Acoplamiento molecular basado en ligando por Complejidad LMC

Mauricio Martínez M., Miguel González-Mendoza

Tecnológico de Monterrey, Estado de México,
México

A00964166@itesm.mx, mgonza@itesm.mx

Resumen El *Acoplamiento molecular* enfrenta problemas asociados al análisis de entidades de alta dimensionalidad y pocas muestras; es el caso de los efectos del fenómeno de *Maldición de dimensionalidad* que se presentan cuando se utilizan algoritmos de *Aprendizaje supervisado* y *Optimización matemática* para acoplar moléculas basados en técnicas basados en *ligando* (*LBVS-Ligand-Based Virtual Screening*). Se propone la utilización del concepto de *Complejidad LMC* como medida de relevancia, para identificar moléculas que por comparación con *Compuestos activos* puedan ser propuestos como *Candidatos a medicamento*. El objetivo es medir la similaridad entre vectores mediante *Complejidad LMC* , y ordenar las comparaciones hechas entre ellos con respecto a ésta característica, para descubrir las moléculas más parecidas a los *Compuestos activos*. Se diseñó un algoritmo de medición de similaridad por *Complejidad LMC*, que comparó dos grupos de vectores de alta dimensionalidad; un grupo de vectores *Candidatos a medicamento* contra un grupo de *Compuestos activos* o *Medicamentos*. Los resultados muestran que la aplicación de este concepto sobre los *Compuestos activos* es más informativa que la búsqueda individual de los mejores *Candidatos a medicamento*. Ya que el grado de similaridad global que mantienen con los *Candidatos a medicamento* permite distinguir que *Compuestos activos* son los mejores vectores con los que coincidiran en alto grado los vectores *Candidatos*. La identificación de vectores por ordenamiento evitó algunos de los efectos del fenómeno de *Maldición de dimensionalidad*.

Palabras clave: Acoplamiento molecular, compuesto activo, candidato a medicamento, complejidad LMC, maldición de dimensionalidad.

Molecular Docking based on *ligand* by LMC Complexity

Abstract. *Molecular Docking* faces problems related to *Curse of dimensionality*, due to the fact that it analyzes data with high dimensionality and few samples. *LBVS-Ligand-Based Virtual Screening* conducts studies of docking among molecules using common attributes registered

in data bases. This branch of *Molecular Docking*, uses *Optimization methods* and *Machine learning* algorithms in order to discover molecules similar to known drugs and can be proposed as drug candidates. Such algorithms are affected by effects of *Curse of dimensionality*. It this paper we propose to use *LMC complexity measure* as similarity measurement among vectors in order to discover the best molecules to be drugs; and present an algorithm, which evaluates the similarity among vectors using this concept. The results suggest that application of this concept on *Drug Example* vectors; in order to classify other vectors as drugs candidates which is more informative than individually searching for vectors. Since the *Drug Examples* show a global similarity degree with drug candidate vectors. The aforementioned similarity degree makes it possible to deduce which elements of the *Drug Examples* show higher degree of similarity with drug candidates. Searching of vectors through individual comparison with *Drug Examples* was less efficient, because their classification is affected by the *Drug Examples* with a higher number of global discrepancies. Finally, the proposed algorithm avoids some of the *Curse of dimensionality* effects by using a ranking process where the best drug candidate vectors are those with the lowest complexity.

Keywords: Molecular docking, active compound, drug candidates, LMC complexity, curse of dimensionality.

1. Introducción

Las técnicas de *Acoplamiento molecular* (o *docking* por su denominación en inglés), buscan encontrar el mejor acoplamiento entre dos o más moléculas de tal forma que la afinidad entre ellas sea óptima, [6]. Tales métodos se aplican principalmente al diseño y descubrimiento de medicamentos, Química computacional, Biología molecular y Remediación ambiental entre otras áreas. Estos métodos se implementan mediante la modelación de uniones geométricas entre las moléculas como son: posición, flexibilidad y rotación, [22]. Lo anterior implica explorar el espacio de posibilidades de las características anteriores y evaluar la relevancia de las propiedades que se busca obtener de las moléculas una vez acopladas, [22].

Virtual screening en *Acoplamiento molecular* agrupa un conjunto de procedimientos basados en algoritmos computacionales que identifican nuevos acoplamientos entre moléculas en base a la similaridad, relaciones de actividad e inactividad, propiedades físicas, químicas, estructurales, funcionales, etc. Tales características, entre muchas otras, son registradas para millones de compuestos en diferentes bases de datos, [15],[7],[11],[24]. La información registrada en ellas es representada con vectores de muy alta dimensionalidad, los cuáles son analizados con *Métodos de optimización* y *Aprendizaje de Máquina*, [22],[1].

La naturaleza de estos datos, provoca que los algoritmos empleados para su análisis presenten algunos de los efectos de *Maldición de dimensionalidad*; como son el *Fenómeno del espacio vacío*, *Hipervolumen de cubos y esferas*, *Hipervolumen de corona esférica* y *Concentración de normas y distancias*, [7,14].

En *Virtual screening* existen dos tendencias en la búsqueda de acoplamiento entre moléculas; Métodos basados en estructuras (*SBVS-Structured Based Virtual* Screening) y basados en *ligando* (*LBVS-Ligand-Based Virtual Screening*). Estos últimos emplean algoritmos como *Máquinas de soporte vectorial*, *Arboles de decisión*, *Redes neuronales*, etc. Cuyo objetivo es clasificar compuestos descritos por etiquetas que registran sus atributos y ordenarlos de acuerdo a la afinidad que manifiesten para interactuar con una *Molécula objetivo*, [12]. Los retos que enfrentan estos algoritmos son grandes montos de datos a procesar, y el descubrimiento de formas de discriminación óptimas entre *Compuestos activos* de *inactivos*, [3].

1.1. Trabajos previos

Los principales usos de algoritmos de *Aprendizaje de Máquina* en *LBVS-Ligand-Based Virtual Screening*, es la Clasificación y Búsqueda de similaridad entre compuestos, además de la Identificación de Patrones de similaridad entre ellos. Las *Máquinas de Soporte Vectorial* son utilizadas en las dos primeras actividades. La idea básica en estos algoritmos, es la identificación de un hiperplano de decisión que separe los vectores de datos más cercanos a él de forma óptima; con el fin de clasificarlos en función de este plano. Estos métodos tienen la desventaja de tener un alto costo computacional, además de necesitar un conjunto de datos de entrenamiento adecuado. Su desempeño baja ante la presencia de ruido y traslape de clases, [12].

Los *Arboles de decisión* tienen como objetivo en ésta área, asociar atributos específicos de las moléculas con alguna actividad o propiedad de interés relacionada con su acoplamiento. Estos métodos están basados en métricas de *Ganancia de información*, *Razón de ganancia de información* e *Indice de Divergencia Gini*. Se construyen los árboles, determinando la separación entre ramas mediante medidas de bifurcación basadas en las métricas anteriormente mencionadas. La construcción puede ser de arriba hacia abajo (*Top-Down*) o de abajo hacia arriba (*Bottom-Up*). Son modelos simples, de fácil interpretación y validación. Sin embargo, sufren de alta varianza, pues cambios pequeños en las mediciones, desencadenan un número alto de bifurcaciones. Su diseño depende del tamaño del conjunto de datos de entrenamiento y pueden sufrir de sobreajuste, [12].

Los *Clasificadores Bayesianos simples* (*Naive Bayes* por su denominación en inglés), determinan la probabilidad de la ocurrencia de un evento B dado que se presenta un evento A; lo que es el Principio del *Teorema de Bayes*. Se aplica en problemas donde se busca la probabilidad de que un compuesto representado por un vector de descriptores sea activo, dado que se conoce el compuesto activo A y un conjunto de compuestos inactivos Z de entrenamiento; , Ecuación 1. Tienen un alto costo computacional cuando la dependencia condicional entre variables es alta, [12]:

$$p(C_A|Z) = \frac{p(C_A)p(Z|C_A)}{p(Z)}. \tag{1}$$

K-nearest neighbors es aplicado en la clasificación de compuestos, ordenamiento y predicción bajo una modalidad de regresión. Se basan en la utilización

de *Medidas de distancia* como las *Euclideanas*, *Manhattan*, *Mahalanobis*, etc. Dependen del tamaño de conjunto de datos de entrenamiento, [24].

1.2. Planteamiento del Problema y preguntas de investigación

Un problema de acoplamiento molecular involucra el diseño de funciones de optimización y métodos de búsqueda eficientes que exploren el espacio de soluciones, [7], [17]. Estos métodos deben ser rápidos y tienen que descubrir atributos relevantes que satisfagan las restricciones impuestas para los acoplamientos buscados entre distintas moléculas, además de satisfacer la o las funciones objetivo propuestas, [6], [20].

En algunos problemas de acoplamiento, se pregunta qué compuestos o moléculas registrados en algunas bases de datos tienen la estructura apropiada para acoplarse a una molécula receptora, tal que el acoplamiento resultante, presente propiedades activas para ser medicamento. Se buscan entonces aquellos compuestos que globalmente tengan un alto grado de similaridad con medicamentos ya conocidos y que tengan alta probabilidad de vincularse a una molécula objetivo.

Los algoritmos dedicados a estas actividades trabajan bajo *Aprendizaje supervisado* tienen procesos de aprendizaje largos y requieren un número de instancias específico para ser entrenados, y validados; además de sufrir algunos de los efectos de *Maldición de dimensionlidad* mencionados anteriormente. En cambio, los métodos de *Aprendizaje no supervisado* son rápidos debido a que su implementación depende del uso de *mediciones de relevancia* aplicadas a los datos, pero tienden a perder precisión debido a su dependencia de *parámetros* y *formas de distribución estadística*, *Factores de escala*, presencia de *outliers*, datos incompletos, etc.

Medidas de relevencia basadas en *Entropía de información* ofrecen cierta independencia respecto a los inconvenientes anteriormente mencionados y requieren únicamente de la identificación en los datos, de los eventos o categorias simples que componen las entidades estudiadas; junto con la frecuencia con que se presentan en ellos, [18]. Uno de los conceptos poco explorados basados en *Entropía de información*, es el de *Complejidad*; término que describe la medida de *Desorden* u *Orden* presente en un conjunto de datos, dada la razón que existe entre las Entropías de información individuales de los distintos eventos reconocidos en ellos y la *Máxima entropía* que sustentan si tuvieran una frecuencia uniforme, [2]. Varios investigadores han propuesto diferentes definiciones para ella, es el caso de *Medición de complejidad* propuesto por Shiner et. al., el de *Complejidad LMC* planteado por Ricardo López-Ruiz et. al. y la *Complejidad de Kolmogorov* entre otras, [19], [13], [5].

Los dos primeros conceptos bajo el punto de vista de *Medidas de relevancia* pueden ser implementadas sobre distintos tipos de datos; lo que no es posible con la definición de Complejidad Kolmogorov, [16]. *Medición de complejidad* y *Complejidad LMC* están basadas en los conceptos de *Orden* y *Desorden*, [15]. *Medición de complejidad* es una cantidad adimensional y describe el comportamiento de los eventos simples de un fenómeno por el producto entre *Desorden*

y *Orden*; Entidades con bajas magnitudes de *desorden* y baja *complejidad* indican un comportamiento predecible o invariante; si tienen altas magnitudes de *desorden* y baja *complejidad*, describen entidades cuya frecuencia de eventos es uniforme entre ellos y por tanto impredecibles, [4]. Altas magnitudes de *Medición de complejidad* supone la presencia de patrones discernibles entre los distintos eventos que componen las entidades analizadas, Ecuación 2, 3, 4, 5 y 6.

$$S = \sum_{i=1}^{n} -P_i log_2 P_i, \tag{2}$$

$$S_{max} = Log_2 N, \tag{3}$$

$$\Delta \equiv S/S_{max}, \tag{4}$$

$$\Omega \equiv 1 - \Delta, \tag{5}$$

$$\Gamma_{\alpha\beta} \equiv \Delta^{\alpha} \Omega^{\beta}. \tag{6}$$

De la *Entropía de información* o de *Shannon*; expresada en la ecuación 2, se deriva el concepto de *Medición de complejidad*. El *Desorden*; denotado por Δ, se define como la razón que existe entre la *Entropía de información* y la *Máxima entropía*; Ecuaciones 4, 2 y 3 respectivamente. Mientras que el *Orden* u Ω es la diferencia entre la unidad y Δ; Ecuación 5. *Orden* y *Desorden* son medidas complementarias. Finalmente, la *Medición de complejidad* o $\Gamma_{\alpha\beta}$; Ecuación 6, se define como el producto de *Orden* y *Desorden* para $\alpha = 1$ y $\beta = 1$ en su expresión más sencilla.

La *Complejidad LMC*, conceptualmente es expresada en términos de *Entropía de información* y *Desequilibrio*. Esta definición de Complejidad, es la ponderación de la *Entropía de información* con la distancia que mantiene la probabilidad simple de cada uno de los eventos que componen una entidad, respecto al inverso del número de ocurrencias total del espacio de eventos; Ecuación 7. En este caso, las mínimas magnitudes de complejidad se presentan cuando el número de eventos en una entidad es único, o existen tantos eventos distintos como datos se tienen de la entidad analizada. El caso de evento único, implica una entropía de magnitud cero y se denomina *Estado de cristal* debido a la predicibilidad del fenómeno. Cuando todos los datos de la entidad implican eventos distintos,el *Desequilibrio* es cero y la entropía adquiere su máxima magnitud, lo cuál implica un *Estado de gas*. Estas cualidades son adecuadas para comparar compuestos por similitud, [13], [4]:

$$C = H * D = -(K) * (\sum_{i=1}^{N} P_i * Log_2 P_i) * (\sum_{i=1}^{N} (P_i - \frac{1}{N})^2). \tag{7}$$

Lo anterior da origen a las siguientes preguntas: ¿Es posible plantear un problema de acoplamiento de moléculas en función del concepto de Complejidad?, ¿La sencillez de los cálculos pueden facilitar la búsqueda de compuestos o moléculas adecuados para acoplar con una molécula objetivo?, ¿Puede ser planteada una función objetivo para evaluar el grado de similaridad mediante el concepto de complejidad?.

1.3. Hipótesis y objetivos

La alta dimensionalidad de vectores para un problema de acoplamiento de moléculas basado en ligando (*LBVS-Ligand-Based Virtual Screening*) puede ser planteado en términos simples e interpretable de acuerdo al concepto de *Complejidad LMC*.

El objetivo entonces es diseñar un algoritmo que destaque la similaridad entre compuestos por medio de la *Complejidad LMC*. Planteando dos eventos simples para la comparación; *Acoplamiento* y *Desacoplamiento* entre atributos de compuestos. Todo lo anterior evitando procedimientos de exploración onerosos del espacio de soluciones y el diseño de funciones objetivo complicadas.

Se propone entonces ponderar el número de acoplamientos de un compuesto con una molécula objetivo o compuesto activo, por los atributos comunes en los que coinciden; o su desacoplamiento cuando tienen estados opuestos para un mismo atributo. La valoración de la afinidad entre ellos, dependerá de la Complejidad LMC que se desprende de la frecuencia entre acoplamientos y desacoplamientos. Un *Ordenamiento ascendente* de las magnitudes calculadas de *Complejidad LMC* para las distintas comparaciones entre *vectores no clasificados* y los distintos compuestos activos permitirán distinguir que vectores tienen el grado más alto de similaridad con ellos al corresponder con valores bajos de *Complejidad LMC* o de disimilitud para valores altos.

2. Materiales y métodos

DuPont Pharmaceuticals liberó para KDD Cup 2001 competition un conjunto de datos que comprenden 1908 compuestos[1] que acoplan con la molécula objetivo Trombina, constituidos de 139,351 atributos con representación binaria, [10] . Tales atributos se consideran activos cuando la posición que les representa dentro de un vector-compuesto tiene un cero, e inactivos cuando tienen un uno.

Estos vectores se dividen en dos conjuntos: 42 vectores considerados como *Medicamentos* o *Compuestos activos* y 1886 vectores que deben postularse como *Candidatos a medicamento* dada la similaridad que presenten con los compuestos activos. La tarea consiste en determinar cuáles son los mejores una vez que son comparados; para lo anterior se empleará el algoritmo 1 que utiliza *Complejidad LMC* para medir la similaridad entre vectores.

El algoritmo de construye en dos secciones: una de preprocesamiento y otra de cálculo de similaridad por *Complejidad LMC*. El preprocesamiento tiene como objetivos, identificar entre los vectores; *vectores activos* y *Candidatos a medicamento*, aquellos cuyas componentes en su totalidad son ceros; para posteriormente almacenar en un arreglo los números de vectores que tienen ésta característica. Sobre los vectores restantes se realiza una búsqueda de las posiciones en cada vector que dispongan de unos. Una Lista de listas recibirá

[1] Agradecemos a DuPont Pharmaceuticals Research Laboratories y KDD Cup 2001 por la disposición de este conjunto de datos mediante UCI Machine Learning Repository

las etiquetas de posición de los unos o una etiqueta de cero si el vector está compuesto únicamente por ceros. Habrá tantas listas como vectores tenga el conjunto de datos.

Una vez llevado a cabo lo anterior se calcula la similaridad por *Complejidad LMC* entre los *Candidatos a medicamento* y *vectores activos*, utilizando el arreglo que registró los vectores con ceros y la lista de listas. El proceso realizará tantas comparaciones como combinaciones posibles existan entre los dos conjuntos de vectores establecidos. Los eventos básicos de cada comparación serán los de *Acoplamiento* y *No Acoplamiento*. En el primer caso, cada acoplamiento puede ser por coincidencia por cero o por uno, el desacoplamiento es la diferencia de estados para un mismo atributo entre dos vectores.

El conteo de la frecuencia de los eventos anteriormente es la base del cálculo de *Complejidad LMC*. Las excepción considerada en este proceso, es cuando se comparan dos vectores compuestos solo de ceros; en este caso se asigna un valor cero a la complejidad pues ambos vectores son iguales. Cuando estos son distintos, se utiliza la *diferencia simétrica de conjuntos* con las etiquetas guardadas para cada lista correspondiente a los vectores comparados. La cardinalidad de ésta diferencia constituye el número de desacoplamientos, y cuando ésta cantidad se resta a la dimensionalidad de los vectores tenemos los acoplamientos. Por cada cálculo de similaridad por *Complejidad LMC* entre comparaciones se registra en una tabla los datos de los *Números de vector* comparados, *Frecuencia de Acoplamientos*, *No acoplamientos* y la *Magnitud de similaridad* por *Complejidad LMC*. Finalmente, se llevaran a cabo distintos ordenamientos sobre ésta para descubrir los mejores acoplamientos entre vectores.

El grado más alto de similaridad entre vectores serán aquellas comparaciones que tengan valores de *Complejidad LMC* muy cercanos a cero; por tanto el ordenamiento de las comparaciones por ésta característica será ascendente. Teniendo los mejores resultados al inicio del ordenamiento y los peores al final.

3. Resultados

Los primeros resultados mostraron la existencia de 593 vectores cuyos componentes están constituidos por ceros en su totalidad. De ellos, 2 pertenecen al grupo de *Compuestos activos*, y el resto a *Candidatos a medicamento*. Esto resultó en la presencia de 1182 valores de complejidad LMC cero, producto de las distintas comparaciones que hubo entre estos vectores. Un análisis por cuartiles de similaridad entre los vectores *Candidatos a medicamento* y cada uno de los *Compuestos activos* se muestra en la Figura 1.

Puede observarse que los *Compuestos activos* con bajas magnitudes de *Complejidad LMC* en su comparación con vectores *Candidatos a medicamento*, mantienen una dispersión alta con ellos. De forma opuesta, si la magnitud de la complejidad es alta, la dispersión de los vectores *Candidatos a medicamento* es baja con respecto al *Compuesto activo*.

La identificación individual de vectores *Candidatos a medicamento* con la más baja magnitud de *Complejidad LMC* con cada uno de los 42 *Compuestos activos*

Entrada: $Comp[1 \ldots N][1 \ldots M]$
Salida: $TabAcopPrHlmc[M][Reng_{Inact}, Reng_{Act}, Acoplan, NoAcoplan, Hlmc]$

inicio
  $i \leftarrow 0,\ j \leftarrow 0,\ k \leftarrow 0,\ V^{(0)}[N]$
  $RengListUnos[[N]] = RengListUnos[[1 \ldots L_1] \ldots [1 \ldots L_N]] \quad 1 \leq L_i \leq N$
  $Reng_{Inact},\ Reng_{Act},\ Acoplan,\ NoAcoplan,\ Hlmc$
  $TabAcopPrHlmc[M][Reng_{Inact}, Reng_{Act}, Acop, NoAcop, Hlmc]$
  para cada $i \in N$ **hacer**
    si $Comp[i][1 \ldots M] == \overrightarrow{0}$ **entonces**
      $V^{(0)}[j] \leftarrow i$
      $j \leftarrow j + 1$
    fin
  fin
  para cada $i \in N$ **hacer**
    si $i \in V^{(0)}$ **entonces**
      Añade$(RengListUnos[[L_i]], 0)$
    en otro caso
      para cada $j \in M$ **hacer**
        si $Comp[i][j] == 1$ **entonces**
          Añade$(RengListUnos[[L_i]], j)$
        fin
      fin
    fin
  fin
  para todo $Reng_{Inact} \in Indices(Comp_i != "Activo")$ **hacer**
    para todo $Reng_{Act} \in Indices(Comp_i == "Activo")$ **hacer**
      si $Reng_{Inact} \in V^{(0)} \vee Reng_{Act} \in V^{(0)}$ **entonces**
        si $Reng_{Inact} \in V^{(0)} \wedge Reng_{Act} \in V^{(0)}$ **entonces**
          $Acoplan = N$
        en otro caso
          si $Reng_{Inact} \in V^{(0)}$ **entonces**
            $Acoplan \leftarrow (N - \mathbf{card}(S_{RengListUnos[[Reng_{Inact}]]}))$
            $NoAcoplan \leftarrow \mathbf{card}(S_{RengListUnos[[Reng_{Inact}]]})$
          en otro caso
            $Acoplan \leftarrow (N - \mathbf{card}(S_{RengListUnos[[Reng_{Act}]]})$
            $NoAcoplan \leftarrow \mathbf{card}(S_{RengListUnos[[Reng_{Act}]]})$
          fin
        fin
      en otro caso
        $NoAcoplan \leftarrow$
        $\mathbf{card}(S_{RengListUnos[[Reng_{Inact}]]} \triangle S_{RengListUnos[[Reng_{Act}]]})$
        $Acoplan \leftarrow (N - NoAcoplan)$
      fin
      si $Acoplan == N$ **entonces**
        $Hlmc \leftarrow 0$
      en otro caso
        $Hlmc \leftarrow (\sum_{i=1}^{N} P_i * Log_2 P_i) * (\sum_{i=1}^{N} (P_i - \frac{1}{N})^2) \Big|_{i \in \{Acoplan, NoAcoplan\}}$
      fin
      $TabAcopPrHlmc_{k+1} \leftarrow c(Reng_{Inact}, Reng_{Act}, Acoplan, NoAcoplan, Hlmc)$
    fin
  fin
  return(TabAcopPrHlmc)
fin

Algoritmo 1: Complejidad LMC H*D

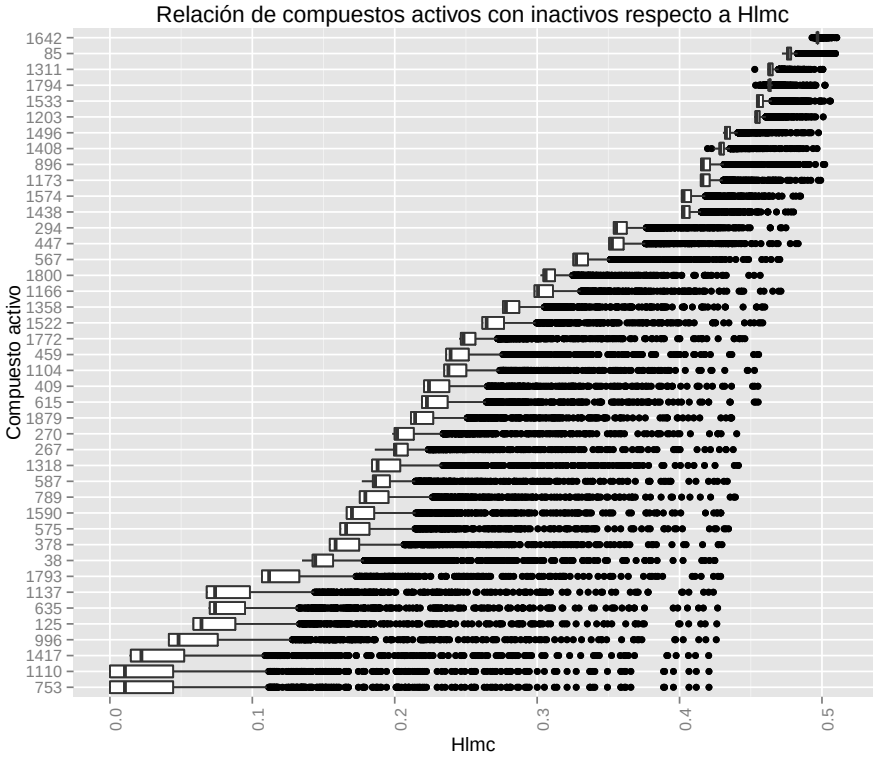


Fig. 1. Similaridad de medicamentos con vectores inactivos por *Complejidad LMC*

muestran que mantienen rangos de complejidad muy amplios en su análsis con boxplots. El ordenamiento de los *Candidatos a medicamento* bajo este criterio muestra valores de complejidad que van desde $3.648\text{x}10^{-4}$ hasta 0.51. La tabla 1 muestra los *Candidatos a medicamento* con más baja complejidad y sus valores máximos y mínimos al ser comparados con los *Compuestos activos*; así como el número de *Compuestos activos* con el que coincidieron bajo esta consideración.

De igual forma fueron identificados los *Candidatos a medicamento* que muestran los más altos niveles de Complejidad LMC, los resultados se muestran en la tabla 2. Se observa, que el número de *Compuestos activos* con los que coinciden es más alto con respecto a los compuestos que ofrecen baja complejidad, pero los grados de similaridad que sustentan son los más bajos.

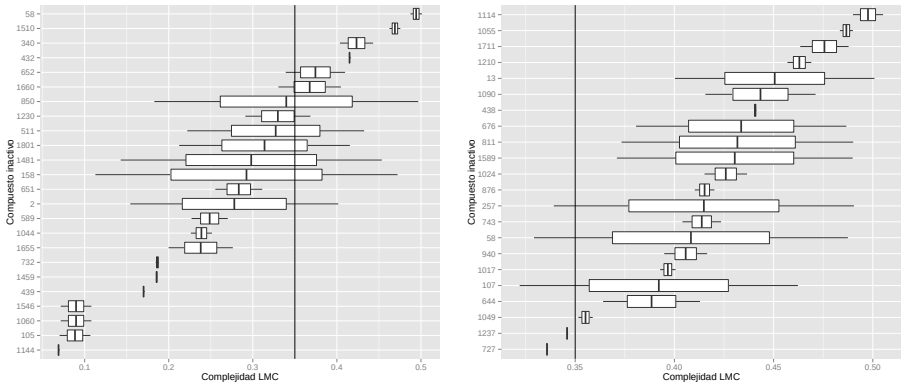
El análisis de similaridad de los *Candidatos a medicamento* con los *Compuestos activos* por boxplots, muestra un comportamiento mixto en la dispersión de estos últimos con respecto a los primeros. Aunque Los *Candidatos a medicamento* con baja complejidad, tienen niveles de dispersión más bajos comparados con los de alta, Figura 2.

Tabla 1. Compuestos Inactivos con baja *Complejidad LMC*

V_{Inact}	$C_{Max.}$	$C_{Min.}$	No. Med.
2	0.041	0.35	13
105	0.0003	0.0003	2
511	0.35	0.4	2
1060	0.18	0.41	3
1459	0.17	0.18	3

Tabla 2. Compuestos Inactivos con alta *Complejidad LMC*

V_{Inact}	$C_{Max.}$	$C_{Min.}$	No. Med.
13	0.051	0.49	4
438	0.47	0.42	19
1024	0.5	0.43	4
1090	0.5	0.42	12
1210	0.5	0.45	3



Vectores cercanos a complejidad cero Vectores con máxima complejidad

Fig. 2. Vectores inactivos

Para el primer recuadro de la figura 2, se observa la coincidencia en posición y ancho de los boxplots para los vectores 105,1060 y 1546. Al verificar el estado de sus atributos, se encontró que muy pocos están en estado inactivo; solo 3 para el vector 105 y de 32 para los últimos dos. Una observación similar se realizó para el segundo recuadro de la figura 2, donde se observa una proximidad alta entre los boxplots de los vectores 676, 811 y 1589; en ellos el número de atributos inactivos es alto, con 12598, 10791 y 12147 respectivamente.

4. Discusión

El acoplamiento de moléculas basados en *ligando* (*LBVS-Ligand-Based Virtual Screening*) utilizando el concepto de *Complejidad LMC* destaca el grado de similaridad globlal que mantienen los *Compuestos activos* con los *Candidatos a medicamento*. Esta característica, ofrece la ventaja de identificar las mejores ejemplos de aprendizaje entre los *vectores activos* para descubrir *Candidatos a medicamento* por similaridad. Sin embargo, la *Identificación individual* por complejidad de los mejores *Candidatos*, tienen el inconveniente de ser afectados por las comparaciones que tienen con los *Compuestos activos* menos eficientes en términos de similitud. Una excepción a las observaciones anteriores, fueron aquellas comparaciones entre vectores que presentaron magnitudes de *Complejidad LMC* cero. Lo que implicó acoples perfectos entre ellos; pudo haber sido interesante si los valores binarios de sus componentes hubieran sido mixtos, pero al revisarlos, se encontró que todos sus componentes tenían ceros.

El Algoritmo propuesto evita algunas de las debilidades asociadas a los métodos utilizados en el acoplamiento de moléculas basados en ligando. La etapa de preprocesamiento sortea la *Exploración del espacio de soluciones* presente en los *Métodos de optimización*, y en el caso de *Aprendizaje de Máquina*, no existe la necesidad de conjuntos de datos balanceados y representativos para procesos de aprendizaje. Los efectos de *Maldición de dimensionalidad* presentes durante el análisis de un número reducido de vectores con alta dimensionalidad, fue compensado al calcular su similaridad mediante el concepto de *Complejidad LMC*. La identificación de eventos simples como los de *Acoplamiento* y *No Acoplamiento* en la comparación de vectores permite la implementación de este concepto en un algoritmo de acoplamiento entre moléculas.

El tiempo de computo se reduce considerablemente al comparar solo las etiquetas de posición de las componentes de los vectores que contienen unos. La medición de similaridad entre dos vectores se implementa utilizando la *diferencia simétrica de conjuntos* con las etiquetas, ya que ésta operación matemática muestra los elementos exclusivos de cada vector; mientras que los elementos restantes son comunes. La cardinalidad de la *diferencia simétrica* es el número de desacoplamientos entre los vectores, y las etiquetas complementarias a la *diferencia simétrica* más las posiciones con valores cero son los acoplamientos.

Finalmente, una característica que se observa en los experimentos es que la naturaleza binaria de la representación de los vectores, determina el costo computacional del algoritmo propuesto. Si la distribución en frecuencia de unos y ceros tiende a ser simétrica, las comparaciones por atributo aumentan y por consecuencia el tiempo de computo. El caso contrario, una distribución asimétrica de estos valores requiere menos tiempo de computo, porque solo se comparan las posiciones de menor frecuencia. Lo anterior implica encontrar vectores *Candidatos a medicamento* de muy baja similaridad con los *Compuestos activos* cuando la frecuencia de unos y ceros tiende a ser simétrica y de alta similaridad para el caso alterno.

5. Conclusiones

El planteamiento de un problema de *Acoplamiento molecular basado en ligando* mediante *Complejidad LMC* está determinado por la identificación de los eventos simples que componen la comparación de dos vectores. El problema planteado en este artículo de medición de similaridad entre vectores *Candidatos a medicamento* y *vectores activos* o ejemplos de medicamento requirió identificar sólo dos eventos; *Acoplamientos* y *No acoplamientos*.

El cálculo de la expresión matemática de *Complejidad LMC* utilizada como *Medida de relevancia* para medir el *Grado de similaridad* entre vectores no requiere la exploración de espacios de soluciones ni procesos de aprendizaje largos, por tanto su implementación en un algoritmo es sencilla y no requiere del diseño de una función objetivo. La evaluación de los resultados para identificar los *Candidatos a medicamento* con el más alto grado de similaridad con los *vectores activos* se realizó por ordenamiento.

la interpretación de la similaridad entre vectores mediante la magnitud de *Complejidad LMC* depende de la *Entropía de información* y el *Desequilibrio*. La *Entropía* determina una magnitud baja de complejidad cuando la frecuencia de alguno de los eventos (*Acoplamientos* o *No acoplamientos*) es preponderante con respecto a los demás. Altas frecuencias de *Acoplamientos*; determinadas por el número de etiquetas comunes en posición y contenido, representan grados de similaridad altos. Si la frecuencia de los eventos es uniforme se obtiene una *Entropía de información* máxima y por lo tanto alta complejidad, lo que significa que la similaridad de los vectores comparados es baja debido a que existe igual número de acoplamiento que de desacoplamientos.

Una complejidad baja por *Desequilibrio* no se presentó en éste análisis, debido a que el número de eventos es mínimo comparado con el número de componentes de los vectores. La interpretación de los resultados permitió deducir que la mejor aproximación de similaridad entre vectores es con respecto a los vectores activos.

los efectos de *Maldición de dimensionalidad* en el algoritmo propuesto son mínimos debido al empleo de ordenamiento por magnitud de complejidad de los distintos acomplamientos existentes entre *Candidatos a medicamento* y *vectores activos*; lo que evita también altos costos computacionales, [23]. Aúnque no es necesario un conjunto de datos entrenamiento como lo exigen los *Métodos de aprendizaje supervisado*, la precisión del algoritmo propuesto es dificíl de determinar, [9]; pues como *Medida de relevancia* la *Complejidad LMC* es un indicador del grado de predicibilidad u orden que sustentan la relación de similaridad entre dos vectores; no se considera un cálculo de error con respecto a un patrón. Aunque, si existe una apreciación global que permite distinguir los *vectores activos* que sustentan el máximo grado de similarida con los *Candidatos a medicamento*, [21]. Consideramos que el concepto de *Complejidad LMC* es una opción flexible en el análisis de *Acoplamiento molecular basado en ligando* y que puede combinar el conocimiento registrados en bases de datos de distintas moléculas junto con sus caracterśticas estructurales, [8].

Referencias

1. Christos A. Nicolaou1, N.B.: Multi-objective optimization methodsin drug design. Drug discovery today 30(20) (2013)
2. Crutchfield, J.P.: Between order and chaos. Nature Physics 8(1), 17–24 (2012)
3. Danishuddin, M., Khan, A.U.: Virtual screening strategies: A state of art to combat with multiple drug resistance strains. MOJ Proteomics Bioinform 2 (2015)
4. Feldman, D.P., Crutchfield, J.P.: Measures of statistical complexity: Why? Physics Letters A 238(4), 244–252 (1998)
5. Grünwald, P.D., Vitányi, P.M.: Kolmogorov complexity and information theory. with an interpretation in terms of questions and answers. Journal of Logic, Language and Information 12(4), 497–529 (2003)
6. Halperin, I., Ma, B., Wolfson, H., Nussinov, R.: Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins: Structure, Function, and Bioinformatics 47(4), 409–443 (2002)
7. Karthikeyan, M., Vyas, R.: Practical chemoinformatics. Springer (2014)
8. Klebe, G.: Virtual ligand screening: strategies, perspectives and limitations. Drug discovery today 11(13), 580–594 (2006)
9. Kurczab, R., Smusz, S., Bojarski, A.J.: Evaluation of different machine learning methods for ligand-based virtual 3(S-1), 41 (2011)
10. Laboratories, D.P.R.: Dorothea data set, https://archive.ics.uci.edu/ml/datasets/Dorothea
11. Lavecchia, A., Di Giovanni, C.: Virtual screening strategies in drug discovery: a critical review. Current medicinal chemistry 20(23), 2839–2860 (2013)
12. Lavecchia, A.: Machine-learning approaches in drug discovery: methods and applications. Drug discovery today 20(3), 318–331 (2015)
13. Lopez-Ruiz, R., Mancini, H., Calbet, X.: A statistical measure of complexity. arXiv preprint nlin/0205033 (2002)
14. Robert Clarke, Habtom W. Ressom, e.a.: The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer 8, 13 (2008)
15. Robert P. Sheridan, S.K.K.: Why do we need so many chemical similarity search methods? Drug Discovery Today 7(17), 903–911 (2002)
16. Seaward, L., Matwin, S.: Intrinsic plagiarism detection using complexity analysis. In: Proc. SEPLN. pp. 56–61 (2009)
17. Shan, S., Wang: Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. Structural and Multidisciplinary Optimization 41(2), 219–241 (Mar 2010), http://dx.doi.org/10.1007/s00158-009-0420-2
18. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal 27, 10–12 (1948)
19. Shiner, J.S., Davison, M., Landsberg, P.T.: Simple measure for complexity. Physical review E 59(2), 1459 (1999)
20. Sousa, S., Ribeiro, A., Coimbra, J., Neves, R., Martins, S., Moorthy, N., Fernandes, P., Ramos, M.: Protein-ligand docking in the new millennium–a retrospective of 10 years in the field. Current medicinal chemistry 20(18), 2296–2314 (2013)
21. Tanrikulu, Y., Krüger, B., Proschak, E.: The holistic integration of virtual screening in drug discovery. Drug Discovery Today 18(7), 358–364 (2013)
22. Teodoro, M.L., Phillips, G.N.: Molecular docking: A problem with thousands of degrees of freedom. In: IEEE International Conference on Robotics and Automation. pp. 960–966 (2001)

23. Zhang, W., Ji, L., Chen, Y., Tang, K., Wang, H., Zhu, R., Jia, W., Cao, Z., Liu, Q.: When drug discovery meets web search: Learning to rank for ligand-based virtual screening. J. Cheminformatics 7, 5 (2015)
24. Zheng, M., Liu, Z., Yan, X., Ding, Q., Gu, Q., platform for ligand-based virtual screening using publicly databases. Molecular diversity 18(4), 829–840 (2014)