Un Equipo Paralelo Multi-Núcleo para Descubrir Motifs en Secuencias de ADN

David L. González-Álvarez [1], Miguel A. Vega-Rodríguez [1], Juan A. Gómez-Pulido [1] y Juan M. Sánchez-Pérez [1]

Resumen—
Este trabajo propone la utilización de un equipo multi-núcleo que explota las características de varios algoritmos cuando resolvemos uno de los problemas biológicos más importantes, el Problema del Descubrimiento de Motifs (PDM). Normalmente, para resolver un determinado problema de optimización tenemos que abordar escenarios muy distintos. Este problema complica el diseño o aplicación de una única técnica o algoritmo que sea capaz de resolver adecuadamente todos estos escenarios. Una posible solución a este problema es combinar de algún modo el comportamiento de estas técnicas o algoritmos utilizando paralelismo en un equipo. En este trabajo definimos una metodología paralela que nos permite combinar las funcionalidades de cuatro algoritmos evolutivos multiobjetivo utilizando OpenMP para resolver diferentes escenarios en el descubrimiento de motifs. Las conclusiones obtenidas muestran como nuestra propuesta multi-núcleo es una técnica flexible y adecuada para este fin, además de representar una importante herramienta para el descubrimiento de motifs.

Palabras clave—
Multi-Núcleo, OpenMP, algoritmos evolutivos, optimización multiobjetivo, ADN.

I. Introducción

Buscar patrones de ADN (PDM - Problema del Descubrimiento de Motifs) es uno de los problemas más importantes de la biología dedicado a descubrir Puntos de Unión de Factores de Transcripción (Transcription Factor Binding Sites, TFBS) en secuencias de ADN [1]. Una de las técnicas que más éxito ha tenido a la hora de resolver el PDM es la optimización multiobjetivo [2]. Debido a esto, en este trabajo hemos decidido abordar este problema como un problema de optimización multiobjetivo (MOP) [3], maximizando simultáneamente tres objetivos: el tamaño, el soporte y la similaridad de los motifs.

En los últimos años, hemos sido capaces de descifrar información genética de una gran cantidad de seres vivos. Este hecho ha producido un incremento considerable de los recursos computacionales necesarios para su procesamiento. El PDM es un problema de optimización de tipo NP-completo, por ello la aplicación de algoritmos o técnicas exactas para resolver instancias reales de este problema no es factible. Por otro lado, abordar este problema utilizando técnicas basadas en búsqueda o metaheurísticas es, sino obligatorio, una de las mejores formas de descubrir motifs de ADN de calidad [4]. Sin embargo, la resolución del PDM se complica cuando debemos diseñar metaheurísticas que resuelvan muchos escenarios, o instancias con propiedades muy distintas. En estos casos, muchas veces, surge la necesidad de implementar más de un algoritmo. Una posible solución a este problema es la aplicación de técnicas paralelas que nos permitan combinar las propiedades de diferentes algoritmos o técnicas en una única heurística. Esta ha sido la motivación principal de este trabajo, diseñar una técnica paralela capaz de combinar el funcionamiento de diferentes técnicas para así, poder resolver escenarios muy diferentes del PDM. Para ello, en este trabajo hemos diseñado un equipo paralelo multi-núcleo de algoritmos evolutivos multiobjetivo utilizando OpenMP [5]. Esta heurística nos permite combinar el funcionamiento de cuatro metaheurísticas, facilitando la resolución de un conjunto de 12 instancias con características muy distintas. Tras llevar a cabo numerosos experimentos, hemos observado como los mejores resultados son obtenidos por los algoritmos basados en Evolución Diferencial (Differential Evolution - DE), Colonia Artificial de Abejas (Artificial Bee Colony - ABC) y Algoritmo de las Luciérnagas (Firefly Algorithm - FA). Por esta razón, hemos seleccionado estos tres algoritmos para formar parte de nuestro equipo paralelo multi-núcleo. Como estamos resolviendo un MOP debemos adaptar también el funcionamiento de nuestros algoritmos a este contexto multiobjetivo. Por lo que hemos tenido que definir tres versiones multiobjetivo de los algoritmos anteriormente mencionados que hemos denominado: Evolución Diferencial con Torneo de Pareto (Differential Evolution with Pareto Tournament - DEPT), Algoritmo Multiobjetivo de las Luciérnagas (Multiobjective Firefly Algorithm - MO-FA) y Colonia Artificial de Abejas Multiobjetivo (Multiobjective Artificial Bee Colony - MOABC). La cuarta metaheurística seleccionada es la conocida como algoritmo SPEA2 [6], un importante algoritmo evolutivo multiobjetivo que ha obtenido muy buenos resultados en muchos problemas de optimización presentes en la literatura.

Cuando optimizamos un problema biológico es también muy importante analizar la relevancia biológica de los resultados obtenidos. Por esta razón, hemos comparado los motifs descubiertos por nuestra propuesta paralela multi-núcleo con las soluciones obtenidos por catorce métodos biológicos bien conocidos.

Este trabajo se organiza del siguiente modo. La Sección II describe el problema biológico abordado en este trabajo, el PDM. En la Sección III explicamos el funcionamiento de los algoritmos evolu-

[1]Dpto. Tecnología de Computadores y Comunicaciones, Grupo de Investigación ARCO, Universidad de Extremadura, e-mails: {dlga, mavega, jangomez, sanperez}@unex.es

tivos multiobjetivo que componen el equipo paralelo multi-núcleo y detallamos el funcionamiento de la heurística paralela. En la Sección IV incluimos los resultados experimentales. La Sección V compara las predicciones realizadas por nuestra propuesta con las realizadas por otros catorce métodos biológicos bien conocidos. Finalmente, la Sección VI presenta las conclusiones y posibles líneas futuras.

II. Problema del Descubrimiento de Motifs

Hemos definido tres objetivos en conflicto que deben maximizarse simultaneamente. Dado un conjunto de secuencias $S = \{S_i | i = 1, 2, ..., D\}$ de nucleótidos definidos en el alfabeto $B = \{A, C, G, T\}$ definimos:

- Una secuencia de nucleótidos como $S_i = \{S_i^j | j = 1, 2, ..., w_i\}$, donde w_i es el tamaño de la secuencia.
- El conjunto de todas las subsecuencias contenidas en S como $\{s_i^{j_i} | i = 1, 2, ..., D, j_i = 1, 2, ..., w_i - l + 1\}$, donde j_i es el punto de unión de un posible motif candidato s_i^j de la secuencia S_i, y l es el tamaño del motif, primer objetivo a maximizar.
- El motif consenso como la cadena resultante de la union de los nucleótidos dominantes en cada posición del motif.

El valor del segundo objetivo, el soporte, se calcula comparando cada motif candidato con el motif consenso. Solo aquellas secuencias que presenten motif candidatos con una cierta calidad con respecto al motif candidato (>=50% de semejanza) se tendrán en cuenta cuando formemos el motif final. El número final de secuencias utilizadas para componer la solución final es indicado por el soporte.

Finalmente, para obtener el valor de la similaridad tenemos que definir:

- La Matriz de Indicadores por Posición (Position Indicator Matrix - PIM) $A = \{A_i | i = 1, 2, ..., D\}$ del motif, donde $A_i = \{A_i^j | j = 1, 2, ..., w_i\}$ es el vector fila de indicadores de la secuencia S_i. A_i^j es 1 si la posición j en S_i es un punto de unión, y 0 en caso contrario.
- La Matriz de Cuentas por Posición (Position Count Matrix - PCM) $N(A)$ con el número de nucleótidos en cada posición de los motifs candidatos (A) $N(A) = \{N(A)^1, N(A)^2, ..., N(A)^l\}$ donde $N(A)^j = \{N(A)_b^j | b \in B\}$ y $N(A)_b^j = |\{S(A)_i^j | S(A)_i^j = b\}|$.
- La Matriz de Frecuencias por Posición (Position Frequency Matrix - PFM) como $\widehat{N} = \frac{N(A)}{|A|}$ generado a partir del nucleótido dominante en cada una de las posiciones.

El valor final de la similaridad se calcula a través de la media aritmética de todos los valores dominantes en cada posición del PFM, tal y como indica la siguiente expresión:

$$Similaridad(Motif) = \frac{\sum_{i=1}^{l} max_b\{f(b,i)\}}{l} \quad (1)$$

donde $f(b, i)$ es el valor obtenido por el nucleótido b en la columna i de la PFM y $max_b\{f(b, i)\}$ es el valor obtenido por el nucleótido dominante de i.

Para guiar la búsqueda de soluciones hacia motifs biológicamente relevantes, hemos incorporado varias restricciones que deben ser satisfechas por todas las soluciones:

- El tamaño de los motifs está restringido en el rango [7,64].
- Establecemos un valor mínimo de soporte igual a 2 en los conjuntos de datos formados por 4 o menos secuencias, y de 3 en las demás (más de 4 secuencias).
- Aplicamos el concepto de complejidad [7] junto con las mejoras propuestas en [8] como una restricción biológica. La complejidad de los motifs candidatos debe ser considerada para evitar la predicción de soluciones poco complejas aplicando la siguiente expresión:

$$Complexity = \log_N \frac{l!}{\prod (n_i)!} \quad (2)$$

  donde $N = 4$ en secuencias de ADN, l es el tamaño del motif y n_i es el número de nucleótidos del tipo $i \in \{A, C, G, T\}$.

III. Equipo Paralelo Multi-Núcleo

Este trabajo presenta un equipo paralelo multi-núcleo compuesto por cuatro algoritmos evolutivos multiobjetivo y controlado por una heurística paralela. El objetivo de esta heurística paralela no es resolver el problema sino gestionar las soluciones obtenidas por cada método. De este modo, cada algoritmo puede aprovechar los descubrimientos realizados por los demás y así evitar o superar posibles situaciones de estancamiento. En la Figura 1 describimos el funcionamiento general del equipo paralelo multi-núcleo. Podemos ver como la heurística paralela gestiona las salidas proporcionadas por los algoritmos, limpiando y organizando las soluciones descubiertas por cada uno de ellos para generar la población correspondiente para la siguiente sincronización. Este diseño está basado en OpenMP (ver Figura 1). Primero, tras generar la población inicial, la heurística paralela abre un bloque paralelo utilizando la directiva '#pragma omp parallel for' y ejecuta los cuatro algoritmos de forma paralela a través de la directiva '#pragma omp task'. Es importante tener en cuenta que, antes la apertura del bloque paralelo, debemos establecer el número de hilos disponibles a través de la función 'omp_set_num_threads(x)'. Además, para evitar que cada tarea se ejecute 'x' veces (cuatro veces en nuestro caso, una por algoritmo), tenemos que incluir cada tarea paralela en un bloque '#pragma omp single nowait'. Cuando todos los algoritmos finalizan su ejecución ('#pragma omp taskwait'), la heurística

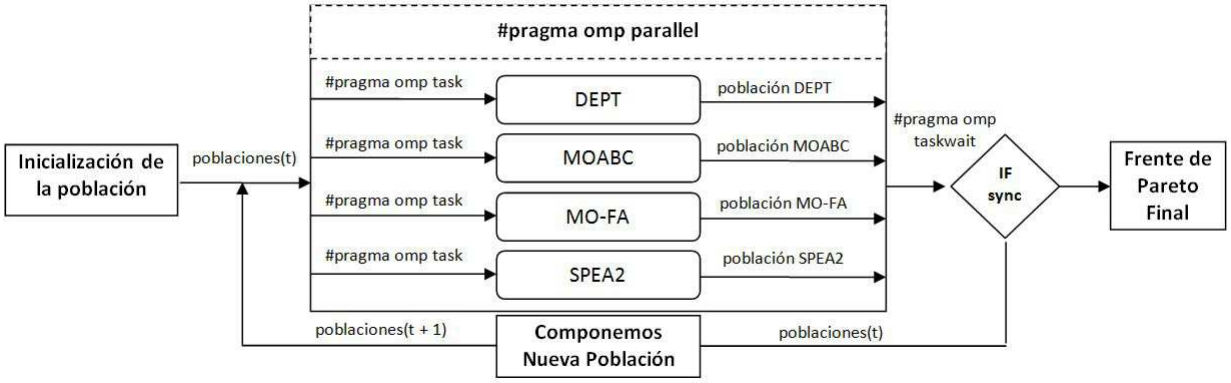

Fig. 1. Diseño del equipo paralelo multi-núcleo.

paralela procesa los frentes de Pareto obtenidos por cada algoritmo para formar la población que se utilizará en la siguiente sincronización. Este proceso se repite hasta que todas las sincronizaciones se hayan realizado. Al final de la ejecución, la heurística paralela retorna el conjunto de soluciones no-dominadas obtenido.

Las especificaciones del problema y las instancias son gestionadas por cada algoritmo, especialmente configurados para resolver el PDM. Las configuraciones utilizadas se detallan en las referencias correspondientes a cada algoritmo. Entre todos estos parámetros destacamos el tamaño de la población, el cual se ha establecido en 200 individuos en todos los algoritmos (por ello el equipo paralelo tiene una población total de 800 soluciones). Tras realizar diferentes pruebas, hemos seleccionado las metaheurísticas que mejores resultados obtenían para este conjunto de instancias: Evolución Diferencial con Torneo de Pareto (DEPT, [9]), Colonia Artificial de Abejas Multiobjetivo (MOABC, [10]), Algoritmo Multiobjetivo de las Luciérnagas (MO-FA) y el algoritmo SPEA2 [6]. El equipo paralelo multi-núcleo combina las propiedades de estos cuatro algoritmos y su funcionamiento se detalla a continuación. Primeramente, el equipo paralelo genera la población inicial (formada por cuatro mini-poblaciones). Tras esto, tal y como ya se mencionó anteriormente, se lanzan las tareas paralelas que ejecutan cada uno de los algoritmos. Cuando los algoritmos finalicen sus ejecuciones se inicia la función más importante: la sincronización de los algoritmos. El equipo paralela multi-núcleo almacena las soluciones generadas por las metaheurísticas y construye una nueva población global compuesta por las 800 soluciones obtenidas (200 por individuo). A continuación, el equipo limpia y organiza la nueva población aplicando el concepto de dominancia [3], obteniéndose un conjunto de soluciones no-dominadas. Es posible que, tras aplicar las funciones de limpieza y organización obtengamos una población menor que 200 (necesitamos 200 individuos por algoritmo porque así lo hemos definido anteriormente). Si esto ocurre, el resto de soluciones de cada población serán las mejores soluciones de cada una de ellas. De este modo cada algoritmo tendrá una población formada por las mejores soluciones obtenidas por los cuatro algoritmos y las mejores soluciones descubiertas por él mismo. Haciendo esto, cada algoritmo no pierde su esencia y puede continuar su proceso evolutivo desde el punto en el que lo dejo. Por último destacar la diferencia existente entre población y frente de Pareto obtenido. Mientras que cada algoritmo posee una población única, todos los algoritmos comparten el frente de Pareto, que se define como un archivo de soluciones compartido por todos los algoritmos.

TABLA I
PROPIEDADES DE LAS INSTANCIAS.

Instancia	Secuencias	#Nucleótidos	Total Nucl.	Tiempo (sg.)
mus09r	2	500	1000	20-80
mus01r	3	500	1500	20-80
mus12m	3	500	1500	20-80
mus06g	3	500	1500	20-80
mus08m	3	1500	4500	20-80
mus05r	4	500	2000	20-80
mus07g	4	1500	6000	20-80
mus03g	5	500	2500	20-80
mus04m	7	1000	7000	20-80
mus02r	9	1000	9000	20-80
mus11m	12	500	6000	30-120
mus10g	13	1000	13000	30-120

IV. Experimentos y Resultados

Todos los experimentos se han realizado sobre un Intel Xeon 2.33 GHz con 8 GB de RAM y 8 núcleos (solo utilizamos 4 de estos núcleos en nuestros experimentos paralelos) sin utilizar opciones de compilación. Además hemos seleccionado doce instancias pertenecientes al ratón (base de datos TRANSFAC [11]) para realizar experimentos. En la Tabla I incluimos la información más relevante de estas instancias: número de secuencias, tamaño de cada secuencia, número total de nucleótidos y tiempos de ejecución fijados. Las instancias las presentamos ordenadas por complejidad, es decir, las instancias más sencillas se incluyen en la parte superior de la tabla y las más complejas en la parte inferior. Dada la naturaleza estocástica de las metaheurísticas y con el objetivo de validar estadísticamente los resultados obtenidos hemos realizado 30 ejecuciones independientes en cada experimento. Para comparar los resultados obtenidos hemos aplicado la métrica multiob-

TABLA II

Comparativa entre los cuatro algoritmos evolutivos multiobjetivo (DEPT, MOABC, MO-FA y SPEA2) ejecutados durante 80 o 120 segundos (hipervolumen mediano y rango intercuartílico).

	DEPT	MO-FA	MOABC	SPEA2	
Instancia	$\bar{HV}_{IQR}$	$\bar{HV}_{IQR}$	$\bar{HV}_{IQR}$	$\bar{HV}_{IQR}$	
mus09r	90,98% 0,32%	90,52% 0,58%	93,52% 0,15%	**93,54%** 0,13%	+
mus01r	81,18% 1,65%	84,40% 1,77%	**87,80%** 0,29%	87,74% 0,33%	- †
mus12m	80,58% 1,69%	82,72% 1,26%	86,67% 0,36%	**86,82%** 0,24%	+
mus06g	80,00% 1,25%	81,79% 0,80%	**85,14%** 0,55%	85,02% 0,29%	- †
mus08m	80,72% 1,66%	83,12% 0,85%	**86,78%** 0,46%	86,48% 0,56%	+
mus05r	78,15% 1,15%	80,32% 0,78%	**82,89%** 0,35%	82,65% 0,33%	+
mus07g	80,89% 2,92%	82,92% 1,27%	**90,98%** 0,66%	90,87% 0,24%	- ‡
mus03g	78,17% 1,39%	79,95% 0,80%	**81,07%** 0,59%	79,12% 0,52%	+
mus04m	75,11% 1,63%	**75,44%** 1,02%	73,67% 1,98%	69,89% 1,31%	- ‡
mus02r	**72,26%** 0,83%	71,70% 1,03%	67,02% 1,32%	63,39% 1,35%	- ‡
mus11m	62,12% 5,13%	**64,14%** 1,52%	55,97% 1,86%	48,17% 1,17%	+
mus10g	59,79% 5,04%	**62,70%** 2,09%	53,80% 2,60%	45,34% 0,57%	+

- † no diferencias significantes entre MOABC y SPEA2. El resto de algoritmos presentan diferencias significativas.

- ‡ no diferencias significantes entre DEPT y MO-FA. El resto de algoritmos presentan diferencias significativas.

+ diferencias significantes entre todos los algoritmos.

TABLA III

Ajuste del número de sincronizaciones del equipo paralelo multi-núcleo (hipervolumen mediano y rango intercuartílico).

	1 sincr.	4 sincr.	9 sincr.
Instance	$\bar{HV}_{IQR}$	$\bar{HV}_{IQR}$	$\bar{HV}_{IQR}$
mus09r	93,61% 0,09%	93,59% 0,11%	**93,62%** 0,07%
mus01r	91,35% 0,28%	91,37% 0,26%	**91,40%** 0,29%
mus12m	90,50% 0,35%	**90,53%** 0,34%	90,44% 0,22%
mus06g	89,50% 0,32%	**89,67%** 0,29%	89,61% 0,43%
mus08m	90,58% 0,15%	**90,71%** 0,28%	90,46% 0,29%
mus05r	84,09% 2,73%	**86,57%** 2,59%	85,41% 2,72%
mus07g	92,26% 0,42%	**92,66%** 0,46%	92,58% 0,28%
mus03g	82,61% 0,56%	**83,17%** 0,40%	82,86% 0,76%
mus04m	77,40% 0,74%	**78,25%** 1,22%	77,72% 2,23%
mus02r	74,08% 0,99%	**76,07%** 1,47%	75,20% 1,99%
mus11m	65,27% 1,63%	**67,20%** 1,74%	67,10% 1,43%
mus10g	61,93% 1,26%	63,27% 1,12%	**64,42%** 1,52%

TABLA IV

Comparativa entre el mejor algoritmo y el equipo paralelo multi-núcleo configurado (hipervolumen mediano y rango intercuartílico).

	Mejor resultado		Equipo multi-núcleo	
Instancia	Mejor algoritmo	$\bar{HV}$	$\bar{HV}_{IQR}$	
mus09r	SPEA2	93,54%	**93,59%** 0,11%	+
mus01r	MOABC	87,80%	**91,37%** 0,26%	+
mus12m	SPEA2	86,82%	**90,53%** 0,34%	+
mus06g	MOABC	85,14%	**89,67%** 0,29%	+
mus08m	MOABC	86,78%	**90,71%** 0,28%	+
mus05r	MOABC	82,89%	**86,57%** 2,59%	+
mus07g	MOABC	90,98%	**92,66%** 0,46%	+
mus03g	MOABC	81,07%	**83,17%** 0,40%	+
mus04m	MO-FA	75,44%	**78,25%** 1,22%	+
mus02r	DEPT	72,26%	**76,07%** 1,47%	+
mus11m	MO-FA	64,14%	**67,20%** 1,74%	+
mus10g	MO-FA	62,70%	**63,27%** 1,12%	- †

- † no diferencias significantes entre MO-FA y el equipo. El resto de algoritmos presentan diferencias significantes con el equipo.

+ diferencias significantes entre los algoritmos y el equipo.

jetivo del hipervolumen [12], calculando el volumen de referencia con los valores máximos de cada objetivo en cada instancia. Finalmente, la representación de los individuos es la misma en todos los algoritmos e incluye la información necesaria para formar un motif de ADN: el tamaño del motif y las posiciones iniciales de cada candidato en cada secuencia.

Una vez definida la metodología experimental podemos proceder a analizar los resultados obtenidos. Como ya dijimos en la introducción, el objetivo principal del equipo paralelo multi-núcleo definido es combinar las propiedades de cuatro algoritmos evolutivos multiobjetivo para resolver un conjunto de instancias con diferentes propiedades. En la Tabla II podemos ver como cada metaheurística es capaz de resolver instancias con ciertas características. En estos experimentos los algoritmos han sido ejecutados durante 20 o 30 segundos dependiendo de sus complejidades (ver Tabla I); pero en la Tabla II hemos incluido los resultados de obtenidos por los cuatro algoritmos ejecutados durante 80 o 120 segundos (cuatro veces más tiempo para así poder comparar de forma justa los resultados con el equipo paralelo multi-núcleo). De este modo estamos comparando un equipo paralelo multi-núcleo compuesto por cuatro algoritmos y ejecutado durante 20 o 30 segundos con cuatro algoritmos que han sido ejecutados durante 80 o 120 segundos. Además, también podemos apreciar cómo dependiendo de la instancia resuelta, el mejor resultado es obtenido por un algoritmo u otro. En las instancias sencillas (las cuatro primeras de la Tabla II) los algoritmos que mejores hipervolúmenes obtienen son SPEA2 y MOABC. En las cuatro instancias siguientes (desde la mus08m a la mus03g) el algoritmo MOABC obtiene los mejores resultados. Estas cuatro instancias pueden considerarse como de complejidad media. Finalmente, los algoritmos DEPT y MO-FA son los que mejores rendimientos presentan en las instancias más complicadas (las cuatro últimas instancias de la Tabla II). Por esta razón es importante definir una única técnica o método que sea capaz de resolver escenarios con diferentes propiedades (como por ejemplo las doce instancias presentadas) combinando el funcionamiento de diferentes algoritmos (en este caso, los cuatro algoritmos seleccionados). Los numerosos experimentos realizados se sostienen gracias al profundo estudio estadístico que hemos llevado a cabo para demostrar que las diferencias existentes entre los datos son estadísticamente relevantes. Para ello hemos aplicado la misma metodología que se aplica en [13], donde los autores analizan las distribuciones de las muestras y las homogeneidades de las vari-

anzas aplicando diferentes test. Los test exitosos se indican con '+' y los test negativos con '-'.

Tal y como explicamos en la Sección III, una de las funciones más importantes de nuestro equipo paralelo multi-núcleo es la sincronización de los algoritmos. Dado que esta función está ya definida, el único punto pendiente es averiguar cuantas sincronizaciones deben realizarse para obtener los mejores resultados. Por esta razón hemos definido un experimento de configuración que analiza esta cuestión, probando tres valores distintos: 1, 4 y 9 sincronizaciones. Hemos escogido estos tres valores porque son los divisores comunes de 20 y 30, que son los tiempos de ejecución establecidos (así realizamos una sincronización cada 10 o 15 segundos, cuatro cada 4 o 6 segundos; y nueve cada 2 o 3 segundos, respectivamente). Los resultados de este experimento se muestran en la Tabla III. Como podemos ver, en nueve de las doce instancias resueltas los mejores resultados se obtienen realizando las sincronizaciones cada 4 o 6 segundos. Por lo que hemos establecido el número de sincronizaciones en 4.

En la Tabla IV comparamos los resultados obtenidos por los cuatro algoritmos ejecutados durante los tiempos anteriormente mencionados (80 o 120 segundos) con los obtenidos por el equipo paralelo multi-núcleo (ejecutado durante 20 o 30 segundos). También incluimos los resultados del estudio estadístico aplicado sobre los datos de la Tabla II. Vemos como en la Tabla IV el equipo paralelo multi-núcleo no solo consigue los mejores resultados en todas las instancias resueltas, sino que también consigue mejores resultados que los mejores algoritmos. Además, los algoritmos logran un hipervolumen medio del 80,80% mientras que el equipo paralelo multi-núcleo consigue una media del 84%, lo que representa un incremento medio del 4%. Esto demuestra que, aunque los algoritmos han sido ejecutados durante un mayor tiempo, combinar las propiedades de diferentes técnicas favorece la evolución global del conjunto de soluciones. Por todas estas razones, podemos decir que nuestra propuesta cumple con el objetivo fijado al inicio de este trabajo: definir una técnica paralela única que combine la capacidad de varios algoritmos para resolver instancias distintas de un problema dado, en este caso el PDM.

TABLA V
COMPARATIVAS ENTRE LOS MOTIFS DESCUBIERTOS POR EL EQUIPO PARALELO MULTI-NÚCLEO (EP) Y LOS RESULTADOS OBTENIDOS POR CATORCE MÉTODOS BIOLÓGICOS BIEN CONOCIDOS.

Sensibilidad (nSn)			
Instancia	Mejor método biológico		EP
mus09r	MEME	0,292683	**0,804878**
mus01r	Improbizer	0,317647	**0,576471**
mus12m	MotifSampler	0,268966	**0,682759**
mus06g	MotifSampler	0,283582	**0,641791**
mus08m	SeSiMCMC	0,365854	**0,634146**
mus05r	MEME3	0,386364	**0,477273**
mus07g	ANN_Spec	0,040000	**0,540000**
mus03g	AlignACE	0,281690	**0,542253**
mus04m	oligodyad & YMF	0,053030	**0,314394**
mus02r	MEME	0,094828	**0,297414**
mus11m	Improbizer	**0,369668**	**0,369668**
mus10g	MEME3	**0,331839**	0,286996
Valor de Predicción Positiva (nPPV)			
Instancia	Mejor método biológico		EP
mus09r	MEME	0,266667	**0,970588**
mus01r	Improbizer	0,500000	**0,907407**
mus12m	MotifSampler	0,367925	**0,846154**
mus06g	oligodyad	0,472222	**0,632353**
mus08m	SeSiMCMC	0,103448	**0,923077**
mus05r	MEME3	0,222222	**0,971429**
mus07g	ANN_Spec	0,020942	**0,894737**
mus03g	AlignACE	0,256410	**0,869565**
mus04m	oligodyad & YMF	0,159091	**0,954023**
mus02r	MEME	0,142857	**0,556452**
mus11m	Weeder	**0,900000**	0,443182
mus10g	Weeder	0,500000	**0,583333**
Coeficiente de Rendimiento (nPC)			
Instancia	Mejor método biológico		EP
mus09r	MEME	0,162162	**0,804878**
mus01r	Improbizer	0,241071	**0,544444**
mus12m	MotifSampler	0,183962	**0,607362**
mus06g	oligodyad	0,197674	**0,467391**
mus08m	SeSiMCMC	0,087719	**0,558140**
mus05r	MEME3	0,155251	**0,394231**
mus07g	ANN_Spec	0,013937	**0,481132**
mus03g	AlignACE	0,155039	**0,416216**
mus04m	oligodyad & YMF	0,041420	**0,309702**
mus02r	MEME	0,060440	**0,240418**
mus11m	Weeder	0,227437	**0,252427**
mus10g	Weeder	**0,213450**	0,189922
Coeficiente de Correlación (nCC)			
Instancia	Mejor método biológico		EP
mus09r	MEME	0,247042	**0,893439**
mus01r	Improbizer	0,370550	**0,711097**
mus12m	MotifSampler	0,253129	**0,737755**
mus06g	oligodyad	0,324567	**0,619989**
mus08m	SeSiMCMC	0,181165	**0,733282**
mus05r	MEME3	0,236944	**0,613003**
mus07g	ANN_Spec	0,006056	**0,671724**
mus03g	AlignACE	0,222480	**0,591621**
mus04m	oligodyad & YMF	0,071890	**0,539597**
mus02r	MEME	0,097480	**0,395813**
mus11m	Weeder	**0,430395**	0,385102
mus10g	Weeder	0,346658	**0,351653**

V. Comparativas con Otros Autores

En esta sección analizamos la relevancia biológica de los motifs descubiertos por el equipo paralelo multi-núcleo configurado. Para realizar esta comparativa primero hemos calculado la relevancia biológica de las predicciones realizadas por los métodos biológicos para a continuación calcular la relevancia de las soluciones obtenidas por nuestra propuesta paralela. Los métodos biológicos son: AlignACE, ANN_Spec, Consensus, GLAM, Improbizer, MEME, MEME3, MITRA, MotifSampler, Oligo/Dyad-Analysis, QuickScore, SeSiMCMC, Weeder y YMF. Una breve descripción de cada uno se incluye en la Tabla I de [14]. Para poder comparar los resultados de todos estos métodos con los nuestros disponemos de los puntos de unión predichos por cada método en cada conjunto de datos y utilizamos cuatro indicadores biológico/estadísticos a nivel de nucleótido como son la Sensibilidad (nSn):

$$nSn = \frac{nTP}{(nTP + nFN)} \tag{3}$$

El Valor de Predicciones Positivas ($nPPV$), proporción de positivos reales correctamente predichos:

$$nPPV = \frac{nTP}{(nTP + nFP)} \tag{4}$$

El Coeficiente de Rendimiento (nPC):

$$nPC = \frac{nTP}{(nTP + nFN + nFP)} \quad (5)$$

Y finalmente, el Coeficiente de Correlación (nCC):

$$nCC = \frac{nTP \cdot nTN - nFN \cdot nFP}{\sqrt{PN \cdot NP \cdot PP \cdot NN}} \quad (6)$$

donde $PN = nTP + nFN$, $NP = nTN + nFP$, $PP = nTP + nFP$ y $NN = nTN + nFN$. Para más información sobre estos cuatro indicadores ver [14].

En la Tabla V incluimos los resultados de esta comparativa. En la primera columna mostramos las instancias procesadas y a continuación el método biológico (de entre los catorce analizados) que mejores resultados obtiene en cada indicador. La tercera columna de las tablas incluye el mejor resultado obtenido por este último método y, finalmente, en la última columna mostramos el resultado obtenido por el mejor motif no-dominado descubierto por nuestro equipo paralelo multi-núcleo. Si analizamos los resultados de la Tabla V, podemos ver como nuestra propuesta paralela es capaz de mejorar los resultados conseguidos por el mejor de los catorce métodos biológicos en la mayoría de las instancias. Además, podemos observar como muchos de los métodos están especialmente diseñados para resolver instancias con ciertas características, por ejemplo: MEME siempre consigue los mejores resultados en la instancias mus09r, MEME3 en mus05r, AlignACE en mus03g... Además, si tenemos en cuenta que estamos analizando los resultados obtenidos por catorce métodos biológicos en doce instancias distintas, podemos llegar a encontrar hasta diez métodos distintos en la Tabla V. Una de las propiedades más destacables de nuestra propuesta paralela es precisamente esta, que es capaz de descubrir motifs biológicamente relevantes en todas las instancias, independientemente de sus propiedades. Por todas estas razones, podemos afirmar que nuestro equipo paralelo multi-núcleo es flexible y adecuado para resolver diferentes escenarios del PDM, resultando ser una herramienta importante para descubrir motifs relevantes.

VI. Conclusiones y Líneas Futuras

En este trabajo presentamos un equipo paralelo multi-núcleo que combina las propiedades de cuatro metaheurísticas para resolver un problema biológico importante, el PDM. Cuando tenemos que abordar escenarios muy distintos en un problema de optimización, se hace muy difícil encontrar una única metaheurística que sea capaz de resolverlos todos de forma adecuada. Si lo que queremos es resolver todos estos escenarios deberemos aplicar diferentes técnicas o algoritmos. Para solventar este punto hemos diseñado un equipo paralelo multi-núcleo que combina el funcionamiento de cuatro algoritmos evolutivos multiobjetivo cuando resolvemos doce instancias reales con diferentes propiedades. Tras configurar la heurística paralela y realizar los experimentos correspondientes hemos demostrado como combinar las capacidades de varios algoritmos favorece la calidad de las soluciones descubiertas, lográndose mejores resultados incluso que cuando ejecutamos los algoritmos durante más tiempo. Además, hemos demostrado también como la heurística paralela es capaz de obtener resultados biológicamente relevantes, hecho demostrado a través de las comparativas con catorce métodos biológicos bien conocidos.

En el futuro tenemos la intención de estudiar el comportamiento de nuestros algoritmos cuando resuelven instancias más distintas aun, incluso de diferentes especies. Además, realizaremos un nuevo estudio donde incorporaremos nuevas metaheurísticas con diferentes propiedades analizando si son o no adecuadas para el equipo. Finalmente, diseñaremos nuevas heurísticas paralelas y nuevas funciones de sincronización que nos permitan mejorar los resultados obtenidos.

Agradecimientos

Gracias a la Fundación Valhondo por el apoyo económico ofrecido a David L. González-Álvarez y hacer posible este trabajo. Gracias también al Gobierno de Extremadura por la ayuda GR10025 concedida al grupo de investigacion TIC015.

Referencias

[1] P. D'haeseleer, *What are DNA sequence motifs?*, Nature Biotechnology, vol. 24, no. 4, pp. 423-425, 2006.

[2] M. Kaya, *MOGAMOD: Multi-objective genetic algorithm for motif discovery*, Expert Systems with Applications, vol. 36, no. 2, pp. 1039-1047, 2009.

[3] K. Deb, *Multi-objective optimization using evolutionary algorithms*, John Wiley & Sons, 2001.

[4] C. Blum y A. Roli, *Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison*, ACM Computing Surveys, vol. 35, pp. 268-308, 2003.

[5] B. Chapman, G. Jost y R. van der Pas, *Using OpenMP Portable Shared Memory Parallel Programming*, Scientific and Engineering Computation, pp. 1-349, 2007.

[6] E. Zitzler, M. Laumanns y L. Thiele, *SPEA2: Improving the strength pareto evolutionary algorithm*, Technical report tik-report 103, Swiss Federal Institute of Technology Zurich, Switzerland, 2001.

[7] G. B. Fogel, et al., *Discovery of sequence motifs related to coexpression of genes using evolutionary computation*, Nucleic Acids Research, vol. 32, no. 13, pp. 3826-3835, 2004.

[8] G. B. Fogel, et al., *Evolutionary computation for discovery of composite transcription factor binding sites*, Nucleic Acids Research, vol. 36, no. 21, pp. e142, 1-14, 2008.

[9] D. L. González-Álvarez, M. A. Vega-Rodríguez, J. A. Gómez-Pulido y J. M. Sánchez-Pérez, *Solving the motif discovery problem by using differential evolution with pareto tournaments*, Proceedings of the 2010 IEEE Congress on Evolutionary Computation (CEC'10), pp. 4140-4147, 2010.

[10] D. L. González-Álvarez, M. A. Vega-Rodríguez, J. A. Gómez-Pulido y J. M. Sánchez-Pérez, *Finding motifs in DNA sequences applying a multiobjective artificial bee colony (MOABC) algorithm*, EVOBIO'11, LNCS, vol. 6623/2011, pp. 89-100, 2011.

[11] E. Wingender, P. Dietze, H. Karas y R. Knuppel, *TRANSFAC: a database on transcription factors and their DNA binding sites*, Nucleic Acids Research, vol. 24, no. 1, pp. 238-241, 1996.

[12] L. While, P. Hingston, L. Barone y S. Huband, *A faster algorithm for calculating hypervolume*, IEEE Transactions on Evolutionary Computation, vol. 10, no. 1, pp. 29-38, 2006.

[13] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, 4th ed. New York: Chapman & Hall/CRC Press, 2007.

[14] M. Tompa, et al., *Assessing computational tools for the discovery of transcription factor binding sites*, Nature Biotechnology, vol. 23, no. 1, pp. 137-144, 2005.