Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

Implementación de la principal referencia metaheurística de Optimización Multiobjetivo NSGA-II aplicada a la Filogenética Computacional

Cristian Zambrano-Vega[1], Antonio J. Nebro[2] y José F. Aldana-Montes[3]

1 Universidad Técnica Estatal de Quevedo, Czambrano@uteq.edu.ec

2 Universidad de Málaga - Centro de Investigación Ada Byron, antonio@lcc.uma.es

3 Universidad de Málaga - Centro de Investigación Ada Byron, jfam@lcc.uma.es

RESUMEN

Uno de los problemas más importantes en la Bioinformática y la Biología Computacional es la búsqueda y reconstrucción de árboles filogenéticos, que describan, lo más realmente posible, la evolución de las especies.La Inferencia filogenética es considerada como un problema de complejidad NP-completo por la exploración del espacio de búsqueda conformado por todas las posibles topologías existentes según el número de especies en análisis, cuyo tamaño incrementa exponencialmente por cada una de ellas, convirtiéndolo en un caso de estudio para ser abordado con técnicas metaheurísticas. El problema de la inferencia filogenética se puede formular en base a dos objetivos a optimizar de forma simultánea (La Máxima Verosimitud y la Máxima Parsimonia).

Por esta razón hemos adaptado una de las técnicas metaheurísticas de mayor referencia en el campo de la optimización multiobjetivo, el algoritmo *Nondominated Sorting Genetic Algorithm-II* (NSGA-II) a la inferencia de árboles filogenéticos incorporando nuevas estrategias de exploración, con el objetivo de conocer cuál es su rendimiento al intentar resolver este tipo de problemas.

Para esta implementación hemos integrado las funcionalidades del framework de optimización multiobjetivo *jMetalCpp*, el conjunto de librerías bioinformáticas *BIO++* y la funciones filogenéticas de la librería *PLL* (Phylogenetic Likelihood Library).

Los resultados obtenidos demuestran un rendimiento competitivo tanto bajo un enfoque biológico como de optimización frente a los resultados publicados en el estado del arte

Palabras Claves: Filogenética, optimización multiobjetivo, Computación Evolutiva.

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

Applying the principal Multiobjective Metaheuristic reference NSGA-II applied to the Computational Phylogenetics

ABSTRACT

One of the more important problems in Bioinformatics and Computational Biology is the search and reconstruction of the best phylogenetic tree that explains, in the best way possible, the evolution of species from a given dataset. Phylogenetic inference is considered as a *NP-hard* problem by the complex exploration of the tree space of topologies possible, that increases exponentially with the number of species in the input dataset, becoming in an ideal study to be addresses with metaheuristics. The phylogenetic inference can be formulated as a bi-objective optimization problem, having two objectives (Maximum Likelihood and Maximum parsimony) to be optimized at the same time.

For this reason we have applied one of the most popular multi-objective metaheuristics, the *Nondominated Sorting Genetic Algorithm-II* (NSGA-II) adapted to phylogenetic inference problem, incorporating new exploration strategies with the aim to evaluate the performance of this algorithm.

We have integrated the features of the multi-objective optimization framework *jMetalCpp*, the set of C++ libraries for Bioinformatics *BIO++* and the phylogenetic functions of the Phylogenetic Likelihood Library (*PLL*).

The obtained results show a high-competitive performance under a biological and optimization perspective.

Keywords: Computational Phylogenetics, Multiobjective Optimization, Evolutionary Computation.

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

1. INTRODUCCIÓN

La historia evolutiva de las especies vivientes y extinguidas en la tierra es una cuestión que ha sido intrigante para la humanidad durante siglos y la construcción del "árbol de la vida" que comprende todas ellas ha sido una idea fascinante y desafiante desde el surgimiento de la teoría de la evolución. Actualmente la inferencia de este "árbol de la vida" es considerado como uno de los "grandes retos" de la Bioinformática (Stamatakis, 2004).

Por lo general, las relaciones evolutivas de los organismos se representan con un árbol evolutivo que indica sus descendencias, y la inferencia filogenética consiste específicamente en eso, en encontrar ese árbol evolutivo que describa lo más exactamente posible las relaciones genealógicas o la historia evolutiva de un conjunto de especies a partir de sus secuencias moleculares.

Entre los criterios de reconstrucción de árboles filogenéticos, existen dos métodos de análisis que son los más presentes en la literatura y objeto de estudio de esta investigación: la máxima parsimonia y la máxima verosimilitud (Felsenstein, 2004). El criterio de la máxima parsimonia es un análisis basado en el razonamiento de Occam, que establece que la explicación más sencilla a un determinado fenómeno tiene preferencia sobre el resto. De acuerdo con esta idea, este método busca inferir un árbol filogenético que minimice el número de cambios de estado de caracteres (o pasos evolutivos) necesario para describir la historia evolutiva entre especies y se prefiere aquel árbol cuya topología implique una menor cantidad de transformaciones a nivel molecular. El criterio de la máxima verosimilitud busca encontrar el árbol que obtenga la probabilidad más alta de predecir la evolución de las especies a partir de los datos observados. Para estimar la verosimilitud de un árbol filogenético es necesario considerar su topología (nodos y hojas), las longitudes de sus ramas y el modelo de evolución molecular que describe las probabilidades con que se puede producir un cambio que implique la sustitución de amino ácidos o nucleótidos entre generaciones.

El problema de estos métodos de inferencia filogenética es la compleja exploración del espacio de búsqueda que deben realizar para encontrar el mejor árbol inferido. Esta exploración se dificulta por cada una de las especies del conjunto en análisis, ya que su tamaño crece de forma exponencial por cada una de ellas. Para el caso de árboles no

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

enraizados, el número de topologías para un número *n* de especies viene dado según la fórmula (Poladian y Jermiin, 2006):

$$\frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Entonces, para tener una idea, se estima que actualmente se pueden generar 10^5 topologías por segundo. Para poder realizar una búsqueda exhaustiva recorriendo por completo todo el espacio de búsqueda que existe en un análisis de tan solo 15 especies, se requerirían aproximadamente 2.51 años, esto sin considerar el tiempo necesario para evaluar la calidad de cada una de ellas.

Por todo esto, muchos científicos coinciden en que la inferencia filogenética es uno de los temas de investigación más importantes en la Bioinformática.

Handl et al. (2007) estudiaron la relación entre la optimización multi-objetivo y la biología computacional, y señalaron a la inferencia filogenética como un problema ideal de la Bioinformática para ser abordado por metaheurísticas de este tipo.

Recientemente se han propuesto algunas metaheurísticas de tipo bio-inspirado para ser aplicadas a la inferencia filogenética: *MOABC*, un algoritmo inspirado en el comportamiento natural de las colonias de abejas (Santander-Jiménez y Vega-Rodríguez, 2013a), y *Mo-FA*, inspirado en el comportamiento natural de las luciérnagas (Santander-Jiménez y Vega-Rodríguez, 2013b), además dos técnicas basadas en algoritmos evolutivos: *PhyloMOEA* (Cancino y Delbem, 2007) y *MO-Phyl* (Santander-Jiménez y Vega-Rodríguez, 2014).

El objetivo de este trabajo es adaptar la metaheurística de optimización multiobjetivo más usada, NSGAII (Deb et al., 2002), a la inferencia de árboles filogenéticos, incorporando nuevas y mejores estrategias de búsqueda, con el objetivo de conocer cuán competitivo es su rendimiento de frente al conjunto de propuestas presentadas hasta ahora.

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

2. MATERIALES Y METODOS

NSGA-II: Algoritmo de Ordenación No-Dominada II

Srinivas y Deb (1994) desarrollaron el Algoritmo Genético de Ordenación No-dominada (*Non-dominated Sorting Genetic Algorithm - NSGA*) basándose en sistemas de clasificación de la población por niveles. A todos los individuos no-dominados se les asigna una categoría con un *fitness* proporcional al tamaño de la población. Para mantener la diversidad de la población, estos individuos son clasificados según este fitness y el resultado obtenido por un parámetro de distribución (parámetro de compartición). A continuación, una vez clasificado este conjunto de individuos, se elimina de la población y se repite el proceso con las soluciones restantes, esto se realiza hasta que toda la población está clasificada. Dado que las primeras soluciones son las de mejor calidad, de estas siempre se realizarán más copias, permitiéndose así una búsqueda más profunda de las regiones no-dominadas. Dada varias falencias de la primera versión del algoritmo: la alta complejidad computacional $O(MN^3)$, donde M es el número de objetivos y N el tamaño de la población; su funcionamiento no elitista, y la necesidad de especificar el parámetro de distribución, Deb et al., (2002) presentaron una versión mejorada del algoritmo denominada NSGA-II a través de un mecanismo de ordenación no-dominada de baja complejidad $O(MN^2)$, un operador de selección para combinar la población padre con la hija, y seleccionando los N mejores individuos teniendo en cuenta su calidad y distribución en el frente de Pareto.

En el cuadro 1 se muestre el pseudocódigo de la adaptación del algoritmo a la inferencia de árboles filogenéticos.

Cuadro 1: Pseudocódigo de la Adaptación filogenética del Algoritmo NSGA-II

Entrada: Parámetros: NP, ECR, CR, PM, M, BLP,LSP
Salida: Conjunto de soluciones P

```
P ← generarPoblacionInicial(NP)
P ← OptParametrosDelModeloSust(P)
P ← OptLongitudesRamas(P, BLP)
P ← evaluarPoblacion(P,NP)
mientras no condición de terminación hacer
        para i = 1 : NP todos individuos hacer
            ind1 ← funcionSeleccion(P,ECR)
            ind2 ← funcionSeleccion(P,ECR) //ind1 ≠ ind2
            Qi ← recombinacion(ind1,ind2,CR)
            Qi ← mutacion(Qi,PM,M)
            Qi ← OptFilogeneticaExhaustivaLS(Qi,LSP)
```

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

```
                    fin para
                    Q ← evaluarPoblacion(Q,NP)
                    R ← P ∪ Q
                    R ← ordenacionNoDominada(R) //R = (F1, F2, ...)
                    P ← ∅
                    i ← 1
                    mientras |P + Fi| < NP hacer
                          P ← P ∪ Fi
                          i ← i + 1
                    fin mientras
                    Fi ← ordenacionPorDistanciaCrowding(Fi)
                    P ← P ∪ Fi[1:(NP - |P| )]

                    Si se cumple intervalo de actualización entonces
                          P ← OptLongitudesRamasExa(P,BLP)
fin mientras
```

Herramientas de software usadas en la Adaptación

Para la adaptación del algoritmo NSGAII a la inferencia de árboles filogenéticos hemos integrado las funcionalidades del framework de optimización multiobjetivo *jMetalCpp* (López-Camacho et al., 2014), el conjunto de librerías bioinformáticas *BIO++* (Dutheil et al., 2006) y las funciones de librería filogenética *PLL* (*Phylogenetic Likelihood Library)* (Flouri et al., 2014).

jMetalCpp: Es la versión C++ del framework de optimización multiobjetivo *jMetal*, dispone entre la lista de sus algoritmos mono y multi objetivo, la implementación original del algoritmo NSGAII. A partir de esta versión se procedió con la implementación de cada uno de los componentes que forman parte de la adaptación filogenética: la representación de las individuos (árboles filogenéticos), las funciones objetivo (máxima parsimonia y máxima verosimilitud), los operadores de cruce y mutación sobre árboles, y una búsqueda local para mejorar la exploración del espacio de búsqueda de árboles filogenéticos.

Bio++: es un conjunto de librerías C++ que proveen funcionalidades para el desarrollo de software en varios campos de la bioinformática: Análisis y Tratamiento de secuencias biológicas, Inferencia Filogenética, Análisis de Evolución Molecular y Genética Poblacional. Se utilizó su librería *SeqLib* (*Sequence Library*) que provee funcionalidades para manipular y analizar secuencias biológicas de ADN, ARN, proteínas y secuencias de codones, para la lectura y manejo de las secuencias de

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

nucleótidos en formato Phylip y Fasta como parámetros de entrada del algoritmo; y se usó la librería filogenética *PhylLib* (*Phylogenetics Library*) que proporciona varias funcionalidades para la manipulación de árboles filogenéticos y algunas utilidades para inferir filogenias para la estimación del score de Máxima Parsimonia, la representación de los individuos como árboles filogenéticos basados en la clase *TreeTemplate* y para la implementación de los operadores genéticos de Cruce y Mutación.

La librería filogenética (PLL): es una librería altamente optimizada para ambientes multi-core y paralelos escrita en C++ para el desarrollo de nuevas herramientas software para la inferencia filogenética, de la cual se utilizó la función PLF (*Phylogenetic Likelihood Fucntion*) para la estimación de la máxima verosimilitud sobre el modelo de sustitución de nucleótidos GTR+Γ, la función de optimización de las longitudes de las ramas por el método de *Newton-Raphson* y la función *pllRearrangeSearch* para la implementación de una de las técnicas LS del algoritmo.

Adaptación Filogenética del algoritmo NSGA-II

El proceso de adaptación del algoritmo NSGA-II (Deb et al., 2002) para ser aplicado a la inferencia de árboles filogenéticos fue el siguiente:

- **Representación de los individuos:** Todos los individuos son representados como árboles filogenéticos, basados en la clase *TreeTemplate,* los cuales están compuesto por un nodo padre y n nodos hijos, cada uno con su respectivo Nombre, ID, longitud de rama e ID del Nodo padre.
- **La Población inicial**: Para definir el punto de partida del algoritmo genético se implementaron tres funcionalidades:
    - A) Aleatoria: árboles cuya topología es generada de forma totalmente aleatoria cuyos valores de longitud de sus ramas es de 0.05. Normalmente estos árboles suelen estar muy alejados de los óptimos globales tanto en parsimonia como verosimilitud, por lo que la convergencia de sus algoritmos podría verse seriamente afectada.
    - B) Definidos por usuario: árboles filogenéticos en formato *newick* previamente inferidos mediante técnicas de bootstrap (Felsenstein, 2004) bajo el criterio de la máxima parsimonia, máxima verosimilitud o combinados; esta estrategia es usada por otras propuestas mono-objetivo (Lemmon & Milinkovitch, 2002;

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

Katoh et al., 2001) y multiobjetivo (Santander-Jiménez y Vega-Rodríguez, 2013a; Santander-Jiménez y Vega-Rodríguez,2013b y Cancino and Delbem, 2007) y

- C) Árboles generados mediante el método Stepwise-Addition: Stamatakis (2004) sugiere empezar con este tipo de árboles parsimoniosos por dos razones: Según Steel y Penny (2000) la parsimonia está relacionada directamente con la verosimilitud, al menos bajo modelos de sustitución simples (ideal en nuestro enfoque), y a diferencia de los métodos basados en distancias como BIONJ, Stepwise-addition es un método no determinista, lo que permite obtener diferentes topologías de inicio en cada ejecución independiente de los algoritmos y poder alcanzar óptimo globales entre varios ejecuciones (Zwickl, 2006). Durante la inicialización de la población las longitudes de las ramas pueden ser optimizadas mediante técnicas de optimización numérica como Newton-Raphson o Gradient (Press, et al. 1992).

- **Operador de Cruce**. De los varios operadores de cruce usados en la literatura (Matsuda, 1995; Congdon, 2002; Lewis, 1998), se ha decido implementar en el software el operador *Prune-Delete-Graft* (PDG) el cual ha demostrar generar mejores resultados sobre los diferentes criterios de optimización.
- **Operadores de Mutación.** Se implementaron tres tipos de operadores de mutación de modificaciones topológicas: *Nearest Neighbour Interchange* (NNI), *Tree Bisection and Reconnection* (TBR) y *Subtree Pruning and Regrafting* (SPR) (Felsenstein, 2004).
- **Funciones Objetivo.** Para estimar la máxima parsimonia se usaron las funcionalidades de la clase *DRTreeParsimonyScore* de la librería *PhylLib* de Bio++ y para la estimación de la máxima verosimilitud se emplearon las funcionalidades de la función PLF(*Phylogenetic Likelihood Fucntion*) de la librería PLL bajo el modelo de sustitución GTR+Γ.
- **Búsquedas Locales para la exploración del espacio de árboles.** Se han incorporado dos técnicas de Búsqueda Local LS (Local Search) específicas para la inferencia de árboles filogenéticos optimizando dos criterios de simultáneamente: la máxima parsimonia y máxima verosimilitud:
  - La primera LS está basada en la alta relación que existe entre la parsimonia y la verosimilitud, en la que bajo una perspectiva teórica definida por (Tuffley, 1997) se establece que minimizar la parsimonia es equivalente a maximizar la verosimilitud

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

bajo ciertos supuestos (ideal en nuestro enfoque), por lo que al igual a una de las técnicas de búsquedas implementadas en el software PhyML (Guindon et al, 2010) ésta técnica explora el espacio de árboles encontrando soluciones multiobjetivo aplicando movimientos topológicos que minimizan la parsimonia y ajusta las longitudes solo de las ramas aplicando el método de optimización Newton-Raphson luego de cada cambio realizado. Para la optimización de la parsimonia se emplea la técnica *Parametric Progressive Tree Neighbourhood* (PPN), propuesta por Göeffon *et* al. (2008), la cual está definida como un conjunto de movimientos topológicos SPR en los que las distancia de la rama del subárbol podado y la rama en la que se realizará el injerto está basada en un valor *d*, cuyo valor inicial es la distancia máxima que hay entre el nodo raíz y las hojas del árbol, permitiendo empezar con movimientos globales, reduciendo progresivamente su valor hasta llegar a un valor mínimo de 1, con el cual los movimientos SPR representan un movimiento local similar a los que se obtienen con la técnica NNI. Solo aquellos movimientos que mejoren la parsimonia son aplicados definitivamente sobre la topología con el fin de ir optimizándola.

- La segunda LS es una combinación parametrizada de dos técnicas enfocadas a mejorar los criterios de optimización por separado: para la parsimonia usa PPN (Göeffon *et* al, 2008), cuyo funcionamiento es el mismo definido en la LS anterior, y por el lado de la verosimilitud una técnica de exploración basada en reordenamiento de topologías de la librería *Phylogenetic Likelihood Library* (PLL) llamada *pllRearrangeSearch*, la cual realiza movimientos topológicos de tipo NNI o SPR desde un nodo específico con todos los demás nodos a su alrededor y dentro de un radio de cobertura. Gracias al esquema genérico de vectorización en el que está desarrollada PLL, se puede recalcular de forma rápida y óptima la verosimilitud de los árboles después de cada movimiento topológico probado.

- **Intervalo de Actualización de Parámetros:** Una misma topología posee un solo valor de parsimonia pero también posee diferentes valores de verosimilitud según las longitudes de sus ramas y los parámetros del modelo de sustitución que se use (Stamatakis, 2004). Es por esto que el algoritmo incluye dos métodos de búsqueda de máximos y mínimos para la optimización de las longitudes de las ramas: Newton-Raphson (Press, et al 1992) y Gradient (Brent, 1973), los cuales pueden ser parametrizados para ser usados antes, durante y al final de la ejecución de los

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

algoritmos. De igual forma se puede parametrizar la optimización de los parámetros del Modelo de Sustitución GTR+Γ: Tasas de Sustitución de nucleótidos μAC, μA, μAT, μCG μCT y μGT y la tasa de sustitución entre sitios alpha (α). El intervalo de actualización está basado en el número de evaluaciones de los algoritmos.

- **Formato de los Resultados:** Los resultados del algoritmo están basados en el formato de salida del framework *jMetalCpp*: se crean dos ficheros de texto planos llamados *FUN+IdExperiment* y *VAR+IdExperiment*. El primero contiene los datos de verosimilitud y parsimonia de las soluciones encontradas, mientras que el segundo contiene los árboles filogenéticos optimizados bi-objetivos en formato newick.

Problemas de Pruebas

Para evaluar el rendimiento de esta adaptación se analizó el siguiente conjunto de secuencias nucleotídicas definidas por las demás propuestas algorítmicas publicadas en el estado del arte (Cuadro 2).

Cuadro 2. Secuencias nucleotídicas definidas para los experimentos

Nombre	Descripción
rbcL_55	55 secuencias de 1314 caracteres del gen rbcL de distintas especies de plantas verdes
mtDNA_186	186 secuencias de 16608 caracteres correspondientes a ADN mitocondrial humano
RDPII_218	218 secuencias de 4182 caracteres de ARN de procariotas
ZILLA_500	contiene 500 secuencias de 1428 caracteres rbcL de plástidos de plantas

Metodología aplicada a los experimentos

Hemos realizado un conjunto de 20 ejecuciones independientes del algoritmo bajo las configuraciones detalladas en el cuadro 3.

Cuadro 3. Parametrización del Algoritmo NSGA-II en la experimentación de su rendimiento resolviendo el conjunto de secuencias: rbcL_55, mtDNA_186, RDPII_218 y ZILLA_500

PARÁMETROS	VALORES
Metaheurística Multiobjetivo	
• Tamaño de la población	**100**
• Número de evaluaciones	**6000**
• Porcentaje de Cruce PDG	**0.8**
• Operador de Mutación	**NNI**
• Porcentaje de Mutación	**0.2**
Optimización Filogenética	
• Población Inicial	**USER –** Repositorio Bootstrap

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

• Optimización inicial de las longitudes de las ramas	**Si**
• Técnica de Optimización inicial de las longitudes de las ramas	**NR**=Newton Raphson
• Optimización inicial de los parámetros Modelo de GTR+ Γ	**Si**
• Técnica de Optimización Topológica	2 = Búsquedas combinadas **PPN y PLL**
• Porcentaje de aplicación Técnicas combinada PPN y PLL	0.5
• Técnica de Optimización longitudes de Ramas	**NR** =Newton Raphson
Búsqueda local PPN	
• Número de Iteraciones	1000
• Número de Cambios aplicados	100
Búsqueda local PLL	
• Porcentaje nodos seleccionados	**0.2**
• Tipo de Movimientos	**SPR**
• Rango de recorrido Mínimo y Máximo	**MinTrav** = 1 **MaxTrav**=20
• Optimización de las ramas afectadas por los movimientos topológico	**Si**

La población inicial de árboles filogenéticos fue generada a partir de un repositorio de árboles obtenidos de un análisis bootstrap creado por los programas *PHYML* y *DNAPars*, la mitad generados bajo el criterio de máxima parsimonia y la otra mitad bajo el criterio de máxima verosimilitud. La selección del modelo de sustitución y sus parámetros fueron establecidos a partir de un análisis estadístico realizado por la herramienta *jModelTest*.

3. RESULTADOS

Los resultados multiobjetivo (Frentes de Pareto) generados por nuestra adaptación del algoritmo NSGAII infiriendo filogenias para cada una de las secuencias nucleotídicas de prueba, se ilustran en la Figura 1, se incluyen dos de los mejores frentes de Pareto generados en base al indicador de calidad Hipervolumen y dos de los mejores publicados por los algoritmos del estado del arte *MOABC*, *MO-FA*, *MO-Phyl* y *PhyloMOEA*. Podemos resaltar que para todos los problemas de prueba los frentes de Pareto generados por nuestra implementación se caracterizan por tener una mejor distribución y un mayor número de soluciones; específicamente podemos resaltar los resultados inferidos sobre las secuencias *rbcL_55* y *mtDNA_186*, los cuales están conformados por un mayor número de soluciones, 14 y 12 soluciones respectivamente, en relación al total de soluciones que conforman los frentes publicados por los algoritmos *MO-Phyl* con 8 y 9 soluciones respectivamente y *MOABC* con 4 y 10

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

soluciones respectivamente. Además podemos observar que los frentes generados por nuestra adaptación presentan una mayor convergencia a un frente de Pareto óptimo en relación a los frentes generados por los demás propuestas algorítmicas, específicamente podemos resaltar a los frentes generados para el análisis de las secuencias *RDPII_218.* Para el caso del conjunto de secuencias *ZILLA_500* los frentes obtenidos por nuestra implementación no logran mejorar los publicados por las demás propuestas, considerando el gran tamaño del espacio de búsqueda que representa analizar un conjunto de 500 especies.

El cuadro 3 muestra los resultados filogenéticos (puntajes de Parsimonia y Verosimilitud) obtenidos por nuestra adaptación frente a los presentados en el estado del arte, en la primera columna se detalle el nombre de la instancia, y luego los mejores valores de Parsimonia y de Verosimilitud para cada uno de ellas. Para el caso de la máxima parsimonia, los valores generados por nuestra implementación son muy relevantes pero no logran mejorar a los publicados por las otras propuestas algorítmicas, a excepción del conjunto de secuencias *rbcL_55* en el cual se logra igualar los resultados. Pero en el caso de la máxima verosimilitud, si se logran mejorar los resultados para los conjuntos de secuencias *rbcL_55* y *mtDNA_186*, para los otros dos conjuntos de secuencias, aunque no se logra mejorarlos, los valores obtenidos son igual de relevantes.

Cuadro 3: Resultados del algoritmo NSGA-II usando el modelo evolutivo GTR + Γ frente a los resultados del estado del Arte.

Instancia	Mejor árbol Parsimonioso		Mejor Árbol Verosímil	
	Pars.	Veros.	Pars.	Veros.
rbcL_55	**4874**	-21800.8	4914	**-21769.2**
mtDNA_186	2433	-39912.1	2445	**-39868.5**
RDPII_218	41573	-135405.9	42778	-134248.3
ZILLA_500	16245	80728.6	16274	-80629.3
Instancia	**Mejor árbol Parsimonioso**		**Mejor Árbol Verosímil**	
	Pars.	**Veros.**	**Pars.**	**Veros.**
rbcL_55	4874	-21816.17	4885	-21791.12
mtDNA_186	**2431**	-40007.28	2453	-39868.94
RDPII_218	**41488**	-136379.87	42837	**-134073.89**
ZILLA_500	**16218**	-81358.20	16320	**-80585.04**

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

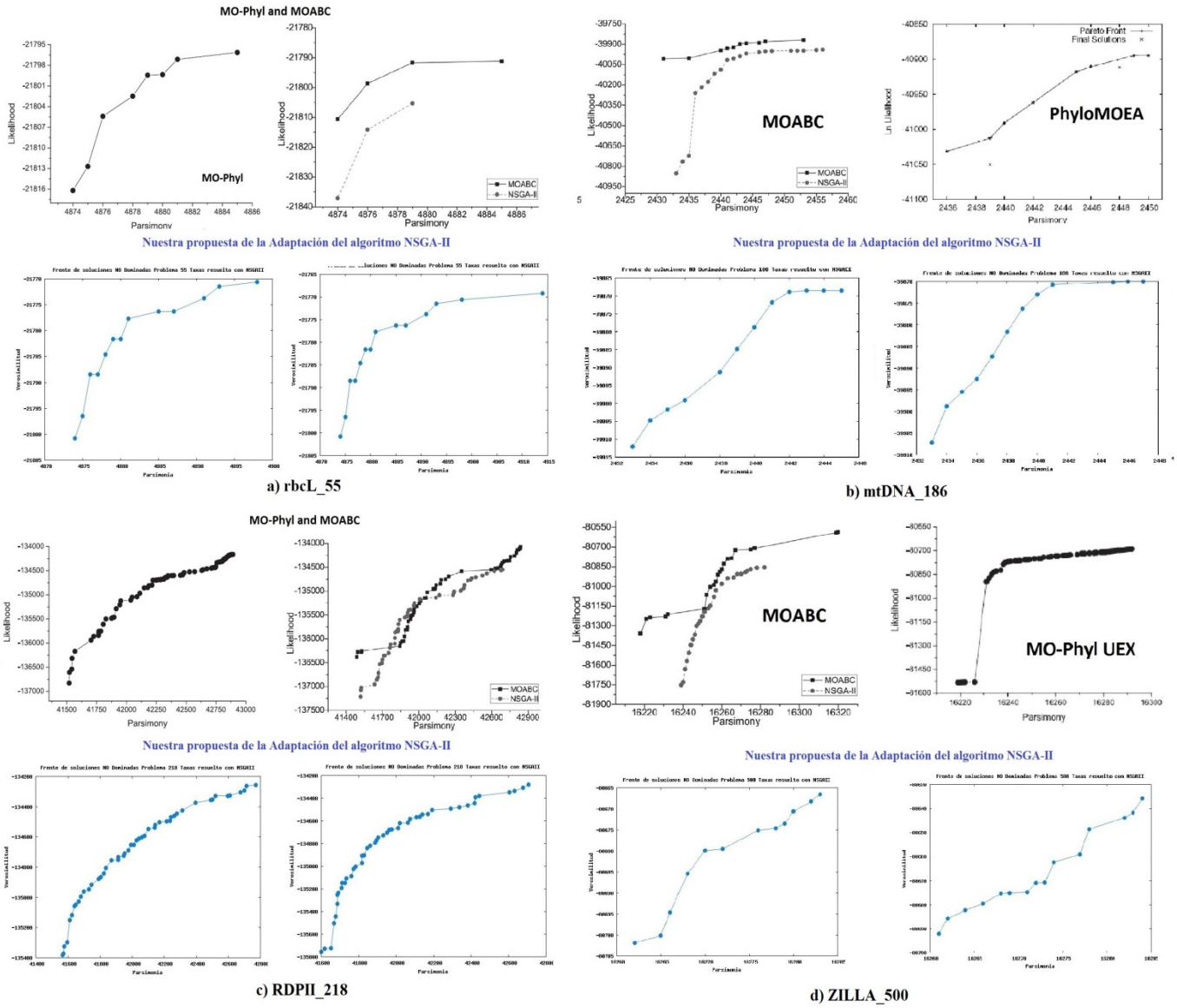

Figura 1: Frentes de Pareto aproximados generados por la implementación del algoritmo NSGAII (color azul – parte inferior) y por otras propuestas algorítmicas del estado del arte (MOABC, Mo-FA, PhyloMOEA y MO-Phyl) resolviendo el conjunto de secuencias a) rbcL_55, b) mtDNA_186, c) RDPII_218 y d) ZILLA_500

Es importante resaltar que en esta discusión no podemos hacer una comparativa justa con los resultados publicados por otros autores del estado del arte, ya que en el caso de los algoritmos *MOABC*, *MO-FA* y *MO-Phyl* no se conocen ni se tienen acceso al repositorio de árboles filogenéticos iniciales con los que los algoritmos empiezan su ejecución, esto depende altamente de los resultados finales. Además no incluimos en la comparativa al algoritmo *PhyloMOEA*, ya que para estimar la máxima verosimilitud de las filogenias inferidas utilizaron el modelo evolutivo HKY85+Γ y nuestra implementación usa el modelo evolutivo GTR+Γ, ya que, según el análisis realizado por

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

el software jModelTest, este es el modelo que mejor se adapta al conjunto de secuencias.

4. CONCLUSIONES

En base a los resultados multiobjetivo (frentes de Pareto aproximados) y filogenético (valores de Parsimonia y Verosimilitud) de las filogenias inferidas a partir de las secuencias nucleotídicas de prueba, podemos concluir que el algoritmo NSGA-II adaptado a la inferencia filogenética permite brindar resultados muy competitivos frente a los resultados publicados por otros algoritmos del estado del arte, bajo las mismas configuraciones y condiciones expuestas por los demás, incluso con una cantidad menor de evaluaciones.

Resaltamos el alto rendimiento de esta adaptación gracias a las estrategias de exploración del espacio de búsqueda incorporadas dentro de su estructura, la implementación de una estrategia híbrida entre las técnicas *Progressive Tree Neighbourhood* (PPN) y *pllRearrangeSearch* de la librería PLL, las cuales de forma independiente han demostrado ser muy reconocidas por su alto rendimiento, y que en nuestro caso nos ha permitido obtener, bajo un enfoque multiobjetivo, resultados muy significativos lo que nos permite ofrecer, no solo una únicamente solución sino un amplio conjunto de hipótesis filogenéticas alternativas.

Con los resultados obtenidos podemos definir que el área de aplicación es muy prometedora, por lo que consideramos implementar otras metaheurísticas de optimización más recientes como SMSPSO, MOEA/D y SMS-EMOA e integrarlas en un solo software de inferencia filogenética multiobjetivo.

5. REFERENCIAS BIBLIOGRÁFICAS

Beume, N., Naujoks, B., and Emmerich, M. (2007). Sms-emoa: Multiobjective selection based on dominated hypervolume. European Journal of Operational Research, 181(3), 1653 – 1669.

Brent, R.P. (1973) Algorithms for minimization without derivatives. Engelwood Clifts, New Jersey.

Cancino, W. and Delbem, A. C. (2007). A multi-objective evolutionary approach for phylogenetic inference. In S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, and T.

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

Murata, editors, Evolutionary Multi-Criterion Optimization, volume 4403 of Lecture Notes in Computer Science, pages 428–442. Springer Berlin Heidelberg.

Congdon, C.B. (2002) Gaphyl: An evolutionary algorithms approach for the study of natural evolution. GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1057-1064.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. Evolutionary Computation, IEEE Transactions on, 6(2), 182–197.

Dutheil, J., Gaillard, S., Bazin, E., Gl´emin, S., Ranwez, V., Galtier, N., and Belkhir, K. (2006). Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC bioinformatics, 7, 188.

Felsenstein, J. (2004). Inferring Phylogenies. Palgrave Macmillan. Flouri, T., Darriba, D., and Aberer, A. J. (2014). The Phylogenetic Likelihood Library.

Flouri, T., Darriba, D. & Aberer, A.J. (2014) The Phylogenetic Likelihood Library.

Göeffon, A., Richer, J.-M., and Hao, J.-K. (2008). Progressive tree neighborhood applied to the maximum parsimony problem. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 5(1), 136–45.

Guindon, S. et al., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology, 59(3), pp.307–21.

Handl, J., Kell, D. B., and Knowles, J. (2007). Multiobjective optimization in bioinformatics and computational biology. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 4(2), 279–92.

Knowles, J.; Corne, D., The Pareto archived evolution strategy: a new baseline algorithm for Pareto multiobjective optimization, Evolutionary Computation, 1999. Proceedings of the 1999 Congress on , vol.1, no., pp.,105 Vol. 1, 1999.

Katoh, K., Kuma, K.i. & Miyata, T. (2001) Genetic algorithm-based maximum-likelihood analysis for molecular phylogeny. Journal of Molecular Evolution, 53, 477-484.

Lemmon, A.R. & Milinkovitch, M.C. (2002) The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation.

Lewis, P. O., A Genetic Algorithm for Maximum-Likelihood Phylogeny Inference Using Nucleotide Sequence Data, pp. 277–283, 1996.

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

López-Camacho, E., García Godoy, M. J., Nebro, A. J., and Aldana-Montes, J. F. (2014). jMetalCpp: optimizing molecular docking problems with a C++ metaheuristic framework. Bioinformatics (Oxford, England), 30(3), 437–8.

Matsuda, H. (1995) Construction of phylogenetic trees from amino acid sequences using a genetic algorithm. Sciences, Computer, p. 560.

Nebro, Antonio J., Durillo, Juan J., Luna, Francisco, Dorronsoro, Bernabé, Alba, Enrique, MOCell: A cellular genetic algorithm for multiobjective optimization, International Journal of Intelligent Systems, vol 24, pp 726--746, 2009.

Poladian, L., Jermiin, L.: Multi-Objective Evolutionary Algorithms and Phylogenetic Inference with Multiple Data Sets. Soft Computing 10(4), p. 359-368. (2006).

Press, W., Flannery, B. Teukolsky, S. y Vetterling, W.T., 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Santander-Jiménez, S. and Vega-Rodríguez, M. A. (2013a). Applying a multiobjective metaheuristic inspired by honey bees to phylogenetic inference. Biosystems, 114(1), 39 – 55.

Santander-Jiménez, S. and Vega-Rodríguez, M. A. (2013b). A multiobjective proposal based on the firefly algorithm for inferring phylogenies. In L. Vanneschi, W. Bush, and M. Giacobini, editors, Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, volume 7833 of Lecture Notes in Computer Science, pages 141–152. Springer Berlin Heidelberg.

Santander-Jiménez, S. & Vega-Rodríguez, M.a. (2014) A hybrid approach to parallelize a fast non-dominated sorting genetic algorithm for phylogenetic inference. Concurrency and Computation: Practice and Experience.

Sheneman, L. (2005) A Survey of Evolutionary Crossover Operators as Applied to Phylogenetic Inferencing. pp. 1-13.

Srinivas, N. and Deb, K. “Multi-objective function optimization using non-dominated sorting genetic algorithms,” Evolutionary Computation, vol. 2, no. 3, pp. 221–248, 1994.

Stamatakis, A. (2004). Distributed and Parallel Algorithms and Systems for Inference of Huge Phylogenetic Trees based on the Distributed and Parallel Algorithms and Systems for Inference of Huge Phylogenetic Trees based on the Maximum Likelihood Method. Ph.D. thesis.

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

Steel, M. and Penny, D. (2000). Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. Molecular biology and evolution, pages 839–850.

Tuffley, C. & Steel, M. (1997) Links between Maximun Likelihood and Maximun Parsimony under a simple model of site substitution. 59, 581-607.

Zhang, Q. and Li, H. (2007). Moea/d: A multiobjective evolutionary algorithm based on decomposition. Evolutionary Computation, IEEE Transactions on, 11(6), 712–731.

Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under maximum likelihood criterion. Ph.D. Thesis.

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

ANEXOS

Cuadro 4. Lista de parámetros del algoritmo NSGAII para la inferencia filogenética implementado en este estudio.

Nombre del Parámetro	Valor Predeterminado
Parámetros de la Metaheurística	
Experimentid	1
DATAPATH*	data
populationsize	100
maxevaluations	2000
Offset	100
Bisections	5
Archivesize	100
intervalupdateparameters	500
Printtrace	false
printbestscores	false
Operadores Genéticos	
selection.method	binarytournament
crossover.method	PGD
crossover.probability	0,8
crossover.offspringsize	2
mutation.method	NNI
mutation.probability	0,2
Parámetros De la Filogenética	
Alphabet	DNA
input.sequence.file*	$(DATAPATH)/sequences/16s.txt
input.sequence.format	Phylip(order=interleaved, type=extended, split=spaces)
input.sequence.print	false
partitionmodelfilepll*	$(DATAPATH)/model/16s.model
init.population	stepwise
init.population.tree.file	$(DATAPATH)/inputusertrees/userinittree.txt
init.population.tree.format	Newick
Modelo Evolutivo	
Model	GTR+G
Base Frequencies	
model.frequences	user
model.piA, model.piC, model.piG, model.piT	0,25 0,25 0,25 0,25
GTR relative rate	
model.AC,model.AG, model.AT, model.CG , model.CT,	1 , 1, 1, 1, 1, 1

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

model.GT	
Distribution Gamma	
rate_distribution	Gamma
rate_distribution.ncat	4
rate_distribution.alpha	0,5
Optimización Filogenética: Búsqueda Local para la exploración del espacio de árboles: h2 y h1	
optimization.method	h2
optimization.method.perc	0,5
Función PPN	
optimization.ppn.numiterations	500
optimization.ppn.maxsprbestmoves	20
Función pllRearrangeSearch	
optimization.pll.percnodes	0,2
optimization.pll.mintranv	1
optimization.pll.maxtranv	20
optimization.pll.newton3sprbranch	true
Optimización de las Longitudes de la ramas	
bl_optimization.starting	false
bl_optimization.starting.method	newton
bl_optimization.starting.maxitevaluation	1000
bl_optimization.starting.tolerance	0,01
Optimización de las Longitudes de la ramas durante una de las técnicas de exploración del árbol-espacio o durante cada intervalo de actualización	
bl_optimization	false
bl_optimization.method	newton
bl_optimization.maxitevaluation	1000
bl_optimization.tolerance	0,01
Optimización de las Longitudes de la ramas de la población final de árboles filogenéticos generadas por el algoritmo	
bl_optimization.final	false
bl_optimization.final.method	newton
bl_optimization.final.maxitevaluation	1000
bl_optimization.final.tolerance	0,01

Revista Publicando, 2(5). 2015, 1-20. ISSN 1390-9304

Optimización del Modelo Evolutivo	
model_optimization	false
model_optimization.method	brent
model_optimization.maxitevaluation	1000
model_optimization.tolerance	0,01
Optimización de la Tasa de distribución	
ratedistribution_optimization	false
ratedistribution_optimization.method	brent