(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 551 250**

(21) Número de solicitud: 201430699

(51) Int. CI.:

G06F 19/00 (2011.01)

(12) SOLICITUD DE PATENTE A2

(22) Fecha de presentación:

13.05.2014

(43) Fecha de publicación de la solicitud:

17.11.2015

(71) Solicitantes:

UNIVERSITAT DE LES ILLES BALEARS (100.0%)
Ctra. de Valldemorsa, km. 7,5. Edifici Son Lledo.
07071 PALMA DE MALLORCA (Illes Balears) ES

(72) Inventor/es:

OLIVER GELABERT, Antoni;
CANALS GUINAND, Vicente José;
MORRO GOMILA, Antoni y
ROSSELLÓ SANZ, José Luis

(74) Agente/Representante:

TEMIÑO CENICEROS, Ignacio

(54) Título: **MÉTODO DE COMPARACIÓN E IDENTIFICACIÓN DE COMPUESTOS MOLECULARES**

(57) Resumen:

Método de comparación e identificación de compuestos moleculares.

La invención se refiere a un método capaz de aislar y comparar geométricamente los núcleos de compuestos moleculares según su polaridad. Preferentemente, el método de comparación de compuestos moleculares de la invención comprende: seleccionar las posiciones de un número *n* de puntos con un determinado valor de al menos una propiedad físico-química asociada a la distribución de la carga eléctrica en dichas moléculas; calcular las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados; establecer una cota máxima d_{max} para las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados; y calcular la similitud entre las moléculas A y B, mediante la comparación de la cota máxima d_{max} con las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados. El método es de aplicación para la identificación de compuestos moleculares, por ejemplo para la identificación de sustancias farmacológicamente activas.

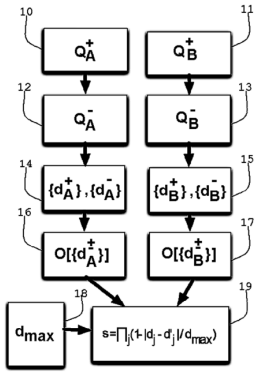

FIG. 1

ES 2 551 250 A2

DESCRIPCIÓN

MÉTODO DE COMPARACIÓN E IDENTIFICACIÓN DE COMPUESTOS MOLECULARES

CAMPO DE LA INVENCIÓN

La presente invención se enmarca en el campo de la química informática o "quimioinformática". Más concretamente, la invención se refiere al uso de metodologías de comparación molecular basadas en su estructura físico-química, orientadas a la búsqueda eficaz en bases de datos moleculares.

ANTECEDENTES DE LA INVENCIÓN

En la actualidad es conocida la utilización de grandes bases de datos moleculares para la detección de compuestos que sean similares a moléculas de actividad biológica conocida, con el objetivo de detectar nuevos compuestos candidatos para su uso como fármacos. La forma molecular tridimensional, junto con la estructura química, son los patrones aceptados habitualmente para discernir dicha actividad, dado que están directamente relacionados con la posible interacción entre la molécula-fármaco y su objetivo macromolecular. Sin embargo, la comparación eficiente entre compuestos, de forma que se tenga en cuenta simultáneamente tanto la forma molecular como la estructura química, sigue suponiendo aún un desafío técnico.

La metodología conocida como "*virtual screening*" es una técnica fundamental para el descubrimiento de fármacos, cuyo objetivo es identificar esas moléculas similares a los fármacos que puedan tener propiedades biológicas beneficiosas. Es una forma efectiva de reducir los costes de caros ensayos biológicos y, asimismo, de hacer frente a la alta tasa de fracasos a la que en la actualidad se enfrenta la industria farmacéutica. En el contexto de esta metodología, por ejemplo aplicada a la técnica basada de acoplamiento molecular, el proceso de acoplamiento de una molécula a un blanco macromolecular (que normalmente suele ser una proteína) se simula para proporcionar una estimación de su energía de enlace, por lo que se proporciona una idea de su actividad biológica. Estas técnicas han estimulado la generación de bases de datos moleculares masivas.

Una metodología alternativa consiste en buscar en una base de datos los compuestos que más se parecen a una determinada molécula que posee una actividad biológica conocida

suponiendo que moléculas con forma y estructura química similar pueden dar lugar a propiedades similares. Dicha similitud puede establecerse con relación a la estructura tridimensional de forma directa, o mediante el desarrollo de descriptores moleculares.

De este modo, los métodos utilizados para comparar la forma molecular se pueden dividir en dos categorías principales: los basados en la superposición y los métodos basados en el uso de descriptores. Los métodos de superposición confían en encontrar una superposición óptima de las moléculas bajo comparación, y los basados en el uso de descriptores (métodos de no superposición) han de ser invariantes a rotaciones y translaciones de las moléculas. Los métodos de superposición están considerados como particularmente eficaces, pero no son tan eficientes computacionalmente como sería deseable. Por su parte, los métodos basados en el uso de descriptores tienen mayor eficiencia computacional, pero se consideran tradicionalmente menos fiables que los métodos de superposición. En cualquiera de los casos, tanto uno como el otro método no tienen en cuenta la estructura química molecular, que es finalmente la que marca la posibilidad de que una molécula-fármaco se pueda adherir a un determinado objetivo molecular. Un método ampliamente utilizado de superposición es el denominado “ROCS” (producto comercializado y en el que se realiza una rápida superposición de estructuras químicas) (ver, por ejemplo, Rush et al., "A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction". J. Med. Chem. 48, 1489-1495 (2005)). ROCS calcula una puntuación de similitud en la superposición volumétrica de las moléculas que se están comparando. La alineación requerida se lleva a cabo a través de, esencialmente, un proceso de optimización local, donde cada una de las iteraciones de dicha optimización implica el cálculo de la superposición volumétrica para la orientación relativa y la posición de las moléculas. Aunque ROCS ha sido promocionado como una metodología mucho más eficiente que los métodos de superposición típica, tiene el inconveniente de que da el mismo valor del radio a todos los átomos de la molécula, lo que puede inducir a errores de cálculo. Igualmente, ROCS no garantiza que la mejor superposición entre moléculas sea la del candidato encontrado y, como método simple de superposición, no tiene en cuenta la estructura química molecular.

Los métodos de comparación basados en el uso de descriptores geométricos codifican la forma de los compuestos, mediante parámetros que son invariantes a rotaciones y translaciones. Estas metodologías son más eficaces computacionalmente que las tradicionales comparaciones tridimensionales. Un ejemplo de técnica basada en el uso de descriptores la encontramos en Zauhar et al., “Shape Signatures, a New Approach to Computer-Aided Ligand- and Receptor Based Drug Design” (1. Med. Chem. 46, 5674-5690

(2003)), en donde cada molécula se describe mediante un histograma de la información derivada de la simulación de un trazado de segmentos dentro del volumen molecular. Dicho histograma relaciona el número de segmentos simulados con la longitud de dichos segmentos. También puede relacionar dicho histograma con la composición química de la molécula, a partir de conocer el potencial de cada punto de colisión de cada segmento con la superficie molecular. Así pues, se genera un histograma bidimensional comparando el número de segmentos con su longitud y el potencial de la superficie donde han colisionado. Finalmente, se comparan las moléculas mediante la superposición de los histogramas. Si bien este método es bastante eficiente, el cálculo de la firma de cada molécula en la base de datos es un procedimiento computacionalmente muy exigente, debido al proceso de propagación de cada segmento. Otro de los defectos de dicha metodología es que se tienen que utilizar un número considerable de parámetros por molécula (unos 50 para el histograma de forma, y unos 250 para el histograma que combina la forma y la estructura química).

Otra técnica de “screening” molecular basada en el uso de descriptores es el método desarrollado por Pedro Ballester ("Shape Recognition Methods and Systems for Searching Molecular Databases", inventor: Pedro J. Ballester, patente num. US 8,244,483 B2, Ago 14, 2012). Dicha metodología, en comparación con el método de Zauhar, proporciona un reducido número de descriptores (12) para cada molécula que lo hace especialmente rápido a la hora de escanear una base de datos extensa. El método se basa en un rápido cálculo de cuatro puntos de la molécula que son invariantes a rotaciones y translaciones (el centroide, el átomo más cercano al centroide, el átomo más lejano al centroide y el átomo más alejado del más lejano al centroide), sobre los que se calculan los tres momentos principales normalizados a la magnitud de longitud. Finalmente, se calcula una métrica de similitud en relación con la distancia Manhatan (esto es, la distancia evaluada en segmentos de camino horizontal y vertical) entre vectores de descripción. El inconveniente principal de dicho método es que no tiene en cuenta la estructura química molecular.

Como metodología de descripción molecular que tenga en cuenta la estructura química, es conocido el trabajo de Christopher A. Hunter ("Quantifying intermolecular interactions: Guidelines for the molecular recognition toolbox" (2004) Angewandte Chemie - International Edition, 43 (40), pp. 5310-5324). En este trabajo se define cada molécula en relación con los máximos y los mínimos de potencial molecular en la superficie de Van der Waals de la misma. Dichos máximos y mínimos se relacionan con las alfas y las betas estequiométricas de interacción molecular. La potencialidad de interacción se relaciona en función de un

cálculo de la energía libre que conlleva el realizar el producto cruzado entre alfas y betas de ambas moléculas. Aunque se tiene en cuenta el potencial de interacción molecular y por tanto la estructura química, no se tiene en consideración la estructura tridimensional del compuesto.

DESCRIPCIÓN BREVE DE LA INVENCIÓN

El objeto de la presente invención es, pues, desarrollar un nuevo método de comparación de compuestos moleculares (por ejemplo, compuestos orgánicos) basados en su estructura físico-química. Dicho objeto se realiza mediante la identificación de las estructuras moleculares tridimensionales con más polaridad, y su comparación partiendo de la base que dichas estructuras darán lugar a las posibles regiones de interacción con otras moléculas, como por ejemplo posibles dianas moleculares para el tratamiento de determinadas enfermedades.

El método propuesto es capaz, por tanto, de aislar y comparar geométricamente los núcleos de las moléculas con más polaridad, comparando las partes biológicamente más activas, además de proporcionar una estimación de las dimensiones de dichos núcleos activos. Así pues, en el método de la invención, una similitud del 100% entre dos moléculas querrá decir que ambas comparten dicho núcleo activo, aunque no necesariamente vayan a ser moléculas idénticas. Dicha estimación de los núcleos activos de las moléculas se realiza, preferentemente, identificando y seleccionando regiones moleculares con determinadas propiedades físico-químicas asociadas a la carga eléctrica de sus átomos, tales como la propia carga eléctrica atómica, el campo eléctrico molecular y/o el potencial eléctrico molecular.

Preferentemente, el método de comparación de compuestos moleculares de la invención comprende, para al menos un par (A, B) de moléculas, los siguientes pasos:

- para ambas moléculas, se seleccionan las posiciones de un número *n* de puntos con un determinado valor de al menos una propiedad físico-química asociada a la distribución de la carga eléctrica en dichas moléculas (en la implementación preferente de la invención serán los n átomos con el envolvente de carga eléctrica máxima);
- para ambas moléculas, se calculan las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados;
- se establece una cota máxima d_{max} para las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados;

- se calcula la similitud entre las moléculas A y B, mediante la comparación de la cota máxima d_{max} con las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados.

Opcionalmente, antes de realizar el cálculo de similitud, se ordenan para cada molécula las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados, empezando por las distancias que unen pares de puntos que poseen mayor polaridad eléctrica, y terminando por las de menor polaridad eléctrica.

En una realización preferente de la invención donde el parámetro físico-químico seleccionado es la carga eléctrica, el método de comparación comprende los siguientes pasos:

- para ambas moléculas (A, B), se seleccionan las posiciones de un número *n* de átomos con un determinado valor de carga positiva Q_A^+ y Q_B^+, y/o las posiciones de un número *n'* de átomos con un determinado valor de carga negativa Q_A^- y Q_B^-;

- para cada molécula, se calculan las distancias atómicas ($\{d_A^+\}$, $\{d_B^+\}$) entre las posiciones de los átomos seleccionados con carga positiva (Q_A^+, Q_B^+) y/o las distancias atómicas ($\{d_A^-\}$, $\{d_B^-\}$) de los átomos seleccionados con carga negativa (Q_A^-, Q_B^-);

- se calcula una cota máxima d_{max} de las distancias atómicas ($\{d_A^+\}$, $\{d_B^+\}$) de los átomos seleccionados con carga positiva (Q_A^+, Q_B^+), y de las distancias atómicas ($\{d_A^+\}$, $\{d_B^+\}$) de los átomos seleccionados con carga negativa (Q_A^-, Q_B^-);

- se calcula la similitud entre las moléculas A y B, mediante la comparación de la cota máxima d_{max} con las distancias atómicas ($\{d_A^+\}$, $\{d_B^+\}$) de los átomos seleccionados con carga positiva (Q_A^+, Q_B^+), y/o con las distancias atómicas ($\{d_A^+\}$, $\{d_B^+\}$) de los átomos seleccionados con carga negativa (Q_A^-, Q_B^-).

La realización anterior de la invención consiste, pues, en identificar aquellos átomos con mayor carga de cada molécula. Para cada compuesto, se calcularán dos grupos de puntos, aquellos que presentan la mayor carga positiva (*n* puntos distintos) y aquellos con la mayor carga negativa (*n'* puntos distintos). Aunque lo habitual será usar un total de cuatro puntos (*n=n'=4*), el método se puede generalizar a cualquier otro número de puntos. Una vez identificados dichos puntos, se estimarán todas y cada una de las distancias entre los mismos para ambos grupos. De esta forma se obtendrán un total de $\binom{n}{2}$ distancias para el grupo de puntos con polaridad positiva y otras $\binom{n'}{2}$ distancias para el grupo de puntos con polaridad negativa, dando lugar a un total de $\binom{n}{2}+\binom{n'}{2}$ distancias que se ordenarán,

preferentemente, con criterios de mayor a menor polaridad de los pares de puntos que se escogen.

Para el cálculo de la similitud, el número *n* y/o *n'* de átomos con un determinado valor de carga positiva o negativa es, preferentemente, 0 ó un número natural superior a 1, donde al menos *n* o *n'* es superior a 0. Para el caso especial de *n*=*n'*=4 se tendrán un total de doce descriptores (o doce componentes para cada vector de descripción molecular). Dicho número resulta adecuado para métodos de comparación basados en bases de datos que comprendan del orden de los mil millones de compuestos.

Preferentemente, el cálculo de similitud *s* se realiza mediante la expresión siguiente:

$$s = \prod_{j=1}^{j=\binom{n}{2}+\binom{n'}{2}} \left(1 - \frac{max(d_{max}, \left|d_{Aj}^{\pm} - d_{Bj}^{\pm}\right|)}{d_{max}}\right).$$

El valor de *s* constituye, pues, una métrica para calcular la similitud entre ambos vectores moleculares, y estará definida entre 1 (características idénticas) y 0 (características completamente distintas).

En otra realización preferente de la invención, una o más de las distancias *d* existentes entre las posiciones de los *n* puntos moleculares seleccionados se obtienen mediante una base de datos de distancias moleculares. En dicha realización, la distancia d_{max} se obtiene, preferentemente, a partir de una distribución de distancias moleculares de la base de datos. Por ejemplo, d_{max} se puede fijar en un valor 3σ por encima del valor medio dμ de la distribución de distancias moleculares de la base de datos (siendo σ la desviación estándar de dicha distribución), de forma que $d_{max}=d\mu+3\sigma$. Dicha realización resulta especialmente adecuada para grandes bases de datos moleculares, que comprendan al menos 10^6 compuestos.

Otro objeto de la presente invención se refiere a un método de identificación de compuestos moleculares que comprende una etapa de comparación de una pluralidad de compuestos, según cualquiera de las realizaciones del método de comparación descrito en el presente documento, junto con una etapa de identificación que comprende seleccionar aquellas comparaciones que poseen una similitud igual o superior a una similitud umbral predeterminada.

Otro objeto de la presente invención es un método de identificación de fármacos que comprende una etapa de comparación según cualquiera de las realizaciones del método de comparación descrito en el presente documento, junto con una etapa de identificación según el método de identificación descrito en el párrafo anterior.

Otro objeto de la presente invención es un sistema para la comparación y/o la identificación de compuestos moleculares que comprende medios físicos de hardware y opcionalmente software, programados con instrucciones para llevar a cabo una o más de las realizaciones de los métodos descritos en el presente documento. Dicho sistema puede comprender, por ejemplo, un ordenador.

DESCRIPCIÓN DE LOS DIBUJOS

En la Figura 1 se muestra un esquema del método de comparación basado en descriptores moleculares, según una realización preferente de la invención.

En la Figura 2 se muestra el resultado del cálculo de similitud para distintas moléculas, según una realización preferente del método de la invención.

En la Figura 3 se muestra una comparación de efectividad de identificación de fármacos entre la realización preferente del método de la invención y el método “Ultra Fast Shape Recognition” (USR) de Pedro Ballester. ("Shape Recognition Methods and Systems for Searching Molecular Databases", inventor: Pedro J. Ballester, patente num. US 8,244,483 B2, Ago 14, 2012)

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

La presente invención tiene por objeto el poder comparar compuestos orgánicos mediante vectores moleculares, cuyos parámetros describen tanto la estructura tridimensional como la composición química de las moléculas orgánicas, con la idea de poder hacer búsquedas rápidas de similitud dentro de grandes bases de datos moleculares. Para ello, primero se describe la estructura 3D de la parte más polar de la molécula, partiendo de la hipótesis de que mediante dicha zona se realizarán las posibles interacciones moleculares. Posteriormente, se comparan los resultados de dicha descripción con las estructuras calculadas de una o más moléculas de referencia (por ejemplo, utilizando bases de datos

moleculares), determinando si existe similitud con determinados compuestos, relacionados con la actividad farmacológica deseada.

Tal y como se muestra en la Figura 1 del presente documento, en el método de la invención se tendrá en cuenta, no solamente la estructura química, sino también la forma molecular de los compuestos. Para las moléculas a comparar (A y B) se calculan las posiciones de un número *n* de átomos con mayor valor de carga positiva (10) y (11). A este conjunto de puntos se le denominará Q_A^+ y Q_B^+. Seguidamente, para ambas moléculas, se buscan los *n*' átomos con mayor valor de carga negativa (12) y (13). A este segundo conjunto de puntos los denominaremos Q_A^- y Q_B^-. Seguidamente para cada molécula se calculan todas las distancias interatómicas dentro de los puntos de Q_A^+ y Q_B^+ (valores que denominaremos $\{d_A^+\}$ y $\{d_B^+\}$). Se hace también lo propio con el conjunto de puntos Q_A^- y Q_B^- a los que definiremos como $\{d_A^-\}$ y $\{d_B^-\}$ respectivamente (referencias (14) y (15) en la Figura 1). Seguidamente, para cada molécula las distancias se ordenan empezando por las parejas de mayor polaridad y acabando por las de menor polaridad (reflejado por la operación "O", en los pasos (16) y (17) de la Figura 1). A partir de una gran base de datos de distancias moleculares $\{d^+\}$ y $\{d^-\}$ se calcula una cota máxima d_{max} de dichas distancias. Finalmente, se calcula la similitud entre las moléculas A y B mediante la expresión:

$$s = \prod_{j=1}^{j=\binom{n}{2}+\binom{n'}{2}} \left(1 - \frac{max(d_{max}, \left|d_{Aj}^{\pm} - d_{Bj}^{\pm}\right|)}{d_{max}}\right) \quad (1)$$

En donde max(x,y) es la función máximo. La expresión (1) vendrá definida entre 0 y 1. De este modo, el valor de *s* proporciona un estimador de la similitud entre ambas moléculas, estando ésta definida entre 0 (ninguna similitud) y 1 (similitud total).

El método propuesto aísla y describe geométricamente el núcleo más polar de cada compuesto. De esta forma, se obtienen descriptores de la parte biológicamente activa de la molécula, lo cual posibilita el poder hacer búsquedas de familias de fármacos que compartan un núcleo activo similar. Entre cada par de moléculas a comparar se estimará, mediante la métrica *s* (expresión (1)), la similitud entre las mismas. Así pues, una similitud de s=1 (100% de similitud) implica que ambas comparten tal cual dicho núcleo activo, aunque no necesariamente vayan a ser moléculas idénticas. Este es el caso, por ejemplo (Figura 2), de la cafeína (24) y la molécula KW-6002 (23), las cuales comparten el mismo núcleo activo y por tanto s=1. De esta forma se pueden comparar moléculas-fármaco por familias, que serán compuestos que comparten el mismo núcleo y que difieren solamente en las zonas más apolares.

Ejemplo de realización de la invención, para n=n'=4:

Para ilustrar el método de la invención, se procede a describir un ejemplo del mismo, donde se procede de la siguiente forma para realizar la comparación molecular. Inicialmente, se estiman los valores de los descriptores de cada par de moléculas a comparar. Para este fin se han de conocer, en cada átomo de ambas moléculas, el valor de la carga eléctrica que se encuentra a su alrededor. Dicha carga puede obtenerse, por ejemplo, a partir de ficheros en formato MOL2. La última columna de dichos ficheros contiene un número que describe la densidad de carga eléctrica alrededor de cada uno de los átomos de la molécula. Existen diferentes métodos elaborados para calcular dichas cargas. Como ejemplo, la base de datos molecular ZINC utiliza el método mecánico-cuántico y semi-empírico AMSOL16. También pueden utilizarse varios programas para la obtención de dichos ficheros, a partir de otros formatos tridimensionales (como PDB o SDF), o bien a partir de simples cadenas de caracteres, como el formato SMILES. Entre muchos otros, está el programa de código abierto Openbabel, muy utilizado en el colectivo de la informática química. Dentro de este programa, se pueden calcular las densidades de carga eléctrica a partir del método GASTEIGER, EEM y MMFF94. Otros como SPARTAN, GAUSSIAN o TORCH tienen incluida también esta opción de conversión a ficheros MOL2 a partir de formatos más simples con varios métodos disponibles. Por tanto, existe la posibilidad de obtener este fichero ya convertido, y disponible en grandes bases de datos moleculares (p. ej. ZINC o LIGANDBOX), o bien calcularlo con alguno de los programas disponibles habilitados para ello y siempre minimizando la energía de la estructura tridimensional como paso previo. También se tiene que conocer la posición de cada átomo dentro de la molécula. Los ficheros de caracterización molecular con el formato MOL2 proporcionan dicha información. A continuación, se calcularán dos grupos de posiciones, la de aquellos átomos que presenten la mayor carga positiva (con un total de *n* puntos distintos, ver Figura 1 (10) y (11)) y aquellos con la mayor carga negativa (otros *n'=n* puntos distintos) (12) y (13). Lo típico será usar un total de cuatro puntos (*n=n'=4*), aunque dicho método se puede generalizar para usar cualquier número de puntos.

Una vez identificados dichos puntos, y para cada componente, se estimarán todas y cada una de las distancias interatómicas dentro de cada grupo. De esta forma se obtendrán un total de $\binom{n}{2}$ distancias para el grupo de puntos con polaridad positiva, y otras $\binom{n}{2}$ distancias para el grupo de puntos con polaridad negativa, dando lugar a un total de $2\binom{n}{2}$ distancias (ver Figura 1, pasos (14) y (15)). Seguidamente, dichas distancias se ordenan por polaridad.

Para cada grupo (Q^+ ó Q^-) primero se seleccionan las distancias entre el átomo más polar con los n-1 átomos restantes. Dichas n-1 distancias se ordenan de mayor a menor polaridad del segundo átomo. Seguidamente se selecciona el segundo átomo más polar del grupo y se cogen las distancias entre dicho átomo con los n-2 átomos restantes (no se tiene en cuenta el átomo más polar, puesto que ya está contemplado en el grupo anterior). Dichas n-2 distancias se ordenan de mayor a menor polaridad del segundo átomo. Este proceso se repite hasta que se emparejen los dos átomos de menor polaridad. En total se habrán ordenado $\binom{n}{2}$ distancias para ambos grupos (el grupo de distancias de carga positiva y el de carga negativa). Así pues, para cada compuesto se obtienen en total $2\binom{n}{2}$ descriptores moleculares estructurados mediante una ordenación de mayor a menor polaridad. Así por ejemplo, para el caso especial de *n=n'=4* (valor típico que se escogerá) se tendrán un total de doce descriptores.

Una vez descrita una base de datos suficientemente extensa mediante dichos descriptores moleculares, se procederá a estudiar los límites de los parámetros que allí se describen. El parámetro d_{max} será definido como aquel que, a partir de la distribución de distancias calculadas, se situará 3σ por encima del valor medio de la distribución (dμ). Siendo σ la desviación estándar (ver en la Figura 1, el paso (18)). Por tanto, se tendrá que $d_{max}=d\mu+3\sigma$. Finalmente, para cada par de moléculas dentro de dicha base de datos se podrá estimar un valor de similitud entre ellas, mediante el uso de la expresión matemática (1) (ver paso (19) de la Figura 1). Así pues, y a partir de los conjuntos de valores $\{d_{Aj}\}$ y $\{d_{Bj}\}$ asociados a ambas moléculas se procederá, mediante el uso de la fórmula (1), a estimar la similitud entre ambos compuestos. El valor de la expresión (1) está acotado entre 0 y 1, correspondiéndose el valor s=1 a la de mayor similitud (o similitud completa).

En la Figura 2 del presente documento se muestra el resultado del cálculo de similitud para tres pares de moléculas distintas. Para la comparación entre la clorpromazina (31) y la tioridazina (32), inhibidores de la dopamina, se obtiene un valor de similitud de s=0.918, muy cercano al 100%. Para la comparación entre la molécula KW-6002 (23) y la cafeína (24), ambas antagonistas de la adenosina A_{2A} y usadas para el tratamiento del Parkinson, se obtiene una similitud completa de s=1 (100%). Finalmente, se muestra el resultado de similitud entre la clomipramina (25) y la promazina (26), ambas inhibidoras de la dopamina, obteniéndose un valor de s=0.806. Como puede apreciase en las distintas figuras, el método propuesto es capaz de aislar el núcleo activo de la molécula mediante el cálculo descrito

para los descriptores moleculares. Así pues, moléculas con el mismo núcleo activo (Schaffold) parecen poseer actividades farmacológicas similares.

En la Figura 3 se muestra una comparativa de la eficiencia de la invención con relación al método USR de Pedro Ballester. En dicha gráfica se muestran los resultados de un experimento de ordenación de compuestos por similaridad a un conjunto de fármacos (en este caso compuestos para el receptor de la encima convertidora de la angiotensina) dentro de una extensa base de datos. Dicha similaridad se calcula tanto por el método USR de Pedro Ballester como por la metodología explicada en la implementación preferente de la invención (en donde n=n'=4 y se utilizan los máximos de la carga eléctrica para referenciar dichos puntos), comparándose ambas con resultados de identificación aleatoria (identificados como “Random”, en la Figura 3). La eficiencia de los métodos se muestra si se colocan a los fármacos conocidos en los primeros puestos del ranking. De esta forma, en la gráfica se muestra, en el eje de las 'X' la relación ordenada de dichos fármacos. Mientras, en el eje 'Y' se muestra qué porcentaje de cobertura hay de los fármacos conocidos. Así pues, como puede observarse, mediante el método propuesto, dentro del primer 1% del ranking aparecen más del 10% de los fármacos conocidos. También se puede observar que en el primer 3% del ranking aparecen aproximadamente el 20% de los compuestos a descubrir. Los resultados son mejores que los proporcionados por el método USR.

Por otra parte, el bajo número de descriptores moleculares utilizados (12 descriptores para *n=n’*=4) implica el poder realizar una comparación muy rápida dentro de una gran base de datos molecular mediante la fórmula (1). Ello hace que el método de la invención resulte especialmente atractivo para aplicaciones quimioinformáticas.

REIVINDICACIONES

1.- Método de comparación de compuestos moleculares que comprende, para al menos un par (A, B) de moléculas, los siguientes pasos:

- para ambas moléculas, se seleccionan las posiciones de un número *n* de puntos con un determinado valor de al menos una propiedad físico-química asociada a la distribución de la carga eléctrica en dichas moléculas;

- para ambas moléculas, se calculan las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados;

- se establece una cota máxima d_{max} para las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados;

- se calcula la similitud entre las moléculas A y B, mediante la comparación de la cota máxima d_{max} con las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados.

2.- Método según la reivindicación anterior, donde la propiedad físico-química utilizada es el campo eléctrico molecular y/o el potencial eléctrico molecular.

3.- Método según cualquiera de las reivindicaciones anteriores donde antes de realizar el cálculo de similitud, para cada molécula, se ordenan las distancias *d* existentes entre las posiciones de los *n* puntos seleccionados, empezando por las distancias que unen pares de puntos que poseen mayor polaridad eléctrica, y terminando por las de menor polaridad eléctrica.

4.- Método según cualquiera de las reivindicaciones anteriores, que comprende los siguientes pasos:

- para ambas moléculas (A, B), se seleccionan las posiciones de un número *n* de átomos con un determinado valor de carga positiva Q_A^+ (10) y Q_B^+ (11), y/o las posiciones de un número *n'* de átomos con un determinado valor de carga negativa Q_A^- (12) y Q_B^- (13);

- para cada molécula, se calculan las distancias atómicas ($\{d_A^+\}$, $\{d_B^+\}$) (14) entre las posiciones de los átomos seleccionados con carga positiva (Q_A^+, Q_B^+) (10, 11) y/o las distancias atómicas ($\{d_A^-\}$, $\{d_B^-\}$) (15) de los átomos seleccionados con carga negativa (Q_A^-, Q_B^-) (12, 13);

- se calcula una cota máxima d_{max} (18) de las distancias atómicas ($\{d_A^+\}$, $\{d_B^+\}$) (14) de los átomos seleccionados con carga positiva (Q_A^+, Q_B^+) (10, 11), y/o de las distancias

atómicas ({d_A^+}, {d_B^+}) (15) de los átomos seleccionados con carga negativa (Q_A^-, Q_B^-) (12, 13);

- se calcula la similitud (19) entre las moléculas A y B, mediante la comparación de la cota máxima d_{max} con las distancias atómicas ({d_A^+}, {d_B^+}) (14) de los átomos seleccionados con carga positiva (Q_A^+, Q_B^+) (10, 11), y/o con las distancias atómicas ({d_A^+}, {d_B^+}) (15) de los átomos seleccionados con carga negativa (Q_A^-, Q_B^-) (12, 13).

5.- Método según la reivindicación anterior, donde el cálculo de similitud (19) se realiza mediante la expresión:

$$s = \prod_{j=1}^{j=\binom{n}{2}+\binom{n'}{2}} \left(1 - \frac{max(d_{max}, \left|d_{Aj}^{\pm} - d_{Bj}^{\pm}\right|)}{d_{max}}\right).$$

6.- Método según cualquiera de las reivindicaciones 4-5, donde antes de realizar el cálculo de similitud para cada molécula, se ordenan (16, 17) las distancias atómicas empezando por las parejas de mayor polaridad y acabando por las de menor polaridad.

7.- Método según cualquiera de las reivindicaciones 4-6, donde el número *n* y/o *n'* de átomos con un determinado valor de carga positiva o negativa es 0 ó un número natural superior a 1, y donde al menos *n* o *n'* es superior a 0.

8.- Método según la reivindicación anterior, donde *n* y/o *n'* es igual a 4.

9.- Método según cualquiera de las reivindicaciones anteriores, donde una o más de las distancias *d* existentes entre las posiciones de los puntos seleccionados se obtienen mediante una base de datos de distancias moleculares.

10.- Método según la reivindicación anterior, donde la distancia d_{max} se obtiene a partir de una distribución de distancias moleculares de la base de datos.

11.- Método según la reivindicación anterior, donde d_{max} se fija en un valor 3σ por encima del valor medio dμ de la distribución de distancias moleculares de la base de datos, siendo σ la desviación estándar de dicha distribución, de forma que d_{max}=dμ+3σ.

12.- Método según cualquiera de las reivindicaciones 7-9, donde la base de datos comprende al menos 10^6 compuestos.

13.- Método según cualquiera de las reivindicaciones anteriores, donde las moléculas son compuestos orgánicos.

14.- Método de identificación de compuestos moleculares que comprende una etapa de comparación según el método de las reivindicaciones 1-13 para una pluralidad de compuestos, y una etapa de identificación que comprende seleccionar aquellas comparaciones que poseen una similitud igual o superior a una similitud umbral predeterminada.

15.- Método de identificación de fármacos que comprende una etapa de comparación según el método de las reivindicaciones 1-13, y/o una etapa de identificación según el método de la reivindicación 14.

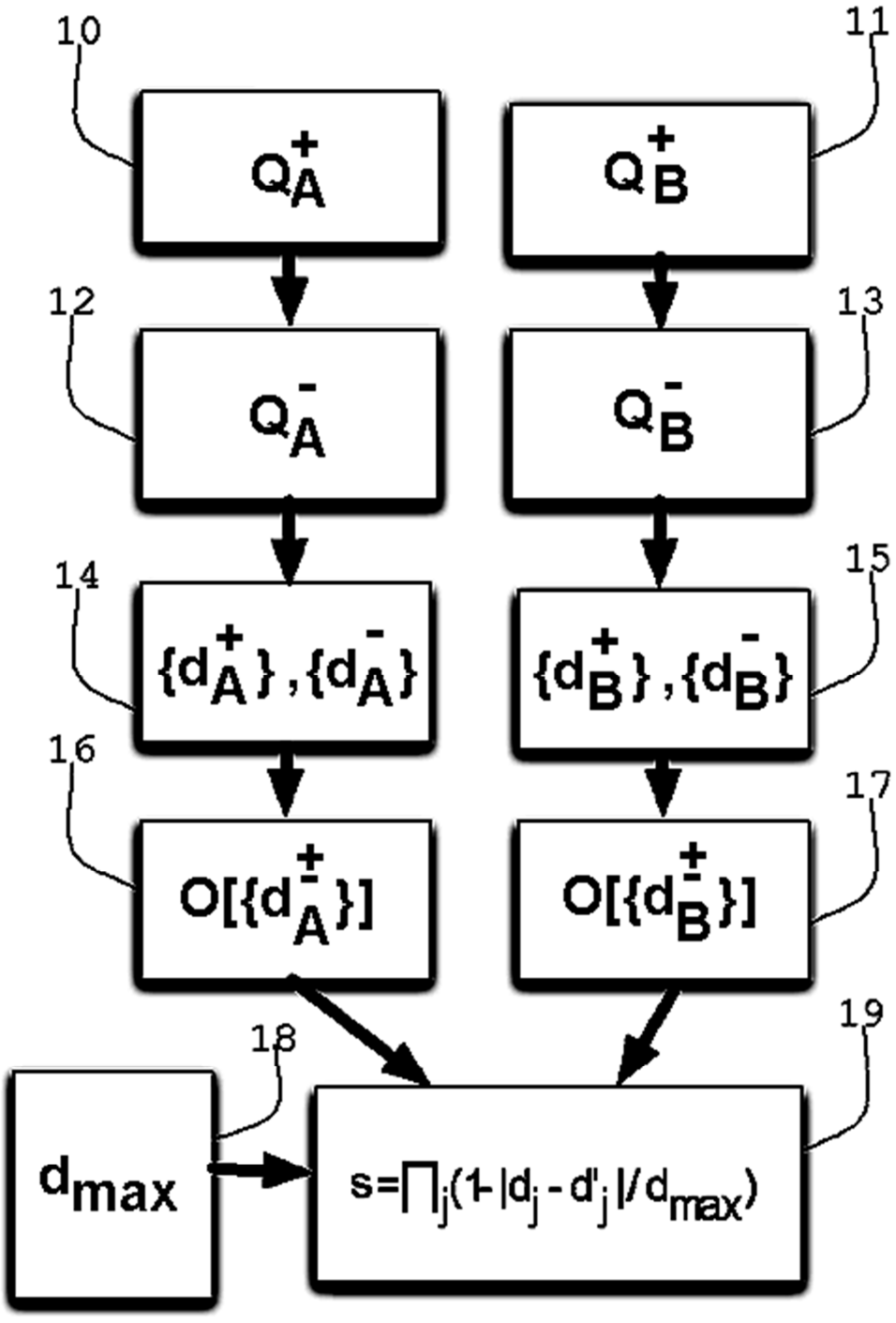
10
11
Q_A^+
Q_B^+
12
13
Q_A^-
Q_B^-
14
15
$\{d_A^+\},\{d_A^-\}$
$\{d_B^+\},\{d_B^-\}$
16
17
$O[\{d_A^\pm\}]$
$O[\{d_B^\pm\}]$
18
19
d_{max}
$s=\prod_j(1-|d_j-d'_j|/d_{max})$

FIG. 1

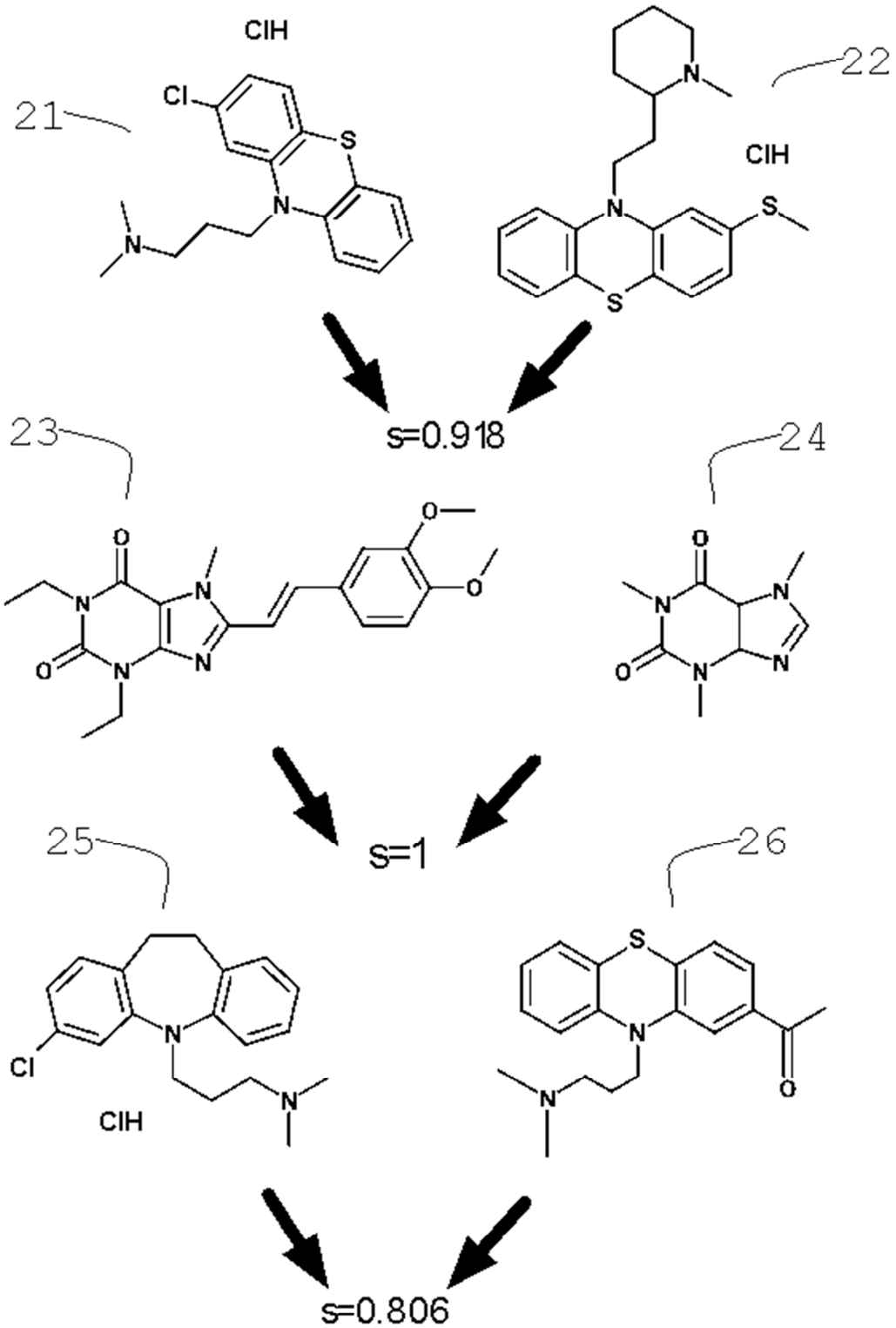
ClH
Cl
21
S
N
N
22
ClH
N
S
S
s=0.918
23
24
O
O
N
N
N
N
O
N
N
s=1
25
26
Cl
N
ClH
S
N
O
N
s=0.806

FIG. 2

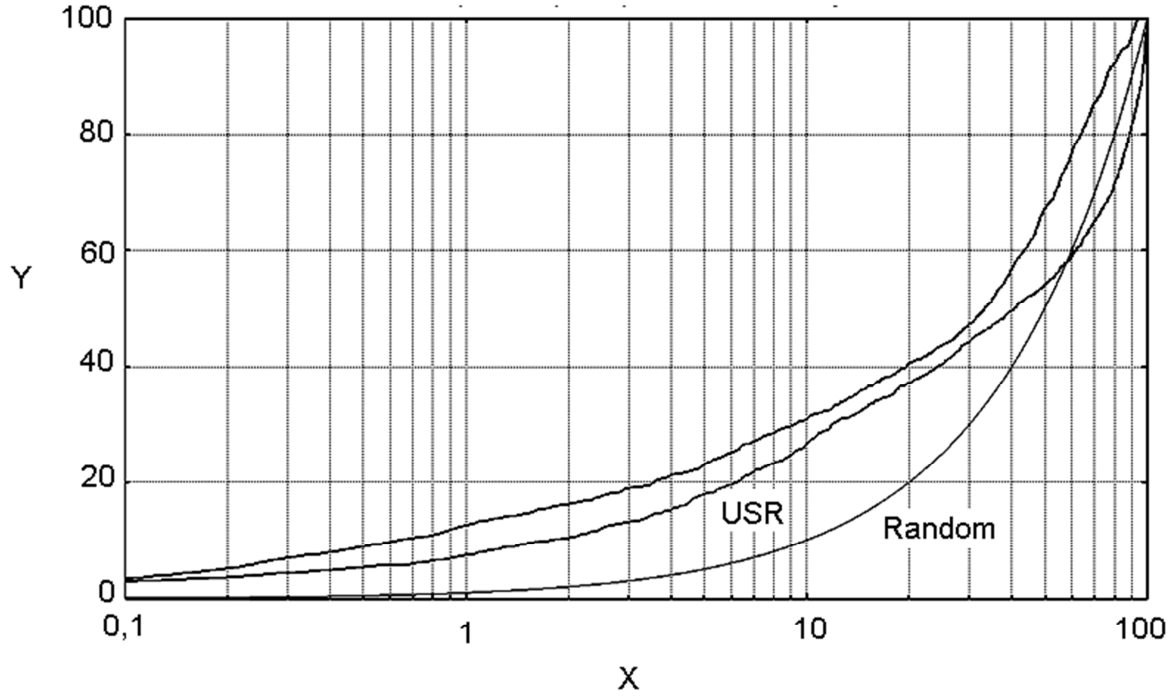

FIG. 3