Docking Inter/Intra-Molecular Mediante Metaheurísticas Multi-objetivo

Esteban López-Camacho, María Jesús García Godoy, José García-Nieto, Antonio J. Nebro y José F. Aldana-Montes

Resumen— **El Acoplamiento Molecular (Molecular Docking) es un problema de optimización de gran complejidad que consiste en predecir la orientación de dos moléculas: el ligando y el receptor, de manera que formen un complejo molecular energéticamente estable. El docking molecular es un problema tradicionalmente tratado con éxito mediante metaheurísticas para la optimización de un objetivo: la mínima energía libre de unión. Sin embargo, en la literatura actual no se encuentran muchos trabajos que traten este problema desde el punto de vista multiobjetivo. En este sentido, todavía no existen estudios comparativos con el fin de dilucidar qué técnica (o qué tipo de ellas) ofrece un mejor rendimiento en general. En este estudio realizamos una comparativa experimental de una serie de algoritmos multiobjetivo representativos del estado del arte actual, para la resolución de instancias complejas de docking molecular. En concreto, los algoritmos evaluados son: NSGA-II, ssNSGA-II, SMPSO, GDE3, MOEA/D y SMS-EMOA. Para ello, hemos seguido un enfoque de optimización basado en las energías inter- e intra-molecular, siendo éstos los dos objetivos a minimizar. En la evaluación de los algoritmos hemos aplicado métricas de rendimiento para medir la convergencia y la diversidad de los frentes de Pareto resultantes, respecto a frentes de referencia calculados. Además, en comparación con soluciones mono-objetivo obtenidas por técnicas de referencia en el problema (LGA), comprobamos cómo los algoritmos multi-objetivo evaluados son capaces de obtener conformaciones moleculares de mínima energía de acoplamiento.**

Palabras clave— **Docking Molecular, Optimización Multi-Objetivo, Comparativa Experimental**

I. Introducción

En el campo de la bioinformática, el docking molecular es un problema de optimización duro que consiste en predecir la orientación del dos moléculas: una pequeña (el ligando) y una macromolécula (receptor), de manera que formen un complejo molecular energéticamente estable. El objetivo principal es encontrar la conformación óptima entre el receptor y el ligando, que resulte con mínima energía de acoplamiento. Este problema viene siendo tradicionalmente abordado mediante técnicas metaheurísticas y otros métodos computacionales inspirados en la naturaleza, obteniendo resultados satisfactorios en la optimización de un único objetivo: la energía mínima de unión de moléculas con cierto grado de flexibilidad [1]. Sin embargo, no se pueden encontrar muchos trabajos en la literatura donde se trate este problema con un enfoque multi-objetivo. En este sentido, un primer intento fue propuesto en 2006 por Oduguwa et al. [2], en el que se evalúan los algoritmos NSGA-II, PAES y SPEA en el docking de tres complejos moleculares. Grosdidier et al. [3] propuso en 2007 una adaptación de un algoritmo evolutivo llamado EADock en el entorno CHARMM de Harvard Macromolecular Mechanics. En 2008, Janson et al. [4] diseñaron un algoritmo multi-objetivo paralelo llamado ClustMPSO, que utiliza un método k-means para guiar la estrategia de migración entre islas. ClustMPSO utiliza el modelo de evaluación de AutoDock 3.05 y fue probado para la conformación óptima de seis complejos moleculares. Este mismo año, Boisson et al. [5] implementaron un modelo bi-objetivo evolutivo mediante la librería ParadisEO y GOLD para la evaluación de seis instancias moleculares. Recientemente, Sandoval-Perez et al. [6] utilizaron la implementación de NSGA-II de la plataforma jMetal [16] para optimizar las energías de enlace como dos objetivos para el docking de cuatro instancias. La mayoría de estos trabajos presentan la aplicación de una técnica aislada a un conjunto limitado de instancias moleculares de complejidad moderada en términos de flexibilidad. Además, aún no se ha realizado una comparativa algorítmica mono/multi-objetivo sobre un conjunto de compuestos moleculares lo suficientemente amplio y aplicando flexibilidad tanto a los ligandos como a los receptores, dando pie a instancias complejas.

En este trabajo, nuestra principal motivación es la realización de una comparativa y análisis exhaustivo sobre el rendimiento de seis metaheurísticas multi-objetivo en el docking molecular, centrándonos además en compuestos moleculares de alta flexibilidad de conformación. Los algoritmos evaluados son: Nondominated Sorting Genetic Algorithm II (NSGA-II) [7] y su versión de estado estacionario (ssNSGA-II) [8], Speed Modulation Multi-Objective Particle Swarm Optimization (SMPSO) [9], Third Evolution Step of Generalized Differential Evolution (GDE3) [10], Multi-Objective Evolutionary Algorithm Based on Decomposition (MOEA/D) [11] y S Metric Evolutionary Multiobjective Optimization (SMS-EMOA) [12]. Estos algoritmos constituyen un conjunto variado de técnicas modernas multi-objetivo (NSGA-II es utilizado como técnica bien conocida de control) del estado del arte, que desarrollan modelos diferentes de aprendizaje e inducen comportamientos heterogéneos en términos de convergencia, diversidad y escalabilidad. En este sentido, todos estos algoritmos vienen mostrando resulta-

Khaos, Dpto. de Lenguajes y Ciencias de la Computación, Universidad de Málaga E-mail: {esteban,mjgarciag,jnieto,antonio,jfam}@lcc.uma.es

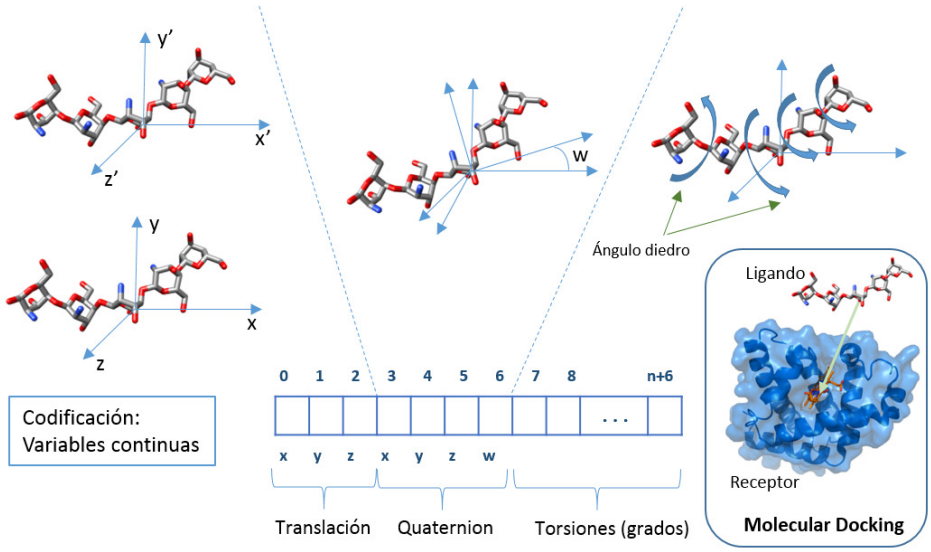

Fig. 1. Esquema de codificación de soluciones para el problema del docking molecular

dos satisfactorios en un gran número de problemas de optimización [13] [14], aunque ninguno de ellos (exceptuando NSGA-II) ha sido utilizado con anterioridad en el docking molecular, lo cual constituye una contribución adicional en este estudio.

Para el estudio experimental hemos utilizado la implementación de los algoritmos seleccionados que provee la plataforma jMetalCpp [15], para el docking de un benchmark de moléculas con diferentes propiedades de tamaño de ligando y resolución cristalográfica. Como estrategia de optimización, hemos seguido la propuesta de Janson et al. [4] mediante el enfoque multi-objetivo de minimizar las energías: inter-molecular e intra-molecular, ambos objetivos contrapuestos como se indica en el citado estudio. Además, los receptores y ligandos han sido configurados con flexibilidad, en lugar de utilizar receptores rígidos en las instancias como se viene haciendo en la literatura actual [4], [5]. De esta manera, abordamos el problema de manera realista y con una mayor complejidad en su resolución. Para la evaluación de las nuevas soluciones generadas, hemos utilizado la popular herramienta de docking molecular AutoDock 4.2 [17], dando pie así a la realización de comparativas adicionales mono-objetivo con el estado del arte y con propuestas futuras.

La metodología de experimentación seguida consiste en realizar un número predefinido de evaluaciones para cada instancia y algoritmo, para más tarde realizar la comparación estadística de los resultados en términos de dos indicadores de calidad diferentes para medir la convergencia y la diversidad de los frentes de Pareto obtenidos. De manera adicional, hemos realizado un análisis sobre algunas soluciones desde el punto de vista biológico, así como en el contexto de una comparativa con la técnica mono-objetivo Lamarckian Genetic Algorithm (LGA) [1], diseñada específicamente para tratar el docking molecular por los desarrolladores de Autodock.

Este trabajo se organiza de la siguiente manera. La Sección II describe el problema del docking molecular y detalla la estrategia multi-objetivo seguida. En la Sección III se presentan los algoritmos evaluados. Seguidamente, la metodología experimental, los resultados obtenidos y los análisis se detallan en la Sección IV. Finalmente, la Sección V contiene las conclusiones y trabajo futuro.

II. El Docking Molecular: Enfoque Multi-objetivo

Como hemos mencionado anteriormente, el docking molecular óptimo consiste en encontrar una conformación, o posiciones de unión, entre el ligando (L) y el receptor (R) que resulte con energía de acoplamiento mínima. Podemos describir la interacción entre el ligando y el receptor mediante una función de energía que se calcula a través del movimiento de tres componentes, en forma de grados de libertad: (1) movimiento de translación del ligando, respecto a los valores de los tres ejes de coordenadas (x, y, z) en el espacio cartesiano; (2) orientación del ligando, modelada mediante la cuaterna que incluye la pendiente del ángulo (θ); (3) la flexibilidad, representada mediante los movimientos de rotación y torsión (ángulos diedros) del ligando y la zona de acoplamiento del receptor.

Codificación. Para la codificación de las soluciones, hemos utilizado un esquema de vector de números reales formado por $7+n$ variables (como se ilustra en Fig. 1) mediante el cual, los tres primeros valores se corresponden con los movimientos de translación del ligando, los cuatro siguientes valores representan la orientación del ligando y/o la macromolécula receptora, y las n restantes variables codifican los ángulos de torsión del ligando. Sin embargo, con la intención de restringir el espacio de búsqueda, hemos

utilizado una metodología en el diseño de la instancia basada en localizar la zona activa de la proteína receptora en una rejilla 3D rectangular, de manera que se calcula y guarda la energía de interacción electrostática y el término de Van der Waals para cada tipo de átomo en cada punto del grid [18]. De esta forma, la contribución de la proteína en cada punto del grid se obtiene mediante la interpolación trilineal en cada celda del grid. Esta interpolación se realiza sobre un rango de variables de translación (x, y, z) de 120 radianes en los límites $[-\pi, \pi]$.

Objetivos. Desde el punto de vista mono-objetivo, la mayoría de propuestas están enfocadas en optimizar la energía libre de unión, tal y como se calcula mediante la Ecuación 1. Sin embargo, siguiendo la propuesta multi-objetivo formulada en [4], podemos evaluar esta energía de manera separada a través de las energías inter-molecular e intra-molecular, E_{inter} y E_{intra} respectivamente, como objetivos distintos y contrapuestos a minimizar, es decir:

- Objetivo 1: E_{intra}, energía intra-molecular entre el ligando y el receptor estimada mediante la diferencia de los estados de enlace (bound) y no enlace (unbound) respecto a ambas moléculas (Ecuación 2).
- Objetivo 2: E_{inter} energía inter-molecular (Ecuación 3) estimada mediante los diferentes estados de enlace (bound) y no enlace (unbound) respecto al complejo (ligando-receptor) completo (Ecuación 3).

$$\Delta G = E_{intra} + E_{inter} + \Delta Z_{conf} \tag{1}$$

$$E_{intra} = (Q^{L-L}_{bound} - Q^{L-L}_{unbound}) + (Q^{R-R}_{bound} - Q^{R-R}_{unbound}) \tag{2}$$

$$E_{inter} = (Q^{R-L}_{bound} - Q^{R-L}_{unbound}) \tag{3}$$

$$\Delta Z_{conf} = W_{conf} N_{tors} \tag{4}$$

$$\begin{aligned} Q = W_{vdw} \sum_{i,j} (\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}) \\ + W_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + \\ + W_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \\ + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2 / 2\sigma^2)} \end{aligned} \tag{5}$$

En estas ecuaciones, cada par de términos incluyen las evaluaciones (Q) de: dispersión/repulsión (vdw), enlaces de hidrógeno $(hbond)$, electrostáticos $(elec)$ y desolvatación (sol). Los pesos W_{vdw}, W_{hbond}, W_{elec} y W_{sol} de la Ecuación 5 son constantes de Van der Waals, enlace de hidrógeno, interacción electrostática y desolvatación, respectivamente. La distancia inter-atómica se representa mediante r_{ij} y los parámetros de Lennard-Jones tomados de la fuerza Amber se representan mediante A_{ij} y B_{ij}. De manera similar, en el segundo término de la Ecuación 5, los parámetros de Lennard-Jones C_{ij} y D_{ij} se utilizan para las energías potenciales entre dos átomos y $E(t)$ define el ángulo de direccionalidad. El tercer término en esta misma ecuación utiliza un enfoque de Coulomb para la electrostática. El cuarto término se calcula a partir del volumen (V) de los átomos que circundan un átomo dado S y el término exponente que representa la distancia entre átomos.

De manera adicional, la Ecuación 1 contiene un término de entropía de conformación (ΔZ_{conf}), directamente proporcional al número de enlaces rotables $(Ntors)$ en la molécula ligando (Ecuación 4). Este término se multiplica además por la constante de fuerza torsional W_{conf}. No obstante, cabe destacar que ΔZ_{conf} se calcula únicamente para el caso de evaluación mono-objetivo (como se verá en la Sección IV-C). Una explicación más detallada de cada uno de estos términos se puede encontrar en [1], donde se explica el modelo de evaluación de Autodock.

III. Algoritmos Evaluados

A continuación, se ofrece una breve descripción de los seis algoritmos de optimización multi-objetivo seleccionados para este estudio experimental:

- NSGA-II: NSGA-II [7] es posiblemente el algoritmo multi-objetivo más utilizado en la literatura relacionada. Se trata de un algoritmo genético que utiliza los operadores genéticos tradicionales (selección, cruce y mutación), además de un procedimiento de ranking para asegurar la convergencia. No obstante, para mejorar la diversidad en el conjunto de soluciones que maneja, utiliza un estimador de densidad basado en distancia de crowding.
- ssNSGA-II: versión de NSGA-II de estado estacionario [8]. Esta propuesta mejora los resultados de NSGA-II para un benchmark de problemas estándar, aunque con el sobre coste de incrementar el tiempo de ejecución, con respecto al algoritmo original.
- SMPSO: propuesto en [9], se trata de una versión multi-objetivo de particle swarm optimization que utiliza una estrategia para limitar la velocidad de las partículas, para reconducirlas hacia posiciones efectivas en aquellos casos en los que la velocidad toma valores más altos que el rango de las variables del problema. SMPSO incluye además un operador de mutación polinomial como factor de turbulencia y un archivo de soluciones no-dominadas a lo largo del proceso de optimización.
- GDE3: Algoritmo de Evolución Diferencial generalizada (GDE) [10] basado en NSGA-II, aunque con operadores diferenciales de cruce y mutación. GDE3 introduce una mejora sobre la distancia de crowding utilizada en NSGA-II para generar un conjunto de soluciones mejor distribuidas.
- MOEA/D: MOEA/D [11] adopta una estrategia de optimización mediante la descomposición de un problema multi-objetivo en varios subproblemas, los

TABLA I

Instancias de coordenadas obtenidas mediante cristalografía de rayos X de la base de datos PDB (Protein Data Bank), junto con los códigos de asociación (en PDB) y la resolución (en Å)

Complejos proteína-ligando	Código PDB	Resolución (Å)
HIV-1 proteasa/AHA006	1AJV	2
HIV-1 proteasa/AHA001	1AJX	2
HIV-1 proteasa/α-D-glucose	1BV9	2,20
HIV-1 proteasa/Macrocyclic peptidomimetic inhibitor 8	1D4K	1,85
HIV-1 proteasa/AHA047	1G2K	1,95
HIV-1 proteasa/U75875	1HIV	2
HIV-1 proteasa/KNI-272	1HPX	2
HIV-1 proteasa/GR126045	1HTF	2,20
HIV-1 proteasa/GR137615	1HTG	2
HIV-1 proteasa/Q8261	1HVH	1,80
HIV-1 proteasa/U100313	2UPJ	3

cuales son entonces optimizados de manera simultanea, utilizando información de los otros subproblemas vecinos. Para este estudio utilizamos la variante MOEA/D-DE [19], que aplica operadores de evolución diferencial en lugar de los originales operadores genéticos de cruce y mutación.

- SMS-EMOA: Se trata de un algoritmo evolutivo de estado estacionario que utiliza un operador de selección basado en la medida del Hipervolumen (HV), en combinación con la ordenación de las soluciones no-dominadas. La principal característica de SMS-EMOA [12] consiste en el guiado de soluciones a conjuntos bien distribuidos de éstas, con lo que es posible centrar la búsqueda en ciertas regiones del frente de Pareto, aunque con el coste adicional en tiempo de computo al tener que recalcular el HV.

De esta forma, hemos incluido en este estudio el algoritmo más utilizado en optimización multiobjetivo (NSGA-II) y su variante de estado estacionario (ssNSGA-II), una técnica de inteligencia colectiva basada en enjambres de partículas (SMPSO), una técnica de evolución diferencial (GDE3), un algoritmo basado en descomposición de problemas (MOEA/D) y una algoritmo de búsqueda guiada mediante indicador (SMS-EMOA).

IV. Experimentos

En este trabajo hemos utilizado un benchmark de 11 instancias de complejos proteína-ligando con flexibilidad en ambas moléculas. Estas instancias fueron obtenidas de la base de datos PDB (Protein Data Base) [1], haciéndolas además accesibles online para facilitar la reproducción experimental. Hemos seleccionado estos complejos moleculares ya que constituyen un conjunto variado con diferentes rangos de cristalografía rayos X y resolución (de 1,8 a 3 Å), además de contener ligandos de estructura heterogénea (urea cíclica, inhibidores de diferentes tamaños, etc). Además, estas instancias han sido utilizadas previamente en trabajos de referencia [1], donde la flexibilidad de la macromolécula está también tenida en cuenta en la función de energía de AutoDock 4.2.

[1]URL: http://www.rcsb.org/pdb/home/home.do

La Tabla I muestra el conjunto de instancias moleculares utilizadas en nuestros experimentos, indicando la cristalografía de rayos X, los códigos de asociación (en PDB) y la resolución (en Å). Para todas estas instancias, los grados de libertad de los movimientos de torsión para ligandos y macromoléculas son 10 y 6, respectivamente, seleccionando aquellas torsiones que permiten considerar el menor número de átomos respecto al núcleo del ligando. Por lo tanto, el número total de variables (n) en las soluciones para estas instancias es de 23 (3 para translación, 4 para rotación y 16 para torsión).

El procedimiento experimental llevado a cabo en este trabajo ha supuesto la realización de 30 ejecuciones independientes por cada algoritmo evaluado e instancia molecular. A partir de las distribuciones resultantes de estas ejecuciones, hemos calculado la mediana y el rango intercuartílico (IQR) como medidas de tendencia central y dispersión estadística, respectivamente. Además, como métricas de rendimiento algorítmico nos hemos centrado en dos indicadores de calidad de solución: Hipervolumen (I_{HV}) y Epsilon Aditivo Unario ($I_{\epsilon+}$) [14]. La primera métrica es indicativa tanto de la convergencia como de la diversidad, mientras que la segunda ($I_{\epsilon+}$) está totalmente enfocada a validar el comportamiento respecto a convergencia. En este sentido, cabe mencionar que estamos tratando un problema real de cual no hay conocimiento del frente óptimo de Pareto que de soporte al calculo de las métricas de calidad. Por tanto, hemos recurrido a la generación de un Frente Pareto de Referencia (Reference Front, RF) para cada instancia del problema mediante el computo de todas las soluciones no-dominadas encontradas a partir de los experimentos realizados con todos los algoritmos evaluados en este estudio.

Como ya se introdujo en la sección anterior, hemos utilizado la implementación de los seis algoritmos estudiados que ofrece el framework jMetalCpp [15], en combinación con la herramienta AutoDock 4.2 de simulación de docking para la evaluación de las soluciones generadas. La plataforma experimental utilizada para las ejecuciones independientes realizadas

TABLA II

MEDIANA Y RANGO INTERCUARTÍLICO DE LAS MÉTRICAS I_{HV} Y $I_{\epsilon+}$, PARA CADA ALGORITMO EVALUADO Y CADA INSTANCIA DE DOCKING. LAS MEJORES Y SEGUNDOS MEJORES RESULTADOS SE INDICAN CON FONDO GRIS Y GRIS CLARO, RESPECTIVAMENTE

I_{HV}	Hipervolumen					
	NSGA-II	ssNSGA-II	SMPSO	GDE3	MOEA/D	SMS-EMOA
1AJV	$4{,}75e{-}2_{1{,}1e-1}$	$7{,}47e{-}4_{3{,}1e-2}$	$2{,}65e{-}1_{9{,}5e-2}$	$1{,}23e{-}1_{9{,}2e-2}$	$7{,}39e{-}2_{1{,}3e-1}$	$1{,}33e{-}2_{1{,}4e-1}$
1AJX	$2{,}91e{-}1_{1{,}2e-1}$	$2{,}96e{-}1_{1{,}3e-1}$	$5{,}69e{-}1_{1{,}1e-1}$	$3{,}44e{-}1_{3{,}2e-2}$	$3{,}22e{-}1_{1{,}2e-1}$	$3{,}07e{-}1_{1{,}6e-1}$
1BV9	$1{,}33e{-}1_{1{,}3e-1}$	$1{,}17e{-}1_{1{,}1e-1}$	$5{,}41e{-}1_{3{,}7e-1}$	$1{,}79e{-}1_{1{,}9e-2}$	$1{,}81e{-}1_{1{,}2e-1}$	$1{,}34e{-}1_{1{,}1e-1}$
1D4K	$2{,}38e{-}1_{1{,}3e-1}$	$2{,}54e{-}1_{1{,}2e-1}$	$5{,}09e{-}1_{1{,}5e-1}$	$3{,}79e{-}1_{2{,}1e-2}$	$2{,}76e{-}1_{7{,}8e-2}$	$2{,}62e{-}1_{9{,}6e-2}$
1G2K	$8{,}06e{-}2_{1{,}5e-1}$	$6{,}11e{-}2_{1{,}0e-1}$	$5{,}74e{-}1_{2{,}0e-1}$	$1{,}62e{-}1_{6{,}4e-2}$	$1{,}13e{-}1_{1{,}2e-1}$	$5{,}69e{-}2_{1{,}0e-1}$
1HIV	$7{,}12e{-}2_{9{,}0e-2}$	$5{,}39e{-}2_{8{,}0e-2}$	$7{,}61e{-}2_{8{,}5e-2}$	$1{,}10e{-}1_{7{,}8e-2}$	$7{,}78e{-}2_{8{,}8e-2}$	$5{,}80e{-}2_{9{,}5e-2}$
1HPX	$5{,}77e{-}2_{1{,}3e-1}$	$2{,}09e{-}2_{6{,}8e-2}$	$3{,}06e{-}1_{3{,}8e-1}$	$9{,}50e{-}2_{4{,}2e-2}$	$8{,}69e{-}2_{9{,}7e-2}$	$4{,}93e{-}2_{1{,}1e-1}$
1HTF	$2{,}80e{-}1_{2{,}8e-1}$	$2{,}85e{-}1_{2{,}0e-1}$	$5{,}29e{-}2_{1{,}3e-1}$	$4{,}03e{-}1_{1{,}3e-1}$	$4{,}98e{-}1_{1{,}8e-1}$	$2{,}95e{-}1_{2{,}1e-1}$
1HTG	$9{,}20e{-}2_{1{,}0e-1}$	$9{,}08e{-}2_{8{,}5e-2}$	$3{,}68e{-}3_{2{,}7e-2}$	$1{,}80e{-}1_{1{,}6e-2}$	$1{,}62e{-}1_{3{,}3e-2}$	$1{,}47e{-}1_{8{,}8e-2}$
1HVH	$2{,}10e{-}1_{1{,}5e-1}$	$9{,}27e{-}2_{1{,}1e-1}$	$5{,}04e{-}1_{2{,}9e-1}$	$3{,}28e{-}1_{1{,}0e-1}$	$1{,}78e{-}1_{1{,}9e-1}$	$9{,}22e{-}2_{2{,}2e-1}$
2UPJ	$3{,}65e{-}1_{7{,}2e-2}$	$3{,}75e{-}1_{9{,}5e-2}$	$4{,}23e{-}1_{1{,}1e-1}$	$5{,}20e{-}1_{1{,}6e-1}$	$4{,}05e{-}1_{9{,}3e-2}$	$3{,}75e{-}1_{1{,}1e-1}$
$I_{\epsilon+}$	Epsilon					
	NSGA-II	ssNSGA-II	SMPSO	GDE3	MOEA/D	SMS-EMOA
1AJV	$7{,}74e{+}0_{2{,}2e+0}$	$8{,}64e{+}0_{1{,}7e+0}$	$1{,}48e{+}0_{3{,}2e-1}$	$6{,}55e{+}0_{8{,}3e-1}$	$6{,}84e{+}0_{1{,}7e+0}$	$8{,}21e{+}0_{2{,}4e+0}$
1AJX	$6{,}65e{+}0_{1{,}5e+0}$	$7{,}04e{+}0_{1{,}4e+0}$	$1{,}64e{+}0_{2{,}6e-1}$	$6{,}44e{+}0_{3{,}9e-1}$	$6{,}18e{+}0_{9{,}5e-1}$	$6{,}51e{+}0_{2{,}0e+0}$
1BV9	$1{,}11e{+}1_{1{,}8e+0}$	$1{,}15e{+}1_{2{,}0e+0}$	$1{,}09e{+}0_{8{,}7e-1}$	$1{,}15e{+}1_{3{,}7e-1}$	$1{,}03e{+}1_{1{,}6e+0}$	$1{,}15e{+}1_{2{,}0e+0}$
1D4K	$1{,}15e{+}1_{2{,}5e+0}$	$1{,}17e{+}1_{2{,}4e+0}$	$2{,}45e{+}0_{6{,}7e-1}$	$1{,}06e{+}1_{1{,}3e+0}$	$1{,}11e{+}1_{1{,}7e+0}$	$1{,}14e{+}1_{2{,}5e+0}$
1G2K	$7{,}55e{+}0_{2{,}4e+0}$	$7{,}81e{+}0_{1{,}7e+0}$	$9{,}30e{-}1_{5{,}1e-1}$	$6{,}13e{+}0_{1{,}0e+0}$	$6{,}99e{+}0_{1{,}7e+0}$	$7{,}88e{+}0_{1{,}8e+0}$
1HIV	$8{,}85e{+}0_{1{,}7e+0}$	$9{,}49e{+}0_{1{,}5e+0}$	$3{,}11e{+}0_{3{,}4e+0}$	$8{,}75e{+}0_{1{,}2e+0}$	$7{,}88e{+}0_{1{,}8e+0}$	$9{,}36e{+}0_{2{,}0e+0}$
1HPX	$6{,}98e{+}0_{1{,}7e+0}$	$7{,}28e{+}0_{1{,}2e+0}$	$2{,}60e{+}0_{6{,}0e+0}$	$6{,}93e{+}0_{3{,}6e-1}$	$6{,}04e{+}0_{1{,}4e+0}$	$7{,}18e{+}0_{1{,}2e+0}$
1HTF	$5{,}26e{+}0_{2{,}0e+0}$	$5{,}23e{+}0_{1{,}5e+0}$	$6{,}93e{+}0_{1{,}4e+0}$	$4{,}37e{+}0_{9{,}4e-1}$	$3{,}65e{+}0_{1{,}3e+0}$	$5{,}16e{+}0_{1{,}6e+0}$
1HTG	$2{,}23e{+}1_{2{,}9e+0}$	$2{,}23e{+}1_{2{,}3e+0}$	$3{,}57e{+}0_{3{,}8e+0}$	$1{,}97e{+}1_{8{,}8e-1}$	$1{,}86e{+}1_{1{,}2e+0}$	$2{,}09e{+}1_{2{,}6e+0}$
1HVH	$8{,}56e{+}0_{2{,}5e+0}$	$9{,}92e{+}0_{2{,}1e+0}$	$3{,}08e{+}0_{2{,}2e+0}$	$7{,}54e{+}0_{1{,}2e+0}$	$7{,}53e{+}0_{2{,}2e+0}$	$1{,}01e{+}1_{3{,}3e+0}$
2UPJ	$8{,}05e{+}0_{1{,}6e+0}$	$7{,}74e{+}0_{2{,}2e+0}$	$1{,}60e{+}0_{2{,}2e-1}$	$5{,}82e{+}0_{2{,}7e+0}$	$6{,}45e{+}0_{1{,}0e+0}$	$7{,}59e{+}0_{1{,}8e+0}$

consiste en un gestor de tareas en paralelo Condor[2], manejando un número máximo de 400 núcleos (cada tarea implica una ejecución independiente).

A. Parámetros

Todos los algoritmos evaluados han sido configurados con un tamaño de población (o cúmulo) de 150 individuos desarrollando un número total de 1.500.000 evaluaciones, siendo este el criterio de parada. Estos valores vienen establecidos por defecto en los algoritmos de AutoDock y dependen de sus configuraciones iniciales. Otros estudios en la literatura [20] donde se trataron las mismas instancias utilizaron también esta configuración, así que hemos decidido mantenerla en aras a establecer comparativas lo más justas posible con el estado del arte.

Respecto a los parámetros específicos de cada algoritmo, hemos utilizado el conjunto de valores recomendado en los trabajos de referencia de cada propuesta (también utilizados como parámetros por defecto en jMetal). En concreto, para los algoritmos genéticos NSGA-II y ssNSGA-II, además de para SMS-EMOA, se han utilizado los operadores de cruce SBX y mutación polinomial. Los índices de distribución para estos operadores son $\eta_c = 20$ para el cruce $\eta_m = 20$ para la mutación. La probabilidad de cruce es $p_c = 0,9$ y la de realizar mutación $p_m = 1/n$, siendo n el número de variables de decisión (longitud de solución). NSGA-II y ssNSGA-II aplican selección por torneo binario, mientras que SMS-EMOA utiliza selección aleatoria. En el caso de GDE3 (en su variante *rand/1/bin*), los dos parámetros principales, es decir, el factor de escala μ y la probabilidad de cruce C_r se inicializaron a 0,5. Para MOEA/D, la constante de mutación μ es 0,5 y la probabilidad de cruce $C_r = 1,0$. Este algoritmo también utiliza el operador de mutación polinomial con valores $\eta_m = 20$ y $p_m = 1/n$. Por último, para SMPSO hemos utilizado los coeficientes de aceleración con valores 1,5, inercia $W = 0,9$ y mutación polinomial aplicada con probabilidad del 6 % y valores de distribución iguales que en los algoritmos anteriores.

[2]URL: http://research.cs.wisc.edu/htcondor/

B. Análisis Comparativo

En la Tabla II podemos observar la mediana y rango intercuartílico de los indicadores de calidad I_{HV} y $I_{\epsilon+}$ referentes a las soluciones obtenidas, por cada una de las once instancias moleculares y los seis algoritmos evaluados: NSGA-II, ssNSGA-II, SMPSO, GDE3, MOEA/D y SMS-EMOA. Para el caso de I_{HV} los valores más altos son los mejores, mientras que para $I_{\epsilon+}$ los valores más bajos son los de mejor calidad. De acuerdo con estos resultados, SMPSO consigue mejores resultados en términos de I_{HV} para 7 de las 11 instancias moleculares utilizadas, seguido por GDE3 que obtiene el segundo mejor resultado también respecto al Hipervolumen, indicando un buen comportamiento de estos algoritmos en cuando a la convergencia y diversidad. Estos resultados se repiten para el indicador $I_{\epsilon+}$ ya que, como se puede ver en la Tabla II, SMPSO obtiene por lo general las mejores medianas de las distribuciones.

No obstante, para dotar estos resultados con soporte estadístico, hemos aplicado una serie de tests estadísticos no-paramétricos, ya que para varias distribuciones resultantes no se cumplieron las condiciones de normalidad y homoscedasticidad [21]. Por tanto, nos centramos en las distribuciones completas para nuestros análisis y comparaciones respecto a las dos métricas de calidad estudiadas. En concre-

TABLA III

RANKING DE FRIEDMAN (AVERAGE) Y p-VALORES AJUSTADOS DE HOLM ($\alpha = 0{,}05$) DE LOS ALGORITMOS COMPARADOS (SMPSO, GDE3, MOEA/D, SMS-EMOA, SSNSGA-II, Y NSGA-II) PARA LAS 11 INSTANCIAS MOLECULARES. EL SÍMBOLO * ESPECIFICA EL ALGORITMO DE CONTROL. LA FILA DE ABAJO MUESTRA EL RANKING DE POSICIONES RESPECTO A I_{HV} Y $I_{\epsilon+}$

Hipervolumen (HV)			Epsilon ($I_{\epsilon+}$)		
Algoritmo	Fri_{Rank}	$Holm_{Ap}$	Algoritmo	Fri_{Rank}	$Holm_{Ap}$
***GDE3**	**1,81**	-	***SMPSO**	**1,45**	-
SMPSO	2,18	6,48e-01	MOEA/D	2,27	3,05e-01
MOEA/D	2,63	6,10e-01	GDE3	2,72	2,21e-01
SMS-EMOA	4,50	2,32e-03	NSGA-II	4,50	4,04e-04
NSGA-II	4,63	1,64e-03	SMS-EMOA	4,63	2,65e-04
ssNSGA-II	5,22	9,62e-05	ssNSGA-II	5,40	3,57e-05
SMPSO (2+1 = 3); GDE3 (1+3 = 4); MOEA/D (3+2 = 5) NSGA-II (5+4 = 9); SMS-EMOA (6+5 = 11); ssNSGA-II (6+6 = 12)					

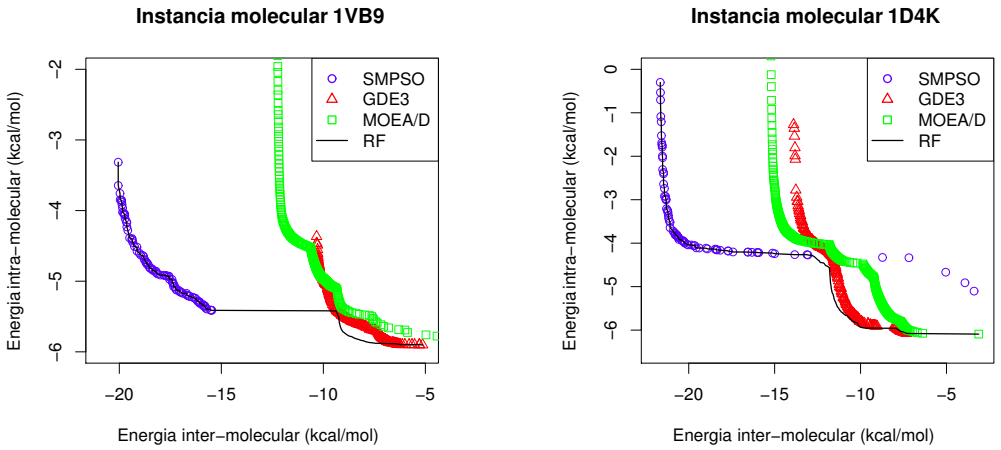

Fig. 2. Ejemplos de frentes obtenidos para las instancias 1VB9 y 1D4K a partir de los algoritmos SMPSO, GDE3 y MOEAD

to, hemos aplicado los tests de ranking de Friedman y multi-comparación post-hoc de Holm [21] a las distribuciones resultantes para conocer qué algoritmos son estadísticamente peores que el algoritmo de control (aquel de mejor ranking).

Tal y como muestra la Tabla III, el algoritmo con mejor ranking (Fri_{Rank}) para el indicador de HV es GDE3 con 1,81, seguido de SMPSO, MOEA/D, SMS-EMOA, NSGA-II y ssNSGA-II. Por tanto, GDE3 es utilizado a partir de ahora como algoritmo de control respecto a HV en el test post-hoc de Holm, siendo de este modo comparado con respecto al resto de algoritmos. Cabe destacar que, a pesar de obtener SMPSO el mayor número de mejores medianas para HV (en Tabla II), tras la aplicación del test de Friedman quedó en segundo lugar después de GDE3, lo cual arroja luz sobre la calidad de las distribuciones para todas las instancias en conjunto. Además, los p-valores ajustados ($Holm_{Ap}$ en Tabla III) resultantes son para los tres últimos algoritmos: SMS-EMOA, NSGA-II y ssNSGA-II, inferiores al nivel de significancia establecido ($\alpha = 0,05$), lo que se interpreta como que GDE3 es estadísticamente mejor que estos algoritmos. En el caso de $I_{\epsilon+}$, SMPSO obtiene mejor ranking que MOEA/D y GDE3, aunque sin diferencia significativa para estos casos. Por el contrario, SMPSO resultó con soluciones estadísticamente mejores que NSGA-II, SMS-EMOA y ssNSGA-II.

Como resumen, sumando las posiciones de ranking en última fila de la Tabla III, podemos observar cómo SMPSO obtiene el mejor balance para los dos indicadores de calidad estudiados. Además, este algoritmo ha generado resultados estadísticamente mejores que NSGA-II, SMS-EMOA y ssNSGA-II. El segundo mejor rendimiento lo obtuvo GDE3, seguido por MOEA/D.

Estos resultados pueden ser visualizados mediante los dos ejemplos mostrados en Fig. 2, que muestran los frentes con mejor hipervolumen obtenidos por SMPSO, GDE3 y MOEA/D para las instancias 1VB9 y 1D4K, respecto al frente de referencia (curva RF). En esta figura podemos observar cómo SMPSO siempre obtiene soluciones en regiones del frente de referencia donde GDE3 y MOEA/D no consiguen converger. Sin embargo, estos dos últimos algoritmos muestran una buena distribución de soluciones en sus frentes de Pareto resultantes, aunque con soluciones que convergen a una región limitada del frente de referencia. En este sentido, una interesante observación es que para todas las instancias moleculares estudiadas, SMPSO converge sobre la región sesgada hacia el objetivo de energía inter-molecular (zona izquierda de RF), mientras que GDE3 y MOEA/D generan sus frentes de soluciones no-dominadas en la zona opuesta, es decir, orientadas a mejorar la energía intra-molecular (zona derecha de RF).

TABLA IV
MEJORES VALORES DE ENERGÍA LIBRE DE UNIÓN DE LAS SOLUCIONES DE SMPSO Y LGA (MONO-OBJETIVO)

Instancia/Algoritmo	SMPSO	LGA
1AJV	-12,56	-7,26
1AJX	-11,22	-6,20
1BV9	-17,07	-7,62
1D4K	-18,67	-11,25
1G2K	-12,83	-7,19
1HIV	-13,89	-9,06
1HPX	-10,34	-5,70
1HTF	-6,83	-7,30
1HTG	-31,66	-31,79
1HVH	-15,59	-9,39
2UPJ	-10,91	-5,90

Por tanto, podemos constatar que las diferentes estrategias de aprendizaje inducidas por SMPSO y GDE3 guían el proceso de optimización de estas técnicas hacia regiones diferentes del espacio de búsqueda del problema, generando así soluciones en partes complementarias del frente de referencia.

C. Comparativa con enfoques multi/mono-objetivo

A la hora de analizar las ventajas de utilizar la formulación multi-objetivo para encarar el docking molecular, una práctica interesante consiste en comparar las soluciones obtenidas con aquellas resultantes de la técnica de referencia mono-objetivo, LGA, provista en AutoDock 4.2. Para esto, hemos recalculado los valores de fitness de las soluciones obtenidas por SMPSO mediante la Ecuación 1.

La Tabla IV contiene la comparativa mono-objetivo de las mejores soluciones obtenidas por SMPSO y LGA para todas las instancias. En general, se puede observar que SMPSO obtiene mejores conformaciones moleculares que LGA para todas las instancias excepto para 1HTF y 1HTG, aunque con valores de energía muy cercanos en estos dos casos. Además, cabe destacar que la estrategia multi-objetivo, en este caso SMPSO, es de propósito general, siendo aún así capaz de proveer al experto con más y mejores soluciones que la propuesta mono-objetivo LGA, la cual esta diseñada con operadores específicos para el docking molecular.

La gráfica en Fig. 3 muestra el frente de mejor HV obtenido por SMPSO para el compuesto molecular 1AJV. Esta instancia implica un inhibidor de urea cíclico y macromolécula HIV-proteasa, un problema complejo dadas las características de tamaño del ligando. La mejor solución obtenida por LGA se presenta como un punto en la línea discontinua vertical calculado mediante la suma de las dos energías, E_{inter} y E_{intra}. El frente de soluciones a la izquierda de la línea vertical domina la mejor solución de LGA, mientras que aquellas soluciones a la derecha de la línea discontinua tiene mejores valores únicamente respecto al objetivo E_{intra}. En este sentido, la inclinación en la selección de un tipo de energía o del otro depende del experto. Por ejemplo, un biomédico puede estar interesado por una solución con baja E_{intra} y conformación de inhibidor estable, o por el contrario en una solución con menor E_{inter} pero más estabilidad en el complejo molecular completo.

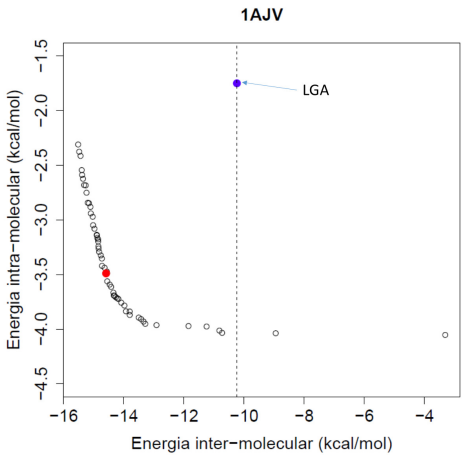

Fig. 3. Frente de soluciones SMPSO vs. solución LGA

En definitiva, desde la perspectiva biológica y considerando el enfoque mono-objetivo, las soluciones de SMPSO son mejores (más estables) que las de LGA. La figura Fig. 4 muestra los complejos del inhibidor acoplado a la macromolécula HIV-proteasa del compuesto 1AJV, obtenidos mediante conformaciones de SMPSO y LGA. El primer complejo (obtenido por SMPSO) es más estable que el segundo (obtenido por LGA) debido a las diferentes energías de unión resultantes: -11,57 kcal/mol y -7,26 kcal/mol, respectivamente. Como se ilustra en Fig. 4, ambos inhibidores (ligandos) se acoplan en el sitio activo de la HIV proteasa, aunque mostrando el de SMPSO una mejor "postura" interna en la hendidura de la macromolécula HIV-proteasa.

V. CONCLUSIONES

La principal motivación de este trabajo es la de realizar un estudio comparativo entre algoritmos multi-objetivo del estado del arte para la resolución del docking molecular. Se han comparado seis algoritmos: NSGA-II, ssNSGA-II, SMPSO, GDE3, MOEA/D y SMS-EMOA, sobre un benchmark de 11 complejos proteína-ligando de alta flexibilidad.

A partir de los resultados obtenidos extraemos las siguientes conclusiones:

(1) Mediante el enfoque multi-objetivo es posible obtener un conjunto de conformaciones moleculares de gran estabilidad energética y que pueden ser seleccionados de acuerdo al peso intra/inter molecular que más interese de cara a un estudio biológico.

(2) SMPSO obtiene en general el mejor rendimiento en convergencia y diversidad, para las instancias estudiadas.

(3) GDE3 y MOEA/D también obtienen resultados satisfactorios para las métricas utilizadas.

(4) Para las instancias estudiadas, SMPSO obtiene sus mejores soluciones sobre el objetivo de energía

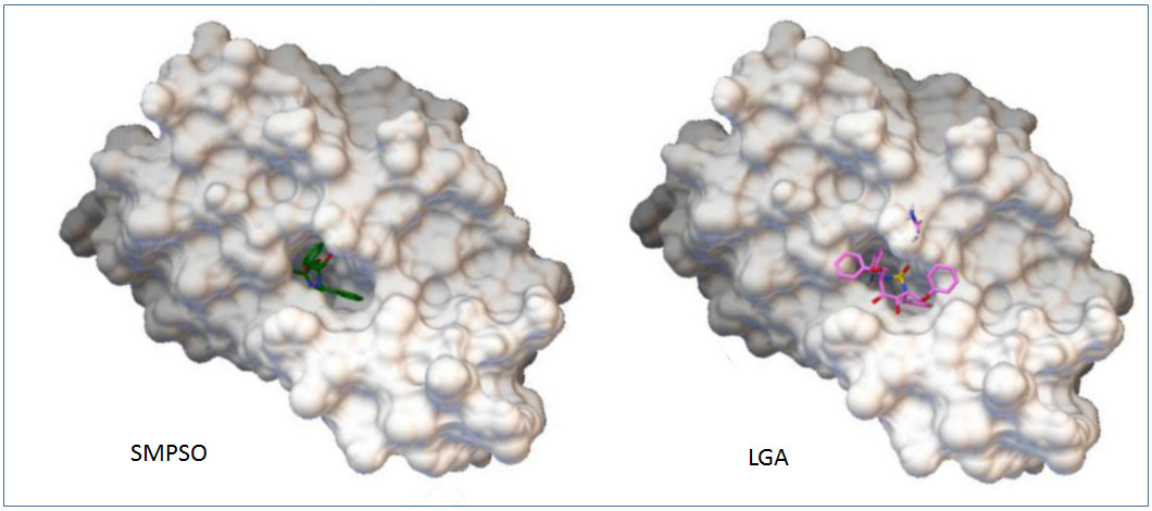

Fig. 4. Conformaciones moleculares obtenidas por SMPSO y LGA para el compuesto 1AJV

inter-molecular, mientras que el comportamiento de GDE3 y MOEA/D está más sesgado hacia la energía intra-molecular.

(5) Respecto al enfoque mono-objetivo, SMPSO encuentra en la mayoría de los casos mejores soluciones que LGA. Este resultado es destacable ya que SMPSO está diseñado con propósito general, mientras que LGA ha sido específicamente adaptado al docking molecular.

Una futura extensión de este trabajo consiste en el diseño de una estrategia híbrida mediante la combinación de los procedimientos de búsqueda de SMPSO y GDE3, con el objetivo de generar frentes de soluciones que cubran todas las regiones del frente de Pareto de referencia. Además, la utilización de un benchmark mayor de instancias moleculares enriquecerá nuestros futuros estudios también desde el punto de vista del impacto biológico.

Agradecimientos

Este trabajo está financiado por las iniciativas TIN2014-58304 (Ministerio de Economía y Competitividad), TIN2011-25840 (Ministerio de Ciencia e Innovación), P11-TIC-7529 y P12-TIC-1519 (Plan Andaluz de Investigación, Desarrollo e Innovación)

Referencias

[1] Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J.: AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Com. Chem. **30** (2009) 2785–2791

[2] Oduguwa, A., Tiwari, A., Fiorentino, S., Roy, R.: Multi-objective optimisation of the protein-ligand docking problem in drug discovery. In Proc. of the 8th Conf. on Genetic and Evolutionary Computation. (2006) 1793–1800

[3] Grosdidier, A., Zoete, V., Michielin, O.: EADock: Docking of small molecules into protein active sites with a multiobjective evolutionary optimization. Proteins **67** (2007) 1010–1025

[4] Janson, S., Merkle, D., Middendorf, M.: Molecular docking with multi-objective particle swarm optimization. Appl. Soft Comput. **8** (2008) 666–675

[5] Boisson, J.C., Jourdan, L., Talbi, E., Horvath, D.: Parallel multi-objective algorithms for the molecular docking problem. In IEEE Symposium on CIBCB (2008) 187–194

[6] Sandoval-Perez, A., Becerra, D., Vanegas, D., Restrepo-Montoya, D., Niño, F.: A multi-objective optimization energy approach to predict the ligand conformation in a docking process. In: EuroGP. (2013) 181–192

[7] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Tran. on Evo. Comp. **6** (2002) 182–197

[8] Durillo, J.J., Nebro, A.J., Luna, F., Alba, E.: On the effect of the steady-state selection scheme in multi-objective genetic algorithms. In: Evolutionary Multi-Criterion Optimization. Volume 5467 of LNCS., Springer Berlin Heidelberg (2009) 183–197

[9] Nebro, A.J., Durillo, J.J., Garcia-Nieto, J., Coello Coello, C.A., Luna, F., Alba, E.: SMPSO: A new PSO-based metaheuristic for multi-objective optimization. In: IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making. (2009) 66–73

[10] Kukkonen, S., Lampinen, J.: GDE3: the third evolution step of generalized differential evolution. In: Evolutionary Computation, 2005. The 2005 IEEE Congress on. Volume 1. (2005) 443–450

[11] Zhang, Q., Li, H.: MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. IEEE T. Evolut. Comput. **11** (2007) 712–731

[12] Beume, N., Naujoks, B., Emmerich, M.: SMS-EMOA: Multiobjective selection based on dominated hypervolume. Eur. J. Oper. Res. **181** (2007) 1653–1669

[13] Coello, C., Lamont, G.B., van Veldhuizen, D.A.: Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley & Sons, Inc. 2nd Ed., NY, USA (2007)

[14] Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley & Sons, Inc. USA (2001)

[15] López-Camacho, E., García-Godoy, M.J., Nebro, A.J., Aldana-Montes, J.F.: jMetalCpp: optimizing molecular docking problems with a C++ metaheuristic framework. Bioinformatics **30** (2014) 437–438

[16] Durillo, J.J., Nebro, A.J.: jmetal: A java framework for multi-objective optimization. Advances in Engineering Software **42** (2011) 760–771

[17] Dallakyan, S., Pique, M.E., Huey, R.: Autodock version 4.2. in url http://autodock.scripps.edu/

[18] Huey, R., Morris, G.M., Olson, A.J., Goodsell, D.S.: A semiempirical free energy force field with charge-based desolvation. J. Comput. Chem. **28** (2007) 1145–1152

[19] Li, H., Zhang, Q.: Multiobjective Optimization Problems With Complicated Pareto Sets, MOEA/D and NSGA-II. IEEE Tran. on Evo. Comp. **13** (2009) 229–242

[20] Norgan, A.P., Coffman, P.K., Kocher, J.P.A., Katzmann, D.J., Sosa, C.P.: Multilevel Parallelization of AutoDock 4.2. J. Cheminform. **3** (2011) 12

[21] Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall (2007)