo que se puede suponer que son similares. Hay que resaltar que todos los representantes de esa rama tienen valores de interacción favorable, al igual que la mayor parte del *cluster 3*; este aspecto no debe ser decisivo en la toma de decisiones, por lo que se debe centrar en la similitud de los representantes.

La mayoría de las referencias tienen forma de H tridimensional, por lo que se dificulta aún más poder encontrar similitud entre los representantes. Sin embargo, al explorar las sub-ramas del *cluster 3*, se van encontrando representantes más pequeños que cubren ciertas partes de la H (ver figura 10.47). La mayor parte del *cluster 3* contiene representantes que cubren todas las partes de las referencias, sin embargo, las ramas que están más próximas a las referencias en el dendrograma empiezan a mostrar las estructuras que comparten similaridad con ellas, como se muestra en las figuras 10.48 y 10.49.

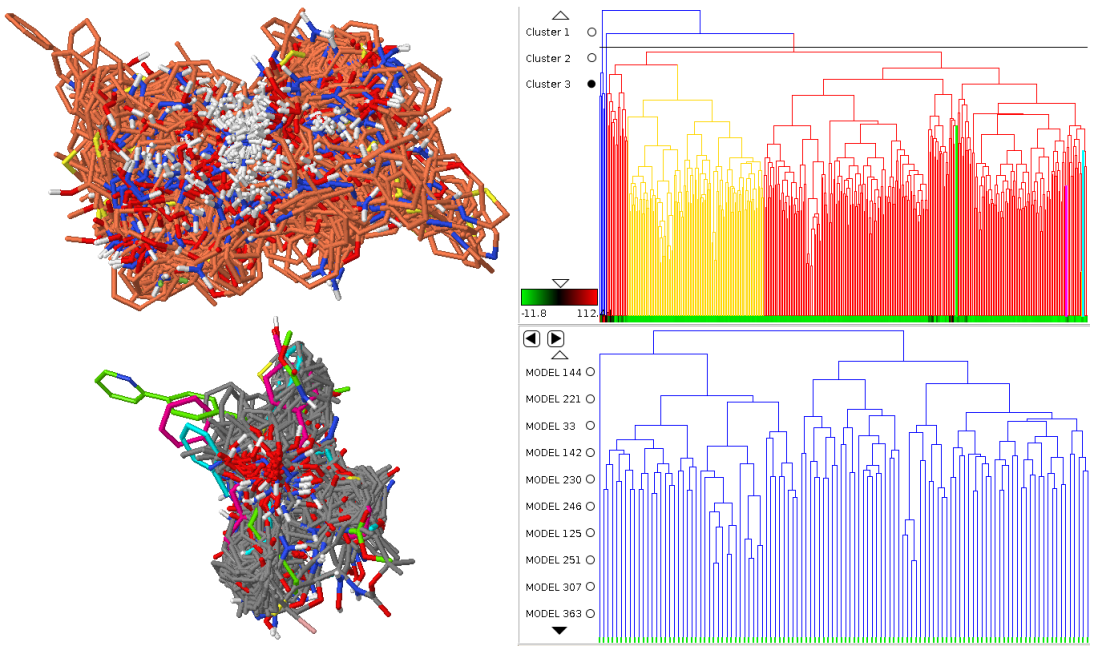


Figura 10.46: *Cluster 3* en la parte superior izquierda. Dendrograma en la derecha con el *cluster* seleccionado en rojo y su rama en amarillo, respectivamente. En el dendrograma detallado, se observa también la ampliación de la rama seleccionada.

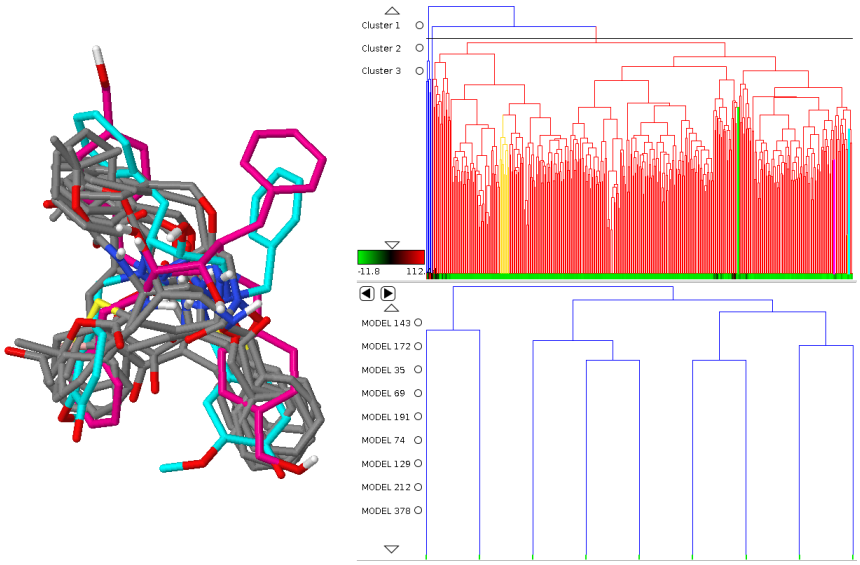


Figura 10.47: Sub-rama del *cluster 3*. Las estructuras cubren la parte inferior y superior izquierda de las referencias.

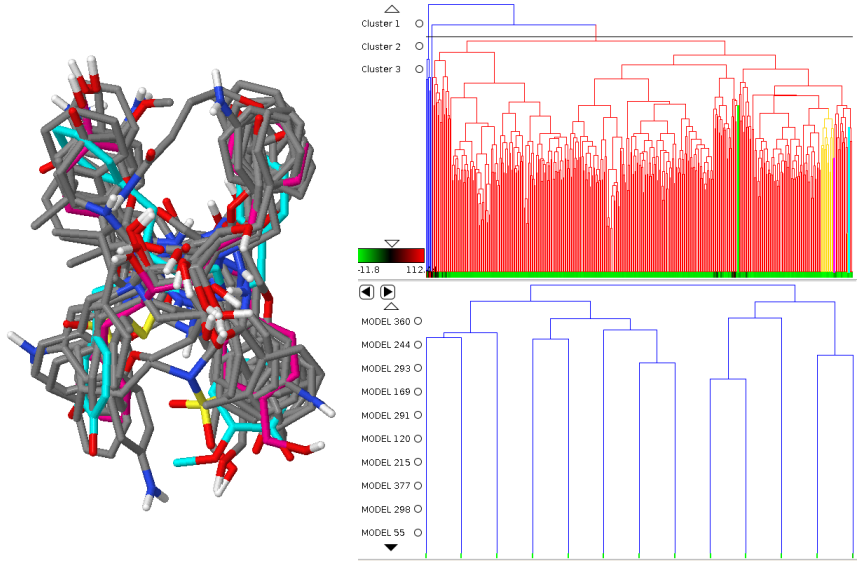


Figura 10.48: Sub-rama del *cluster 3* cercana a las referencias. Las estructuras cubren la forma de H de las referencias.

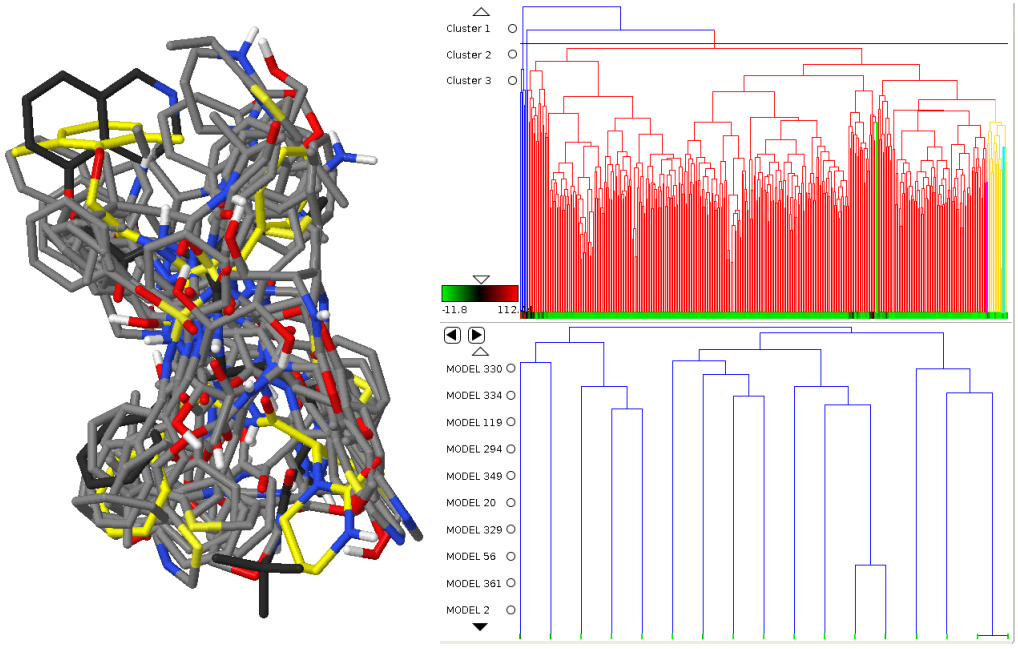


Figura 10.49: Sub-rama del *cluster 3* contiene las estructuras predichas como similares por el programa AuPosSOM.

Por último, en la tabla 10.1 se muestra la comparación de los métodos empleados para analizar los representantes de la proteasa de VIH-1. En la segunda columna se muestran los representantes encontrados por el programa AuPosSOM; en la tercera columna los encontrados por JADOPPT; finalmente, en la cuarta columna, se pueden ver los representantes que suponen una nueva aportación debida a nuestro enfoque. JADOPPT agrupa los mismos representantes que el programa AuPosSOM como se muestra en la tercera columna (para diferenciarlos, aquellos procedentes de JADOPPT aparecen en color rosa), y encuentra otros representantes que son similares. Se aprecia claramente que el aporte que hace a los *clusters* obtenidos con AuPos-SOM es significativo en la segunda y tercera columna de la tabla. Por otra parte, esas sub-ramas contienen representantes muy grandes que son difíciles de agrupar de acuerdo a su similaridad, lo que permite validar la efectividad de nuestra propuesta.

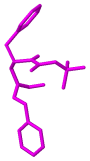
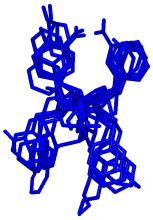
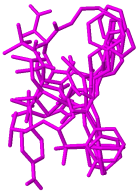
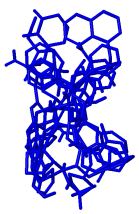
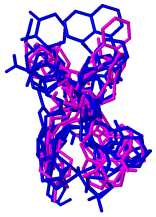
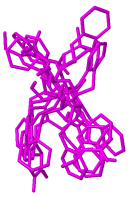
Cluster 3	AuPosSOM	JADOPPT	Aporte
Sub-rama figura 10.47			
Sub-rama figura 10.48			
Sub-rama figura 10.49			

Tabla: 10.1: Comparación de *clusters* entre AuPosSOM y JADOPPT.

10.5. Resultados de *Docking* con Ficheros *Map* Modificados

A continuación se presentan los resultados obtenidos del experimento de *docking* empleando ficheros *maps* creados en JADOPPT. Estos resultados fueron contrastados con otros previos de un experimento de *docking* sobre los siguiente ligandos: colchicina, podofilotoxina, HKC, HKE, NDG y NDK.

El análisis partió considerando que no existía la estructura de la diana, sin embargo, es posible emplear las coordenadas de otra diana que sea muy similar, por lo tanto partiendo de la información ya existente, en este caso de los experimentos previos de *docking*, se llevó a cabo el siguiente diseño de los ficheros *map*: A, C, F, NA, O, Br, d, HD, N, SA, Cl, e, I, OA, y S. En otras palabras, se tomaron las coordenadas del centro del *grid*, número de puntos, espaciado y otra información necesaria para realizar el *docking* –lo más importante es el centro, número de puntos y espaciado-.

Una vez cargada la información de los ficheros *map*, así como las referencias se procedió a diseñar las nuevas zonas de interacción (figura 10.50). Se diseñaron para los siguientes ficheros *map* las mismas esferas: BR, CL, F, I, NA, OA, S, y SA. Para el resto hubo variación en la colocación de las esferas, la figura 10.50 visualiza todos los puntos existentes para el experimento de *docking*. Los valores máximos y mínimos para todas las esferas es de -1.0, lo cual es un valor artificial pues generalmente el programa autogrid –programa que genera los ficheros *map*- no definiría una zona con los mismos valores, al contrario, estos valores son calculados de acuerdo a la diana y por lo tanto varían.

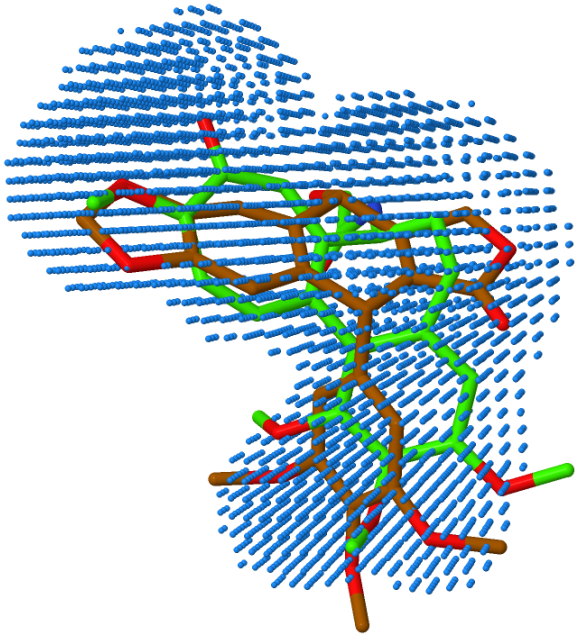


Figura 10.50: Todos los puntos generados por JADOPPT de manera artificial para el experimento de *docking* con el fin de hacer una búsqueda por farmacóforos.

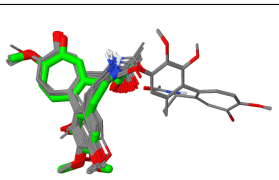
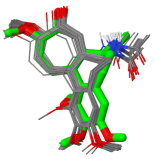
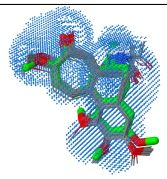
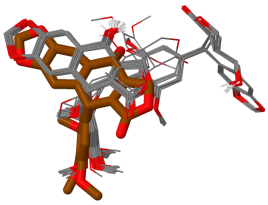
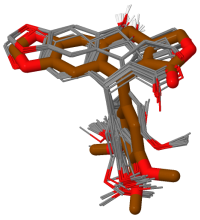
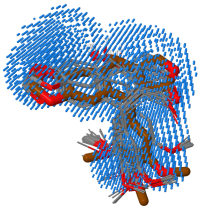
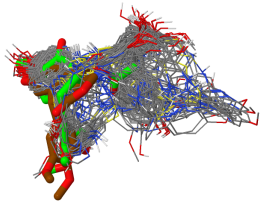
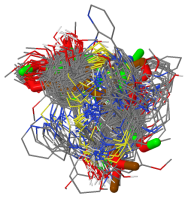
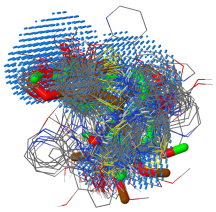
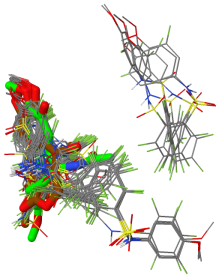
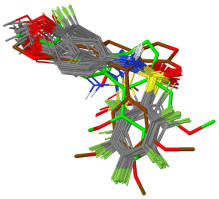
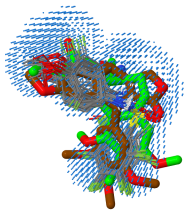
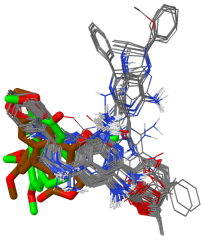
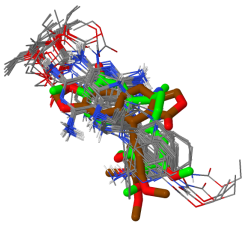
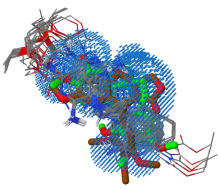
En la tabla 10.2 se muestran los resultados obtenidos cuando existe la diana y cuando se carece de ella. La primera columna es el ligando a ensayar, la segunda es el resultado de Autodock cuando se tiene la diana, la tercera es en ausencia de ésta y en la cuarta se muestra el resultado anterior, pero con las zonas diseñadas por los químicos.

En un ensayo en condiciones normales, obtendríamos el resultado de la segunda columna, el objetivo principal de un ensayo de *docking* es encontrar la mejor conformación que interaccione con la diana, esto es, que tenga mayor afinidad con la diana; en otras palabras, que el ligando –en este caso el fármaco- sea más atractivo químicamente para interactuar que cualquier otra cosa alrededor. Para determinar la afinidad –su pose en el espacio- podemos utilizar un ligando que de antemano se sabe que tiene afinidad con la diana, en este caso se cuenta con dos ligandos de referencia, la colchicina y la podofilotoxina. Por lo tanto, el ensayo de *docking* se centra en obtener una conformación lo suficientemente parecida a esos ligandos de referencia, aclarando que los ligandos de referencia no forman parte del ensayo de *docking*, sino, como su término indica, se trata únicamente de un punto de referencia. Sin embargo, lo que sí se puede hacer para obtener los resultados esperados es tomar a la referencia y crear el *grid* en base al tamaño de la referencia y realizar varios ensayos de *docking* previos hasta obtener una conformación lo más parecida a la pose de la referencia. En otras palabras, se trata de tomar a la referencia y ensayarla hasta reproducir su pose original, esto implica modificar varias veces el tamaño y centro del *grid*. Y, una vez obtenida la pose similar, se toman esos valores y se ensayan en ese *grid* los otros ligandos que se deseen probar.

Regresando a la segunda columna, podemos apreciar que para colchicina posicionó en dos zonas distintas las cien conformaciones que por lo general se generan. Esto quiere decir que dado que autogrid genera valores de acuerdo a unos cálculos de interacción, marca dos zonas como posibles candidatos de mayor interacción o afinidad con la diana. Sin embargo al contrastar las conformaciones con la referencia podemos observar que una de ellas es la que se asemeja a su pose. El caso de la podofilotoxina es el contrario, en la que encuentra zonas de interacción, sin embargo, podemos observar que sólo una zona es la que tiene mayor similitud con ella. En el resto de los ligandos es evidente que la mayoría no comparten la misma zona que cualquiera de las dos referencias, por lo que sería aconsejable realizar una modificación de los ficheros *map*, con el fin de refinar los resultados.

Los resultados de la tercera columna deben ser interpretados de la siguiente manera. Partiendo de información previa de interacción de grupo de átomos y careciendo de la estructura de la diana, el químico define zonas de interacción. En este caso, se tomó la información necesaria (centro, espaciado y número de puntos) para construir un *grid* artificial, el cual consta de ceros salvo las zonas que él definió (figura 10.50). Lo cual obliga a Autodock a considerar solamente esas zonas para generar las conformaciones, y el resultado como podemos ver es bastante satisfactorio para colchicina, podofilotoxina y HKE, dado que la orientación en el espacio es bastante similar a las referencias. Más aun, en el caso de colchicina y podofilotoxina están reproduciendo casi a la perfección la pose de sus referencias, lo cual es uno de los principales objetivos de los ensayos de *docking*. Por otra parte, se cumple el objetivo de convertir a Autodock en una herramienta de búsqueda por farmacóforos.

Una vez expuestos los resultados de este trabajo de tesis, en el capítulo siguiente expondremos las conclusiones a las que hemos llegado.

Ligando	Con diana	Sin diana	Nuevas zonas
COL			
PODO			
HKC			
HKE			
NDG			

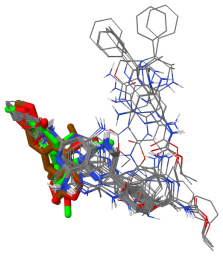
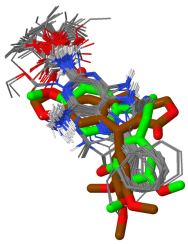
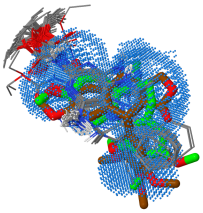
NDK			

Tabla: 10.2: Comparación de los experimentos de *docking* con presencia y ausencia de la diana. La cuarta columna muestra los resultados del *docking* empleando las zonas diseñadas por los químicos.

Parte VII

Conclusiones

Capítulo 11

Conclusiones

A continuación presentamos las conclusiones a las que hemos llegado después de realizar este trabajo de tesis. En la primera sección se abordan las conclusiones a las que llegamos después de automatizar el proceso de selección de representantes. La segunda sección trata las conclusiones sobre el *clustering* de representantes y se exponen las líneas de trabajo futuro sobre el mismo. Finalmente, en la tercera sección abordaremos las conclusiones a las que hemos llegado sobre el diseño de farmacóforos en relación a convertir a Autodock en un programa de búsqueda por farmacóforos.

11.1. Automatización y Selección de Resultados de *Docking*.

La imposibilidad de conocer a priori el número de *clusters* para un determinado experimento de *docking* hace que sea inviable la agrupación de representantes a través de métodos de *clustering* tradicionales que necesitan esta información de entrada, como es el caso de *k-means*. Por otro lado, un enfoque que presuma un número predeterminado de *clusters* no alcanza resultados válidos, como comprobamos en los primeros experimentos realizados en este trabajo. Por tanto, los algoritmos RMSD y jerárquico, mostraron ser las mejores opciones para automatizar la agrupación y selección de representantes.

Los ficheros de *docking* que se emplearon en este trabajo de tesis, la mayoría presentaban una distribución muy dispersa de las conformaciones, sin mostrar patrones definidos. Por lo que el algoritmo jerárquico, con el método de distancia máxima, obtuvo el mejor resultado en el *clustering* de los resultados de *docking*, a diferencia de los otros dos métodos jerárquicos (distancia promedio y distancia mínima). Por otro lado, el RMSD, es más restrictivo en su método de agrupación, y, por tanto, al aplicarlo sobre los mismos datos, se obtuvieron una mayor cantidad de representantes. Por otro lado, esto puede significar que no redujo de manera satisfactoria el número de conformaciones. No obstante, esto aporta información valiosa al químico sobre posibles zonas en el espacio de los ficheros *map*, que podrían rediseñarse con la herramienta. Finalmente, es claro que el algoritmo jerárquico, mostró ser eficaz en tiempo, agrupación y reducción de conformaciones que el RMSD. Sin embargo, ambos algoritmos son eficaces en el filtrado de representantes.

Podemos afirmar que la selección automática de *clusters*, permite agilizar el proceso de análi-

sis de resultados de *docking*. En específico para estos resultados de *docking*, JADOPPT agrupó y seleccionó representantes de 168 ficheros dlg con 100 conformaciones en 1 minuto, por medio del algoritmo jerárquico, comparado con 15 minutos en promedio que tardaría un químico en agrupar un fichero con 100 conformaciones y después seleccionar un representante de cada *cluster*, estaríamos hablando de 42 horas en promedio para realizar la misma tarea.

En conclusión, podemos afirmar que la implementación de distintas herramientas visuales como el *treemp*, el dendrograma, la barra que lo corta, la comparación y unión de *clusters*, el código de colores de energía de unión en el *treemap*, así como la integración interactiva del dendrograma y *treemap* con el visualizador molecular Jmol, y sobretodo, el enfoque de Analítica Visual, permitieron la automatización y selección de representantes de cada resultado de *docking* de la misma forma en la que la realizaría un experto –en este caso, el químico-, pero en un tiempo mucho más corto.

11.2. Agrupamiento de Representantes

Los resultados obtenidos en esta parte, provienen de la selección de conformaciones en el paso anterior. Después de probar con diversas formas para comparar estructuras de átomos diferentes, tales como: seleccionar los primeros veinte átomos de cada molécula, los veinte átomos de mejor energía de unión, y descriptores moleculares; observamos que la mejor opción era realizar la comparación mediante la comparación de descriptores moleculares –que nosotros llamamos observadores-. Porqué, nos permitió comprimir la información de la molécula en un valor para compararla con otras diferentes.

Por otra parte, los observadores son los puntos del *grid* que emplea Autodock para realizar el *docking*, nos permitió ver aquellas conformaciones que prefieren un determinado sitio. Por lo que, podemos afirmar que esta información es útil al diseño de fármacos, debido a que nos muestra a los átomos que interactúan con la diana. Así mismo, llegamos a la conclusión de que, dependiendo del tipo de interacción, esta zona podría removerse o darle más peso en futuros ensayos de *docking*. Por lo tanto, podemos concluir que los observadores no sólo tuvieron una aportación como puntos evaluadores para comparar estructuras diferentes, sino que además aportan información para el diseño de nuevos fármacos.

Por otro lado, la visualización de los resultados del *clustering* en el dendrograma, nos permitió de forma rápida, ver todo el contexto de la información evaluada. Por otra parte, las referencias nos ayudaron a realizar un análisis específico en primera instancia; en segunda el bloque de colores de la parte inferior mostró ser de gran ayuda para el caso de los representantes de RMSD. Por otro lado, el bloque de colores nos permitió ver que la mayoría de los resultados de *docking* estaban en un rango de energía de unión medio a bajo. Sin embargo, también, nos ayudó a desvelar que algunos representantes con energía de unión favorable no eran similares a su referencia, lo cual indica un error en los cálculos del *docking*. Por lo tanto, la suma de cada elemento visual en el dendrograma, esto es, los códigos de colores, las referencias resaltadas, así como la interacción con el visualizador molecular de Jmol permitieron darnos una idea general para discriminar conformaciones, y con esto nos referimos a las conformaciones que no son similares en la orientación a sus referencias o que no están cerca de ellas. En otras palabras, podemos concluir que las ayudas visuales y la interacción con visualizador molecular, son imprescindibles en el análisis de grandes cantidades de información, como lo son los resultados de *docking*.

De los tres tipos de filtrado –RMSD, energía de unión y filtrado por los primeros-, concluimos que, el filtrado de energía de unión, puede ser aplicado como segundo paso después de filtrar por RMSD. En otras palabras, es filtrar por RMSD, evaluar los resultados y refinar si es necesario, finalmente filtrar de nuevo por la energía de unión. Por otro lado, también podemos concluir que, para una reducción rápida en la que no sea prioritario refinar los filtrados, basta con aplicar el filtrado de los dos primeros representantes de cada resultado de *docking*.

A la conclusion final que llegamos, es que, las preguntas de investigación y los objetivos han sido respondidos en relación a la automatización de selección de representantes y el *clustering* de representantes, sin embargo, no todo es idílico, estamos conscientes de las limitaciones de JADOPPT. Y por consiguiente se desprenden los objetivos del trabajo futuro para esta parte de nuestra propuesta:

- Incorporar la lectura de ficheros mol2 y procesar ficheros de resultados de vina.
- Explorar con otros algoritmos de *clustering*, por ejemplo, los mapas auto-organizados, redes neuronales, algoritmos de clasificación, etc.
- Implementar otras visualizaciones que apoyen el análisis de comparación de representantes.
- Diseñar e implementar nuevas visualizaciones, tales como: visualizar la secuencia del receptor, visualizar puentes de hidrogeno, etc.
- Mejorar la integración con el visualizador molecular Jmol.

11.3. Diseño de farmacóforos

La visualización de los resultados de *docking* en las dos etapas de análisis, continuamente sobresalían posibles zonas a mejorar o eliminar. En los casos de colchicina y podofilotoxina, los resultados de *docking* mostraron que el programa Autodock consideró distintas zonas en el espacio como posibles sitios de interacción. Sin embargo, las referencias muestran un único sitio. Después de rediseñar los *maps*, y realizar todo el proceso de análisis de resultados de *docking*, podemos concluir que el rediseño de los ficheros *maps* -colocando nubes en ciertas zonas que cubren a las referencias- mejora los resultados de *docking*.

Por otra parte, las herramientas visuales, y, sobretodo la interacción entre las distintas visualizaciones aceleraron el proceso de diseño de farmacóforos. Más aun, los resultados mostraron que con el rediseño los ficheros *map* es posible convertir el programa Autodock, en un programa para realizar búsquedas por farmacóforos. Finalmente, podemos concluir que actualmente, no existe una herramienta visual e interactiva, enfocada en el diseño farmacóforos.

Este trabajo de tesis ha desarrollado una herramienta que aplica la analítica visual llamada JADOPPT para el análisis y generación de conocimiento a partir de resultados de experimentos de *docking*. La combinación de los métodos de visualización con los de minería de datos ayuda a los químicos a analizar de manera rápida y concisa, la gran cantidad de información generada del *Virtual Screening*. Más aun, hemos podido demostrar que Autodock puede servir para realizar búsqueda por farmacóforos, lo cual es un aporte importante de este trabajo ya que Autodock no fue considerado para realizar dicha función.

Anexo A

Glosario

Afinidad (Affinity). Afinidad es la tendencia de una molécula de asociarse con otra. La afinidad de un fármaco es la habilidad de unirse con la diana biológica (receptor, enzima, etc.).

Análogo (Analog). Un análogo es un fármaco en el que su estructura está relacionada con otro fármaco pero que sus propiedades químicas y biológicas pueden ser diferentes.

Diseño de fármacos asistido por ordenador (CADD). Incluye todas las técnicas asistidas por ordenador para descubrir, desarrollar, diseñar y optimizar bilógicamente a compuestos activos como un uso posible de fármaco.

Química computacional (Computational chemistry). Utiliza métodos matemáticos para el cálculo de propiedades moleculares o para la simulación del comportamiento de una molécula.

Congénere (congener). A congener is a substance literally con-(with)generated or synthesized by essentially the same synthetic chemical reactions and the same procedures. Analogs are substances that are analogous in some respect to the prototype agent in chemical structure.

De novo design. Es el diseño de compuesto bioactivos por construcción incremental de un modelo de ligando dentro de un sitio activo de un modelo de receptor o enzima, cuya estructura es conocida a partir datos de rayos-X o resonancia magnética nuclear (NMR).

Estudios de docking (Docking studies). Son estudios de modelado molecular con el objetivo de encontrar ajuste entre el ligando y la diana.

Fármaco (Drug). Un fármaco es una sustancia que se emplea para el tratamiento, prevención, o curación de una enfermedad en humanos o animales. Un fármaco, también puede ser utilizado para hacer un diagnostico medico o para restaurar, corregir o modificar funciones fisiológicas.

Efficacy. Describes the relative intensity with which agonists vary in the response they produce even when they occupy the same number of receptors and with the same affinity. Efficacy is not synonymous to Intrinsic activity.

Hydrophilicity. Es la tendencia de una molécula para ser disuelta en agua.

Hydrophobicity. Es la asociación de grupos no-polares o moléculas en un ambiente acuoso que

surge de la tendencia del agua para excluir a las moléculas no-polares.

Actividad intrínseca (Intrinsic activity). Es la máxima respuesta de estimulación inducida por un compuesto en relación a un compuesto de referencia.

Descubrimiento de líderes (Lead discovery). Es el proceso de identificar entidades químicas nuevas, que por medio de modificaciones puede ser transformada en un fármaco de prescripción.

Generación de líderes (Lead generation). El término se aplica a las estrategias desarrolladas para identificar compuestos que posean una actividad biológica deseada pero aun por optimizar.

Optimización de líderes (Lead optimization). Es la modificación sintética de un compuesto biológicamente activo, para cumplir con todos los requerimientos stereoelectronic, fisicoquímicos, farmacocinéticas, y toxicológicos para la utilidad clínica.

Bibliografía

[1] Stewart A. Adcock and J. Andrew McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106(5):1589–1615, 2006. PMID: 16683746.

[2] Christopher Ahlberg and Ben Shneiderman. The alphaslider: a compact and rapid selector. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, CHI 94, pages 365–371, New York, NY, USA, 1994. ACM.

[3] Rommie Amaro, Riccardo Baron, and J. McCammon. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of Computer-Aided Molecular Design*, 22:693–705, 2008. 10.1007/s10822-007-9159-2.

[4] C. R. Aragon, S. S. Poon, G. S. Aldering, R. C. Thomas, and R. Quimby. Using visual analytics to maintain situation awareness in astrophysics. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 27–34, 2008.

[5] Cecilia R. Aragon, Stephen J. Bailey, Sarah Poon, Karl J. Runge, and Rollin C. Thomas. Sunfall: A collaborative visual analytics system for astrophysics. In *VAST 07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 219–220, Washington, DC, USA, 2007. IEEE Computer Society.

[6] Wojciech Basalaj. Proximity visualization of abstract data. http://www.pavis.org/essay/.

[7] Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[8] D. P. Berrar, W. Dubitzky, and M. Granzow. *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers, New York, Boston, Dordrecht, London, Moscow, 2003.

[9] Nadia Boukhelifa and Peter J. Rodgers. A model and software system for coordinated and multiple views in exploratory visualization. *Information Visualization*, 2(4):258–269, 2003.

[10] Guillaume Bouvier, Nathalie Evrard-Todeschi, Jean-Pierre Girault, and Gildas Bertho. Automatic clustering of docking poses in virtual screening process using self-organizing map. *Bioinformatics*, 26:53–60, January 2010.

[11] Dominique Brodbeck, Riccardo Mazza, and Denis Lalanne. Human machine interaction. chapter Interactive Visualization - A Survey, pages 27–46. Springer-Verlag, Berlin, Heidelberg, 2009.

[12] Patrick Bultinck, Tom Kuppens, Xavier Gironés, and Ramon Carbó-Dorca. Quantum similarity superposition algorithm (qssa): A consistent scheme for molecular alignment and molecular similarity based on quantum chemistry. *Journal of Chemical Information and Computer Sciences*, 43(4):1143–1150, 2003. PMID: 12870905.

[13] S.K. Card and J. Mackinlay. The structure of the information visualization design space. *infovis*, 00:92, 1997.

[14] Stuart K. Card, Jock D. MacKinlay, and George G. Robertson. Readings in intelligent user interfaces. chapter A morphological analysis of the design space of input devices, pages 597–609. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[15] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[16] Marion E. Cass, Henry S. Rzepa, David R. Rzepa, and Charlotte K. Williams. The use of the free, open-source program jmol to generate an interactive web site to teach molecular symmetry. *Journal of Chemical Education*, 82(11):1736, 2005.

[17] José A. Castellanos-Garzón, Carlos Armando García, and Luis A. Miguel-Quintales. An evolutionary hierarchical clustering method with a visual validation tool. In *IWANN 09: Proceedings of the 10th International Work-Conference on Artificial Neural Networks*, pages 367–374, Berlin, Heidelberg, 2009. Springer-Verlag.

[18] Remco Chang, Caroline Ziemkiewicz, Tera Marie Green, and William Ribarsky. Defining insight for visual analytics. *IEEE Comput. Graph. Appl.*, 29(2):14–17, 2009.

[19] Chaomei Chen. An information-theoretic view of visual analytics. *IEEE Computer Graphics and Applications*, 28:18–23, 2007.

[20] Min Chen, D. Ebert, H. Hagen, R.S. Laramee, R. van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. *Computer Graphics and Applications, IEEE*, 29(1):12–19, jan-feb 2009.

[21] Vincent B. Chen, Ian W. Davis, and David C. Richardson. King (kinemage, next generation): A versatile interactive molecular and scientific visualization program. *Protein Science*, 18(11):2403–2409, 2009.

[22] Luca Chittaro. Information visualization and its application to medicine. *Artificial Intelligence in Medicine*, 22(2):81–88, 2001. Information Visualization in Medicine.

[23] Kris Cook, Rae Earnshaw, and John Stasko. Guest editors introduction: Discovering the unexpected. *IEEE Comput. Graph. Appl.*, 27(5):15–19, 2007.

[24] Geoffrey Ellis Daniel Keim, Joern Kohlhammer and Florian Mansmann. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, November 2010.

[25] Helmut Doleisch and Helwing Hauser. Smooth brushing for focus+context visualization of simulation data in 3d. volume 10, pages 147–154, 2001.

[26] Charles Dumontet and Mary Ann Jordan. Microtubule-binding agents: a dynamic field of cancer therapeutics. *Nat Rev Drug Discov*, 9(10):790–803, Oct 2010.

[27] Brynn Evans and Stuart Card. Augmented information assimilation: social and algorithmic web aids for the information long tail. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 989–998, New York, NY, USA, 2008. ACM.

[28] David A. Evans, Michael J. Bodkin, S. Richard Baker, and Gary J. Sharman. Janocchio - a java applet for viewing 3d structures and calculating nmr couplings and noes. *Magnetic Resonance in Chemistry*, 45(7):595–600, 2007.

[29] Clifton Forlines and Ryan Lilien. Adapting a single-user, single-display molecular visualization application for use in a multi-user, multi-display environment. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '08, pages 367–371, New York, NY, USA, 2008. ACM.

[30] Ben Fry. *Visualizing data - exploring and explaining data with the processing environment*. O'Reilly, 2008.

[31] Benjamin Fry. Investigating exhibits. In *Symposium on Designing Interactive Systems*, pages 32–36, 2002.

[32] David C. Y. Fung, Marc R. Wilkins, David Hart, and Seok-Hee Hong. Using the clustered circular layout as an informative method for visualizing protein-protein interaction networks. *PROTEOMICS*, 10(14):2723–2727, 2010.

[33] Carlos García, José Castellanos-Garzón, and Carlos Blanco. Analyzing gene expression data on a 3d scatter plot. In Emilio Corchado, Václav Snásel, Javier Sedano, Aboul Hassanien, José Calvo, and Dominik Slezak, editors, *Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011*, volume 87 of *Advances in Intelligent and Soft Computing*, pages 349–356. Springer Berlin / Heidelberg, 2011.

[34] Carlos Armando García, Roberto Therón, Rafael Peláez, José Luis López-Pérez, and Gustavo Santos-Garcia. Visual evaluation of clustered molecules in the process of new drugs design. In *SG 09: Proceedings of the 10th International Symposium on Smart Graphics*, pages 3–14, Berlin, Heidelberg, 2009. Springer-Verlag.

[35] Eleanor J. Gardiner, Valerie J. Gillet, Peter Willett, and David A. Cosgrove. Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *Journal of Chemical Information and Modeling*, 47(2):354–366, 2007.

[36] Alexander G. Gee, Georges G. Grinstein, Yu Min, and Li Hongli. Dynamic and interactive dimensional anchors for spring-based visualizations. Technical report, University of Massachusetts Lowell, Dept. of Computer Science, Lowell, 2005.

[37] Alexander G. Gee, Hongli Li, Min Yu, Mary Beth Smrtic, Urska Cvek, Howie Goodell, Vivek Gupta, Christine Lawrence, Jainping Zhou, Chih-Hung Chiang, and Georges G. Grinstein. Universal visualization platform. In R. F. Erbacher, J. C. Roberts, M. T. Gröhn, & K. Börner , editor, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 5669 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 274–283, March 2005.

[38] J. M. Geoffrey, K. A. Do, and C. Ambroise. *Analyzing Microarray Gene Expression Data*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2004.

[39] T.M. Green, W. Ribarsky, and B. Fisher. Visual analytics for complex concepts using a human cognition model. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium*, pages 91–98, 2008.

[40] Jenny Gu and Philip E. Bourne. *Structural Bioinformatics, 2nd Edition*, chapter 9, pages 246–248. Wiley-Blackwell, 2009.

[41] Robert M Hanson. Jmol - a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography*, 43(5-2):1250–1260, Oct 2010.

[42] Helwig Hauser. Generalizing focus+context visualization. Technical report, Scientific Visualization: Extracting Information and Knowledge from Scientific Data Sets, 2003.

[43] Jeffrey Heer, Stuart K. Card, and James A. Landay. prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '05, pages 421–430, New York, NY, USA, 2005. ACM.

[44] R. J. Hendley, N. S. Drew, A. M. Wood, and R. Beale. Case study: Narcissus: visualising information. In *INFOVIS 95: Proceedings of the 1995 IEEE Symposium on Information Visualization*, page 90, Washington, DC, USA, 1995. IEEE Computer Society.

[45] Ivan Herman, Guy Melancon, and Scott Marshall. Graph visualization and navigation in information visualisation: a survey. *IEEE Transactions on Visualisation and Computer Graphics*, 6(1):24–43, 2000.

[46] Angel Herráez. Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255–261, 2006.

[47] Richard J. Heuer. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, 1999.

[48] John D. Holliday, Sarah L. Rodgers, Peter Willett, Min-You Chen, Mahdi Mahfouf, Kevin Lawson, and Graham Mullier. Clustering files of chemical structures using the fuzzy k-means clustering method. *Journal of Chemical Information and Computer Sciences*, 44(3):894–902, 2004.

[49] Donald H. House, Alethea S. Bair, and Colin Ware. An approach to the perceptual optimization of complex visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 12:509–521, 2006.

[50] Eva Hudlicka. To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies*, 59(1-2):1–32, 2003. Applications of Affective Computing in Human-Computer Interaction.

[51] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[52] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 01:69–91, 1985.

[53] R. J. K. Jacob. New human-computer interaction techniques. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, 129:131, 1994.

[54] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall Englewood Cliffs, New Jersey 07632, 1998.

[55] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[56] Ajay N. Jain. Morphological similarity: A 3d molecular similarity method correlated with protein-ligand recognition. *Journal of Computer-Aided Molecular Design*, 14:199–213, 2000. 10.1023/A:1008100132405.

[57] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16:1370–1386, 2004.

[58] S. Kaisler, F. Armour, J.A. Espinosa, and W. Money. Big data: Issues and challenges moving forward. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 995–1004, 2013.

[59] Maria Kavallaris. Microtubules and resistance to tubulin-binding agents. *Nat Rev Cancer*, 10(3):194–204, Mar 2010.

[60] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. pages 154–175. 2008.

[61] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[62] Daniel A. Keim, Daniel A. Keim, R. Daniel Bergeron, R. Daniel Bergeron, Ronald M. Pickett, and Ronald M. Pickett. Test data sets for evaluating data visualization techniques. In *in: Perceptual Issues in Visualization*, pages 9–22. Springer, 1994.

[63] Daniel A. Keim and Hans-Peter Kriegel. Visualization techniques for mining large databases: A comparison. *IEEE Trans. on Knowl. and Data Eng.*, 8(6):923–938, 1996.

[64] Daniel A. Keim, Florian Mansmann, Daniela Oelke, and Hartmut Ziegler. Visual analytics: Combining automated discovery with interactive visualizations. In *DS 08: Proceedings of the 11th International Conference on Discovery Science*, pages 2–14, Berlin, Heidelberg, 2008. Springer-Verlag.

[65] Daniel A. Keim, Florian Mansmann, Jorn Schneidewind, and Hartmut Ziegler. Challenges in visual data analysis. In *IV 06: Proceedings of the conference on Information Visualization*, pages 9–16, Washington, DC, USA, 2006. IEEE Computer Society.

[66] Daniel A. Keim, Florian Mansmann, Andreas Stoffel, and Hartmut Ziegler. *Visual Analytics*. Springer, 2009.

[67] Nam Wook Kim, Stuart K. Card, and Jeffrey Heer. Tracing genealogical data with timenets. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 241–248, New York, NY, USA, 2010. ACM.

[68] R. Kincaid and K. Dejgaard. Massvis: Visual analysis of protein complexes using mass spectrometry. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 163 –170, 2009.

[69] S. Konecni, Jianping Zhou, and G. Grinstein. A visual analytics model applied to lead generation library design in drug discovery. In *Information Visualisation, 2009 13th International Conference*, pages 345 –352, 2009.

[70] Robert Kosara, Silvia Miksch, and Helwig Hauser. Focus+context taken literally. *IEEE Computer Graphics and Applications*, 22(1):22–29, 2002.

[71] Sven Krasser, Gregory Conti, Julian Grizzard, and Jeff Gribschaw andHenrry Owen. Real-time and forensic network data analysis using animated and coordinated visualization. *Systems, Man and Cybernetics (SMC) Information Assurance Workshop, 2005. Proceedings from the Sixth Annual IEEE*, pages 42–49, 2005.

[72] Edgar Eduardo Lara-Ramírez, Aldo Segura-Cabrera, Xianwu Guo, Gongxin Yu, Carlos Armando García-Pérez, and Mario A. Rodríguez-Pérez. New implications on genomic adaptation derived from the helicobacter pylori genome comparison. *PLoS ONE*, 6(2):e17300, 02 2011.

[73] Weizhong Li. A fast clustering algorithm for analyzing highly similar compounds of very large libraries. *Journal of Chemical Information and Modeling*, 46(5):1919–1923, 2006.

[74] Thy-Hou Lin, Yih-Shiang Yu, and Hong-Jih Chen. Classification of some active compounds and their inactive analogues using two three-dimensional molecular descriptors derived from computation of three-dimensional convex hulls for structures theoretically generated for them. *Journal of Chemical Information and Computer Sciences*, 40(5):1210–1221, 2000.

[75] Jeroen Logtenberg. Multi-user interaction with molecular visualizations on a multi-touch table. Master's thesis, University of Twente, 2009.

[76] John MacCuish, Christos Nicolaou, and Norah E. MacCuish. Ties in proximity and clustering compounds. *Journal of Chemical Information and Computer Sciences*, 41(1):134–146, 2001.

[77] Dharmesh M. Maniyar, Ian T. Nabney, Bruce S. Williams, and Andreas Sewing. Data visualization during the early stages of drug discovery. *Journal of Chemical Information and Modeling*, 46(4):1806–1818, 2006.

[78] Ana B. S. Maya, Concepción Pérez-Melero, Carmen Mateo, Dulce Alonso, José Luis Fernández, Consuelo Gajate, Faustino Mollinedo, Rafael Peláez, Esther Caballero, and Manuel Medarde. Further naphthylcombretastatins. an investigation on the role of the naphthalene moiety. *Journal of Medicinal Chemistry*, 48(2):556–568, 2005.

[79] Brian McMahon and Robert M. Hanson. A toolkit for publishing enhanced figures. *Journal of Applied Crystallography*, 41(4):811–814, Aug 2008.

[80] Jamel Eddine Meslamani, Francžois Andre, and Michel Petitjean. Assessing the geometric diversity of cytochrome p450 ligand conformers by hierarchical clustering with a stop criterion. *Journal of Chemical Information and Modeling*, 49(2):330–337, 2009.

[81] Torsten Moller. A parallel coordinates style interface for exploratory volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 11(1):71–80, 2005.

[82] Rachel E. Morgan, Sunjoo Ahn, Sandra Nzimiro, Jean Fotie, Mitch A. Phelps, Jeffrey Cotrill, Adam J. Yakovich, Dan L. Sackett, James T. Dalton, and Karl A. Werbovetz. Inhibitors of tubulin assembly identified through screening a compound library. *Chemical Biology and Drug Design*, 72(6):513–524, 2008.

[83] Garrett M Morris, David S Goodsell, Robert S Halliday, Ruth Huey, William E Hart, Richard K Belew, and Arthur J Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, 1998.

[84] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 2009.

[85] Asher Mullard. 2012 fda drug approvals. *Nat Rev Drug Discov*, 12(2):87–90, February 2013.

[86] Henrick R. Nagel. scientific visualization versus information visualization, 2006.

[87] Chris North. Multiple views and tight coupling in visualization: A language, taxonomy, and system. In *Proc. CSREA CISST 2001 Workshop of Fundamental Issues in Visualization*, pages 626–632, 2001.

[88] Chris North, Nathan Conklin, Kiran Indukuri, and Varun Saini. Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases. *Information Visualization*, 1(3/4):211–228, 2002.

[89] Chris North and Ben Shneiderman. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In *AVI 00: Proceedings of the working conference on Advanced visual interfaces*, pages 128–135, New York, NY, USA, 2000. ACM.

[90] Seán I. O'Donoghue, David S. Goodsell, Achilleas S. Frangakis, Fabrice Jossinet, Roman A. Laskowski, Michael Nilges, Helen R. Saibil, Andrea Schafferhans, Rebecca C. Wade, Eric Westhof, and Arthur J. Olson. Visualization of macromolecular structures. *Nature methods*, 7(3 Suppl):42–55, March 2010.

[91] Frank Oellien, Wolf-Dietrich Ihlenfeldt, and Johann Gasteiger. Infvis - platform-independent visual data mining of multidimensional chemical data sets. *Journal of Chemical Information and Modeling*, 45(5):1456–1467, 2005. PMID: 16180923.

[92] Kristina A. Paris, Omar Haq, Anthony K. Felts, Kalyan Das, Eddy Arnold, and Ronald M. Levy. Conformational landscape of the human immunodeficiency virus type 1 reverse transcriptase non-nucleoside inhibitor binding pocket: Lessons for inhibitor design from a cluster analysis of many crystal structures. *Journal of Medicinal Chemistry*, 52(20):6413–6420, 2009. PMID: 19827836.

[93] Rafael Peláez, Roberto Therón, Carlos Armando García, José Luis López, and Manuel Medarde. Design of new chemoinformatic tools for the analysis of virtual screening studies: Application to tubulin inhibitors. 49:189–196, 2008.

[94] Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, AVI 04, pages 109–116, New York, NY, USA, 2004. ACM.

[95] Matthew D. Plumlee and Colin Ware. Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. *ACM Trans. Comput.-Hum. Interact.*, 13:179–209, June 2006.

[96] Nicolay Postarnakevich and Rahul Singh. Global-to-local representation and visualization of molecular surfaces using deformable models. In *Proceedings of the 2009 ACM symposium on Applied Computing*, SAC '09, pages 782–787, New York, NY, USA, 2009. ACM.

[97] Ramana Rao and Stuart K. Card. Exploring large tables with the table lens. In *Conference companion on Human factors in computing systems*, CHI '95, pages 403–404, New York, NY, USA, 1995. ACM.

[98] Matt Rasmussen and George Karypis. gcluto-an interactive clustering, visualization, and analysis system. 2004. CiteSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States) ER.

[99] W.C. Ray. A visual analytics approach to identifying protein structural constraints. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 249 –250, 2010.

[100] Casey Reas and Ben Fry. *Getting Started with Processing - a Hands-On Introduction to Making Interactive Graphics*. O'Reilly, 2010.

[101] Casey Reas and Benjamin Fry. Processing: programming for the media arts. *AI Soc.*, 20(4):526–538, 2006.

[102] Charles H. Reynolds, Ross Druker, and Lori B. Pfahler. Lead discovery using stochastic cluster analysis (sca): A new method for clustering structurally similar compounds. *Journal of Chemical Information and Computer Sciences*, 38(2):305–312, 1998.

[103] Theresa-Marie Rhyne. Does the difference between information and scientific visualization really matter? *IEEE Comput. Graph. Appl.*, 23:6–8, May 2003.

[104] Daniel M. Russell, George Furnas, Mark Stefik, Stuart K. Card, and Peter Pirolli. Sensemaking. In *CHI '08 extended abstracts on Human factors in computing systems*, CHI EA '08, pages 3981–3984, New York, NY, USA, 2008. ACM.

[105] Daniel M. Russell, Peter Pirolli, George Furnas, Stuart K. Card, and Mark Stefik. Sensemaking workshop chi 2009. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, CHI EA '09, pages 4751–4754, New York, NY, USA, 2009. ACM.

[106] J.D. Saffer, V.L. Burnett, Guang Chen, and P. van der Spek. Visual analytics in the pharmaceutical industry. *Computer Graphics and Applications, IEEE*, 24(5):10–15, Sept.-Oct. 2004.

[107] Rodrigo Santamaría, Roberto Therón, and Luis Quintales. Bicoverlapper: A tool for bicluster visualization. *Bioinformatics*, 24(9):1212–1213, 2008.

[108] Rodrigo Santamaría, Roberto Therón, and Luis Quintales. A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics*, 9, 2008.

[109] Jinwook Seo and Ben Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. *Information Visualization, IEEE Symposium on*, 0:65–72, 2004.

[110] Jinwook Seo and Ben Shneiderman. Interactively exploring hierarchical clustering results. *IEEE Computer society*, 35(7):80–86, july 2002.

[111] Ron Shamir and Roded Sharan. Click: A clustering algorithm for gene expression analysis. 2000.

[112] Jianyin Shao, Stephen W. Tanner, Nephi Thompson, and Thomas E. Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334, 2007.

[113] Jian Shen. Had: An automated database tool for analyzing screening hits in drug discovery. *Journal of Chemical Information and Computer Sciences*, 43(5):1668–1672, 2003. PMID: 14502501.

[114] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. *Visual Languages, IEEE Symposium on*, 0:336, 1996.

[115] Brian K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432:862–865, 2004.

[116] M.M. Shovman, A. Szymkowiak, J.L. Bown, and K.C. Scott-Brown. Changing the view: Towards the theory of visualisation comprehension. In *Information Visualisation, 2009 13th International Conference*, pages 135 –138, july 2009.

[117] Yedendra Babu Shrinivasan and Jarke J. van Wijk. Supporting the analytical reasoning process in information visualization. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1237–1246, New York, NY, USA, 2008. ACM.

[118] Harri Siirtola. Combining parallel coordinates with the reorderable matrix. In *CMV '03: Proceedings of the conference on Coordinated and Multiple Views In Exploratory Visualization*, page 63, Washington, DC, USA, 2003. IEEE Computer Society.

[119] Parminder Singh, Krishnan Rathinasamy, Renu Mohan, and Dulal Panda. Microtubule assembly dynamics: An attractive target for anticancer drugs. *IUBMB Life*, 60(6):368–375, 2008.

[120] Andrew Smellie. Accelerated k-means clustering in metric spaces. *Journal of Chemical Information and Computer Sciences*, 44(6):1929–1935, 2004.

[121] Robert Spence. *Interpretation of Quantitative Data*, chapter 3, pages 33–51. Addison-Wesley, 1 edition, 2001.

[122] Robert Spence. *Information Visualization: Design for Interaction (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007.

[123] Martin Stahl and Harald Mauser. Database clustering with a combination of fingerprint and maximum common substructure methods. *Journal of Chemical Information and Modeling*, 45(3):542–548, 2005.

[124] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The chemistry development kit (cdk): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500, 2003. PMID: 12653513.

[125] John E. Stone, Axel Kohlmeyer, Kirby L. Vandivort, and Klaus Schulten. Immersive molecular visualization and interactive modeling with commodity hardware. In *Proceedings of the 6th international conference on Advances in visual computing - Volume Part II*, ISVC'10, pages 382–393, Berlin, Heidelberg, 2010. Springer-Verlag.

[126] Ying Tao, Yang Liu, Carol Friedman, and Yves A. Lussier. Information visualization techniques in bioinformatics during the postgenomic era. *Drug Discovery Today: BIOSILICO*, 2(6):237 – 245, 2004.

[127] James J. Thomas and Kristin A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[128] Edward R. Tufte. *Envisioning Information*. Graphics Pr, 1990.

[129] Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Pr, 1997.

[130] Edward R. Tufte. *Beautiful Evidence*. Graphics Pr, 2006.

[131] Johannes Jansen van Vuuren, Michelle Kuttel, and James Gain. Visualization of solution sets from automated docking of molecular structures. In *Proceedings of the 7th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, AFRIGRAPH '10, pages 111–120, New York, NY, USA, 2010. ACM.

[132] Vivek VYAS, Anurekha JAIN, Avijeet JAIN, and Arun GUPTA. Virtual screening : A fast tool for drug design. *Scientia pharmaceutica*, 76:333–360, 2008.

[133] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[134] Colin Ware. *Visual Thinking: for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[135] Colin Ware and Peter Mitchell. Visualizing graphs in three dimensions. *ACM Trans. Appl. Percept.*, 5:2:1–2:15, January 2008.

[136] Brian White, Azmin Kahriman, Lois Luberice, and Farhia Idleh. Evaluation of software for introducing protein structure. *Biochemistry and Molecular Biology Education*, 38(5):284–289, 2010.

[137] Leland Wilkinson and Anushka Anand. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006. Member-Grossman,, Robert.

[138] E. L. Willighagen and M. Howard. Fast and scriptable molecular graphics in web browsers without java3d. 2007.

[139] Silvia Miksch Wolfgang Aigner, Alessio Bertone. Tutorial: Introduction to visual analytics. Volume 4799:453–456, 2007.

[140] Han-Ming Wu, Yin-Jing Tien, and Chun houh Chen. Gap: A graphical environment for matrix visualization and cluster analysis. *Computational Statistics & Data Analysis*, 54(3):767–778, 2010. Second Special Issue on Statistical Algorithms and Software.

[141] Fumiyoshi Yamashita, Hideto Hara, Takayuki Ito, and Mitsuru Hashida. Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: Application to structure activity relationship analysis of cytochrome p450 metabolism. *Journal of Chemical Information and Modeling*, 48(2):364–369, 2008. PMID: 18211048.

[142] Jing Yang, Anilkumar Patro, Shiping Huang, Nishant Mehta, Matthew O. Ward, and Elke A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization*, pages 73–80, Washington, DC, USA, 2004. IEEE Computer Society.

[143] Ke-Bing Zhang, Mehmet Orgun, Yanchang Zhao, and Abhaya Nayak. The discovery of hierarchical cluster structures assisted by a visualization technique. In Kok Wong, B. Mendis, and Abdesselam Bouzerdoum, editors, *Neural Information Processing. Theory and Algorithms*, volume 6443 of *Lecture Notes in Computer Science*, pages 703–711. Springer Berlin / Heidelberg, 2010.

[144] JIANPING ZHOU, Shawn KONECNI, and Georges GRINSTEIN. Visually comparing multiple partitions of data with applications to clustering. 7243:1, 2009. Anglais.