

UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
Departamento de Ingeniería Informática



**INFERENCIA FILOGENÉTICA MULTI-OBJETIVO CONSIDERANDO
FENÓMENOS RETICULARES**

Manuel José Villalobos Cid

Profesor guía: PhD, Mario Inostroza Ponta

Tesis de grado presentada en
conformidad a los requisitos
para obtener el grado de Doctor en
Ciencias de la Ingeniería mención
Informática.

SANTIAGO – CHILE

2017

© Manuel José Villalobos Cid - 2017



• Algunos derechos reservados. Esta obra está bajo una Licencia Creative Commons Atribución-Chile 3.0. Sus condiciones de uso pueden ser revisadas en:
<http://creativecommons.org/licenses/by/3.0/cl/>.

RESUMEN

El avance en las técnicas de secuenciación molecular y la gran cantidad de evidencia biológica disponible, ha convertido el proceso de inferencia filogenética en uno de los problemas actuales de la bioinformática. Este proceso racional permite proponer una hipótesis para explicar las relaciones evolutivas entre un conjunto de organismos. Estas relaciones comúnmente son representadas por medio de árboles filogenéticos. Para ello, se debe asumir que la transferencia de la información evolutiva tiene un carácter vertical, herencia exclusiva entre padres e hijos, no permitiendo el modelamiento de mecanismos evolutivos más complejos, como transferencia horizontal de genes, hibridación, recombinación, entre otros. Trabajos recientes proponen la representación de estos fenómenos por medio de redes filogenéticas. Para su construcción se plantea diferentes estrategias que abordan el proceso de inferencia como un problema de optimización basado en un criterio específico, como parsimonia o verosimilitud. Sin embargo, las topologías resultantes de estas aproximaciones presentan conflicto en sí, y varían dependiendo de diferentes fuentes de sesgo: (1) el criterio escogido para la selección de una topología determinada, (2) la elección de un modelo evolutivo asociado al cálculo de verosimilitud, (3) el tipo de evidencia evolutiva empleada para la construcción de una hipótesis, (4) el paradigma aplicado para la combinación de evidencia biológica, y (5) el sentido topológico de la red.

Esta tesis tiene como objetivo el desarrollo de una estrategia para inferir filogenia representando fenómenos reticulares, reduciendo el sesgo asociado a la elección de criterios específicos de construcción de topologías, asegurando la obtención de soluciones de calidad en términos de los criterios optimizados. Con este fin, se ha propuesto un algoritmo basado en optimización multi-objetivo capaz de inferir redes filogenéticas incorporando las diferentes fuentes de sesgo en la construcción de hipótesis evolutivas. Para su desarrollo se aplicó el método heurístico de Polya, extrayendo conocimiento de los problemas de: (1) inferencia multi-objetivo de árboles filogenéticos, (2) combinación de evidencia biológica, e (3) inferencia de redes filogenéticas.

La propuesta es evaluada empleando conjuntos de datos de la literatura relacionada, obteniendo topologías reticulares que representan un compromiso entre tres criterios de inferencia divergentes no dependientes de modelos evolutivos. Esta es comparada con otra estrategia de la literatura basada en la optimización de un único objetivo, presentando soluciones de mejor calidad en términos de dominancia. Esta investigación es la primera en desarrollar el problema de inferencia de redes filogenéticas considerando optimización multi-objetivo.

Palabras Claves: inferencia filogenética; redes filogenéticas;optimización multi-objetivo; evidencia biológica.

ABSTRACT

Owing to the recent advances in sequencing technology and the large volume of data available, the phylogenetic inference problem has become one of the most important research topics in the field of bioinformatics. Phylogenetic inference allows building a hypothesis about the evolutionary relationships between a group of species, which is usually represented as a phylogenetic tree. It assumes that the evolutionary process has a “*vertical sense*”, transmitting the genetic information exclusively from parents down to the offspring. More complex evolutionary mechanisms which involve a “*horizontal sense*” (eg. horizontal gene transfer, hybridisation, recombination, among others) can not be modelled using this representation. Phylogenetic networks provide us an alternative to model phylogeny considering these reticulated events. The phylogenetic inference process based on networks has been treated as optimisation problem considering a single criterion, such as parsimony or likelihood. The strategies designed to infer phylogeny considering reticular events have resulted in conflicting evolutionary hypotheses, including several bias factors: (1) dependency of the optimal criterion defined to choose a specific topology, (2) the evolutionary model related to the likelihood score calculation, (3) the available biological evidence, (4) the paradigm selected to perform a simultaneous analysis of multiple data sets, and (5) the network sense.

The goal of this thesis is to propose a new strategy to infer phylogenetic networks considering reticulated phenomena, reducing the bias associated to the selection of specific inference criteria, and obtaining good solutions in terms of optimisation. To this purpose, we have proposed a new multi-objective optimisation algorithm adapted to tackle the multi-objective phylogenetic network inference problem, by considering the different sources of bias in the evolutionary hypothesis inference. We applied the Polya method, using three well-known subproblems: (1) multi-objective phylogenetic inference based on trees, (2) the combination of multiple biological evidence, (3) phylogenetic inference based on networks.

The proposed algorithm was evaluated using data sets from the related literature. It obtained reticulated topologies which represent good compromises between three divergent criteria, without depending of an evolutionary model. The proposal was contrasted with a current state of the art single-objective optimisation tool, finding better solutions in terms of dominance. This is the first research work that deal with the multi-objective phylogenetic inference problem based on networks.

Keywords: phylogenetic inference; phylogenetic networks; multi-objective optimisation; biological evidence.

AGRADECIMIENTOS

Antes de escribir los agradecimientos, quisiera hacer un ejercicio algo alocado, basado en un texto que alguna vez leí. Por cierto, admito que puede ser una aproximación grosera, sin embargo, esto no influye en la idea que se quiere plantear.

*Al nacer, una mujer posee una carga de 750.000 óvulos y un hombre producirá a lo largo de la vida alrededor de $4 * 10^{12}$ espermatozoides. Bajo condiciones ideales, la probabilidad de que uno de ellos se encuentre con el óvulo específico para generar un individuo particular es de 1 en $3 * 10^{18}$. Nuevamente, asumiendo condiciones ideales, en que un individuo haya alcanzado una condición adulta reproductiva saludable a lo largo de diferentes generaciones durante la evolución, suponiendo 150.000 generaciones, es de 1 en $3 * 10^{18 * 150,000}$, lo que es equivalente a 1 en $3 * 10^{2,700,000}$. Ahora bien, la probabilidad de que dos individuos interactúen al mismo tiempo en el transcurso de la vida es de 1 en $10^{5,400,000}$. Solicito de favor ignorar el número 3, así como puede ignorar la probabilidad de nacer en el mismo país (1/192), la probabilidad de que un hombre conozca a una determinada mujer a lo largo de la vida (1/20000), o incluso tener algún tipo de relación amorosa (1/2000). La cifra $10^{5400000}$ quizás no tenga sentido usted, sin embargo, puedo entregar algunas referencias para su entendimiento. Por ejemplo, la probabilidad de obtener la lotería en Chile es del orden de 1 en $4 * 10^6$, el número promedio de microbios que está atacando su cuerpo en este momento bordea los 10^{14} , el número de metros cúbicos del Océano Atlántico es del orden de 10^{17} (misma cantidad de hormigas en la actualidad sobre nuestro planeta), y el número de átomos en el universo es del orden de 10^{80} . Todas estas cifras resultan insignificantes al lado de la probabilidad de que usted me conozca, o haya conocido a cualquier persona en particular a lo largo de su vida (1 entre $10^{5,400,000}$). Incluso, la probabilidad de que usted haya tenido a su hijo o hija tal como la conoce, si es que es padre o madre.*

Desde un punto de vista religioso, usted podría considerar este evento como un milagro, ya que la probabilidad de que dos personas específicas interactúen a lo largo de su existencia es prácticamente nula, y el número obtenido es básicamente una aproximación vulgar que omite infinitos eventos. Por otro lado, bajo la lógica científica, la interacción de dos personas puede ser considerada numéricamente como un evento anómalo o una singularidad, dado que, por ejemplo, durante el desarrollo de una investigación, una hipótesis puede ser validada considerando una probabilidad condicional (p-value) de 0,001; valor que está completamente fuera de escala en relación a la cifra sugerida.

Bajo esta lógica, y luego de haber intentado explicar mi visión respecto a lo valioso que me resulta el conocer a una persona específica en el transcurso de mi vida, empleo un sentido más tradicional para agradecer infinitamente a: mi familia, por su apoyo incondicional, a mi orientador Mario Inostroza, por cumplir este rol a nivel académico y a nivel personal, a mis amig@s, por su valiosa compañía y apoyo. También, particularmente, a todas aquellas personas que han aportado o van a aportar en las diferentes etapas de mi vida, dado que la probabilidad de que este evento ocurra es casi nula.

En este espacio también debo efectuar agradecimientos formales: a CeBIB (FB00001), CITIAPS (PMI USA1204), DICYT-VRIDEI, USACH (061619IP), la Asociación de Universidades Grupo Montevideo (AUGM) y la Universidad Federal de Rio Grande del Sur (UFRGS). Los experimentos computacionales fueron apoyados por la infraestructura de supercómputo del NLHPC (National Laboratory for High Performance Computing) (ECM-02) y la infraestructura Microsoft Azure for Research Award.

TABLA DE CONTENIDO

1	Introducción	1
1.1	Antecedentes y motivación	1
1.2	Descripción del problema	3
1.3	Hipótesis y objetivos	4
1.3.1	Hipótesis	4
1.3.2	Objetivo general	4
1.3.3	Objetivos específicos	4
1.3.4	Alcances y limitaciones	5
1.4	Propuesta de investigación	5
1.5	Organización del documento	9
2	Inferencia multi-objetivo de árboles filogenéticos	11
2.1	Inferencia filogenética basada en árboles	13
2.1.1	Árboles filogenéticos	13
2.1.2	Métodos para inferencia de árboles filogenéticos	14
2.1.3	Problema de inferencia multi-objetivo de árboles filogenéticos	16
2.2	Aproximación basada en optimización multi-objetivo	19
2.2.1	Descripción del algoritmo	20
2.2.1.1	Criterios de optimalidad	20
2.2.1.2	Inicialización de la población	21
2.2.1.3	Operadores de cruzamiento	21
2.2.1.4	Operador de mutación	22
2.2.1.5	Estrategia de búsqueda local	22
2.2.2	Evaluación de desempeño	23
2.2.2.1	Operadores de cruzamiento	24
2.2.2.2	Operadores de mutación	24
2.2.2.3	Configuración global	24
2.2.2.4	Aproximaciones basadas en optimización de un único objetivo	25
2.2.2.5	Aproximaciones basadas en optimización de múltiples objetivos	25
2.2.2.6	Estudio de conjunto de datos de aminoácidos	26
2.3	Resultados	26
2.3.1	Operadores de cruzamiento	27
2.3.2	Operadores de mutación	28
2.3.3	Configuración global	29
2.3.4	Comparación con métodos basados en optimización de objetivo único	29
2.3.5	Comparación con métodos basados en optimización de objetivo múltiples	30
2.3.6	Experimentación con secuencias de aminoácidos	32
2.4	Conclusiones	37
3	Combinación de evidencia biológica en inferencia filogenética	39
3.1	Antecedentes	41
3.1.1	Paradigmas de evidencia total y congruencia taxonómica	41
3.1.2	Estrategias para combinar fuentes biológicas en inferencia filogenética	42
3.1.3	Métodos para comparación de árboles filogenéticos	43
3.1.4	Problema multi-objetivo de combinación de evidencia biológica en filogenia	44
3.2	Aproximación basada en optimización multi-objetivo	45
3.2.1	Descripción de algoritmo	46
3.2.1.1	Conjunto de datos de entrada	46
3.2.1.2	Inicialización de la población	47
3.2.1.3	Criterios de optimalidad	47
3.2.1.4	Operación de cruzamiento	47

3.2.1.5	Operación de mutación	47
3.2.1.6	Ordenamiento no dominado	48
3.2.2	Parametrización del algoritmo	48
3.2.3	Evaluación de capacidad para reconstruir hipótesis evolutivas integrales	48
3.2.4	Comparación entre aproximaciones multi-objetivo	49
3.2.5	Aplicación de conjuntos de datos sin hipótesis evolutiva de referencia	51
3.3	Resultados	51
3.3.1	Parametrización del algoritmo	51
3.3.2	Evaluación de capacidad para reconstruir hipótesis evolutivas integrales	54
3.3.3	Comparación con aproximación multi-objetivo basado en verosimilitud	54
3.3.4	Aplicación de conjuntos de datos sin hipótesis evolutiva de referencia	55
3.4	Conclusiones	55
4	Reducción de espacio de búsqueda y tomadores de decisiones	61
4.1	Reducción de espacio de búsqueda	63
4.1.1	Descripción del experimento	63
4.1.2	Resultados	63
4.2	Tomadores de decisiones	65
4.2.1	Descripción del experimento	66
4.2.2	Resultados	66
5	Inferencia filogenética multi-objetivo basado en redes	70
5.1	Inferencia filogenética basada en redes	72
5.1.1	Redes filogenéticas	72
5.1.2	Métodos para inferencia de redes filogenéticas	74
5.1.3	El problema de inferencia multi-objetivo de redes filogenéticas	75
5.2	Aproximación basada en optimización multi-objetivo	77
5.2.1	Descripción de algoritmo	77
5.2.1.1	Conjunto de datos de entrada	78
5.2.1.2	Inicialización de la población	78
5.2.1.3	Criterios de optimalidad	78
5.2.1.4	Operación de cruzamiento	79
5.2.1.5	Operación de mutación	79
5.2.1.6	Ordenamiento no dominado	79
5.2.2	Parametrización del algoritmo	79
5.2.3	Evaluación de reconstrucción de hipótesis evolutivas reticuladas	80
5.2.4	Espacio de soluciones y espacio objetivo	80
5.2.5	Comparación con otras propuestas	80
5.3	Resultados	81
5.3.1	Evaluación de reconstrucción de hipótesis evolutivas reticuladas	81
5.3.2	Espacio de soluciones y espacio objetivo	85
5.3.3	Comparación con otras propuestas	87
5.4	Conclusiones	88
6	Conclusiones y trabajo futuro	90
6.1	Conclusiones	90
6.2	Trabajo futuro	92
Listado de acrónimos		95
Referencias bibliográficas		108
Anexos		108
A Número de publicaciones por año pertinentes al área de investigación		109

B Definiciones de originalidad para un trabajo de investigación	110
C Métodos para reconstrucción de árboles filogenéticos	111
C.1 Métodos de reconstrucción filogenética basados en distancia	111
C.1.1 Algoritmos para generación de árboles ultramétricos	111
C.1.2 Algoritmos para generación de árboles aditivos	113
C.1.3 Métodos basados en optimización	114
C.1.3.1 Mínimos cuadrados	114
C.1.3.2 Evolución mínima	114
C.2 Métodos de reconstrucción filogenética basados en caracteres	115
C.2.1 Máxima parsimonia	115
C.2.2 Máxima verosimilitud	116
C.2.3 Aproximación bayesiana	116
C.3 Modelos evolutivos	117
D Algoritmo memético multi-objetivo para inferencia de árboles filogenéticos	119
D.1 Parámetros para estrategias de búsqueda local	119
D.1.1 Número de iteraciones para algoritmo de búsqueda local	119
D.1.2 Parámetros para algoritmo SA	120
D.2 Herramientas basadas en optimización de objetivo único	122
D.3 Herramientas basadas en optimización de múltiples objetivos	124
D.4 Condición de balance en operadores de cruzamiento	127
D.5 Comparación de operadores de cruzamiento y mutación	127
E Métricas de rendimiento para estrategias multi-objetivo	129
E.1 Hipervolumen	129
E.2 Representatividad de las soluciones en la Frontera de Pareto	130
E.3 Cobertura	130
F Tomadores de decisiones multi-objetivo	132
G Tiempo de ejecución MO-PhyNet	133

ÍNDICE DE TABLAS

Tabla 2.1	Conjuntos de datos empleados en experimentos I	27
Tabla 2.2	Métrica de Robinson-Foulds para operadores de cruzamiento	28
Tabla 2.3	Métrica de Robinson-Foulds para operadores genéticos	28
Tabla 2.4	Mejores configuraciones MO-MA según métrica de hipervolumen	30
Tabla 2.5	Peores configuraciones MO-MA según métrica de hipervolumen	31
Tabla 2.6	Comparación de MO-MA y otras propuestas: objetivo único	32
Tabla 2.7	Comparación de MO-MA y otras propuestas: multi-objetivo	33
Tabla 3.1	Conjuntos de datos empleados en experimentos I	52
Tabla 3.2	Conjuntos de datos empleados en experimentos II	52
Tabla 3.3	Parametrización MO-CS	53
Tabla 3.4	Comparación métodos para integración de datos en inferencia filogenética II	58
Tabla 3.5	Criterios de MO-CS y métrica Kendall-Colijn II	59
Tabla 3.6	MO-CS y algoritmo genético multi-objetivo (verosimilitud)	60
Tabla 4.1	Agrupamientos criterios optimización - topología: Robinson-Foulds	64
Tabla 4.2	Agrupamientos criterios optimización - topología: Diferencia de caminos	65
Tabla 4.3	Agrupamientos criterios optimización - topología: Kendall-Colijn	65
Tabla 5.1	Conjuntos de datos empleados en experimentos I	81
Tabla 5.2	Conjuntos de datos empleados en experimentos II	81
Tabla 5.3	Evaluación MO-PhyNet I	82
Tabla 5.4	Evaluación MO-PhyNet II	82
Tabla 5.5	Evaluación MO-PhyNet III	82
Tabla 5.6	Espacio de soluciones y espacio objetivo	85
Tabla 5.7	Resultados ConsensusNet	88
Tabla A.1	Número de publicaciones relacionada en Pubmed: 2007-2017.	109
Tabla A.2	Número de publicaciones relacionada en ScienceDirect: 2007-2017.	109
Tabla A.3	Número de publicaciones relacionada en ProQuest: 2007-2017.	109
Tabla D.1	Parametrización MO-MA: herramientas optimización objetivo único	123
Tabla D.2	Tiempo de ejecución NSGA-II EM	123
Tabla D.3	Parametrización MO-MA: herramientas optimización multi-objetivo	125
Tabla G.1	Tiempo de ejecución MO-PhyNet.	133

ÍNDICE DE ILUSTRACIONES

Figura 1.1	Áreas de conocimiento en inferencia filogenética	6
Figura 1.2	Áreas de conocimiento y desarrollo de tesis	9
Figura 2.1	Áreas de conocimiento y desarrollo de tesis	13
Figura 2.2	Tipos de árboles filogenéticos	15
Figura 2.3	Clasificación de métodos para inferencia de árboles filogenéticos	16
Figura 2.4	Problema multi-objetivo de inferencia de árboles filogenéticos	17
Figura 2.5	Operadores de cruzamiento	23
Figura 2.6	Comparación MO-MA con herramientas mono-objetivas	34
Figura 2.7	Comparación MO-MA con herramientas multi-objetivas	35
Figura 2.8	Inferencia filogenética de ureasas	36
Figura 3.1	Áreas de conocimiento y desarrollo de tesis	41
Figura 3.2	Métodos para combinación de evidencia biológica	44
Figura 3.3	Evaluación de métodos para combinación de evidencia biológica	50
Figura 3.4	Espacio de árboles para métodos de combinación de datos	57
Figura 3.5	Espacio de árboles para métodos de combinación de datos multi-objetivo	60
Figura 4.1	Áreas de conocimiento y desarrollo de tesis	61
Figura 4.2	Tomadores de decisiones I	67
Figura 4.3	Tomadores de decisiones II	67
Figura 4.4	Tomadores de decisiones III	68
Figura 4.5	Tomadores de decisiones IV	68
Figura 4.6	Tomadores de decisiones V	69
Figura 5.1	Áreas de conocimiento y desarrollo de tesis	71
Figura 5.2	Tipos de redes filogenéticas	73
Figura 5.3	Esquema con métodos para inferencia de redes filogenéticas	75
Figura 5.4	Fronteras de Pareto - inferencia de redes filogenéticas	83
Figura 5.5	Fronteras de Pareto - inferencia de redes filogenéticas	84
Figura 5.6	Matrices de distancia - inferencia de redes filogenéticas	86
Figura 5.7	MO-PhyNet y ConsensusNet	87
Figura C.1	Métodos aglomerativos	112
Figura C.2	Métodos aditivos	113
Figura C.3	Mínimos cuadrados y evolución mínima	115
Figura C.4	Máxima parsimonia y verosimilitud	117
Figura C.5	Modelos evolutivos	118
Figura D.1	Parámetrización de búsqueda local en MO-MA	121
Figura D.2	Comparación entre MO-MA y NSGA-II EM	126
Figura D.3	Operadores de cruzamiento y métrica Robinson-Foulds	127
Figura D.4	Operadores genéticos y métrica Robinson-Foulds	128
Figura E.1	Métrica de hipervolumen	129
Figura E.2	Métrica de Representatividad de las soluciones en Frontera de Pareto	130
Figura E.3	Métrica de Cobertura	131

ÍNDICE DE ALGORITMOS

Algoritmo 2.1	Algoritmo memético multi-objetivo (MO-MA)	20
Algoritmo 3.1	Algoritmo genético multi-objetivo (MO-CS)	46
Algoritmo 5.1	Algoritmo genético multi-objetivo (MO-PhyNet)	77
Algoritmo D.1	Algoritmo Pareto local search	119
Algoritmo D.2	Algoritmo Simulated Annealing	122

CAPÍTULO 1. INTRODUCCIÓN

1.1 ANTECEDENTES Y MOTIVACIÓN

Desde que aparecieron los primeros organismos hace aproximadamente 4.000 millones de años, la vida ha explorado múltiples caminos. A pesar de que solo algunos de ellos tuvieron éxito y perduraron, consiguieron generar la gran diversidad de organismos existentes en la actualidad: aproximadamente 8,7 millones de especies eucariontes (Mora et al., 2011) y probablemente cientos de millones de especies de procariontes (Sadava et al., 2011). Todo el proceso de cambios y adaptaciones ha quedado registrado en el genoma de las especies actuales (Abascal et al., 2014), permitiendo explorar la historia de la vida según el método científico.

El conjunto de metodologías que posibilita establecer una hipótesis o aproximación a la historia evolutiva entre un conjunto de organismos se denomina inferencia filogenética. Los avances en las técnicas de secuenciamiento durante las últimas décadas, junto a la creciente disponibilidad de grandes volúmenes de datos, han permitido inferir filogenia considerando un gran número de organismos (Hinchliff et al., 2015) y diferente evidencia biológica: información morfológica, conductual, ecológica y datos moleculares (alineamientos múltiples de nucleótidos, aminoácidos y otros marcadores moleculares) (Grechko, 2002; Wilgenbusch et al., 2017).

El proceso de inferencia filogenética ha aportado al conocimiento de diversas áreas como: biología evolutiva, ecología, biomedicina, paleontología, antropología y bioquímica (Eguiarte, 2007; Santander-Jiménez & Vega-Rodríguez, 2013a). Particularmente, la reconstrucción filogenética se ha convertido en un componente indispensable en estudios comparativos que tratan de establecer, por ejemplo, si las características de un organismo son producto de la selección natural o de otro proceso evolutivo, conocer la secuencia de cambios que ha tenido una característica de un organismo en el tiempo, determinar si organismos relacionados ecológicamente constituyen un ejemplo de co-evolución, hacer estudios de biogeografía o filogeografía, determinar relaciones de homología al comparar genomas, estimar tiempos de divergencia, reconstruir e identificar proteínas ancestrales, detectar puntos de recombinación en virus, identificar mutaciones asociadas a enfermedades y patógenos emergentes (Yang & Rannala, 2012; Eguiarte, 2007). Estudios más recientes han utilizado inferencia filogenética para clasificar secuencias de meta-genomas, identificar genes, reconstruir genomas ancestrales, elementos reguladores y secuencias de ARN no codificante (Yang & Rannala, 2012). Chile no ha estado ajeno a estos avances, aplicando inferencia filogenética en diversas áreas como: microbiología (Thiel et al., 2010; Venegas et al., 2011; Torres et al., 2016), ecología (Rodríguez et al., 2014; Tapia et al., 2015), filogeografía (Vila et al., 2013; Ulloa et al., 2017), entre otras.

La inferencia filogenética se ha convertido en uno de los problemas computacional-

mente complejos de tratar en el campo de la bioinformática (Zararsiz & Coşgun, 2014). Desde los primeros algoritmos exhaustivos desarrollados para efectuar inferencia filogenética por medio de árboles binarios (Fitch & Margoliash, 1967), hasta los últimos modelos basados en optimización multi-objetivo mediante meta-heurísticas, se han realizado grandes avances en términos de desempeño y calidad de resultados. Sin embargo, el aporte propio del conocimiento ha generado nuevos desafíos en el modelamiento, como la reducción del sesgo asociado a la elección de diferentes criterios de calidad para escoger una hipótesis evolutiva, el desarrollo de modelos que sean capaces de integrar evidencia biológica, o la consideración de diferentes mecanismos de transferencia de material genético.

La creciente tasa de publicaciones en el área, los desafíos pendientes a nivel biológico y computacional, y el potencial campo de aplicación en el contexto nacional e internacional, hacen en general del área una fuente importante de desarrollo e investigación. Particularmente, la motivación para abordar esta temática como investigación de tesis doctoral se fundamenta en los siguientes aspectos:

- **Contemporaneidad y actualidad.** El problema de inferencia filogenética y el modelamiento basado en optimización multi-objetivo son áreas de interés para el desarrollo de investigación, reflejado en un considerable número de publicaciones por año (Anexo A).
- **Potencial contribución de conocimiento.** Mejorar el modelamiento de inferencia filogenética, sobretodo a nivel algorítmico, puede llevar a contribuciones prácticas a nivel multi-disciplinario.
- **Generalización de modelos.** Debido a que se trata de un problema computacionalmente complejo, la inferencia filogenética es un problema ideal para el diseño y evaluación de modelos, permitiendo que las metodologías desarrolladas y los algoritmos propuestos puedan ser utilizados para tratar problemas de igual o menor complejidad en bioinformática u otros campos de investigación.
- **Originalidad.** Abordar la problemática mediante la propuesta de solución desarrollada en este trabajo, satisface al menos tres de las nueve definiciones de originalidad para una investigación propuestas por Phillips (1992)(Anexo B).
- **Formación:** La combinación de inferencia filogenética con el modelamiento basado en optimización multi-objetivo enlaza perfectamente las áreas de la biología e ingeniería, permitiendo continuar el desarrollo de la profesión del autor como ingeniero biomédico, manteniendo su línea de investigación dentro de las áreas de bioinformática y biología computacional.

1.2 DESCRIPCIÓN DEL PROBLEMA

En la actualidad la mayoría de las hipótesis evolutivas son representadas por medio de árboles filogenéticos (Hinchliff et al., 2015). Para esto se debe asumir que las especies han evolucionado a partir de un ancestro común a través de un proceso de ramificación simple, transmitiendo exclusivamente información entre padres e hijos. Sin embargo, el proceso evolutivo es más complejo, por lo que este tipo de representación omite una serie de fenómenos biológicos significantes (Hinchliff et al., 2015). Específicamente, imposibilita el estudio de la evolución considerando diferente evidencia biológica, y descarta mecanismos de transferencia complejos como: paralogías ocultas, hibridación, ordenamiento incompleto de linaje debido a divergencia, recombinación, transferencia horizontal de genes, entre otros (Smith et al., 2015). Para abordar estos fenómenos se ha desarrollado diferentes estrategias que permiten obtener representaciones reticulares de hipótesis evolutivas, reemplazando los árboles por topologías en forma de red (Makarenkov & Legendre, 2004; Horiike et al., 2011; Yu & Nakhleh, 2015; Wheeler, 2015; Albrecht, 2015). No obstante, independiente de cual de ellas se seleccione para efectuar inferencia filogenética, la topología resultante se encuentra condicionada por diferentes factores que sesgan la hipótesis evolutiva obtenida (Swofford et al., 2001; Rokas et al., 2003), como por ejemplo:

1. El criterio de optimización aplicado para la selección de una determinada topología.
2. La evidencia biológica empleada para inferir una hipótesis evolutiva.
3. El modelo evolutivo seleccionado para el cálculo de verosimilitud.
4. El paradigma usado para la combinación de la evidencia biológica.
5. La interpretación del sentido de una topología reticulada.
6. La selección de una topología determinada, cuando múltiples soluciones satisfacen un criterio previamente definido.

A raíz de estos antecedentes se puede formular la siguiente pregunta de investigación, ¿Es posible modelar inferencia filogenética considerando fenómenos reticulares, reduciendo fuentes de sesgo, e integrando diferentes criterios de optimalidad?

1.3 HIPÓTESIS Y OBJETIVOS

1.3.1 Hipótesis

Un modelo computacional basado en optimización multi-objetivo permite efectuar inferencia filogenética considerando fenómenos reticulares, reduciendo el sesgo asociado a la elección de modelos basados en optimización de objetivo único, encontrando soluciones de calidad en términos de dominancia.

1.3.2 Objetivo general

Diseñar un modelo computacional para efectuar inferencia filogenética en organismos, considerando: representación de fenómenos reticulares, integración de evidencia biológica conflictiva entre sí, y múltiples criterios de optimalidad.

1.3.3 Objetivos específicos

1. Modelar computacionalmente fenómenos biológicos reticulares para ser incorporados en el proceso de inferencia filogenética basada en topología de red.
2. Diseñar e implementar un modelo computacional multi-objetivo para efectuar inferencia filogenética mediante árboles y construir una línea base respecto al estado del arte.
3. Evaluar cuantitativamente la capacidad de los algoritmos multi-objetivos para combinar evidencia biológica conflictiva y efectuar inferencia filogenética.
4. Diseñar e implementar mecanismos para reducción de espacio de búsqueda de soluciones y toma de decisiones, evaluando su rendimiento en los algoritmos propuestos en el estado del arte.
5. Diseminar el conocimiento generado por medio de publicaciones.

1.3.4 Alcances y limitaciones

La investigación propuesta en este documento plantea el modelamiento de inferencia filogenética considerando las características de solución descritas en la Sección 1.4. Sin embargo, no es parte de esta investigación:

1. La comparación funcional de herramientas, entornos, lenguajes de programación, y bibliotecas disponibles para efectuar inferencia filogenética o modelamiento multi-objetivo.
2. La implementación de algoritmos paralelos o distribuidos para cada uno de los modelos.
3. El desarrollo de una herramienta computacional, comercial o no, que considere los modelos desarrollados.
4. La contrastación de resultados mediante validación experimental *in vivo* o *in vitro*.
5. El desarrollo de modelos matemáticos para ser empleados como criterios en la generación de topologías reticulares.

1.4 PROPUESTA DE INVESTIGACIÓN

La propuesta desarrollada a lo largo de esta investigación consiste en un nuevo modelo basado en optimización multi-objetivo, que permite efectuar inferencia filogenética representando fenómenos reticulares bajo diferentes paradigmas para la combinación de evidencia biológica. Específicamente, se propone un modelo basado en un algoritmo genético multi-objetivo capaz de emplear conjuntos de datos provenientes de diferentes fuentes de evidencia biológica (almacenados como secuencias de caracteres, matrices de distancia, o árboles filogenéticos binarios), para construir hipótesis evolutivas reticuladas que satisfacen diferentes criterios de optimización conflictivos entre sí. Estos criterios son independientes de la elección de un modelo evolutivo particular.

Desarrollar una estrategia para comprobar la hipótesis planteada no es una tarea trivial, debido a que se trata de enfrentar un problema que pertenece a un área de conocimiento que no ha sido explorada previamente por la literatura. Las áreas más cercanas se limitan a la investigación del proceso de inferencia de árboles filogenéticos, donde la gran mayoría de las propuestas resultantes emplean un criterio exclusivo para construir una hipótesis evolutiva. Avances más recientes reconocen la existencia de sesgo asociado a estos modelos, y han propuesto el uso de nuevas estrategias basadas en optimización multi-objetivo dependiendo de

una fuente biológica en particular (Sección ??). Paralelamente, un área diferente de investigación se ha encargado del diseño de estrategias para inferir árboles filogenéticos empleando diferentes paradigmas de combinación de evidencia biológica (Sección 3.1). Otros autores han señalado que la representación por medio de árboles es insuficiente para modelar mecanismos evolutivos complejos, desarrollando el área relacionada a la inferencia de redes filogenéticas (Sección 5.1.2). La Figura 1.1 representa por medio de un cladograma las relaciones entre las principales publicaciones vinculadas con el diseño de estrategias para inferencia filogenética. El desarrollo de esta investigación aparece como una nueva arista que integra las áreas de combinación de evidencia biológica y el desarrollo de estrategias para representar fenómenos reticulares.

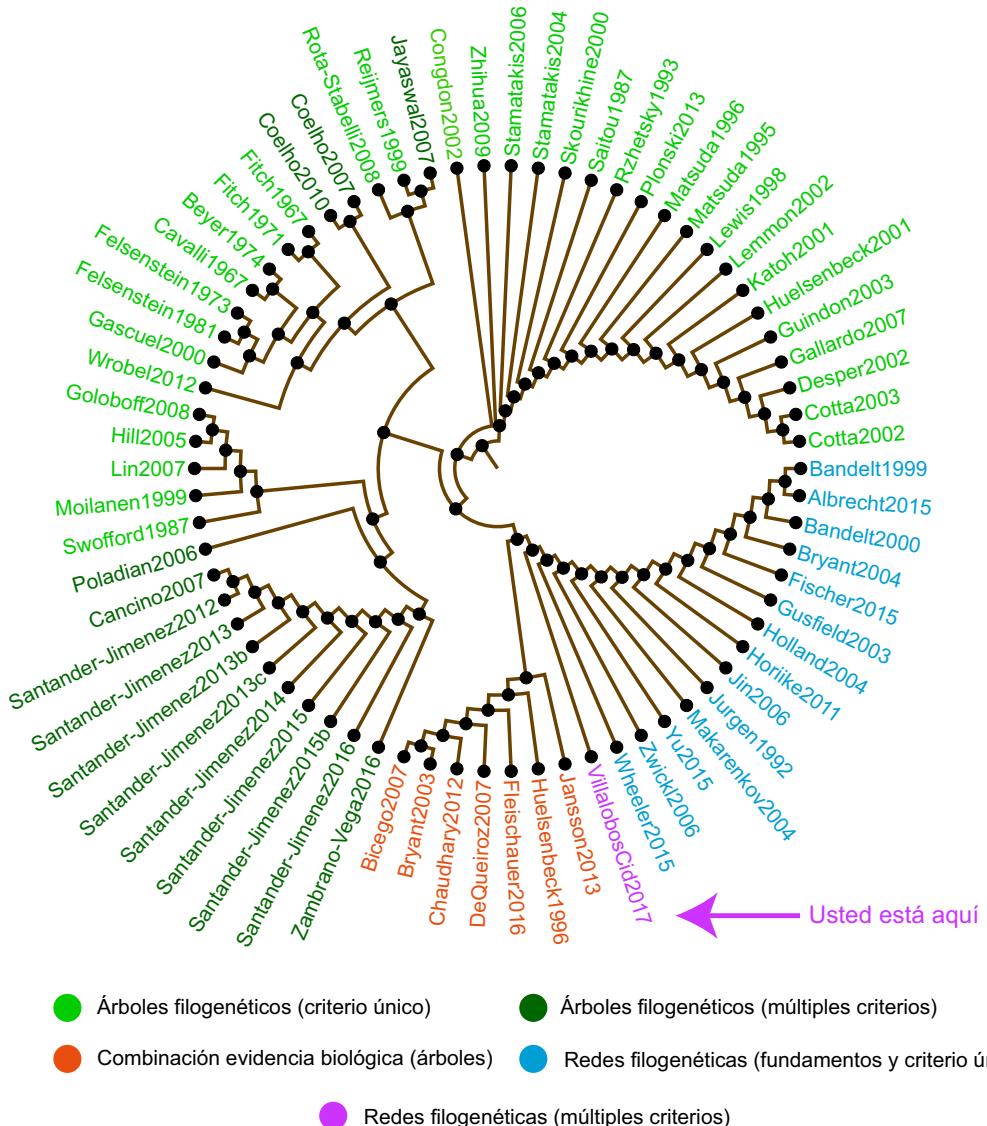


Figura 1.1: Áreas de conocimiento y principales publicaciones relacionadas al problema de inferencia filogenética.

Fuente: Elaboración propia, 2017.

Ante la incertidumbre asociada al enfrentar un problema en una nueva área de investigación, asociado al tiempo limitado que se dispone para su desarrollo en el contexto de un trabajo de tesis doctoral, se opta por emplear una estrategia que permita el desarrollo de la propuesta, maximizando la probabilidad de éxito en la comprobación de la hipótesis de investigación acorde al método científico empírico analítico: el método heurístico de Polya (Pólya, 2004). Esta estrategia plantea que, si un problema es muy complejo o no se consigue hallar una solución, este puede ser dividido en subproblemas o resolverse indirectamente mediante otros problemas relacionados de menor complejidad. Bajo esa lógica, el problema de inferencia multi-objetivo de redes filogenéticas puede ser descompuesto en los subproblemas relacionados a las áreas del conocimiento mostradas en la Figura 1.1, siendo abordado por esta investigación bajo los siguientes enfoques:

- **Inferencia multi-objetivo de árboles filogenéticos.** Las recientes estrategias diseñadas para inferir árboles filogenéticos basándose en optimización multi-objetivo han conseguido reducir el sesgo asociado a la dependencia del criterio de optimización empleado para la selección de una determinada topología. Sin embargo, no se ha evaluado las diferentes configuraciones y operadores que componen estas propuestas, generando dificultad a la hora de evaluar su desempeño. Por otro lado, la aplicación del proceso de inferencia ha sido limitado al análisis de secuencias de nucleótidos, excluyendo el uso de otra evidencia biológica. A raíz de ello, esta investigación propone contrastar el desempeño de las diferentes propuestas caracterizando sus operadores y componentes. Con los resultados de esta evaluación se desarrolla una estrategia multi-objetivo propia, que es comparada con los actuales modelos propuestos en la literatura: optimización de objetivo único y múltiple.
- **Combinación de evidencia biológica.** La integración de diferente evidencia biológica y/o hipótesis evolutivas se puede efectuar considerando diferentes paradigmas de combinación. Las estrategias derivadas de cada uno de ellos poseen topologías resultantes sesgadas, que dependen del paradigma de combinación empleado y el criterio de optimización usado para el proceso de inferencia. Por otro lado, algunas estrategias presentan inconsistencias en la consideración de la información biológica a lo largo de las ramas de los árboles filogenéticos construidos. Para abordar estas dificultades, la presente investigación propone un modelo basado en optimización multi-objetivo que integra estos paradigmas considerando cualquier tipo de evidencia biológica, presentando soluciones en base a múltiples objetivos, e incluir así información evolutiva sobre la topología total del árbol resultante. La capacidad de la propuesta para inferir hipótesis evolutivas integrales es comparada con otros métodos propuestos en la literatura.

- **Selección de soluciones y reducción de espacio de búsqueda.** Si bien, esta no es un área de investigación propia de la inferencia filogenética como las presentadas en la Figura 1.1, el modelamiento de inferencia filogenética basado en optimización multi-objetivo da origen a dos nuevas aristas en la investigación: (1) la identificación de mecanismos para reducir el espacio de búsqueda de soluciones con el fin de mejorar el desempeño de las aproximaciones, y (2) el estudio de estrategias para reducir el sesgo asociado a la elección de una solución representativa desde una Frontera de Pareto. A raíz de ello, este trabajo estudia la relación entre el espacio de soluciones y los diferentes criterios empleados para inferencia filogenética, determinando la factibilidad de aplicar topologías inferidas previamente para mejorar la convergencia de las propuestas. También se aplican diferentes métodos para la toma de decisiones que han sido usados exitosamente en otras áreas de investigación, proponiendo una nueva estrategia que selecciona topologías desde el espacio de soluciones para obtener árboles filogenéticos representativos. Finalmente se analiza la relación entre las topologías de los árboles filogenéticos seleccionados, y su representación en el espacio objetivo.
- **Inferencia multi-objetivo de redes filogenéticas.** Con las herramientas y conocimiento obtenido al tratar los subproblemas previamente definidos, finalmente se aborda el problema relacionado a la hipótesis de esta investigación. Para ello se desarrolla un modelo basado en optimización multi-objetivo para inferencia de fenómenos reticulares, que integra los operadores topológicos que son estudiados en las áreas anteriores de conocimiento. Esta propuesta considera diferentes criterios que no dependen de la selección de un modelo evolutivo. Además, es capaz de integrar evidencia biológica empleando diferentes paradigmas de combinación, incluyendo diversas interpretaciones para la topología de una red. La estrategia es evaluada usando diferentes tipos de evidencia biológica proveniente de conjuntos de datos reales, mientras que su desempeño es comparado con otras propuestas de la literatura.

La Figura 1.2 presenta las relaciones entre las diferentes áreas de conocimiento abordadas en este trabajo. Para su desarrollo, se exploran paralelamente tres áreas del conocimiento: inferencia multi-objetivo de árboles filogenéticos, combinación de evidencia biológica e inferencia filogenética considerando fenómenos reticulares. El conocimiento adquirido de las dos primeras de ellas es integrado, y se emplea para estudiar la relación entre espacio de soluciones y espacio objetivo en inferencia filogenética. Conjuntamente se evalúa el comportamiento de diferentes tomadores de decisiones. Finalmente, todas las áreas son incorporadas, desarrollando el enfoque multi-objetivo de inferencia filogenética considerando fenómenos reticulares.

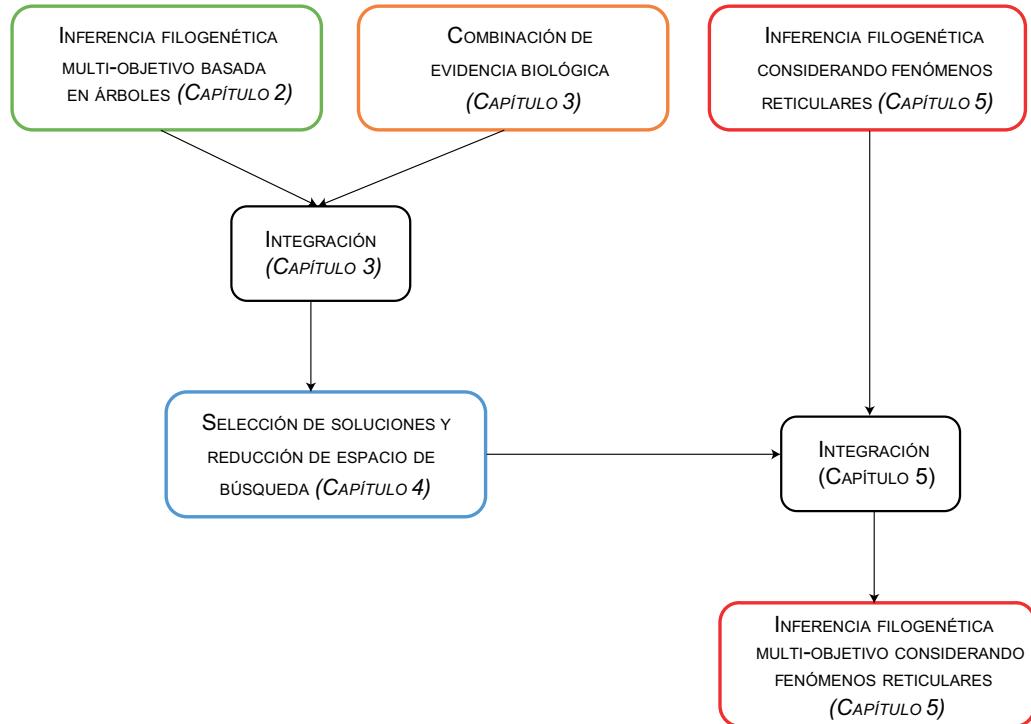


Figura 1.2: Relaciones entre áreas del conocimiento y desarrollo de tesis.
Fuente: Elaboración propia, 2017.

1.5 ORGANIZACIÓN DEL DOCUMENTO

A excepción del capítulo final, la estructura de este documento está asociada al estudio de cada una de las áreas del conocimiento descritas en la sección anterior (subproblemas según la definición formal de Polya):

- Capítulo 2: Inferencia multi-objetivo de árboles filogenéticos.
- Capítulo 3: Combinación de evidencia biológica en inferencia filogenética.
- Capítulo 4: Reducción de espacio de búsqueda y tomadores de decisiones.
- Capítulo 5: Inferencia filogenética multi-objetivo considerando fenómenos reticulares.

Cada capítulo presenta una organización común que describe:

- Una introducción al problema abordado.
- Los conceptos básicos requeridos para entender cada uno de los problemas.
- La formulación del problema.

- Una definición detallada de las estrategias propuestas para su modelamiento o estudio.
- Las herramientas y recursos empleados.
- La descripción de los diferentes experimentos realizados para evaluar el desempeño de las propuestas.
- La presentación de resultados y un resumen de las conclusiones obtenidas.

El capítulo final (Capítulo 6) expone las principales conclusiones asociadas al proceso general de investigación, determina el alcance de los resultados y establece las nuevas áreas de investigación para trabajos futuros. Al término del documento se incluyen anexos que complementan el contenido de los capítulos. Estos son:

- **ANEXO A:** muestra el número de publicaciones asociadas a diferentes palabras claves relacionadas a inferencia filogenética.
- **ANEXO B:** presenta diferentes aristas que definen la originalidad y contribución en un proceso de investigación.
- **ANEXO C:** incluye el detalle de diferentes métodos para reconstrucción de árboles filogenéticos.
- **ANEXO D:** describe los detalles de un algoritmo memético multi-objetivo diseñado para inferir filogenia basándose en árboles filogenéticos.
- **ANEXO E:** detalla las diferentes métricas de rendimiento empleada para comparar el desempeño de las estrategias multi-objetivo en esta investigación.
- **ANEXO F:** describe los principales tomadores de decisiones multi-objetivo propuestos en la literatura.
- **ANEXO G:** especifica los tiempos de ejecución empleados por MO-PhyNet.

CAPÍTULO 2. INFERENCIA MULTI-OBJETIVO DE ÁRBOLES FILOGENÉTICOS

Un árbol filogenético es una representación de una hipótesis que busca explicar las relaciones evolutivas entre un conjunto de especies (Sección 2.1.1). Desde un punto de vista de optimización, el problema de inferencia filogenética consiste en encontrar el árbol que satisface en mayor medida un determinado criterio de selección entre todas las posibles topologías de árboles que se pueden construir dado un conjunto de organismos. Este criterio de selección puede estar basado en diferentes principios: evolución mínima, mínimos cuadrados, máxima parsimonia, y verosimilitud. Cada uno de ellos puede resultar en diversas topologías de árbol para un mismo conjunto de especies estudiadas (Sección 2.1.2).

La aplicación de aproximaciones exhaustivas para la búsqueda de un árbol filogenético óptimo ha resultado ser infactible, debido al elevado número de árboles filogenéticos que se puede construir al combinar las diferentes especies de un conjunto de datos. Este problema ha sido clasificado en teoría computacional como un problema NP-duro (Poladian & Jermiin, 2006). Las primeras aproximaciones prácticas consideraron solo un criterio de optimización y se basaron en algoritmos evolutivos o bio-inspirados (Gascuel, 2000; Santander-Jiménez & Vega-Rodríguez, 2013a). Esto permitió encontrar soluciones aproximadas para instancias de problemas que incluían grandes volúmenes de datos (Santander-Jiménez & Vega-Rodríguez, 2013a). Sin embargo, algunos puntos permanecieron sin resolver como el sesgo asociado a la elección de un criterio determinado (Swofford et al., 2001), dependencia de la evidencia biológica estudiada y del modelo evolutivo seleccionado (Cancino & Delbem, 2007). El desarrollo de algoritmos probabilísticos entregó solución a estos problemas proponiendo árboles de consenso. Sin embargo, estos requirieron de conocimiento previo respecto a la ponderación de cada una de las fuentes de incongruencia, además de criterios no conflictivos entre sí. El principal problema de estas metodologías es que pueden descartar soluciones biológicamente significativas (Poladian & Jermiin, 2006).

Handl et al. (2007) señalaron las ventajas de emplear optimización multi-objetivo en bioinformática y biología computacional respecto a aproximaciones basadas en objetivo único: minimización de la probabilidad de estancamiento en mínimos locales y zonas sin gradientes, reducción del efecto de ruido en los datos, e incorporación de múltiples fuentes u objetivos que pueden presentar conflicto entre sí (Sección 2.1.3). Los métodos más recientes desarrollados para inferir árboles filogenéticos se basan en optimización multi-objetivo (Poladian & Jermiin, 2006; Coelho & Zuben, 2007; Cancino & Delbem, 2007; Jayaswal et al., 2007; Coelho et al., 2010; Santander-Jiménez & Vega-Rodríguez, 2013a,c,b, 2014, 2016; Zambrano-Vega et al., 2016). Estas propuestas consideran una amplia gama de meta-heurísticas, diferentes métodos

para inicializar poblaciones, y diversos operadores genéticos para cruzamiento y mutación (Sección ??). Sin embargo, pocos trabajos han evaluado el rendimiento de estas estrategias de reordenamiento topológico como operadores genéticos (Santander-Jiménez et al., 2012; Santander-Jiménez & Vega-Rodríguez, 2013c,b), así como la parametrización de los algoritmos. Esto genera incertidumbre al momento de proponer nuevos modelos para inferencia filogenética, imposibilitando establecer conclusiones al comparar desempeño, ya que se desconoce si los resultados se deben a la estructura general de los modelos, las meta-heurísticas empleadas como base, o a la aplicación de un determinado operador.

La inferencia multi-objetivo de árboles filogenéticos corresponde a un problema particular de la representación multi-objectivo de fenómenos reticulares (Figura 2.1). Su entendimiento permite extraer conocimiento para el modelamiento de fenómenos evolutivos complejos, aportando al objetivo final de esta tesis doctoral. Bajo esta premisa, con el propósito de tener una mayor certidumbre respecto a las características de los modelos actuales y sus operadores, se propone un algoritmo memético multi-objetivo (MO-MA) basado en el Algoritmo genético de ordenamiento no dominado (*Non dominated Sorting Genetic Algorithm II, NSGA-II* (Deb et al., 2002)) (Sección 2.2). Este permite la inferencia de árboles filogenéticos considerando dos criterios: máxima parsimonia y verosimilitud. Su diseño implica la comparación y evaluación de diferentes configuraciones y operadores: cuatro métodos basados en distancia para construir las topologías iniciales, múltiples estrategias de reordenamiento de árboles, y dos algoritmos de búsqueda local (Sección 2.3.1 a 2.3.3). Por otro lado, empleando conjuntos de datos clásicos de la literatura relacionada, se compara MO-MA con otros métodos basados en optimización de objetivo único y múltiple (Sección 2.3.4 y 2.3.5). Finalmente se presenta una evaluación de MO-MA usando conjuntos de datos de aminoácidos (Sección 2.3.6). Esto resulta relevante ya que todas las propuestas actuales trabajan exclusivamente con secuencias de nucleótidos, a pesar de que sus modelos evolutivos son diferentes. Las principales contribuciones de esta investigación en relación a los trabajos previamente publicados son:

- Una robusta evaluación de una estrategia basada en NSGA-II adaptada para enfrentar el problema de inferencia filogenética de árboles, considerando diferentes algoritmos de búsqueda local y múltiples operadores de cruzamiento y mutación.
- Una caracterización de diferentes estrategias de cruzamiento, mutación y búsqueda local aplicadas al problema de inferencia multi-objetivo de árboles filogenéticos, y la determinación de sus efectos en el proceso de búsqueda de soluciones.
- La propuesta de un nuevo operador de cruzamiento que combina los parámetros de modelos evolutivos empleados en el cálculo de verosimilitud.
- La obtención de nuevas soluciones para conjuntos de datos de la literatura, mejorando

métricas clásicas del estado del arte para evaluar estrategias que optimizan uno y múltiples objetivos. Esta contribución hace de MO-MA una real alternativa para el campo.

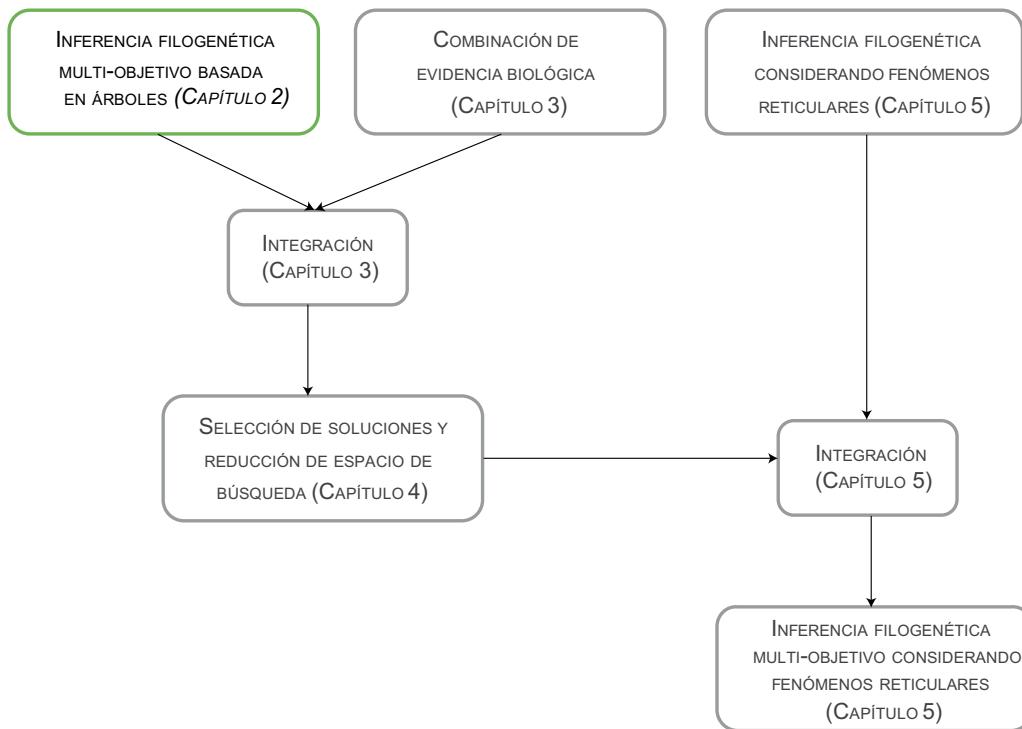


Figura 2.1: Relaciones entre áreas del conocimiento y desarrollo de tesis.

Fuente: Elaboración propia, 2017.

2.1 INFERENCIA FILOGENÉTICA BASADA EN ÁRBOLES

2.1.1 Árboles filogenéticos

Un árbol filogenético es una representación de una hipótesis que explica las relaciones evolutivas entre un conjunto de especies (o en casos específicos, entre moléculas, por ejemplo, evolución de proteínas, identificación de meta-genomas y genes). Los nodos de un árbol filogenético corresponden a los diferentes organismos estudiados, mientras que la longitud de las ramas representa tiempo o distancia evolutiva. Dependiendo de esto último, un árbol filogenético puede ser enraizado o sin enraizar (Figura 2.2a y Figura 2.2b). El primero asume la existencia de un antepasado común y a partir de él se establece una dirección del proceso evolutivo. Por el contrario, un árbol sin enraizar muestra la relación evolutiva entre los organismos sin un ancestro común (Strachan & Andrew, 2011). El número de posibles topologías de árboles enraizados (nr) o

sin enraizar (nu) a inferir dado el número n de especies puede ser calculado según las Ecuaciones 2.1 y 2.2. Estas permiten comprender la complejidad asociada a la combinación de especies y el número de topologías posibles; por ejemplo, para sólo 20 especies se puede construir $8,2 \times 10^{21}$ topologías de árboles enraizados.

El problema de encontrar la mejor topología ante un determinado criterio ha sido categorizado como NP-duro (Day, 1987), demostrándose su equivalencia en complejidad con el problema de Triangulación de grafos coloridos (*Triangulating Colored Graphs, TCG*) (Wernicke, 2003).

$$nr(n) = \frac{(2n - 3)!}{(n - 2)!2^{n-2}} \quad (2.1)$$

$$nu(n) = \frac{(2n - 5)!}{(n - 3)!2^{n-3}} \quad (2.2)$$

En un árbol filogenético la suma de la longitud de las ramas que separan dos organismos se conoce como distancia filética o distancia patrística. Se denomina árbol filogenético aditivo cuando la distancia observada entre dos organismos es igual a la suma de las ramas que separan estos mismos a nivel topológico (Figura 2.2c). Si la distancia entre todos los organismos estudiados resulta aditiva, el árbol es considerado como un árbol filogenético ultramétrico (Figura 2.2d). Según su estructura, los árboles filogenéticos pueden ser clasificados como n-arios (multifurcados) o binarios (bifurcados), siendo estos últimos los más usados debido a su simple interpretación.

2.1.2 Métodos para inferencia de árboles filogenéticos

La inferencia de árboles filogenéticos puede ser efectuada usando aproximaciones basadas en distancias o en caracteres (Figura 2.3). Las primeras construyen árboles filogenéticos usando como entrada una matriz de distancia entre cada organismo (derivada de datos moleculares, anatómicos, entre otros). Dentro de las aproximaciones basadas en distancia se encuentra estrategias que permiten construir árboles filogenéticos ultramétricos, como los métodos *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA) y *Weighted Pair Group Method with Arithmetic Mean* (WPGMA). También son parte de esta aproximación, estrategias golosas que permiten inferir árboles filogenéticos aditivos, como *Neighbor joining* (NJ) y su versión modificada *Bio-neighbor joining* (BioNJ) (de Bruyn et al., 2014). Otras aproximaciones basadas en distancia consideran la optimización de un criterio como el principio de evolución mínima (Kidd & Sgaramella-Zonta, 1971; Rzhetsky & Nei, 1993) o los mínimos cuadrados (Cavalli-Sforza &

Edwards, 1967; Fitch & Margoliash, 1967). La primera de ellas busca el árbol cuya topología minimiza la sumatoria del largo total de sus ramas, mientras que la segunda busca el árbol con menor sumatoria de la diferencia entre los pares de distancia observadas y patrísticas.

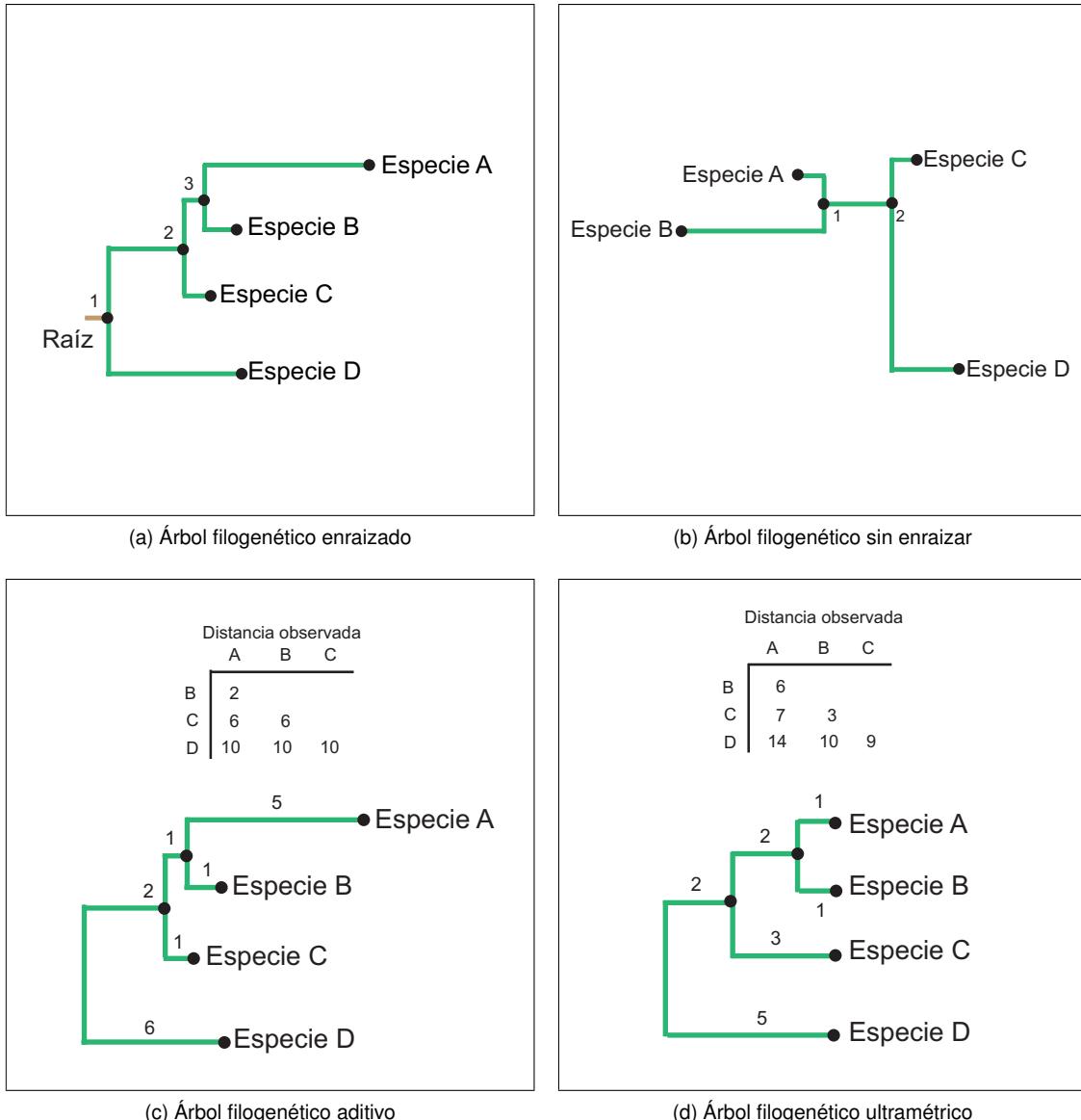


Figura 2.2: Esquema de tipo de árboles filogenéticos
Fuente: Elaboración propia, 2017.

Las aproximaciones basadas en caracteres usan como entrada un conjunto de secuencias alineadas y la información evolutiva de sus caracteres. Para ello se ha propuesto diferentes criterios de optimización como: máxima parsimonia, máxima verosimilitud y métodos bayesianos (Felsenstein, 2004). Máxima parsimonia busca el árbol filogenético que minimiza el número de cambios necesarios para explicar los datos de entrada. En cambio, máxima

verosimilitud selecciona el árbol más probable en relación a los datos observados respecto a un modelo evolutivo previamente determinado (de Bruyn et al., 2014). Por último, los métodos bayesianos se basan en la probabilidad posterior que es obtenida con el producto entre verosimilitud y la probabilidad previa. El Anexo C contiene una explicación detallada para cada una de las aproximaciones presentadas en esta subsección.

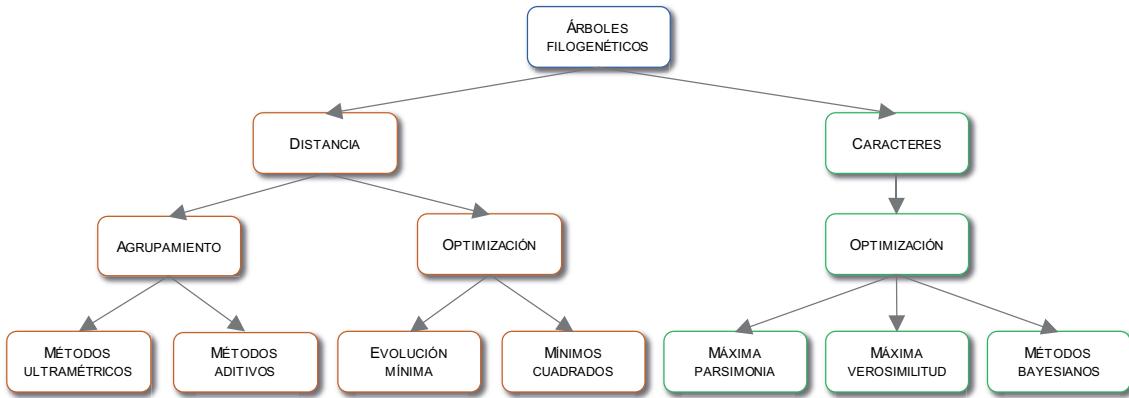


Figura 2.3: Clasificación de métodos para inferencia de árboles filogenéticos.

Fuente: Elaboración propia, 2017.

2.1.3 Problema de inferencia multi-objetivo de árboles filogenéticos

Un problema de optimización basado en un único objetivo implica la maximización (o minimización) de solo una función objetivo. No obstante, un problema de optimización multi-objetivo involucra múltiples criterios, los que pueden ser conflictivos entre sí (Handl et al., 2007). En el contexto de inferencia filogenética, como es posible apreciar en el Anexo C, dos criterios pueden resultar en diferentes topologías para el mismo conjunto de datos. Esta relación entre estos criterios no ha sido establecida, por ejemplo, la relación entre parsimonia y verosimilitud. El problema multi-objetivo (bi-objetivo) de inferencia de árboles filogenéticos puede ser definido por la Ecuación 2.3:

$$\text{maximizar } \vec{z} = \vec{f}(x) = (f_1(x), f_2(x)), x \in X \quad (2.3)$$

Donde x es una solución correspondiente a un árbol filogenético en el conjunto de todas las posibles topologías o soluciones X , y $z = \vec{f}(x)$ es el vector de objetivos, donde f_1 es la función de parsimonia y f_2 es la función de verosimilitud (Anexo C). Es necesario considerar que la maximización de parsimonia implica una disminución de su magnitud. El conjunto de soluciones de Pareto considera aquellos árboles filogenéticos en que es imposible mejorar un

objetivo sin empeorar otro (dominancia). Los puntos en el espacio objetivo correspondientes al conjunto de Pareto óptimo se denominan soluciones no dominadas y conforman la Frontera de Pareto. El problema puede ser representado usando la Figura 2.4, donde las soluciones óptimas se posicionan en la zona superior izquierda del espacio objetivo.

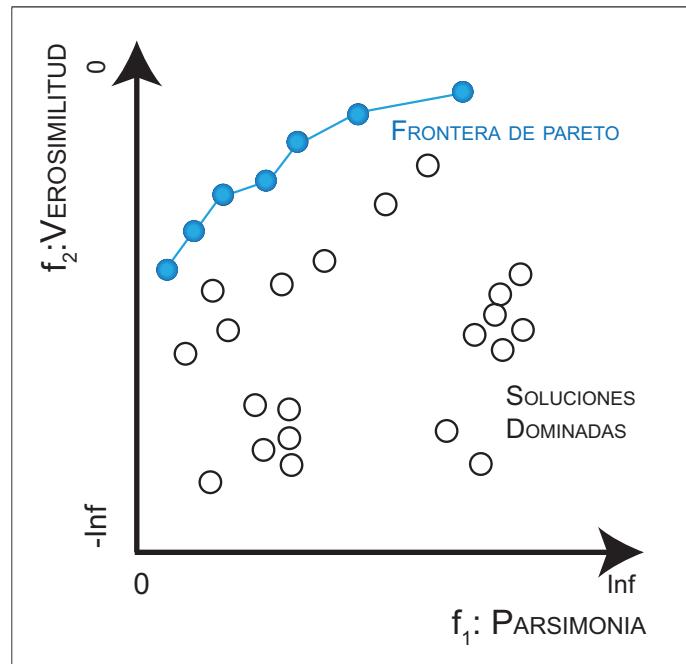


Figura 2.4: Frontera de Pareto en el problema de inferencia multi-objetivo de árboles filogenéticos.

Fuente: Elaboración propia, 2017.

Los primeros métodos publicados para inferir árboles filogenéticos se basaron en la optimización de un solo criterio (Cavalli-Sforza & Edwards, 1967; Fitch & Margoliash, 1967; Fitch, 1971; Felsenstein, 1973, 1981; Swofford & Maddison, 1987; Saitou & Nei, 1987; Matsuda, 1995, 1996; Lewis, 1998; Moilanen, 1999; Gascuel, 2000; Skourikhine, 2000; Katoh et al., 2001; Huelsenbeck & Ronquist, 2001; Desper & Gascuel, 2002; Cotta & Moscato, 2002; Lemmon & Milinkovitch, 2002; Cotta & Moscato, 2003; Guindon & Gascuel, 2003; Stamatakis, 2004, 2006; Zwickl, 2006; Lin et al., 2007; Gallardo et al., 2007; Goloboff et al., 2008; Rota-Stabelli & Telford, 2008; Zhihua et al., 2005; Wróbel et al., 2012; Plonski & Radomski, 2013). Una acabada revisión de estas estrategias es efectuada en Santander-Jiménez & Vega-Rodríguez (2013a). La investigación abordada en este capítulo se centra en el problema de inferencia multi-objetivo de árboles filogenéticos.

Una de las técnicas más usadas en la actualidad para resolver problemas multi-objetivos es el algoritmo NSGA propuesto por Srinivas & Deb (1994), aunque la mayoría de las aplicaciones han empleado su modificación, NSGA-II (Deb et al., 2002), en la que se reduce su complejidad computacional (Yijie & Gongzhang, 2008). Paralelamente, diferentes estrategias

de búsqueda local han sido exitosamente integradas en aproximaciones multi-objetivo (Ishibuchi et al., 2003; Ochoa et al., 2010; Dubois-Lacoste et al., 2012; Drugan & Thierens, 2012). Los algoritmos meméticos también han sido aplicados exitosamente para resolver problemas de optimización de único y múltiples objetivos en bioinformática (Clark & Kalita, 2015; Rubio-Largo et al., 2016) y en otras áreas de investigación (Mariano et al., 2010; Li et al., 2013; Wang & Zhu, 2013), combinando las ventajas de las estrategias evolutivas con las de búsqueda local.

El primer trabajo con un enfoque multi-objetivo que abordó inferencia filogenética fue desarrollado por Poladian & Jermiin (2006). Este usó un algoritmo evolutivo multi-objetivo para maximizar verosimilitud e inferir diferentes árboles filogenéticos empleando dos fuentes biológicas en conflicto: información génica mitocondrial y nuclear. Un año después, Jayaswal et al. (2007) aplicaron la misma metodología usando fuentes biológicas en conflicto de secuencias de nucleótidos provenientes de simios para obtener diferentes hipótesis evolutivas. El mismo año, Cancino & Delbem (2007) propusieron un algoritmo evolutivo multi-objetivo: PhyloMOEA. Este usó los criterios de parsimonia y verosimilitud para evaluar la hipótesis evolutivas de cuatro conjuntos de datos de nucleótidos. Se emplearon estrategias de reordenamiento de árboles para las etapas de cruzamiento (Lewis's operator (Sheneman & Foster, 2006)) y de mutación (*Nearest-neighbor interchange*, NNI). Estas estrategias modifican la topología de un árbol empleando operaciones de poda e inserción.

Otras aproximaciones bioinspiradas han sido estudiadas para tratar el problema de inferencia filogenética. Coelho et al. (2010) propusieron un algoritmo multi-objetivo inspirado en el sistema inmunológico para inferir árboles empleando criterios basados en distancia: mínima evolución y mínimos cuadrados. Las operaciones genéticas fueron aplicadas directamente sobre matrices de distancias, sin usar estrategias de reordenamiento. Otra aproximación bioinspirada fue presentada por Santander-Jiménez & Vega-Rodríguez (2013a). En ella se usó una adaptación de un algoritmo inspirado en colonias de abejas (*Artificial Bee Colony*, MO-ABC) para maximizar parsimonia y verosimilitud empleando conjuntos reales de secuencias de nucleótidos. Su algoritmo incorporó NNI y fue comparada con NSGA-II usando el operador *Prune-Delete-Graft* (PDG). Además, la propuesta fue contrastada con métodos clásicos de la literatura basados en optimización de objetivo único. En un trabajo previo se compararon diferentes operadores para esta meta-heurística (Santander-Jiménez et al., 2012). Los mismos autores propusieron un algoritmo multi-objetivo inspirado en luciérnagas (*Multi-objective Firefly Algorithm*, MO-FA) para inferir árboles filogenéticos usando parsimonia y verosimilitud (Santander-Jiménez & Vega-Rodríguez, 2013c). Además evaluaron el comportamiento de diversos métodos de iniciación para esta aproximación (Santander-Jiménez & Vega-Rodríguez, 2013b).

Versiones paralelas de estas estrategias también han sido estudiadas (Santander-Jiménez & Vega-Rodríguez, 2015a,b). Por ejemplo, Santander-Jiménez & Vega-Rodríguez (2016)

abordaron la reconstrucción filogenética empleando un algoritmo evolutivo paralelo basado en indicadores usando la métrica de hipervolumen (*Parallel indicator-based evolutionary algorithm using the hypervolume metric*). Recientemente, Zambrano-Vega et al. (2016) también propusieron una herramienta para inferir árboles filogenéticos maximizando parsimonia y verosimilitud: MO-Phylogenetic, disponiendo de múltiples operadores genéticos para seleccionar.

El diseño de diferentes métodos para inferir filogenia ha requerido el uso de diversas estrategias de reordenamiento para buscar topologías óptimas de árboles. Algunos de ellos han sido ampliamente usados en la literatura como operadores de mutación en algoritmos evolutivos (Felsenstein, 2004): (1) NNI, (2) *Sub-tree Pruning and Re-grafting*, (SPR), y (3) *Tree Bisection and Reconnection* (TBR). El operador NNI intercambia subárboles desde una rama interna seleccionada aleatoriamente para obtener un nuevo árbol. Por otro lado, SPR selecciona un subárbol aleatorio desde un árbol inicial, lo remueve y luego lo injerta en una posición aleatoria de un segundo árbol para obtener una nueva topología. TBR combina ambas técnicas (Santander-Jiménez & Vega-Rodríguez, 2013a). Otras estrategias también han sido usadas como operadores de cruzamiento en algoritmos genéticos (Matsuda, 1995; Lewis, 1998; Reijmers et al., 1999; Congdon, 2002). Sin embargo, los trabajos más recientes han empleado PDG como operador. Este operador selecciona un subárbol aleatorio desde uno de los padres y lo injerta en el otro en un punto de inserción aleatorio, eliminando las especies duplicadas desde el árbol receptor (Santander-Jiménez & Vega-Rodríguez, 2013a). Se ha reportado que esta estrategia de cruzamiento posee sesgo, resulta destructiva y desbalanceada, ya que tiende a conservar las características de solo uno de los padres (Sheneman & Foster, 2006; Pirkwieser & Raidl, 2008).

La aplicación de estrategias de optimización multi-objetivo en inferencia filogenética para construcción de árboles es un foco actual de investigación, principalmente centrado en el desarrollo, optimización y evaluación de nuevas estrategias algorítmicas. A pesar de estos avances y las ventajas que han demostrado tener las aproximaciones multi-objetivo, la mayoría de las actuales herramientas usadas para inferir árboles filogenéticos se basan en la optimización de solo un objetivo.

2.2 APROXIMACIÓN BASADA EN OPTIMIZACIÓN MULTI-OBJETIVO

A lo largo de esta sección se describe el desarrollo de un algoritmo memético multi-objetivo, diseñado para abordar el problema de inferencia de árboles filogenéticos. Se detallan diferentes pruebas desarrolladas para la evaluación y caracterización de las estrategias usadas en las operaciones genéticas y búsqueda local.

2.2.1 Descripción del algoritmo

Se ha modificado el algoritmo NSGA-II, integrando estrategias de búsqueda local y de reordenamiento propias del problema de inferencia filogenética basada en árboles. El Algoritmo 2.1 muestra el pseudo-código de la propuesta, donde D corresponde a un conjunto de datos en formato PHYLIP (Felsenstein, 2005), ps es el tamaño de la población, cr y mr son la tasa de cruce y mutación, ls corresponde al número de iteraciones realizadas por el algoritmo de búsqueda local. Las siguientes subsecciones describen la estructura de la propuesta.

Algoritmo 2.1: Algoritmo memético multi-objetivo (MO-MA)

Input: D, ps, cr, mr, ls
Output: Población P de árboles (Frontera de Pareto)

```
/* Inicialización de población */  
1  $P \leftarrow INICIALIZACION\_POBLACION(D, ps);$   
2 while condición de término no es alcanzada do  
3   for each  $p$  en  $P$  do  
4     /* Operaciones genéticas */  
5      $[T_1, T_2] \leftarrow SELECCION\_TORNEO\_BINARIO(P);$   
6      $Q[p] \leftarrow CRUZAMIENTO(T_1, T_2, cr);$   
7      $Q[p] \leftarrow MUTACION(Q[p], mr);$   
8   end  
9   /* Actualización de la Frontera de Pareto */  
10   $P \leftarrow ORDENAMIENTO\_NO\_DOMINADO(P, Q, ps);$   
11  /* Aplicación de operador de búsqueda local */  
12   $P \leftarrow ESTRATEGIA\_BUSQUEDA\_LOCAL(P, ls);$   
13 end  
14 return  $P;$ 
```

2.2.1.1 Criterios de optimidad

Las aproximaciones basadas en caracteres infieren árboles filogenéticos usando como entrada secuencias alineadas. Estos métodos han resultado ser estadísticamente más consistentes que las aproximaciones basadas en distancia, debido a la pérdida inevitable de información evolutiva asociada a la construcción de matrices de distancia, y al trabajar con grandes volúmenes de datos (Zhihua et al., 2005). En consecuencia, la mayoría de

las aproximaciones basadas en optimización multi-objetivo propuestas en la literatura infieren filogenia considerando los criterios de parsimonia y verosimilitud (Sección ??). En MO-MA el criterio de parsimonia es calculado usando el algoritmo de Fitch (Felsenstein, 2004), y la verosimilitud es obtenida usando el algoritmo estocástico de búsqueda propuesto por Nguyen (Schliep, 2011). Ambos criterios son calculados usando la biblioteca *phangorn* de R (Sección 2.3).

2.2.1.2 Inicialización de la población

La función *INICIALIZAR_POBLACION* construye un primer árbol T_i aplicando un método de distancia definido previamente (UPGMA, WPGMA, NJ, o BioNJ). El modelo evolutivo asociado es estimado empleando el criterio de información de Akaike (Anexo C). Usando T_i como base, el algoritmo construye un nuevo árbol T_p optimizando parsimonia, y estima la longitud de ramas según el método ACCTRAN (Swofford & Maddison, 1987). El modelo evolutivo correspondiente es calculado aplicando el mismo criterio del árbol inicial. A continuación un tercer árbol T_l es inferido desde T_i maximizando verosimilitud, para ello se emplea el mismo modelo evolutivo definido para T_i . Finalmente, la población inicial P con p_s individuos es creada usando estrategias de reordenamiento (NNI, SPR o TBR) sobre los árboles iniciales T_p y T_l .

2.2.1.3 Operadores de cruzamiento

Una vez que la población inicial es construida, una segunda población Q es creada aplicando operadores genéticos. Con objetivo de combinar las características de diferentes soluciones por medio de una operación de cruzamiento, el algoritmo selecciona dos padres (T_1 y T_2) desde P aplicando un torneo binario aleatorio (*SELECCION_TORNEO_BINARIO*). Posteriormente, estos padres son combinados por la función *CRUZAMIENTO* empleando estrategias de reordenamiento. Para este propósito se propusieron cuatro operadores: PDG (Cotta & Moscato, 2002), una modificación de PGD (PDGm), *Branch Exchange* (BE) y un operador de consenso (CS). El operador PDG fue implementado según la literatura respectiva (Cotta & Moscato, 2002). Con objetivo de construir un segundo operador, se propuso una pequeña modificación del algoritmo de PDG (PDGm) en que la selección de un subárbol aleatorio desde uno de los padres, es reemplazada por la selección del subárbol de menor tamaño con una restricción para un mínimo de dos especies. El operador BE poda una rama aleatoria desde

uno de los padres y lo re-inserta en una posición aleatoria del otro. Las especies duplicadas de este último son eliminadas. Finalmente, un operador de consenso (CS) goloso fue diseñado para buscar árboles que conserven las características de ambos padres. Este operador aplica la estrategia NNI en el padre T_1 hasta que el descendiente alcance una determinada distancia aleatoria r (métrica Robinson-Foulds) entre ambos padres (Robinson & Foulds, 1981). La Figura 2.5, inspirada en el trabajo de Gallardo et al. (2007), muestra un esquema de los cuatro operadores propuestos. Luego de aplicar una de estas cuatro estrategias, un operador de cruzamiento uniforme aleatorio combina los parámetros del modelo evolutivo de los padres para ser aplicados en el cálculo de verosimilitud.

2.2.1.4 Operador de mutación

La función *MUTACION* permite usar tres heurísticas diseñadas para reordenamiento topológico de árboles como operadores de mutación: NNI, SPR, y TBR. Los dos primeros están disponibles en la biblioteca *phangorn* de R, mientras que el tercero fue implementado como una nueva función.

2.2.1.5 Estrategia de búsqueda local

Mediante la función *ESTRATEGIA_BUSQUEDA_LOCAL*, el algoritmo propuesto estudia la vecindad de la población usando los mismos operadores de mutación señalados en la Sección 2.2.1.4. Una de estas estrategias es aplicada en cada solución $p \in P$ para generar un nuevo árbol p' . Si p no domina a p' , este último es agregado a la población P para que un algoritmo de ordenamiento no dominado que incluye distancia de aglomeración (*crowding distance*) sea aplicado y descarte una solución, al igual que el algoritmo NSGA-II. Posteriormente, el algoritmo aplica una de dos estrategias de búsqueda local previamente definida: (1) *Pareto local search* (PLS) (Drugan & Thierens, 2012; Dubois-Lacoste et al., 2012)), o (2) *Simulated annealing* (SA) (Bandyopadhyay et al., 2008)). Debido a su carácter exhaustivo, ambas estrategias fueron limitadas a un número fijo de ls iteraciones. Por otro lado, SA debió ser parametrizado para determinar los valores del coeficiente de temperatura α y la temperatura inicial. El Anexo D.1 muestra los detalles de los algoritmos y la parametrización de ambas estrategias de búsqueda local.

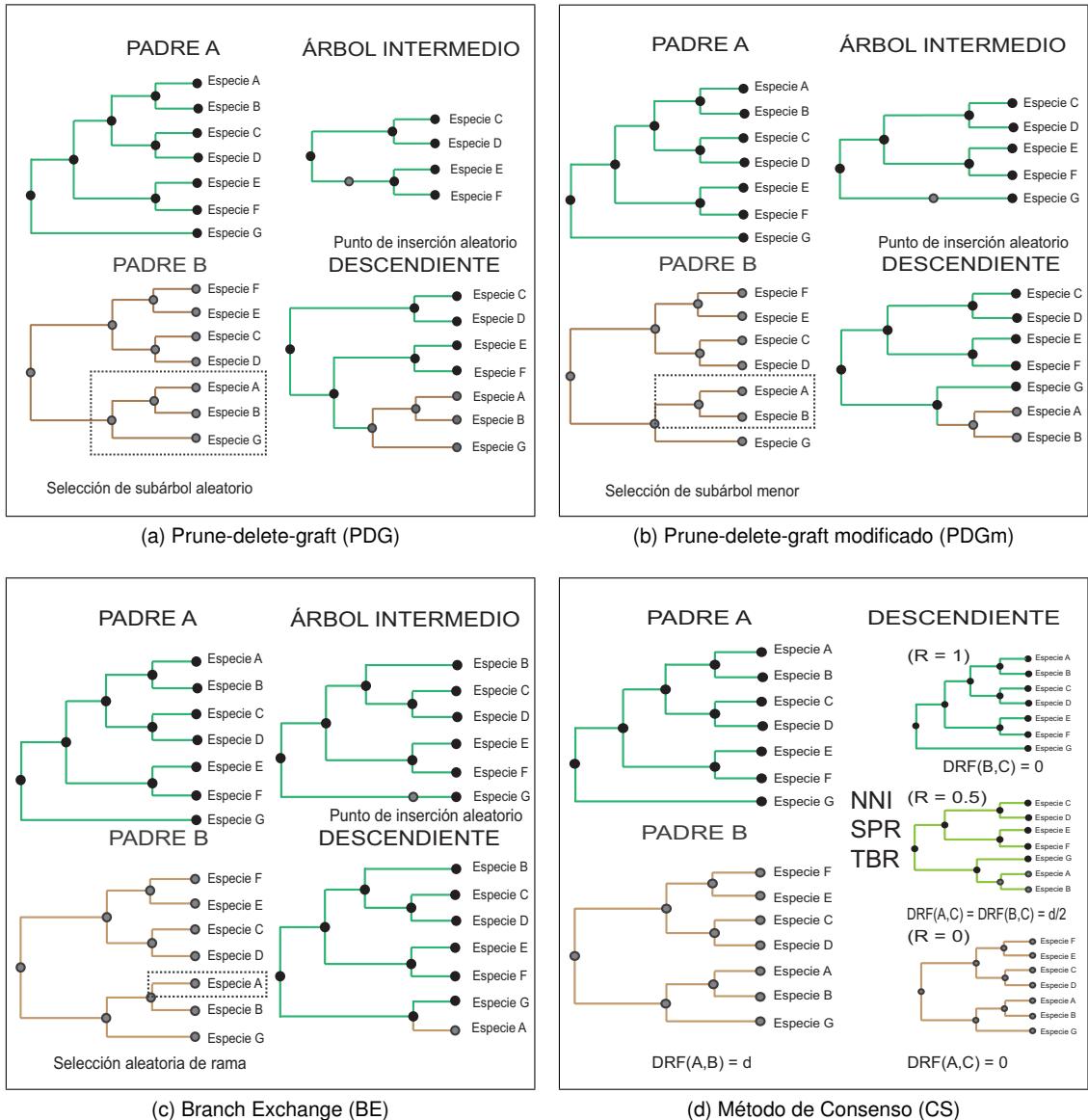


Figura 2.5: Esquema de operadores de cruzamiento propuestos para MO-MA
Fuente: Elaboración propia, 2017.

2.2.2 Evaluación de desempeño

Con objetivo de identificar la configuración que maximiza el desempeño de MO-MA, se evaluaron y caracterizaron diferentes estrategias de reordenamiento como operadores genéticos. Estas fueron integradas en la estructura del algoritmo propuesto para comparar múltiples configuraciones (Sección 2.2.2.3). Con el fin de medir el desempeño de MO-MA frente a otras estrategias se usaron parámetros definidos previamente en la literatura para NSGA-II

(Santander-Jiménez & Vega-Rodríguez, 2013a): tamaño de población ps , tasa de cruzamiento cr y mutación mr . Los parámetros para ambas estrategias de búsqueda local fueron definidos experimentalmente (Anexo D.1).

2.2.2.1 Operadores de cruzamiento

Se evaluaron dos condiciones de los operadores de cruzamiento: (1) el sesgo asociado a la producción de descendientes al incluir características de ambos padres (condición de balance), y (2) la habilidad de visitar el espacio de búsqueda. Para evaluar la primera condición se efectuaron mil operaciones de cruzamiento usando cada estrategia de reordenamiento sobre las topologías iniciales T_p y T_l . Posteriormente, se calculó la distancia de los descendientes a cada parente en forma individual usando la métrica de Robinson-Foulds (Robinson & Foulds, 1981). Esta fue normalizada por la distancia entre padres.

Con objetivo de medir la capacidad de cada estrategia de reordenamiento para revisar el espacio de búsqueda y así obtener nuevas soluciones, se midió la distancia media normalizada entre descendientes y ambos padres (métrica Robinson-Foulds).

2.2.2.2 Operadores de mutación

También se midió la habilidad de cada operador de mutación para visitar el espacio de búsqueda de soluciones usando la métrica de Robinson-Foulds. Para ello se generaron mil mutaciones sobre las topologías iniciales, siendo evaluadas con el mismo método explicado para caracterizar la operación de cruzamiento.

2.2.2.3 Configuración global

Se evaluó el desempeño de MO-MA usando diferentes configuraciones: cuatro métodos de distancia para generar topologías iniciales (UPGMA, WPGMA, NJ y BioNJ), cinco alternativas de cruzamiento (PDG, PDGm, BE, CS y sin aplicación de operador de cruzamiento), cuatro estrategias de mutación (NNI, SPR, TBR y sin estrategia de mutación), y tres estrategias de búsqueda local (PLS, SA, y sin aplicación de búsqueda local). El número de combinaciones

posibles es de 240 ($4 \times 5 \times 4 \times 3$). Cada configuración fue ejecutada 30 veces, resultando en 7200 ejecuciones para cada conjunto de datos: *primates_14*, *rbcL_55* y *HIV1_192*. Este número de ejecuciones ha sido suficiente para brindar soporte estadístico a experimentos similares expuestos en la literatura relacionada (Cantú-Paz & Goldberg, 2003).

Para evaluar la calidad de las soluciones se utilizó la métrica de hipervolumen (Jiang et al., 2014). El detalle de esta y otras métricas se encuentra en el Anexo F. Con el fin de comparar la diferencia estadística significante entre el desempeño de las configuraciones, se empleó la prueba de Kruskal-Wallis por categoría. El análisis post-hoc fue realizado usando la prueba de Dunn con un nivel de significancia del 1 %. Para complementar la métrica de hipervolumen se calculó la cobertura de las soluciones (Jiang et al., 2014), y el porcentaje de representatividad en la Frontera de Pareto global. Para ello se usó el conjunto de soluciones de Pareto que corresponden al valor de la mediana del hipervolumen alcanzado por cada configuración (solución representativa). Este método ha sido aplicado en trabajos previos de la literatura (Santander-Jiménez & Vega-Rodríguez, 2013a). La Frontera de Pareto global considera las soluciones no dominadas entre todas las configuraciones y no necesariamente corresponde a la solución óptima del problema.

2.2.2.4 Aproximaciones basadas en optimización de un único objetivo

Para mostrar los beneficios de la aplicación de un enfoque multi-objetivo al problema de inferencia filogenética respecto a la versión de objetivo único, se comparó MO-MA con otras herramientas ampliamente utilizadas en la literatura basadas en la optimización de un único criterio: PHYLML (parsimonia de Sankoff), DNAPARS, RAxML (Guindon & Gascuel, 2003) y MEGA (Sudhir et al., 2016) (criterios de parsimonia y verosimilitud). Cada uno de ellos fue ejecutado 31 veces para cada conjunto de datos. La métrica de hipervolumen, cobertura y la representatividad sobre la Frontera de Pareto global fueron calculadas usando los criterios expuestos en la Sección 2.2.2.3. El proceso de parametrización se detalla en el Anexo D.2.

2.2.2.5 Aproximaciones basadas en optimización de múltiples objetivos

Con el fin de evaluar el desempeño de MO-MA con las nuevas propuestas de la literatura, se efectuó una comparación entre métodos basados en optimización multi-objetivo: NSGA-II (Santander-Jiménez & Vega-Rodríguez, 2013a), PhyloMOEA (Cancino & Delbem, 2007),

MO-ABC (Santander-Jiménez & Vega-Rodríguez, 2013a), MO-FA (Santander-Jiménez & Vega-Rodríguez, 2013c), MO-Phyl (Santander-Jiménez & Vega-Rodríguez, 2015a) y MO-phylogenetics (Zambrano-Vega et al., 2016). Además de estas estrategias, se propuso un NSGA-II que incluye el operador de cruzamiento que combina los parámetros del modelo evolutivo (NSGA-II EM). Este último y MO-MA fueron ejecutados 31 veces para cada conjunto de datos. La mediana del hipervolumen fue calculada para hallar el conjunto de soluciones representativas entre ejecuciones, y así realizar la comparación de los métodos. La Frontera de Pareto representativa de las otras propuestas fue extraída desde los gráficos publicados por cada referencia usando la herramienta *WebPlotDigitizer tool 3.11* (Rohatgi & ZlatanStanojevic, 2017; Drevon et al., 2017). A fin de realizar una justa comparación, se limitó el número de generaciones de MO-MA al tiempo requerido por NSGA-II para iterar 100 generaciones (Anexo D.2).

2.2.2.6 Estudio de conjunto de datos de aminoácidos

Las ureasas tienen especial interés en biología, ya que la historia evolutiva de estas enzimas multi-funcionales es difícil de estimar. Esto se debe especialmente a su variada organización estructural (Carlini & Ligabue-Braun, 2016). Una hipótesis evolutiva aceptada para estas proteínas ha sido previamente inferida por la literatura usando la herramienta MEGA (Ligabue-Braun et al., 2013). Dado que todos los conjuntos de datos explorados por las propuestas multi-objetivo en la literatura corresponden exclusivamente a secuencias de nucleótidos, se evaluó el significado biológico de las soluciones obtenidas por MO-MA usando secuencias de aminoácidos. El propósito de ello es validar la aplicación de diferentes tipos de modelos evolutivos. Para este propósito se ha aplicado la prueba de Shimodaira-Hasegawa (Cancino & Delbem, 2007), evaluando si la diferencia entre la solución de mayor verosimilitud y la topología propuesta en Ligabue-Braun et al. (2013) es estadística significante.

2.3 RESULTADOS

Los algoritmos, experimentos y análisis estadísticos desarrollados en esta investigación fueron efectuados usando el entorno R 3.3.2 y RStudio 0.99.491. Particularmente se emplearon las bibliotecas *phangorn* (Schliep, 2011), *phytools* (Revell, 2012), y *emoa* (Mersmann, 2011). Los experimentos fueron realizados usando tres procesadores MS Azure Standard DS5 v2 con 16 núcleos Xeon 56-2673v3 (Haswell), 2.4 GHz, 56 GB Ram y 112 GB de disco duro. La

Tabla 2.1 presenta los detalles de las secuencias utilizadas y sus correspondientes referencias.

Tabla 2.1: Conjuntos de datos y referencias empleados en experimentos.

Datos	#Secuencias	#Caracteres	Fuente
<i>primates_14</i>	14	232	(Felsenstein, 2005)
<i>rbcL_55</i>	55	1314	(Coelho & Zuben, 2007)
<i>HIV2_72</i>	72	828	(Santander-Jiménez & Vega-Rodríguez, 2013a)
<i>membra_81</i>	81	3321	(Santander-Jiménez & Vega-Rodríguez, 2013a)
<i>ureases_126</i>	126	609	(Carlini & Ligabue-Braun, 2016)
<i>mtDNA_186</i>	186	16608	(Santander-Jiménez & Vega-Rodríguez, 2013a)
<i>HIV1_182</i>	192	817	(Coelho & Zuben, 2007)
<i>RDP II_218</i>	218	4182	(Santander-Jiménez & Vega-Rodríguez, 2013a)
<i>ZILLA_500</i>	500	759	(Coelho & Zuben, 2007)

Fuente: Elaboración propia, (2017)

2.3.1 Operadores de cruzamiento

La Tabla 2.2 muestra la métrica Robinson-Foulds obtenida entre descendientes y cada uno de los padres empleando los operadores de cruzamiento. En las estrategias PDG, PDGm y BE esta diferencia resultó ser estadísticamente significante, lo que implica que estos operadores generan hijos que tienden a conservar la topología de uno de los padres en relación al otro, produciendo sesgo en la descendencia (condición de desbalance). En contraste, el operador de consenso presenta una condición balanceada debido a la variable aleatoria que controla la similaridad de los descendientes hacia ambos padres.

El segundo experimento midió la distancia promedio entre hijos y ambos padres (Tabla 2.3). Los menores valores fueron obtenidos por PDG, PDGm y BE, cuya diferencia resultó estadísticamente significante respecto a CS. Esto significa que los hijos producidos por estos operadores pertenecen a una vecindad cercana a sus padres en comparación al método basado en consenso. Los resultados indican que este último operador es capaz de generar soluciones que difieren significativamente de sus padres, permitiendo recorrer regiones más lejanas del espacio de soluciones. Los detalles de las comparaciones se encuentran en el Anexo D.4.

Tabla 2.2: Métrica de Robinson-Foulds entre descendientes y cada padre (P1, P2) obtenida usando diferentes operadores de cruzamiento. Se presentan valores promedios y desviación estándar. Los valores en negrita indican que existe diferencia significativa entre la distancia hacia ambos padres.

Cruzamiento	Padres	<i>rbcL_55</i>	<i>HIV1_192</i>	<i>ZILLA_500</i>
PDG	P1	1.7 (0.4)	1.0 (0.1)	1.1 (0.0)
	P2	1.0 (0.3)	0.2 (0.1)	0.2 (0.0)
PDGM	P1	2.5 (0.2)	1.0 (0.0)	1.1 (0.0)
	P2	1.8 (0.5)	0.2 (0.0)	0.2 (0.0)
CS	P1	0.1 (1.9)	0.1 (0.7)	3.3 (1.7)
	P2	1.1 (1.3)	1.0 (0.2)	1.4 (1.2)
BE	P1	1.9 (0.3)	1.0 (0.0)	1.0 (0.0)
	P2	1.1 (0.4)	0.1 (0.0)	0.1 (0.0)

Fuente: Elaboración propia (2017).

Tabla 2.3: Métrica Robinson-Foulds entre descendientes y padres luego de la aplicación de diferentes operadores genéticos: cruzamiento y mutación. Se presentan valores promedios y desviación estándar.

Cruzamiento	<i>rbcL_55</i>	<i>HIV1_192</i>	<i>ZILLA_500</i>
PDG	1.6 (0.4)	0.5 (0.5)	0.6 (0.5)
PDGM	2.2 (0.4)	0.6 (0.4)	0.7 (0.5)
CS	3.6 (1.4)	1.3 (0.2)	1.0 (1.7)
BE	1.6 (0.4)	0.6 (0.5)	0.6 (0.5)
Mutación	<i>rbcL_55</i>	<i>HIV1_192</i>	<i>ZILLA_500</i>
NNI	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)
SPR	18.0 (6.0)	26.1 (8.4)	37.7 (10.3)
TBR	16.8 (5.3)	30.9 (8.1)	40.3 (9.3)

Fuente: Elaboración propia, (2017).

2.3.2 Operadores de mutación

La Tabla 2.3 muestra la métrica Robinson-Foulds luego de la aplicación de los diferentes operadores de mutación. Naturalmente los descendientes del operador NNI tienen una menor distancia en relación a SPR y TBR. Esto se debe a su propia definición en que se intercambia la rama interna de un árbol para generar otro diferente (dos ediciones de reconstrucción). No hubo diferencia estadística significante entre SPR y TBR. Los detalles adicionales pueden ser encontrados en el Anexo D.5.

2.3.3 Configuración global

Un total de 240 configuraciones de MO-MA fueron evaluadas empleando tres conjuntos de datos (*rbcL_55*, *HIV1_192* y *primates_14*). Las Tablas 2.4 y 2.5 muestran las cinco mejores y peores configuraciones según las métricas de hipervolumen, cobertura y representatividad en la Frontera de Pareto.

En la Tabla 2.4 es posible observar que la configuración compuesta por el operador de población NJ, PDG, NNI y la estrategia de búsqueda local golosa PLS (G), posee la mayor métrica de hipervolumen para los tres conjuntos de datos. Particularmente en el caso del conjunto de datos *rbcL_55*, esta configuración alcanzó mayores métricas de hipervolumen que otras configuraciones siendo significante en un 54 % de los casos. En los otros dos conjuntos de datos, a pesar de que esta configuración también alcanzó los mayores valores de hipervolumen, solo fue significante en un 6 % y 1 % de las comparaciones. Esto se debe a la alta desviación estándar alcanzada para el conjunto de datos *HIV1_192*, y al pequeño tamaño del conjunto de datos *primates_14*, en que muchas configuraciones convergieron a la misma solución. De acuerdo a la métrica de cobertura, las soluciones encontradas por la configuración NJ-PDG-NNI-G resultaron ser no dominadas por las otras configuraciones, y contribuyeron con la mayor parte de la Frontera de Pareto global (31 %, 25 %, y 55 % respectivamente).

En el caso de configuraciones con menor métrica de hipervolumen (Tabla 2.5) no fue posible identificar una configuración común considerando todos los conjuntos de datos; sin embargo, WPGMA estuvo presente en las peores configuraciones. Basado en estos resultados, la configuración NJ-PDG-NNI-G fue seleccionada como estructura base para MO-MA, permitiendo su comparación con métodos alternativos propuestos en la literatura.

2.3.4 Comparación con métodos basados en optimización de objetivo único

La Tabla 2.6 muestra las métricas de hipervolumen, cobertura y representatividad en la Frontera de Pareto global para MO-MA y otras herramientas basadas en optimización de objetivo único. Los valores de parsimonia y verosimilitud fueron normalizados entre 0 y 1. Para calcular la métrica de hipervolumen se usó como punto de referencia (2,2). Los valores en negrita representan las mayores magnitudes de hipervolumen que resultaron estadísticamente significantes comparados con los otros métodos.

MO-MA posee la mayor métrica de hipervolumen en siete de los ocho conjuntos de datos estudiados, diferencia que es estadísticamente significante comparada con los otros

métodos. La métrica de cobertura en los mismos conjuntos de datos muestran que MO-MA produce una Frontera de Pareto con soluciones que no son dominadas por las otras herramientas, a excepción de *RDPII_218* y *ZILLA_500*, existiendo 38 % y 50 % de soluciones dominadas (Figura 2.6) . En relación al número de soluciones en la Frontera de Pareto global, MO-MA contribuye con al menos el 64 % de las soluciones, promediando un 80 % al considerar todos los conjuntos de datos. Un caso particular corresponde al conjunto de datos *mtDNA_186*, para el cual RAxML obtuvo los mejores resultados. Esta herramienta alcanza los mejores valores en las tres métricas evaluadas. Es interesante ver en la Figura 2.6 que MO-MA generó una Frontera de Pareto con una única solución. Ambos métodos alcanzan niveles similares de parsimonia; sin embargo, la verosimilitud de la solución producida por RAxML es mayor que cualquier otra solución.

Tabla 2.4: Mejores cinco configuraciones de algoritmo memético multi-objetivo según métrica de hipervolumen (hip). Se incluyen valores promedios y desviación estándar. La métrica de cobertura (cob) representa el porcentaje de soluciones no dominadas. La representatividad de la Frontera de Pareto (% Par) corresponde a la relación entre el número de soluciones no dominadas obtenidas para un método específico, y el número total de soluciones en la Frontera de Pareto global. (*ini*: inicialización, *cr*: cruzamiento, *mut*: mutación, y *ls*: búsqueda local).

Datos	ini	cr	mut	ls	hip.	cob.	% Par.
<i>rbcL_55</i>	NJ	PDG	NNI	G	3.67(0.1)	0 %	31 %
	UPGMA	CS	NNI	G	3.63(0.1)	0 %	14 %
	BIONJ	CS	NNI	G	3.52(0.3)	0 %	2 %
	NJ	NONE	NNI	SA	3.45(0.2)	5 %	1 %
	BIONJ	BI	NNI	-	3.34(0.2)	0 %	1 %
<i>HIV1_192</i>	NJ	PDG	NNI	G	3.09(0.7)	0 %	25 %
	UPGMA	PDG	NNI	SA	3.00(0.0)	0 %	10 %
	UPGMA	CS	NNI	-	2.72(0.9)	0 %	2 %
	NJ	CS	NNI	SA	2.58(1.1)	0 %	1 %
	WPGMA	PDG	NNI	SA	2.43(1.3)	100 %	0 %
<i>primates_14</i>	WPGMA	PDG	NNI	G	3.39(0.0)	1 %	5 %
	UPGMA	PDG	NNI	G	3.39(0.0)	2 %	20 %
	NJ	PDG	NNI	G	3.39(0.0)	0 %	55 %
	BioNJ	CS	NNI	G	3.37(0.1)	98 %	1 %
	BioNJ	CS	TBR	SA	3.27(0.1)	98 %	1 %

Fuente: Elaboración propia, (2017).

2.3.5 Comparación con métodos basados en optimización de objetivo múltiples

La comparación de las aproximación multi-objetivo es mostrada en la Tabla 2.7. Con el fin de calcular la métrica de hipervolumen, los valores de parsimonia y verosimilitud fueron

Tabla 2.5: Peores cinco configuraciones de algoritmo memético multi-objetivo según métrica de hipervolumen (hip). Se incluyen valores promedios y desviación estándar. (*ini*: inicialización, *cr*: cruceamiento, *mut*: mutación, y *ls*: búsqueda local)

Datos	ini	cr	mut	ls	hip.	cob.	% Par.
<i>rbcL_55</i>	WPGMA	NONE	TBR	-	0.60(1.3)	50 %	0 %
	WPGMA	BI	TBR	SA	0.33(1.0)	50 %	0 %
	BIONJ	PDGM	SPR	-	0.30(0.9)	50 %	0 %
	WPGMA	-	SPR	-	0.30(0.9)	100 %	0 %
	WPGMA	PDG	SPR	-	0.30(0.9)	100 %	0 %
<i>HIV1_192</i>	WPGMA	BI	TBR	G	0.00(0.0)	100 %	0 %
	WPGMA	BI	TBR	SA	0.00(0.0)	100 %	0 %
	WPGMA	-	TBR	SA	0.00(0.0)	100 %	0 %
	WPGMA	PDGM	TBR	SA	0.00(0.0)	100 %	0 %
	WPGMA	PDG	TBR	G	0.00(0.0)	100 %	0 %
<i>primates_14</i>	BioNJ	PDG	TBR	SA	3.08(0.1)	98 %	1 %
	BioNJ	PDGM	TBR	G	3.06(0.1)	98 %	1 %
	NJ	PDG	TBR	G	3.00(0.0)	98 %	1 %
	NJ	PDG	TBR	SA	3.00(0.0)	98 %	2 %
	WPGMA	PDGM	TBR	SA	3.00(0.0)	65 %	0 %

Fuente: Elaboración propia, (2017).

normalizados entre 0 y 1 para cada conjunto de datos. El punto de referencia utilizado fue definido como (2,2).

MO-MA obtuvo el mayor hipervolumen para siete de los ocho conjuntos de datos. Cuando se comparó MO-MA con NSGA-II EM hubo una diferencia estadísticamente significante en el 50 % de los casos. Con el resto de las propuestas MO-MA alcanzó soluciones con mejores métricas, siendo significantes en todos los conjuntos de datos. La única excepción es MO-ABC, que obtiene mejores valores en el conjunto de datos *ZILLA_500*.

La métrica de cobertura demuestra que las soluciones obtenidas por MO-MA son no dominadas por los otros métodos en todos los conjuntos de datos a excepción de *ZILLA_500*. En este caso MO-ABC, MO-FA, y MO-Phyl también tienen soluciones no dominadas. Siguiendo la misma linea, MO-MA constituye la mayor parte de la Frontera de Pareto global para todos los conjuntos de datos. Sin embargo, nuevamente cuando el conjunto de datos *ZILLA_500* es considerado, MO-Phyl y MO-ABC también tienen soluciones que conforman parte de esta frontera global (Figuras 2.7). Estos árboles mejoran el valor de parsimonia en relación a MO-MA.

Los resultados demuestran que las soluciones encontradas por MO-MA son nuevas, y contribuyen en promedio con más del 90 % de la Frontera de Pareto global en siete conjuntos de datos. La única excepción es *ZILLA_500*, en que los algoritmos MO_Phyl y MO_ABC tienen una mayor contribución a la Frontera de Pareto global.

Tabla 2.6: Hipervolumen, cobertura y representatividad en la Frontera de Pareto para soluciones de MO-MA y otros métodos basados en optimización de objetivo único. El nombre del conjunto de datos *membracidae1_81* fue abreviado.

Métrica de hipervolumen						
Datos	MO-MA	DNAPARS	MEGALIK	MEGAPAR	PHYML	RAXML
<i>primates_14</i>	3.14 (0.16)	3.04 (0.01)	2.48 (0.00)	2.48 (0.00)	2.00 (0.00)	1.75 (0.01)
<i>rbcL_55</i>	3.92 (0.04)	1.53 (0.02)	1.00 (0.00)	2.45 (0.01)	1.90 (0.00)	2.99 (0.01)
<i>HIV2_72</i>	3.62 (0.16)	1.51 (0.10)	1.47 (0.04)	2.24 (0.05)	1.64 (0.08)	1.90 (0.10)
<i>membra1_81</i>	3.98 (0.01)	1.76 (0.01)	1.75 (0.00)	1.98 (0.00)	3.33 (0.00)	3.32 (0.00)
<i>mtDNA_186</i>	2.17 (0.08)	2.00 (0.00)	1.52 (0.00)	3.27 (0.01)	1.02 (0.01)	4.00 (0.00)
<i>HIV1_192</i>	3.79 (0.01)	1.97 (0.01)	3.28 (0.01)	1.67 (0.00)	1.00 (0.01)	3.49 (0.01)
<i>RDP II_218</i>	3.49 (0.01)	3.04 (0.01)	2.86 (0.00)	2.54 (0.00)	1.00 (0.00)	3.06 (0.00)
<i>ZILLA_500</i>	3.22 (0.01)	2.33 (0.01)	2.00 (0.01)	1.57 (0.00)	2.52 (0.00)	2.14 (0.00)
Métrica de cobertura						
Datos	MO-MA	DNAPARS	MEGALIK	MEGAPAR	PHYML	RAXML
<i>primates_14</i>	0 %	100 %	100 %	100 %	100 %	100 %
<i>rbcL_55</i>	0 %	100 %	100 %	100 %	100 %	100 %
<i>HIV2_72</i>	0 %	100 %	0 %	100 %	100 %	0 %
<i>membra1_81</i>	0 %	100 %	100 %	100 %	100 %	100 %
<i>mtDNA_186</i>	100 %	100 %	100 %	100 %	100 %	0 %
<i>HIV1_192</i>	0 %	100 %	100 %	100 %	100 %	0 %
<i>RDP II_218</i>	38 %	0 %	100 %	100 %	100 %	0 %
<i>ZILLA_500</i>	50 %	0 %	0 %	100 %	0 %	100 %
% Soluciones en Frontera de Pareto global						
Datos	MO-MA	DNAPARS	MEGALIK	MEGAPAR	PHYML	RAXML
<i>primates_14</i>	100 %	0 %	0 %	0 %	0 %	0 %
<i>rbcL_55</i>	100 %	0 %	0 %	0 %	0 %	0 %
<i>HIV2_72</i>	83 %	0 %	0 %	0 %	0 %	17 %
<i>membra1_81</i>	100 %	0 %	0 %	0 %	0 %	0 %
<i>mtDNA_186</i>	0 %	0 %	0 %	0 %	0 %	100 %
<i>HIV1_192</i>	67 %	0 %	0 %	0 %	0 %	33 %
<i>RDP II_218</i>	86 %	7 %	0 %	0 %	0 %	7 %
<i>ZILLA_500</i>	64 %	12 %	12 %	0 %	12 %	0 %

Fuente: Elaboración propia, (2017).

2.3.6 Experimentación con secuencias de aminoácidos

La Frontera de Pareto obtenida por MO-MA es mostrada en la Figura 2.8. Esta considera tres puntos en el espacio objetivo y cinco diferentes árboles (A, B, C, D, y F). En este caso, los árboles con igual valor de calidad resultaron en diferentes topologías como (A,D) y (C,E). La mayor métrica de Robinson-Foulds fue de 16 ediciones. Este valor corresponde a la distancia entre el árbol más parsimonioso (B), y uno de los árboles con mayor verosimilitud (A). Esta diferencia demuestra el conflicto existente entre criterios.

La prueba de Shimodaira-Hasegawa no reportó diferencia significante entre la topología presentada en Ligabue-Braun et al. (2013) y las topologías de los árboles de mayor verosimilitud obtenidos en las diferentes ejecuciones al utilizar MO-MA. Esto demuestra que MO-MA es capaz de producir soluciones que no solo tienen un buen rendimiento en término de valores numéricos, sino que también en significancia biológica, encontrando soluciones equivalentes a hipótesis evolutivas aceptadas por la literatura.

Tabla 2.7: Hipervolumen, cobertura y representatividad en la Frontera de Pareto para soluciones de MO-MA y otros métodos basados en optimización de múltiples objetivos.

Métrica de hipervolumen									
Datos	MO-MA	NSGA-II EM	MO-ABC	MO-FA	MO-Phyl	MO-Phylo	NSGA-II	PhyloMOEA	
<i>primates_14</i>	2.20 (0.33)	2.14 (0.05)	-	-	-	-	-	-	-
<i>rbcL_55</i>	4.00 (0.00)	3.99 (0.00)	2.13	2.11	2.12	2.13	2.12	1.10	
<i>HIV2_72</i>	3.96 (0.02)	3.96 (0.02)	2.59	-	-	-	2.58	-	
<i>membra1_81</i>	3.99 (0.01)	3.97 (0.00)	1.68	-	-	-	1.57	-	
<i>mtDNA_186</i>	4.00 (0.00)	4.00 (0.00)	1.90	1.89	1.89	-	1.83	1.13	
<i>HIV1_192</i>	4.00 (0.00)	3.99 (0.00)	-	-	3.27	-	-	-	
<i>RDPII_218</i>	4.00 (0.00)	3.99 (0.00)	2.29	2.28	2.28	2.27	2.26	1.24	
<i>ZILLA_500</i>	3.46 (0.01)	3.45 (0.01)	3.96	3.80	3.88	-	3.54	2.03	
Métrica de cobertura									
Datos	MO-MA	NSGA-II EM	MO-ABC	MO-FA	MO-Phyl	MO-Phylo	NSGA-II	PhyloMOEA	
<i>primates_14</i>	0 %	0 %	-	-	-	-	-	-	-
<i>rbcL_55</i>	0 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
<i>HIV2_72</i>	0 %	100 %	86 %	-	-	-	100 %	-	
<i>membra1_81</i>	0 %	100 %	100 %	-	-	-	100 %	-	
<i>mtDNA_186</i>	0 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
<i>HIV1_192</i>	0 %	100 %	-	-	100 %	-	-	-	
<i>RDPII_218</i>	0 %	0 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
<i>ZILLA_500</i>	60 %	100 %	73 %	89 %	49 %	-	100 %	100 %	
% Soluciones en Frontera de Pareto global									
Datos	MO-MA	NSGA-II EM	MO-ABC	MO-FA	MO-Phyl	MO-Phylo	NSGA-II	PhyloMOEA	
<i>primates_14</i>	75 %	25 %	-	-	-	-	-	-	-
<i>rbcL_55</i>	100 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
<i>HIV2_72</i>	71 %	0 %	29 %	-	-	-	0 %	-	
<i>membra1_81</i>	100 %	0 %	0 %	-	-	-	0 %	-	
<i>mtDNA_186</i>	100 %	0 %	0 %	0 %	0 %	-	0 %	0 %	
<i>HIV1_192</i>	100 %	0 %	-	-	0 %	-	-	-	
<i>RDPII_218</i>	84 %	16 %	0 %	0 %	0 %	0 %	0 %	0 %	
<i>ZILLA_500</i>	9 %	0 %	14 %	7 %	70 %	-	0 %	0 %	

Fuente: Elaboración propia, (2017).

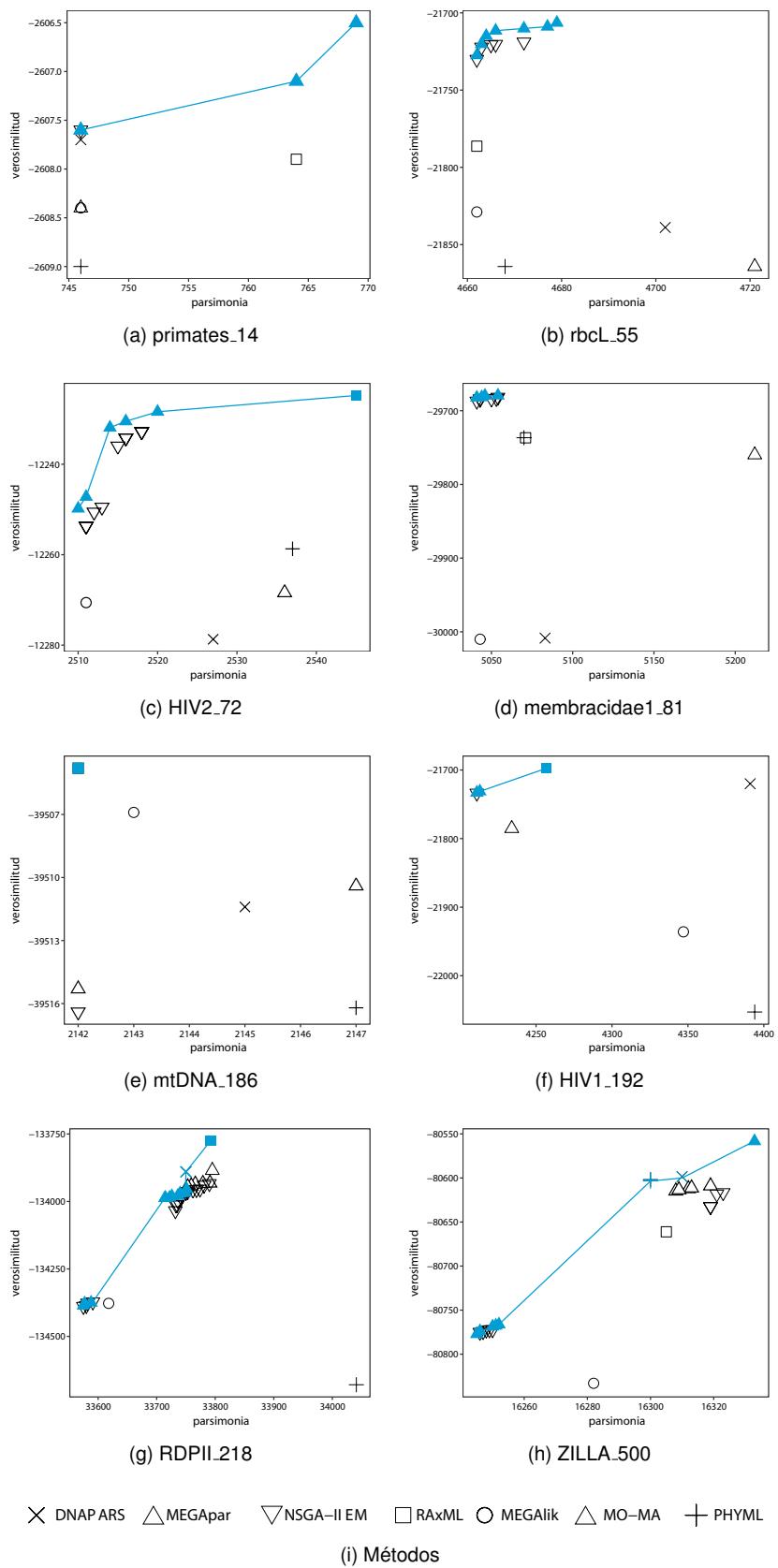


Figura 2.6: Comparación MO-MA con herramientas mono-objetivas. FP global en color.
Fuente: Elaboración propia, 2017.

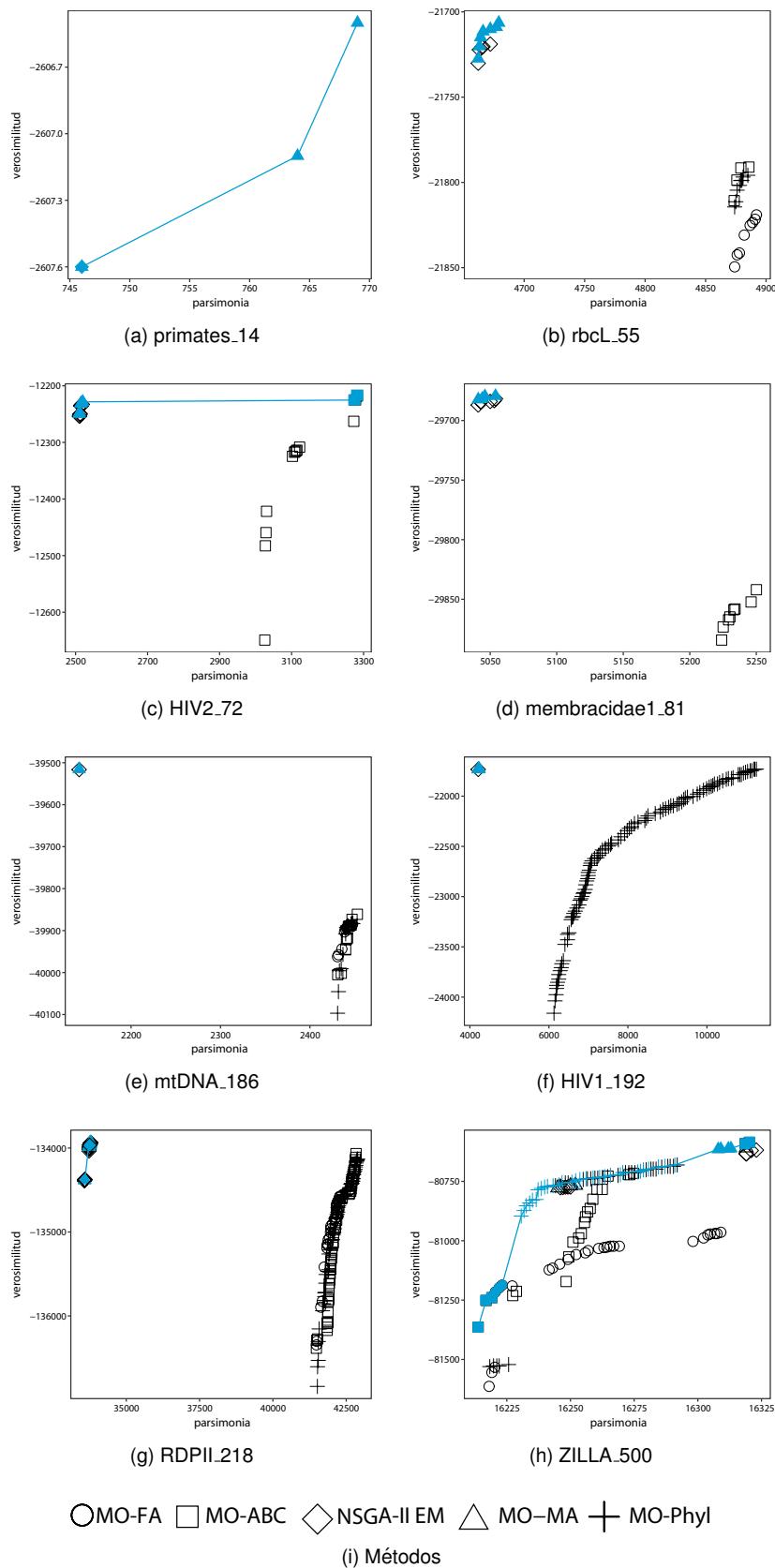


Figura 2.7: Comparación MO-MA con herramientas multi-objetivas. FP global en color.
Fuente: Elaboración propia, 2017.

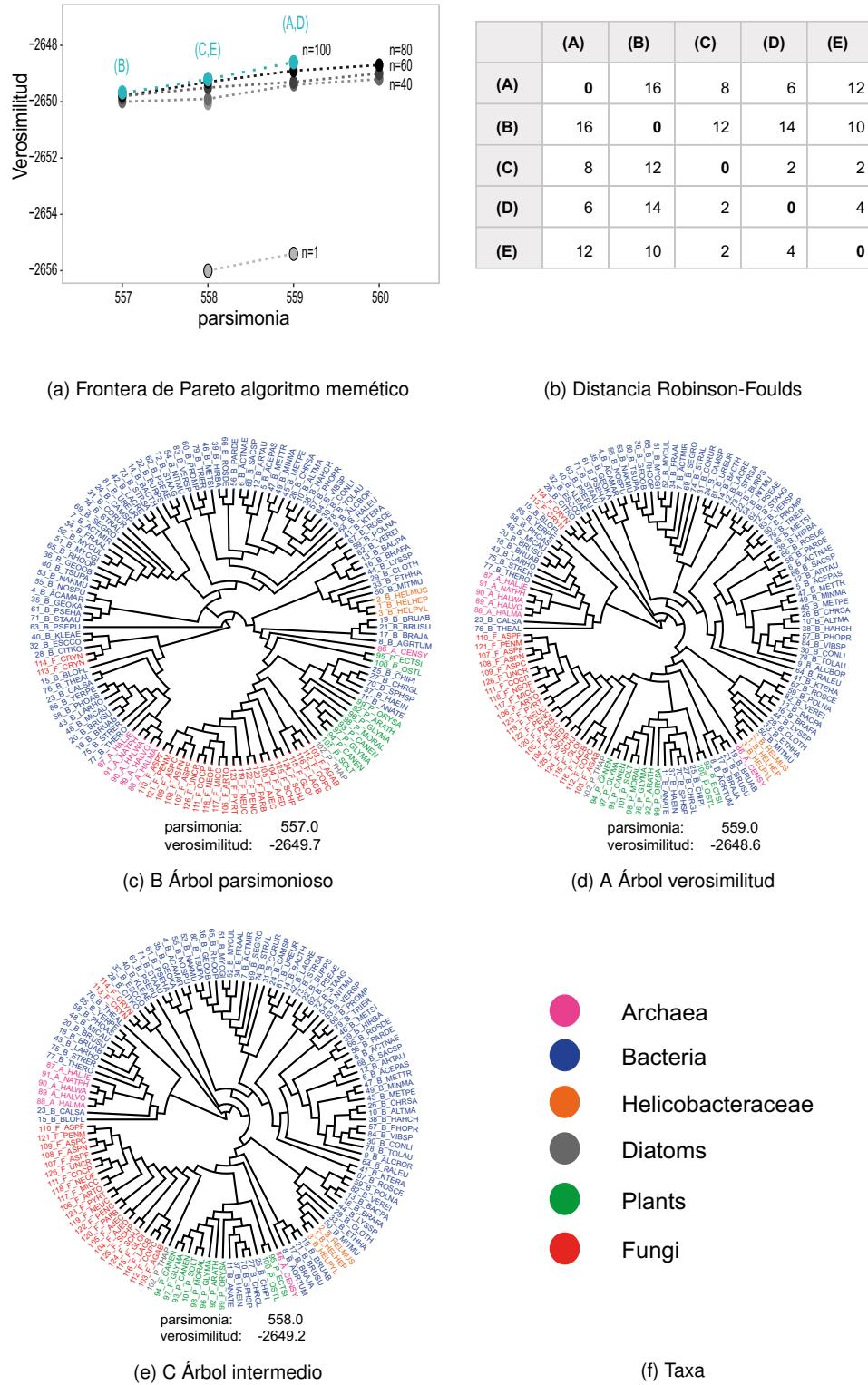


Figura 2.8: Resultados obtenidos por MO-MA usando secuencias de amino ácidos (ureas).
Fuente: Elaboración propia, 2017.

2.4 CONCLUSIONES

En este trabajo se propone una nueva aproximación para enfrentar el problema de inferencia filogenética basada en árboles empleando técnicas evolutivas. La propuesta es alcanzada después de una exhaustiva evaluación de diferentes operadores durante el diseño. En general, los árboles obtenidos con MO-MA tienen mayores valores en los criterios de optimalidad en relación a propuestas de la literatura para la mayoría de los conjuntos de datos estudiados.

Basándose en los resultados presentados en la sección previa, no es posible identificar una buena configuración con una mayor métrica de hipervolumen estadísticamente significante en todos los conjuntos de datos. Sin embargo, la configuración NJ-PDG-NNI-G resulta una estructura con buenas métricas de hipervolumen, cobertura y representación en Frontera de Pareto global.

También características individuales de las configuraciones pueden ser identificadas. Por ejemplo, el operador NNI es parte de las estructuras con mejor rendimiento, mientras que WPGMA es el método para inicialización de población que está presente en las estructuras con peores métricas en todos los conjuntos de datos.

La condición de balance de los operadores de cruzamiento es importante debido a su dependencia del método de selección aplicado sobre la población. Su combinación puede afectar las características de los descendientes. Por ejemplo, cuando un método de selección no estocástico es usado junto a un operador de cruzamiento no balanceado, los descendientes en cada generación recibirán las características de solo uno de los padres, aumentando el riesgo de estancamiento en mínimos locales. Sin embargo, en generaciones avanzadas de la heurística, esta configuración podría preservar las características generales de la población.

El operador de Consenso forma parte de las mejores configuraciones para los tres conjuntos de datos empleados. Cuando este es estudiado individualmente resulta balanceado, recibiendo en igual forma las características de ambos padres considerando la métrica de Robinson-Foulds. Además, este presenta la mayor distancia entre padres y descendientes en relación a los otros operadores de cruzamiento. Lo anterior permite inferir que este operador puede ser usado para recorrer rápidamente el espacio de búsqueda de soluciones en etapas tempranas de la aplicación de una meta-heurística. Por otro lado, cuando sea necesario conservar las características de una población, y al mismo tiempo minimizar el riesgo de estancamiento en mínimos locales, otros operadores como PDG, PDGm y BE pueden ser usados.

En relación a las estrategias de reordenamiento de árboles en la etapa de mutación, SPR y TBR tienen la mayor métrica de Robinson-Foulds comparados a NNI. Sin embargo, no se encuentra diferencia estadística significante entre estos dos operadores. Estos resultados indican

que el uso de TBR no posee ventajas en relación a SPR, lo que es importante debido a que TBR requiere un mayor número de movimiento que SPR para ejecutar una operación (Giribet, 2007), resultando ser más complejo computacionalmente.

Cuando MO-MA considerando una configuración NJ-PDG-NNI-G es comparado con propuestas de la literatura, este obtiene un mejor desempeño que los otros métodos considerando la métrica de hipervolumen en la mayor parte de los conjuntos de datos estudiados. En los otros casos, MO-MA infiere árboles que no son cubiertos por los otros métodos y representan regiones diferentes de la Frontera de Pareto global. Esto significa que MO-MA es capaz de entregar nuevas soluciones al problema. En cuatro conjuntos de datos, no se halla diferencia estadísticamente significante en la métrica de hipervolumen para la comparación de MO-MA y NSGA-II EM. Considerando estos resultados y la comparación con NSGA-II sin el operador de cruzamiento relacionado al modelo evolutivo, es posible inferir que este operador ayuda a maximizar considerablemente el valor de la verosimilitud, mejorando las métricas de evaluación multi-objetivo. La estrategia de búsqueda local efectuó un refinamiento de las soluciones, convirtiendo a MO-MA en un método competitivo comparada a las otras propuestas.

En relación al estudio que incluye secuencias de aminoácidos, no se encuentran diferencias significativas al emplear la prueba de Shimodaira-Hasegawa. Esto demuestra que MO-MA tiene un rendimiento competitivo no sólo desde una perspectiva algorítmica, sino que también considerando significado biológico, construyendo una hipótesis evolutiva validada por la literatura.

A pesar de que los resultados de este trabajo son prometedores, aún existen aspectos importantes por mejorar a nivel algorítmico, como el estudio de búsqueda local considerando diferentes estrategias (operadores, tiempo, condiciones de término, definición de vecindad, entre otros). Por otro lado, se requiere aplicar diferentes técnicas para reducir el espacio de búsqueda usando conocimiento previamente adquirido. Esta idea ha sido aplicada exitosamente en otras áreas como bionformática estructural (Borguesan et al., 2015). Además, diferentes métricas (dispersión, cobertura o hipervolumen) pueden ser incluidas como funciones objetivo, mejorando la calidad de la Frontera de Pareto e incrementando la velocidad de convergencia. Para generar una futura herramienta aplicable se deben explorar diferentes técnicas multi-objetivo para toma de decisiones en el contexto filogenético.

El resultado de esta investigación ha sido publicado en Villalobos-Cid et al. (2017a).

CAPÍTULO 3. COMBINACIÓN DE EVIDENCIA BIOLÓGICA EN INFERENCIA FILOGENÉTICA

Los avances en las técnicas de secuenciamiento durante las últimas décadas y el creciente número de datos disponibles, ha permitido inferir filogenia considerando diversa evidencia biológica: datos morfológicos, sociales, conductuales, ecológicos, y moleculares (alineamiento múltiple de nucleótidos, aminoácidos, y otros marcadores moleculares) (Grechko, 2002; Wilgenbusch et al., 2017). Dependiendo de la evidencia seleccionada para el proceso de inferencia, se pueden obtener diferentes topologías de árboles filogenéticos o hipótesis evolutivas.

Dos paradigmas han sido propuestos para construir hipótesis integrales que combinen la información evolutiva proveniente de diferente evidencia biológica: congruencia taxonómica (*Taxonomic congruence*, TC) (Li & Lecointre, 2009; Yassin et al., 2010; Lobo et al., 2016; Borges et al., 2016) y evidencia total (*Total evidence*, TE) (Zrzavý et al., 2009; Zhang et al., 2015; Gavryushkina et al., 2017; Pyron, 2017). El primero divide la evidencia en conjuntos de datos separados infiriendo hipótesis evolutivas independientes, para luego combinarlos en un único árbol filogenético. Evidencia total combina todos los conjuntos de datos previo al proceso de inferencia filogenética.

La integración de evidencia en inferencia filogenética ha sido tratada como un problema de optimización usando cada paradigma. Este ha sido clasificado en teoría computacional como un problema NP-duro (Chaudhary et al., 2012). Los métodos para construir árboles de consenso trabajan bajo el paradigma de TC (Bryant, 2003). Ellos combinan un conjunto de árboles filogenéticos que incluyen las mismas especies en una única topología usando un criterio específico, ejemplo, matriz de representación por parsimonia (*Matrix representation with parsimony*, MRP) (Levasseur & Lapointe, 2006), árbol promedio basado en distancias de edición (*Average tree based on edition distances*) (Levasseur & Lapointe, 2006), o mezclador de árboles de consenso goloso (*Greedy consensus merger*) (Fleischauer & Böcker, 2016), entre otros. Por otro lado, el paradigma TE combina conjuntos de datos usando diferentes estrategias: concatenación de secuencias, métodos multi-modales (Bicego et al., 2007) y estrategias basadas en super matrices (Queiroz & Gatesy, 2007). En este último caso el proceso de inferencia filogenética también involucra la optimización de un criterio: evolución mínima, mínimos cuadrados (Kidd & Sgaramella-Zonta, 1971; Rzhetsky & Nei, 1993), máxima parsimonia (Cavalli-Sforza & Edwards, 1967; Fitch & Margoliash, 1967), o verosimilitud (Felsenstein, 2004) (Sección 2.1.2). Las estrategias usadas para combinar la evidencia bajo cada paradigma pueden resultar en diferentes hipótesis evolutivas para el mismo conjunto de datos. Sin embargo, ninguno de ellos ha demostrado ser el método definitivo para la combinación de datos. Esto se debe a que cada uno efectúa simplificaciones en etapas diferentes del procesamiento de datos (von Haeseler, 2012).

Modelos basados en optimización multi-objetivo han sido exitosamente aplicados para resolver problemas en bioinformática y biología computacional, demostrando ventajas en relación a los métodos basados en la optimización de objetivo único (Handl et al., 2007). Por ello, como se vió en el capítulo anterior, diferentes estrategias multi-objetivo han sido propuestas para inferir árboles filogenéticos considerando como entrada datos desde una evidencia biológica exclusiva y dos criterios de optimización (Coelho & Zuben, 2007; Cancino & Delbem, 2007; Coelho et al., 2010; Santander-Jiménez & Vega-Rodríguez, 2013a,c,b, 2014, 2016; Zambrano-Vega et al., 2016). Sin embargo, ninguno de ellos ha abordado el problema de combinación de fuentes biológicas.

Uno de los objetivos de esta tesis es la integración de diferentes tipos de evidencia biológica para inferir hipótesis evolutivas reticulares. Es por ello que extraer conocimiento respecto a esta área del conocimiento particular es fundamental. Es necesario comprender el funcionamiento de los paradigmas y estrategias de combinación, su capacidad de constituir hipótesis evolutivas integrales, sin la dependencia de un criterio particular de inferencia. En base a ello, este capítulo propone una aproximación basada en NSGA-II (Deb et al., 2002) (MO-CS) que complementa MO-MA en la inferencia de filogenia considerando múltiples conjuntos de datos y criterios relacionados con los paradigmas TC y TE (Figura 3.1). La propuesta es comparada con los métodos más usados para combinar datos filogenéticos, evaluando su capacidad de generar hipótesis evolutivas integrales. Las principales contribuciones de este trabajo en relación a la actual literatura son:

- Una aproximación multi-criterio basada en NSGA-II para combinar datos en el contexto del problema de inferencia filogenética, considerando:
 - Múltiples entradas (árboles o secuencias alineadas) obtenidas desde cualquier tipo de evidencia biológica.
 - Dos criterios de optimalidad basados en los paradigmas TC y TE con independencia de la elección de un modelo evolutivo.
 - Consideración de información evolutiva a lo largo de ramas.
- Una evaluación de la habilidad de diferentes estrategias para reconstruir una hipótesis evolutiva integral, usando como entrada conjuntos de datos parciales.
- Una comparación entre MO-CS y una estrategia multi-objetivo basada en verosimilitud para combinar conjuntos de datos moleculares.

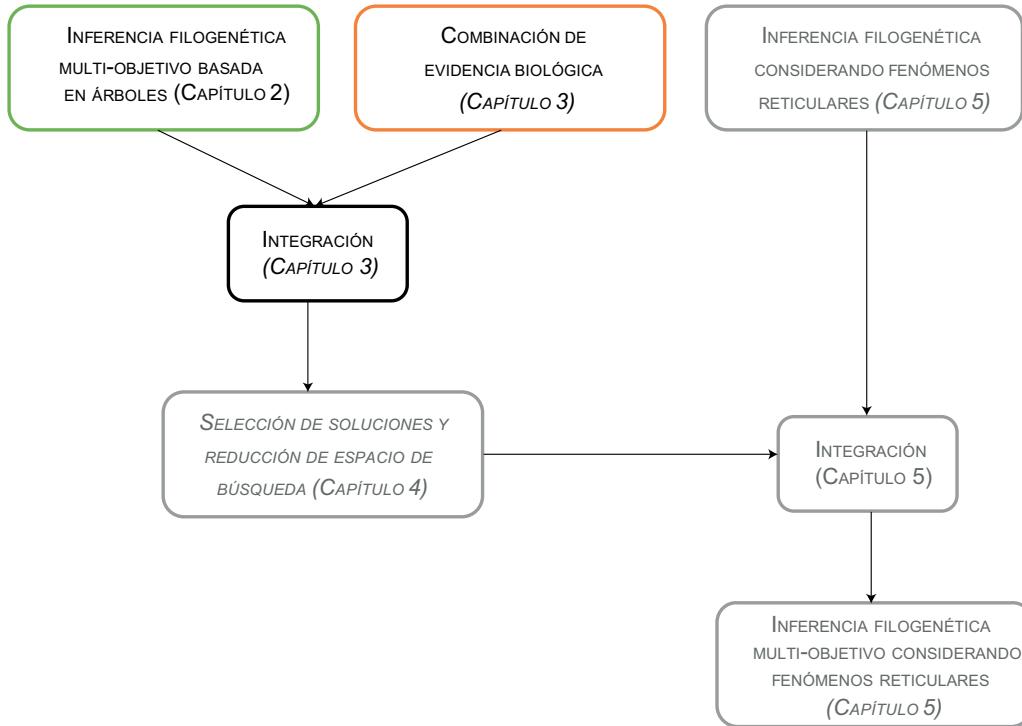


Figura 3.1: Relaciones entre áreas del conocimiento y desarrollo de tesis.

Fuente: Elaboración propia, 2017.

3.1 ANTECEDENTES

3.1.1 Paradigmas de evidencia total y congruencia taxonómica

En la literatura se ha propuesto diferentes paradigmas para realizar inferencia filogenética combinando diferentes fuentes de datos o evidencia biológica. El paradigma de evidencia total (Kluge, 1989) considera que una hipótesis evolutiva para un conjunto específico de especies debe ser efectuada combinando toda la evidencia biológica disponible antes del proceso de inferencia filogenética. Existen varios argumentos en favor de este paradigma (Huelsenbeck et al., 1996; Eernisse & Kluge, 1993), por ejemplo, se ha demostrado que diferentes genes que evolucionan a diferente tasa pueden interactuar positivamente para resolver diferentes niveles de un árbol filogenético (Huelsenbeck et al., 1996). En contraste, el paradigma de congruencia taxonómica considera que la evidencia debe ser estudiada en forma separada infiriendo hipótesis evolutivas independientes, para luego generar una hipótesis unificada. Este paradigma usa la diferencia entre las hipótesis evolutivas independientes para corroborar y dar soporte a la relación entre especies.

Se ha demostrado que la calidad de las hipótesis evolutivas obtenidas bajo el paradigma de TE es directamente proporcional a la homogeneidad de los datos (tasas evolutivas relativas). En cambio, el paradigma de TC trabaja bajo condiciones bajas de homogeneidad (Huelsenbeck et al., 1996). Una tercera alternativa, llamada Combinación condicional de datos (*Conditional data combination*), recomienda la aplicación de TC o TE dependiendo de la homogeneidad de los datos. Esta puede ser evaluada empleando diferentes estrategias: *bootstrapping*, distancias de edición, o índices específicos de homogeneidad (Huelsenbeck et al., 1996; Kumar & Gadagkar, 2001; Leigh et al., 2008).

3.1.2 Estrategias para combinar fuentes biológicas en inferencia filogenética

La literatura dispone de diferentes alternativas para combinar conjuntos de datos usando el paradigma TE. El más simple de ellos involucra la directa concatenación de las secuencias moleculares en un único arreglo o secuencia de caracteres (Dikow, 2009). Este requiere que otros tipos de evidencia biológica sean codificados previamente. Las aproximaciones basadas en súper matrices combinan todos los conjuntos de datos de caracteres en una única matriz filogenética (Figura 3.2a). Esta matriz puede ser construida usando distancias de edición sobre las secuencias concatenadas, o por medio de la unión de múltiples matrices. También se ha propuesto otros métodos de súper matrices más complejos basados en parsimonia o probabilidad Bayesiana. El trabajo desarrollado por Queiroz & Gatesy (2007) efectúa una revisión acabada de estos. Los métodos multi-modales (Bicego et al., 2007) combinan matrices individuales derivadas de cada conjunto de datos usando el promedio (MMEAN), el producto (MPROD), el mínimo (MMAX), o el valor máximo entre cada posición de las matrices (MMIN). Estas matrices también pueden ser combinadas usando pesos. Los métodos multi-modales permiten estudiar conjuntos de datos con diferente tipo de información evolutiva y reloj molecular (Figura 3.2b).

Los métodos de consenso trabajan considerando el paradigma TC. Estos combinan fuentes de árboles obtenidos desde la literatura (mismas especies) en un único árbol (von Haeseler, 2012). Las estrategias más antiguas diseñadas para generar árboles de consenso se basan en divisiones y agrupaciones (*splits and clusters*), e ignoran la información evolutiva de las ramas. Dentro de estas estrategias existen diferentes métodos para combinar las topologías (Bryant, 2003). Por ejemplo, el método de consenso estricto de árboles (*strict consensus tree*) considera exclusivamente las divisiones comunes en todos los árboles de entrada. El método con la regla de mayoría (*majority rule tree*) contiene las divisiones que están presentes en más de la mitad de los árboles de entrada. El método semi-estricto o de loose (*loose o semi-strict consensus tree*) usa exactamente las divisiones que son compatibles con todos los árboles de entrada. Por

último, el árbol de consenso de Adams (*Adam's consensus tree*) considera las divisiones comunes para árboles enraizados, ubicando las especies conflictivas en un nodo aislado en la topología resultante. Estos métodos pueden ser generalizados usando ponderaciones de los árboles de entrada, y frecuentemente pueden resultar en árboles multifurcados (Figura 3.2c).

El método más usado para construir árboles de consenso es la matriz de representación por parsimonia (Levasseur & Lapointe, 2006). Los árboles de entrada son codificados en una única matriz usando una representación binaria, para luego inferir un árbol de consenso empleando el criterio de parsimonia (Figura 3.2d). Otra estrategia usada es el método del árbol de consenso medio o de promedio. Este transforma cada árbol de entrada una matriz de pares de distancia entre especies (matriz cofenética), para luego combinar las soluciones en una única matriz que representa el árbol de consenso. Para este propósito se usa la misma lógica de los métodos multi-modales. Otras alternativas aplican un operador goloso para encontrar el árbol con distancia equivalente entre árboles de entrada usando propiedades geométricas (Bryant, 2003; Jombart et al., 2017). Este operador fue propuesto como estrategia de cruzamiento en la Sección 2.2.1.3.

3.1.3 Métodos para comparación de árboles filogenéticos

La literatura dispone de diferentes métricas para realizar comparaciones entre dos árboles filogenéticos (Wróbel et al., 2012). La más usada de ellas corresponde a la métrica de Robinson-Foulds (Robinson & Foulds, 1981). Esta se define como el mínimo número de ediciones (combinación o separación de nodos) necesarias para transformar un árbol en otro. Otra distancia comúnmente empleada se denomina diferencia de caminos (*path difference*) (Steel & Penny, 1993). Esta cuantifica el número de ramas entre pares de hojas que diferencian a dos árboles filogenéticos. Una versión diferente de esta distancia, denominada diferencia cuadrática de caminos (*quadratic path difference*), incluye la longitud evolutiva de las ramas. El puntaje de rama (*branch score*) (Kuhner & Felsenstein, 1994) calcula la raíz cuadrada de la suma de las diferencias cuadráticas del largo de las ramas considerando las mismas divisiones (*splits*) entre ambos árboles. Una de las últimas métricas propuestas corresponde a la métrica de Kendall-Colijn (Kendall & Colijn, 2016). Esta es capaz de combinar la estructura topológica de dos árboles con la información evolutiva de las ramas.

La diferencia entre múltiples árboles filogenéticos puede ser estudiada mediante el espacio de árboles (*tree-space*) que representa las distancias pares entre árboles. El espacio de árboles puede ser construido aplicando propiedades geométricas (Billera et al., 2001), o por medio del método de escala métrica multidimensional (*Metric Multidimensional Scaling method*,

MMS) (Jombart et al., 2017).

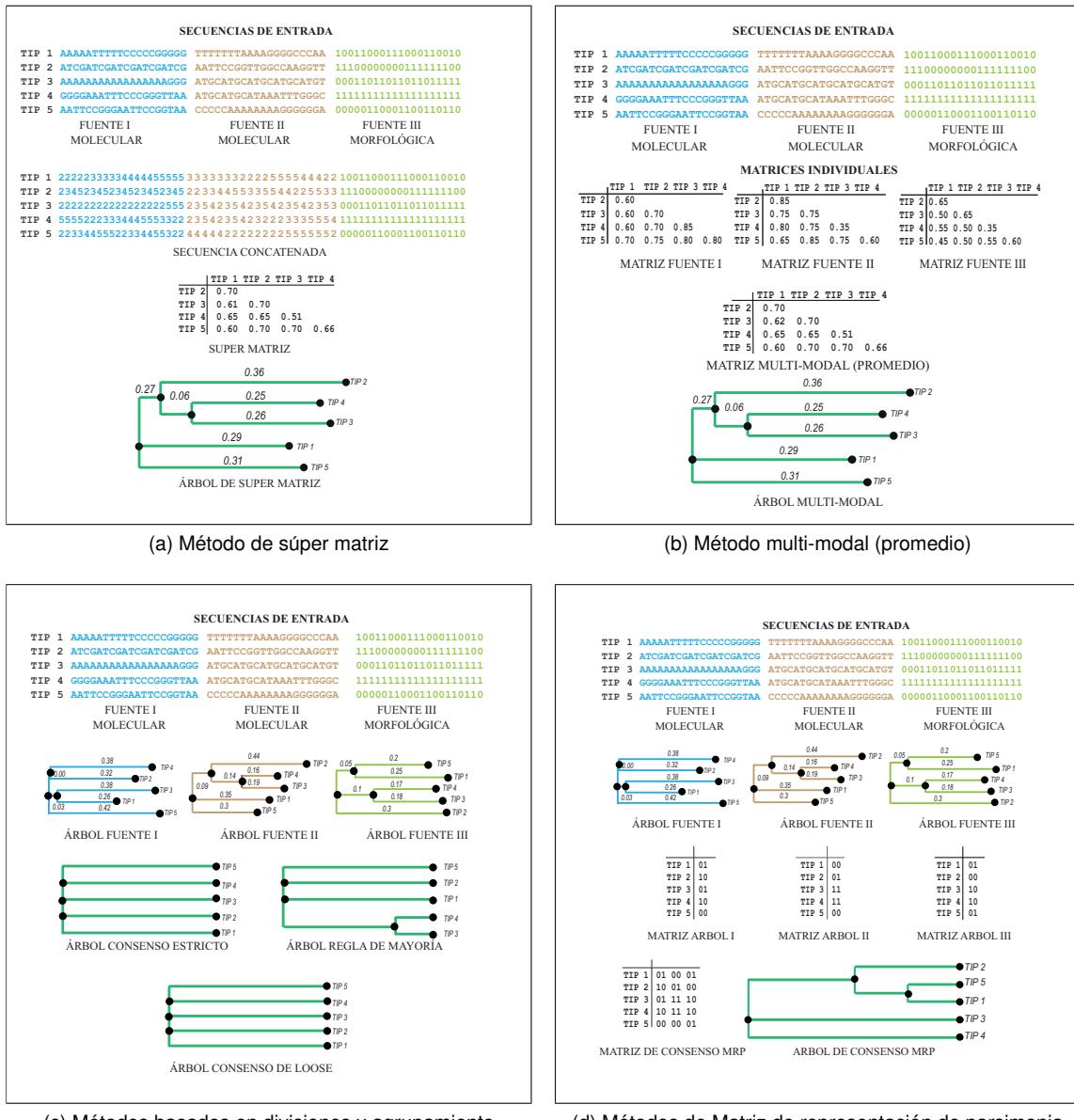


Figura 3.2: Métodos para combinación de evidencia biológica en inferencia filogenética.
Fuente: Elaboración propia, 2017.

3.1.4 Problema multi-objetivo de combinación de evidencia biológica en filogenia

El problema multi-objetivo de combinación de datos puede ser definido empleando la Ecuación 2.3. En este caso x representa una solución (una hipótesis evolutiva asociada a un árbol filogenético) en el conjunto de todas las posibles topologías X , y $z = \vec{f}(x)$ es un vector

objetivo, donde f_1 y f_2 corresponden a criterios asociados a TC y TE (Sección 3.2.1.3). Eernisse & Kluge (1993) demostró que estos paradigmas son conflictivos entre sí, y pueden resultar en diferentes hipótesis evolutivas. Se debe recordar que en un problema de optimización multiobjetivo el conjunto de soluciones óptimas de Pareto corresponde a aquellas soluciones en que es imposible mejorar un objetivo sin empeorar otro (Handl et al., 2007), y que las soluciones no dominadas conforman la Frontera de Pareto.

El algoritmo NSGA-II (Srinivas & Deb, 1994; Deb et al., 2002) ha sido aplicado exitosamente para resolver problemas en bioinformática y otras áreas (Sengupta & Bandyopadhyay, 2012; Ortuno et al., 2012; Parraga-Alava & Inostroza-Ponta, 2016; Hasnat & Molla, 2016). Poladian & Jermini (2006) propusieron una aproximación que combina conjuntos de datos filogenéticos desde diferente evidencia en conflicto. Esta propuesta basada en una estrategia evolutiva fue probada combinando información génica mitocondrial y nuclear empleando el criterio de máxima verosimilitud. También se efectuaron pruebas combinando diferentes secuencias de simios usando el mismo criterio (Jayaswal et al., 2007). A pesar de que estos autores usaron una aproximación multi-objetivo, solo uno de estos criterios fue optimizado (máxima verosimilitud) considerando dos tipos de evidencia biológica. Debido a la elección de este criterio, esta propuesta resultó limitada a la combinación de datos moleculares y a la especificación de un modelo evolutivo.

En este punto de la investigación se propone un algoritmo basado en NSGA-II que optimiza dos criterios considerando los paradigmas de TC y TE, denominado MO-CS. Este es capaz de combinar cualquier tipo de evidencia biológica y no requiere la especificación de un modelo evolutivo.

3.2 APROXIMACIÓN BASADA EN OPTIMIZACIÓN MULTI-OBJETIVO

Esta sección describe el algoritmo multi-objetivo propuesto para abordar el problema de combinación de evidencia biológica en inferencia filogenética. Se detallan diferentes pruebas desarrolladas para su evaluación considerando los métodos más usados en la literatura (Sección 3.1.2).

3.2.1 Descripción de algoritmo

Se ha adaptado el algoritmo NSGA-II para abordar el problema de combinación de evidencia biológica en inferencia filogenética. El pseudo-código del algoritmo es mostrado en el Algoritmo 3.1, donde D corresponde a los conjuntos de entrada, ps es el tamaño de la población, cr y mr son la tasa de cruce y mutación respectivamente. Las siguientes subsecciones describen la estructura de la propuesta.

Algoritmo 3.1: Algoritmo multi-objetivo para combinación de evidencia biológica - MO-CS

Input: D, ps, cr, mr

Output: Población P de árboles (Frontera de Pareto)

```
/* Inicialización de población */  
1  $P \leftarrow INICIALIZAR\_POBLACION(D, ps);$   
2 while condición de término no es alcanzada do  
3   for each  $p$  en  $P$  do  
4     /* Operaciones genéticas */  
5      $[T_1, T_2] \leftarrow SELECCION\_TORNEO\_BINARIO(P);$   
6      $Q[p] \leftarrow CRUZAMIENTO(T_1, T_2, cr);$   
7      $Q[p] \leftarrow MUTACION(Q[p], mr);$   
8   end  
9   /* Actualización de la Frontera de Pareto */  
10   $P \leftarrow ORDENAMIENTO\_NO\_DOMINADO(P, Q, ps);$   
11 end  
12 return  $P;$ 
```

3.2.1.1 Conjunto de datos de entrada

MO-CS fue diseñado para trabajar usando como entradas múltiples secuencias alineadas de caracteres o árboles filogenéticos inferidos previamente. Las secuencias de caracteres deben estar almacenadas en diferentes archivos usando el formato PHYLIP (Felsenstein, 2004), mientras que los árboles filogenéticos deben usar el formato NEWICK (Huson et al., 2011).

3.2.1.2 Inicialización de la población

La función *INICIALIZAR_POBLACION* es responsable de generar la población inicial para cada uno de los paradigmas. TE involucra la combinación de conjuntos de datos previo al proceso de inferencia filogenética. Si la entrada a MO-CS corresponde a un conjunto de secuencias, estas son concatenadas para construir una secuencia unificada. Por otro lado si la entrada es un conjunto de árboles filogenéticos, estos son concatenados usando su matriz de representación por parsimonia (Figura 3.2d). Para ambos casos, un árbol inicial (T_i) es inferido usando el algoritmo NJ sobre el conjunto de datos inicial aplicando la distancia Hamming. La población inicial P es creada usando ps veces el algoritmo NNI sobre el T_i .

3.2.1.3 Criterios de optimalidad

El algoritmo emplea un criterio de optimalidad para cada paradigma. TE es aplicado infiriendo filogenia mediante la maximización del criterio de parsimonia basado en el algoritmo de Fitch. Para ello se emplea como referencia el conjunto de entrada concatenado. El paradigma TC es aplicado por medio de la minimización de los cuadrados entre la matriz cofenética que representa cada individuo, y las matrices cofenéticas que se relacionan con cada árbol de entrada. Ambos criterios son independientes de la especificación de un modelo evolutivo.

3.2.1.4 Operación de cruzamiento

La operación de *CRUZAMIENTO* comienza usando un torneo binario para seleccionar dos padres T_1 y T_2 desde la población P (*SELECCION_TORNEO_BINARIO*). Posteriormente ellos son combinados empleando el operador PDG. Esta operación es aplicada ps veces según la probabilidad cr para construir la población Q .

3.2.1.5 Operación de mutación

La función *MUTACION* modifica los árboles filogenéticos que componen la población Q usando el operador NNI, considerando una probabilidad mr .

3.2.1.6 Ordenamiento no dominado

Las soluciones de P y Q son comparadas usando un algoritmo de ordenamiento no dominado que incluye distancia de aglomeración (*ORDENAMIENTO_NO_DOMINADO*). Las soluciones son ordenadas considerando los criterios a optimizar y dominancia, seleccionando los primeros ps individuos para actualizar la población P .

3.2.2 Parametrización del algoritmo

Parámetros para el algoritmo NSGA-II en el contexto de inferencia filogenética fueron explorados y presentados en el capítulo anterior. Sin embargo, estos parámetros no son aplicables al problema de combinación de evidencia biológica debido a la diferencia de criterios empleados. Con objetivo de parametrizar MO-CS se efectuaron 30 ejecuciones para diferentes valores de cr y mr usando tres conjuntos de datos diferentes, seleccionando aquellos parámetros que maximizan la métrica de hipervolumen (Jiang et al., 2014). Para el cálculo de esta métrica los valores de las funciones objetivos fueron normalizados entre 0 y 1, y el punto de referencia se definió como (2,2).

3.2.3 Evaluación de capacidad para reconstruir hipótesis evolutivas integrales

Se infirió hipótesis evolutivas para diferentes conjuntos de datos usando el criterio de máxima verosimilitud. Estos árboles filogenéticos representan las hipótesis de referencia para cada conjunto de datos (Figura 3.3a y Figura 3.3b).

Con el fin de evaluar la capacidad de cada método para combinar la evidencia biológica disponible, y así inferir hipótesis evolutivas integrales, se dividió cada conjunto de datos en subconjuntos menores compuestos por secuencias parciales con longitud equivalente (3, 5 y 10 subsecciones) (Figura 3.3c). Posteriormente se infirieron los árboles filogenéticos para cada uno de estos subconjuntos de datos, aplicando nuevamente el criterio de máxima verosimilitud (Figura 3.3d).

Los subconjuntos de datos fueron usados por los métodos de combinación de datos filogenéticos. Específicamente, los subconjuntos almacenados como secuencias fueron aplicados como entrada para los métodos multi-modales (Figura 3.3e). Mientras que los árboles filogenéticos inferidos de estos subconjuntos fueron empleados para las estrategias de árboles de consenso: método de consenso estricto (CSSTR), el método de la mayoría (CSMAJ), la matriz de

representación de parsimonia, el árbol promedio (CSAVE), y MO-CS (Figura 3.3f y Figura 3.3g). Debido a que las estrategias basadas en divisiones y agrupamiento no consideran la información evolutiva de las ramas, su longitud se estimó usando el método de ACCTRAN (Swofford & Maddison, 1987).

Para considerar una solución multi-objetivo representativa de MO-CS, se seleccionó el conjunto de Pareto que corresponde a la ejecución que posee la mediana de la métrica de hipervolumen en cada conjunto de datos. Con objetivo de considerar la comparación de la información topológica y evolutiva, se aplicó la métrica de Kendall-Colijn para medir la diferencia entre los árboles resultantes y el árbol de referencia en cada conjunto de datos (Figura 3.3h). También se empleó el MMS para visualizar el espacio de árboles e identificar sus relaciones (Figura 3.3i). El método que construye el árbol con la menor diferencia en relación al árbol filogenético de referencia en cada conjunto de datos, genera la hipótesis evolutiva más exacta.

3.2.4 Comparación entre aproximaciones multi-objetivo

Se ha efectuado una comparación de MO-CS con la estrategia multi-objetivo basada en verosimilitud propuesta por Jayaswal et al. (2007). Esta, por definición, se limita a la combinación de dos conjuntos de datos moleculares, requiriendo de la especificación de un modelo evolutivo. Para su cálculo se empleó el criterio de información de Akaike.

Las secuencias de los conjuntos de datos fueron divididas en dos subconjuntos de igual longitud y ambas aproximaciones multi-objetivas fueron evaluadas aplicando 31 ejecuciones. Los algoritmos fueron parametrizados con el método explicado en la Sección 3.2.2 considerando 100 generaciones (Santander-Jiménez & Vega-Rodríguez, 2013c). La Frontera de Pareto que representa la mediana de la métrica de hipervolumen fue seleccionada, y las diferencias topológicas fueron estudiadas por medio de la métrica de Kendall-Colijn. Las hipótesis evolutivas de referencia fueron inferidas maximizando el criterio de verosimilitud sobre cada conjunto de datos.

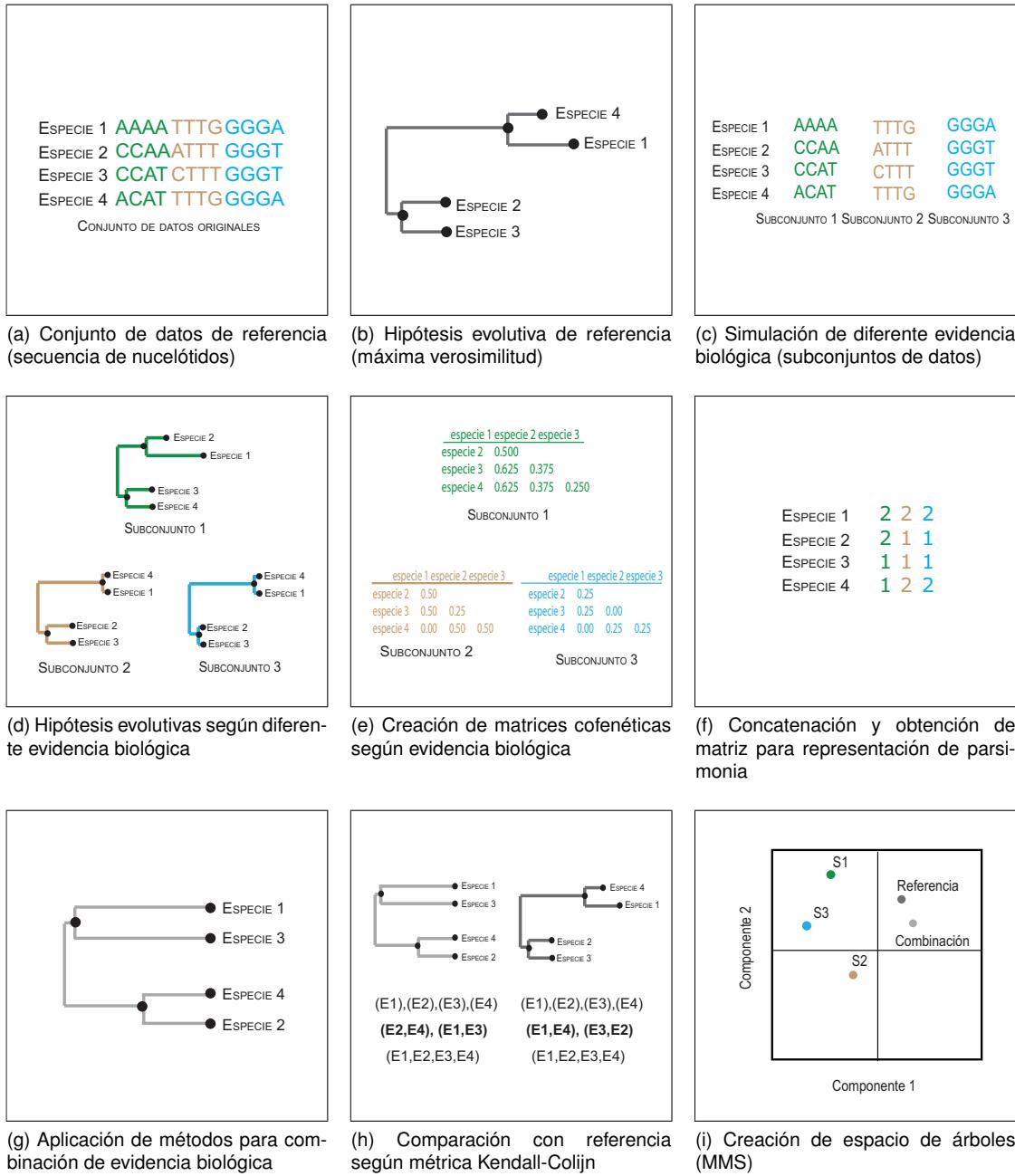


Figura 3.3: Evaluación de la capacidad de generación de hipótesis evolutivas integrales empleando diferentes estrategias para combinación de evidencia biológica (3 subconjuntos).

Fuente: Elaboración propia, 2017.

3.2.5 Aplicación de conjuntos de datos sin hipótesis evolutiva de referencia

MO-CS también fue evaluado usando cuatro conjuntos de datos que no han sido previamente subdivididos (conjuntos de datos sin referencia), ya que estos han sido estudiados directamente en la literatura por divergir en las hipótesis evolutivas resultantes: *dengue_17*, *flu_165*, *fungi_20*, y *haemoglobin_210* (Tablas 3.1 y 3.2). Este último conjunto de datos incluye secuencias de nucleótidos, aminoácidos, y estructura secundaria. Las secuencias de aminoácidos fueron alineadas usando Clustal Omega (Sievers & Higgins, 2014). Mediante el empleo de *tblastn* (Altschul et al., 1990) sobre estas secuencias alineadas se derivaron las secuencias de nucleótidos respectivas. Finalmente la estructura secundaria fue estimada usando la herramienta *Consensus* (Rost et al., 1994; King & Sternberg, 1996; Guermeur et al., 1999). Las soluciones obtenidas para cada conjunto de datos fueron representadas en el espacio de árboles.

3.3 RESULTADOS

Los algoritmos, experimentos y análisis estadísticos desarrollados en esta investigación fueron efectuados usando el entorno R 3.3.2 y RStudio 0.99.491. Particularmente se emplearon las bibliotecas *phangorn* (Schliep, 2011), *phytools* (Revell, 2012) y *emoa* (Mersmann, 2011). Los experimentos fueron realizados usando tres procesadores MS Azure Standard DS5 v2 con 16 núcleos Xeon 56-2673v3 (Haswell), 2.4 GHz, 56 GB Ram y 112 GB de disco duro. Los conjuntos de datos fueron obtenidos desde la literatura relacionada. Las Tablas 3.1 y 3.2 muestran detalles de las secuencias y su correspondiente referencia.

3.3.1 Parametrización del algoritmo

La Tabla 3.3 muestra la métrica de hipervolumen obtenida empleando diferentes tasas de cruzamiento y mutación para diferentes números de generaciones en tres conjuntos de datos. Solo tres combinaciones de parámetros resultaron con un buen desempeño estadísticamente significante, y corresponden a la mejor solución en todos los conjuntos de datos. Específicamente, la mayor métrica de hipervolumen se obtuvo usando 0.75 y 1.00 como tasa de cruzamiento y mutación. Además, se puede observar que la métrica de hipervolumen incrementa proporcionalmente con número de generaciones. Para la aplicación de MO-CS se ha

seleccionado los dos parámetros que maximizaron el hipervolumen, definiendo un máximo de 100 generaciones.

Tabla 3.1: Conjuntos de datos empleados en experimentos

Datos	Tipo	#Especies	#Número
<i>primates_14</i>	secuencias nucleótidos	14	232
<i>dengue_17</i>	árboles	17	500
<i>fungi_20</i>	secuencias nucleótidos	20	1927
<i>haemoglobin_20</i>	secuencias nuc, ami, struc.	20	200
<i>rbcl_55</i>	secuencias nucleótidos	55	1314
<i>HIV2_72</i>	secuencias nucleótidos	72	828
<i>membracidae1_81</i>	secuencias nucleótidos	81	3321
<i>ureasas_126</i>	secuencias aminoácidos	126	609
<i>flu_165</i>	árboles	165	200
<i>haemoglobin_210</i>	secuencias nuc, ami, struc.	210	4214

Fuente: Elaboración propia, (2017).

Tabla 3.2: Referencia conjuntos de datos empleados en experimentos

Datos	Referencias
<i>primates_14</i>	(Felsenstein, 2005)
<i>dengue_17</i>	(Jombart et al., 2017)
<i>fungi_20</i>	(Mahe et al., 2012)
<i>haemoglobin_20</i>	(Berman et al., 2000)
<i>rbcl_55</i>	(Coelho & Zuben, 2007)
<i>HIV2_72</i>	(Santander-Jiménez & Vega-Rodríguez, 2013a)
<i>membracidae1_81</i>	(Santander-Jiménez & Vega-Rodríguez, 2013a)
<i>ureasas_126</i>	(Carlini & Ligabue-Braun, 2016)
<i>flu_165</i>	(Jombart et al., 2017)
<i>haemoglobin_210</i>	(Berman et al., 2000)

Fuente: Elaboración propia, (2017).

Tabla 3.3: Mejores cinco métricas de hipervolumen obtenidas empleando diferentes parámetros para MO-CS. (*hip*: métrica de hipervolumen, *cr*: tasa de cruzamiento, *mr*: tasa de mutación). Las configuraciones que resultaron con diferencias estadísticamente significantes se han marcado en negrita. La mejor configuración ha sido destacada con color gris.

Data set	Generación 1			Generación 30			Generación 60			Generación 100		
	cr	mr	hyper	cr	mr	hyper	cr	mr	hyper	cr	mr	hyper
Haemoglobin_126	0.25	0.25	2.42 (0.06)	0.75	1.00	3.24 (0.22)	0.75	1.00	3.43 (0.27)	0.75	1.00	3.61 (0.31)
	0.00	0.75	2.39 (0.06)	1.00	1.00	3.31 (0.21)	1.00	1.00	3.27 (0.30)	1.00	1.00	3.41 (0.33)
	0.75	0.50	2.38 (0.13)	0.25	1.00	3.00 (0.13)	0.75	0.75	3.16 (0.18)	0.75	0.75	3.31 (0.01)
	0.75	1.00	2.35 (0.10)	1.00	0.75	2.99 (0.03)	0.25	1.00	3.11 (0.01)	0.25	1.00	3.22 (0.01)
	1.00	1.00	2.32 (0.04)	0.50	1.00	2.98 (0.02)	0.00	0.75	3.10 (0.01)	0.00	1.00	3.22 (0.01)
dengue_17	0.00	0.00	2.22 (0.00)	0.75	1.00	3.96 (0.00)	0.75	1.00	3.96 (0.00)	0.75	1.00	3.96 (0.00)
	0.25	1.00	2.22 (0.00)	0.50	0.75	3.96 (0.00)	0.50	0.75	3.96 (0.00)	0.50	0.75	3.96 (0.00)
	0.50	0.00	2.22 (0.00)	0.50	1.00	3.96 (0.00)	0.50	1.00	3.96 (0.00)	0.50	1.00	3.96 (0.00)
	0.50	0.25	2.22 (0.00)	0.25	1.00	0.96 (0.00)	0.25	1.00	3.96 (0.00)	0.25	1.00	3.96 (0.00)
	0.50	0.50	2.22 (0.00)	1.00	1.00	0.95 (0.02)	1.00	1.00	3.95 (0.02)	1.00	1.00	0.95 (0.02)
Haemoglobin_20	0.50	0.00	2.10 (0.03)	0.75	1.00	3.10 (0.05)	0.00	1.00	3.34 (0.08)	1.00	1.00	3.75 (0.09)
	0.50	0.75	2.08 (0.03)	0.00	0.75	3.04 (0.09)	0.25	1.00	3.39 (0.08)	0.00	1.00	3.72 (0.03)
	1.00	0.00	2.07 (0.09)	0.00	1.00	3.04 (0.08)	0.75	1.00	3.39 (0.07)	0.75	1.00	3.71 (0.07)
	0.00	0.00	2.06 (0.05)	0.25	1.00	2.99 (0.08)	1.00	1.00	3.36 (0.11)	0.25	1.00	3.65 (0.08)
	0.75	0.50	2.06 (0.05)	1.00	1.00	2.97 (0.01)	0.00	0.75	3.34 (0.97)	0.25	0.75	3.61 (0.08)

Fuente: Elaboración propia, (2017).

3.3.2 Evaluación de capacidad para reconstruir hipótesis evolutivas integrales

La Tabla 3.4 muestra la métrica de Kendall-Coljin que compara las hipótesis evolutivas de referencia y los árboles filogenéticos inferidos por cada método para combinación de datos. En el caso de *rbcL_55* subdividido en tres subconjuntos de datos, las mejores reconstrucciones fueron efectuadas por el método multi-modal basado en el mínimo valor y MO-CS. En todos los otros conjuntos de datos los mínimos valores de la métrica de Kendall-Coljin fueron obtenidos por MO-CS. Esto significa que las hipótesis evolutivas obtenidas por esta aproximación resultan más parecidas a la hipótesis de referencia respecto a los otros métodos, en términos de estructura e información evolutiva. Esto es independiente del número de subconjuntos de datos.

La Figura 3.4 muestra el espacio de árboles obtenido para cada conjunto de datos considerando tres subconjuntos. Las soluciones más parecidas a la hipótesis evolutiva de referencia fueron obtenidas por MO-CS, siendo parte del mismo cuadrante del espacio de árboles en todos los conjuntos de datos estudiados. Esto significa que las soluciones multi-objetivo comparten la misma relación con los conjuntos de entrada que la hipótesis integral. Las peores reconstrucciones fueron obtenidas usando la regla de consenso estricto y los métodos multi-modales basados en el valor mínimo y de producto. Es importante destacar que en todos los casos las soluciones obtenidas y la hipótesis de referencia no corresponden con el punto equidistante de los conjuntos de entrada.

La Tabla 3.5 muestra el valor de mínimos cuadrados y el puntaje de parsimonia para diferentes conjuntos de datos empleando tres subdivisiones de conjuntos. Los valores obtenidos confirman la relación entre estos criterios y la métrica de Kendall-Coljin, disminuyendo en forma proporcional.

3.3.3 Comparación con aproximación multi-objetivo basado en verosimilitud

La Tabla 3.6 resume la métrica de Kendall-Colijn obtenida para las aproximaciones multi-objetivo considerando cada conjunto de datos. En el caso del conjunto *primates_14*, la solución más similar a la hipótesis de referencia fue obtenida por la aproximación basada en verosimilitud. Sin embargo, el valor de la métrica de Kendall-Colijn difiere solo en una unidad comparada a la solución obtenida por MO-CS. En relación al conjunto de datos *rbcL_55*, ambos métodos coinciden en la misma solución. Sin embargo, la aproximación basada en verosimilitud tiene peor promedio de la métrica Kendall-Coljin entre soluciones no dominadas. En todos los

otros conjuntos de datos MO-CS obtiene soluciones más similares en relación a la hipótesis de referencia, teniendo menor valor de la métrica Kendall-Coljin.

3.3.4 Aplicación de conjuntos de datos sin hipótesis evolutiva de referencia

La Figura 3.5 muestra el espacio de árboles obtenido por MO-CS empleando conjuntos de datos reales sin referencia. Como en los experimentos previos, las soluciones obtenidas no coinciden con los árboles filogenéticos de entrada, o con el centro geométrico del espacio de árboles. Sin embargo, no es posible evaluar la diferencia entre las soluciones obtenidas y la hipótesis de referencia, ya que esta última no ha sido propuesta o validada por la literatura.

3.4 CONCLUSIONES

Ante la decisión de qué paradigma usar para combinar diferente evidencia biológica e inferir hipótesis evolutivas integrales: evidencia total o congruencia taxonómica, se propone una aproximación multi-objetivo (MO-CS) que involucra las ventajas de cada uno de ellos. Ambos paradigmas son aplicados combinando los criterios de máxima representación por parsimonia y los mínimos cuadrados (métodos multi-modales).

La evaluación de diferentes métodos para combinación de datos demuestra que MO-CS es capaz de obtener soluciones empleando conjuntos de datos parciales (múltiple evidencia biológica) que resultan similares a las hipótesis evolutivas de referencia obtenidas por conjuntos de datos completos, maximizando verosimilitud. Las soluciones que componen la Frontera de Pareto tienden a ser similares a la hipótesis de referencia en términos de topología e información evolutiva de las ramas según la métrica de Kendall-Colijn. Estos resultados son independientes del tamaño de las secuencias y el número de subconjuntos. Además, la comparación de los criterios de mínimos cuadrados y parsimonia muestran ser proporcionales a la métrica Kendall-Colijn. Cuando el espacio de árboles es estudiado, las soluciones más cercanas a la hipótesis de referencia también corresponden a los árboles inferidos por MO-CS. En todos los casos, la hipótesis de referencia resulta ser diferente a las hipótesis evolutivas que representan los conjuntos parciales de entradas. Además, ninguno de ellos coincide con el centro geométrico que considera igual distancia a los árboles de cada subconjunto de datos. Esto es importante debido a que algunas estrategias emplean este principio como parte del proceso de inferencia: el

árbol consenso promedio, el árbol consenso mediano, métodos multi-modales, entre otros.

También se compara MO-CS con otra aproximación multi-objetivo basada en verosimilitud diseñada específicamente para combinar dos entradas de datos moleculares. MO-CS, sin depender de un modelo evolutivo, es capaz de encontrar soluciones con menor métrica de Kendall-Colijn en cinco de los seis conjuntos de datos estudiados. Además, los experimentos aplicados sobre conjuntos de datos reales sin referencia demuestran que MO-CS es capaz de trabajar considerando como entradas múltiples marcadores genéticos almacenados como secuencias o árboles filogenéticos.

A pesar de que los resultados son favorables, existen varios puntos por mejorar a nivel algorítmico, por ejemplo, el uso de diferentes técnicas para inicializar poblaciones, efectuar las operaciones genéticas o efectuar búsquedas locales. Así mismo, considerando la actual área de investigación en computación evolutiva en que se proponen nuevos algoritmos que permiten trabajar con más de dos objetivos, es posible diseñar un modelo que incluya nuevos criterios u otras métricas para guiar la búsqueda a través del espacio de soluciones. En este contexto, con el objetivo de generar una herramienta funcional, se hace necesaria la exploración de métodos para toma de decisiones. Otra tarea, desde un punto de vista biológico, es la evaluación de la relación entre el soporte de árboles (bootstrapping, jackknife, entre otros) y los diferentes métodos para combinación de datos. Sin embargo, esta evaluación requiere el apoyo de validación experimental *in vivo* o *in vitro* empleando diferentes marcadores moleculares.

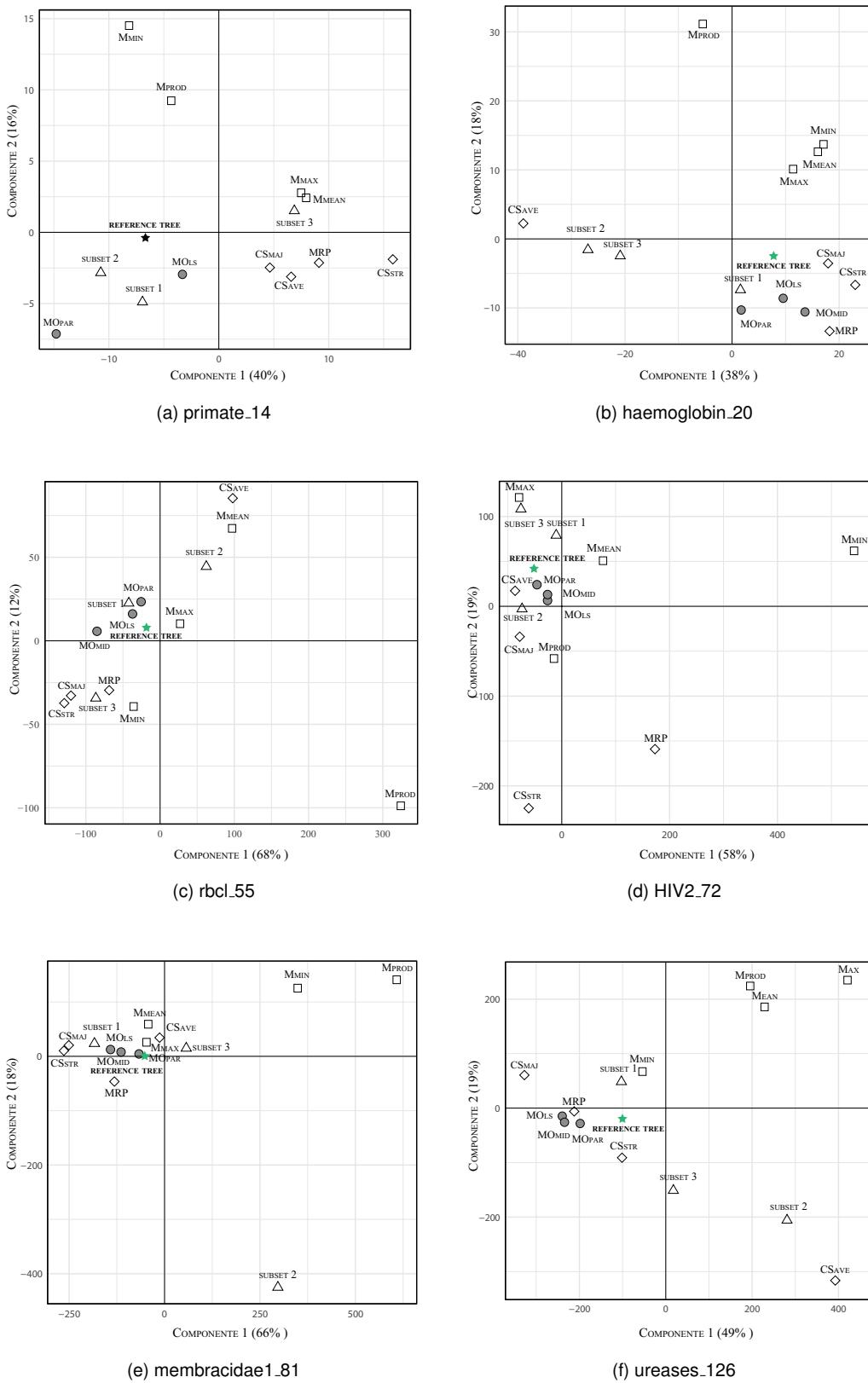


Figura 3.4: Espacio de árboles para diferentes conjuntos de datos obtenidos usando MMS y la métrica de Kendall-Colijn. La hipótesis evolutiva de referencia es mostrada en color verde.
Fuente: Elaboración propia, 2017.

Tabla 3.4: Métrica de Kendall-Coljin entre la hipótesis evolutiva de referencia para cada conjunto de datos, y los árboles obtenidos mediante la aplicación de los métodos para combinación de datos. Los menores valores han sido marcados en color gris. (*MOLS* y *MOPAR* corresponden a las soluciones extremas obtenidas por *MO-CS*, mientras que *MO-MID* es una solución intermedia.)

Data set	# Subsets	Métodos								NSGA-II		
		MRP	MMEAN	MMAX	MMIN	MPROD	CSSTR	CSMAJ	CSAVE	MOLS	MOpar	MO-MID
primates_14	3	19	18	19	21	15	26	15	22	14	9	14
	5	18	17	20	33	34	28	17	17	16	16	17
	10	18	34	40	33	67	56	27	34	16	17	17
haemoglobin_20	3	28	32	43	35	46	35	31	47	26	28	25
	5	39	38	35	36	55	37	35	30	32	13	32
	10	41	41	34	37	58	38	34	29	29	22	31
rbcL_55	3	154	125	110	126	312	191	191	140	110	114	140
	5	165	164	142	402	197	131	142	124	93	56	93
	10	158	145	142	408	530	138	145	164	81	83	82
HIV2_72	3	363	132	129	607	168	295	166	110	127	107	119
	5	324	124	111	230	584	283	255	149	86	75	101
	10	307	148	111	346	675	285	254	119	91	100	97
membracidae1_81	3	239	181	168	466	708	278	277	173	167	194	162
	5	205	213	234	376	530	202	205	187	129	101	236
	10	238	197	215	544	496	205	206	214	182	189	189
ureases_126	3	252	633	479	311	452	308	312	604	220	196	216
	5	237	245	232	781	674	253	379	378	114	254	123
	10	262	193	316	2034	1937	249	272	388	153	153	117

Fuente: Elaboración propia, (2017).

Tabla 3.5: Valores de mínimos cuadrados y parsimonia considerando tres conjuntos de datos con tres subdivisiones cada uno. Los mínimos valores son mostrados en gris.(*MOLS* y *MOPAR* corresponden a las soluciones extremas obtenidas por *MO-CS*, mientras que *MOMID* es una solución intermedia.)

Data set		Métodos								NSGA-II		
		MRP	MMEAN	MMAX	MMIN	MPROD	CSSTR	CSMAJ	CSAVE	MOLS	MOPAR	MOMID
primates_14	min cuadrados	61	28	39	27	30	111	78	56	27	30	27
	parsimonia	37	36	36	36	36	43	42	43	37	35	37
haemoglobin_20	min cuadrados	86	69	75	78	89	77	99	107	89	116	93
	parsimonia	56	98	99	102	97	85	74	57	69	52	66
rbcL_55	min cuadrados	561	427	364	384	832	1712	534	443	481	459	389
	parsimonia	250	247	258	275	251	356	353	259	221	223	231
HIV2_72	min cuadrados	1304	1465	1547	1130	1286	1169	908	1519	856	1017	877
	parsimonia	302	290	307	346	367	811	437	298	307	280	294
membracidae1_81	min cuadrados	1247	1097	1043	1208	1211	3962	1415	2438	1019	1214	1118
	parsimonia	328	326	324	399	382	499	447	293	303	271	295
ureases_126	min cuadrados	6808	5417	4496	5416	3829	10280	8747	4198	1982	2206	2041
	parsimonia	914	918	885	910	925	979	1001	943	569	540	554

Fuente: Elaboración propia, (2017).

Tabla 3.6: Métrica de Kendall-Colijn para comparación entre MO-CS y la aproximación multiobjetivo basada en verosimilitud.

Datos	NSGA-II verosimilitud			MO-CS		
	L1	L2	MID	MOLS	MOPAR	MO-MID
<i>primates_14</i>	18	2	3	16	3	6
<i>haemoglobin_20</i>	29	22	22	40	12	27
<i>rbcL_55</i>	95	95	56	56	65	69
<i>HIV2_72</i>	99	63	80	87	57	72
<i>membracidae_81</i>	184	101	119	162	91	107
<i>ureases_126</i>	242	102	158	213	92	143

Fuente: Elaboración propia, (2017).

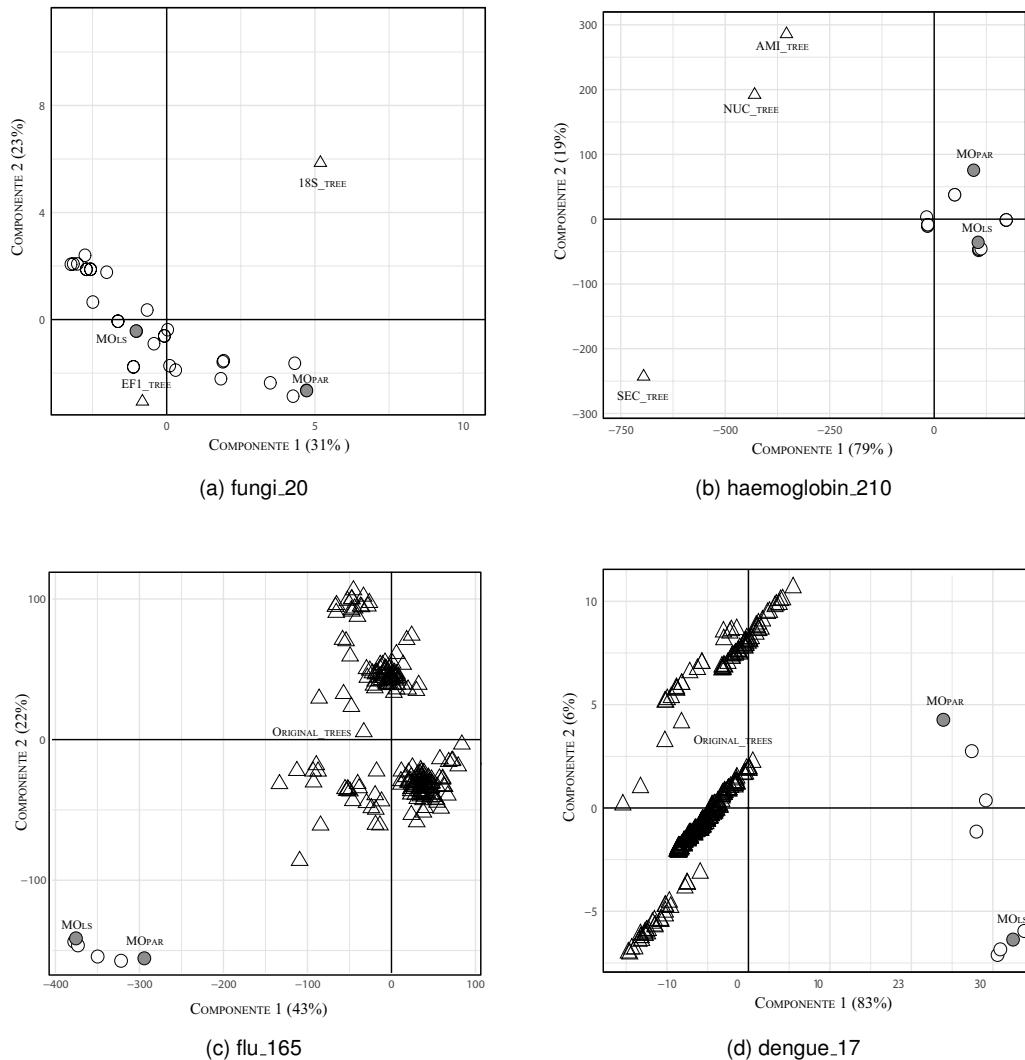


Figura 3.5: Espacio de árboles para diferentes conjuntos de datos obtenidos usando MMS y la métrica de Kendall-Colijn. La hipótesis evolutiva de referencia es mostrada en color negro.

Fuente: Elaboración propia, 2017.

CAPÍTULO 4. REDUCCIÓN DE ESPACIO DE BÚSQUEDA Y TOMADORES DE DECISIONES

Los resultados presentados en los capítulos anteriores evidencian que, con objetivo de construir una herramienta funcional que integre los modelos multi-objetivos empleados en inferencia filogenética, es necesario: (1) la búsqueda de estrategias para reducción de espacio de búsqueda que mejoren la velocidad de convergencia de las meta-heurísticas garantizando soluciones de calidad, y (2) la aplicación de estrategias para toma de decisiones que reduzcan el número de árboles obtenidos desde las Fronteras de Pareto. En este capítulo se exploran ambas áreas, empleando el conocimiento adquirido en los capítulos previos como se muestra en la Figura 4.1.

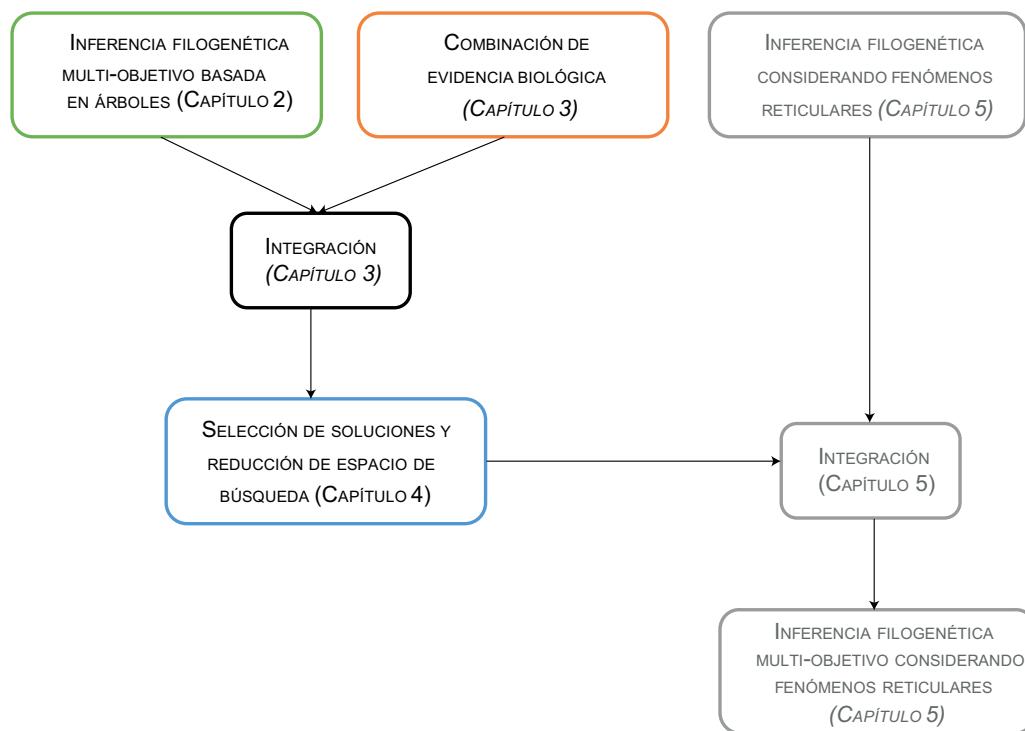


Figura 4.1: Relaciones entre áreas del conocimiento y desarrollo de tesis.
Fuente: Elaboración propia, (2017).

Algunos problemas de bioinformática han sido abordados por la literatura empleando estrategias que orientan la búsqueda de soluciones mediante probabilidad, apoyando la convergencia de los algoritmos con información obtenida experimentalmente. Por ejemplo, algoritmos genéticos diseñados para predecir estructura tridimensional de las proteínas usan información previa sobre la preferencia de ángulos de rotación de su cadena principal en sus operaciones genéticas. Esta es obtenida mediante experimentación no *in silico* (Borguesan et al., 2015). En el contexto del problema de inferencia filogenética las hipótesis evolutivas no pueden

ser validadas directamente mediante esta vía de experimentación, limitando en los algoritmos el uso de topologías previamente construidas que no necesariamente aportan en el mejoramiento de una solución.

Los actuales modelos basados en optimización multi-objetivo en inferencia filogenética presentan estrategias que permiten reducir el espacio de soluciones considerando los criterios a optimizar. La mayor parte de ellas se enfocan en el proceso de inicialización de poblaciones. Por ejemplo, en vez del uso de una población inicial de árboles con una topología aleatoria que puede resultar lejana a las soluciones óptimas, se ha propuesto la utilización de árboles previamente inferidos bajo los criterios de máxima parsimonia o verosimilitud (Coelho et al., 2010; Santander-Jiménez & Vega-Rodríguez, 2013a,c; Zambrano-Vega et al., 2016). Esta estrategia también fue empleada previamente por propuestas basadas en optimización de objetivo único (Lemmon & Milinkovitch, 2002; Katoh et al., 2001), y ha sido implementada en la estrategia descrita en el Capítulo 2: MO-MA. Asumiendo la existencia de una relación entre los criterios de parsimonia y verosimilitud, Stamatakis (2004) ha propuesto otra alternativa de inicialización de poblaciones usando árboles parsimoniosos inferidos por un método no determinista llamado Adición paso a paso (*Stepwise-Addition*). Otras estrategias propuestas para reducir el espacio de búsqueda trabajan optimizando métricas sobre el espacio objetivo (Santander-Jiménez & Vega-Rodríguez, 2016). Sin embargo, los resultados no han demostrado mejoras frente a otros métodos que no incluyen esta clase de estrategias (MO-MA, MO-Phylogenetics, entre otros).

Otro punto que se debe considerar en los actuales modelos multi-objetivo diseñados para inferencia filogenética es la inclusión de tomadores de decisiones. Ello se debe a lo poco práctico que resulta entregar como resultado múltiples topologías y al sesgo asociado a la elección de una de ellas. Diferentes tomadores de decisiones han sido aplicados en problemas de optimización multi-objetivo para resolver problemas en otras áreas del conocimiento: Método del punto de referencia (*Reference Point Method*), Método de utilidad marginal (*Marginal Utility Method*) y métrica L2 (*L2-metric*) (Padhye & Deb, 2011) (Anexo F). No obstante, ninguna de estas estrategias ha sido aplicada en el contexto de inferencia filogenética.

En este punto de la investigación se ha desarrollado:

- Un estudio de la factibilidad del uso de topologías de árboles para apoyar el proceso de inferencia filogenética y reducir el espacio de búsqueda.
- Aplicación y comparación de diversos operadores para toma de decisiones en el problema multi-objetivo de inferencia filogenética de árboles.
- Un nuevo tomador de decisiones basado en topología de árboles.

Las siguientes secciones describen los detalles de cada uno de estos puntos.

4.1 REDUCCIÓN DE ESPACIO DE BÚSQUEDA

El siguiente experimento fue diseñado con fin de conocer la relación entre topología y los criterios empleados en aproximaciones multi-objetivos para inferencia filogenética de árboles. El objetivo es evaluar si las características topológicas de los árboles filogenéticos pueden ser incorporadas en nuevas estrategias para reducir el espacio de búsqueda de soluciones.

4.1.1 Descripción del experimento

Se han inferido 100 árboles filogenéticos usando los conjuntos de datos detallados en la Tabla 2.1. Para cada uno de estos árboles se ha calculado su valor de parsimonia y verosimilitud según los métodos presentados en la Sección 2.2.1.1. Los árboles fueron agrupados empleando el método de Particionando basado en medoides (*Partitioning Around Medoids, PAM*) (Baser & Saini, 2015) y distancia Euclídea. El número de grupos fue estimado optimizando el índice de Silhouette (Baser & Saini, 2015). Paralelamente, se construyeron tres matrices de distancias topológicas empleando (1) la métrica de Robinson-Foulds, (2) Diferencia de caminos y (3) Kendall-Colijn (Sección 3.1.3). Usando cada una de las matrices resultantes se agruparon los árboles con PAM. Los agrupamientos obtenidos al considerar criterios de optimización y topologías fueron comparados mediante el índice de Jaccard (IJ) (Real & Vargas, 1996).

4.1.2 Resultados

La Tabla 4.1 muestra los resultados al comparar las agrupaciones generadas por los criterios de parsimonia y verosimilitud, y las agrupaciones construídas aplicando la métrica de Robinson-Foulds. Se debe recordar que esta última solo considera información topológica, ignorando la información evolutiva de las ramas. Según el IJ no existe relación entre topología y las agrupaciones obtenidas al emplear los dos criterios de optimalidad, alcanzado un máximo valor de 0.6. La mayor relación se obtuvo para el conjunto de datos *primates_14* con un valor de 0.8.

Los resultados presentados en la Tabla 4.3 demuestran que la relación entre agrupaciones se mantiene al considerar la métrica de diferencia de caminos, variando levemente el número de grupos (k). Esta métrica ignora la topología y considera exclusivamente la información evolutiva de las ramas. Al considerar la métrica Kendall-Colijn (Tabla 4.2) que

combina información evolutiva y topológica, aumenta levemente la relación entre agrupaciones. Por ejemplo, el IJ en el conjunto *membracidae1_81* aumenta desde 0.4 a 0.8. Sin embargo, el valor general de IJ continúa bajo, sin establecer una relación entre topología y objetivos.

Los resultados del experimento anterior demuestran que no existe relación entre los criterios de parsimonia y verosimilitud, y las características topológicas de los árboles filogenéticos al emplear métricas de diferenciación. Esto quiere decir que la generación de estrategias para mejorar el recorrido del espacio de búsqueda en un modelo, no debe considerar árboles previos de la literatura a menos que hayan sido inferidos empleando los mismos criterios de optimización. Esto explica el éxito de las estrategias para inicialización de poblaciones en las actuales propuestas. La diferencia entre topologías y criterios también evidencia la inexistencia de reglas para las soluciones de una Frontera de Pareto obtenida por un modelo basado en optimización multi-objetivo. Por ejemplo, un mismo punto de la Frontera de Pareto o un punto del espacio objetivo, podrá representar topologías de árboles totalmente diferentes. Por el contrario, dos regiones lejanas del espacio objetivo pueden representar topologías similares. Esto tiene relevancia a nivel algorítmico al efectuar descarte de árboles basado en el espacio de soluciones (distancia de aglomeración), ya que, si bien se puede estar descartando dos soluciones equivalentes desde el punto de vista de los criterios de optimización, estas pueden diferir en la hipótesis evolutiva, desestimando soluciones con significado biológico. Debido a que los árboles corresponden a un caso particular de las redes filogenéticas, es altamente probable que este fenómeno se repita en este tipo de representación.

Tabla 4.1: Comparación entre agrupamiento de árboles obtenido empleando parsimonia y verosimilitud, y agrupamiento generado considerando la métrica de Robinson-Foulds.

Datos	parsimonia		verosimilitud		par. \cap ver.	
	<i>k</i>	IJ	<i>k</i>	IJ	<i>k</i>	IJ
<i>primates_14</i>	4.0	0.8	4.0	0.2	4.0	0.5
<i>rbcL_55</i>	12.0	0.1	21.0	0.0	21.0	0.1
<i>HIV2_72</i>	10.0	0.2	2.0	0.6	2.0	0.6
<i>membracidae1_81</i>	12.0	0.4	20.0	0.1	2.0	0.5
<i>HIV1_192</i>	7.0	0.2	29.0	0.3	45.0	0.1
<i>RDP II_218</i>	2.0	0.5	2.0	0.5	2.0	0.5
<i>ZILLA_500</i>	2.0	0.6	2.0	0.6	2.0	0.6

Fuente: Elaboración propia, (2017).

Tabla 4.2: Comparación entre agrupamiento de árboles obtenido empleando parsimonia y verosimilitud, y agrupamiento generado considerando la métrica de Diferencia de caminos.

Datos	parsimonia		verosimilitud		par. \cap ver.	
	k	IJ	k	IJ	k	IJ
<i>primates_14</i>	4.0	0.8	4.0	0.2	4.0	0.5
<i>rbcL_55</i>	12.0	0.1	21.0	0.0	21.0	0.1
<i>HIV2_72</i>	10.0	0.2	2.0	0.4	2.0	0.4
<i>membracidae1_81</i>	12.0	0.4	22.0	0.0	2.0	0.4
<i>HIV1_192</i>	7.0	0.2	29.0	0.2	44.0	0.1
<i>RDP II_218</i>	2.0	0.5	2.0	0.5	2.0	0.5
<i>ZILLA_500</i>	2.0	0.6	2.0	0.6	2.0	0.6

Fuente: Elaboración propia, (2017).

Tabla 4.3: Comparación entre agrupamiento de árboles obtenido empleando parsimonia y verosimilitud, y agrupamiento generado considerando la métrica de Kendall-Colijn.

Datos	parsimonia		verosimilitud		par. \cap ver.	
	k	IJ	k	IJ	k	IJ
<i>primates_14</i>	4.0	0.8	4.0	0.2	4.0	0.5
<i>rbcL_55</i>	12.0	0.1	21.0	0.0	21.0	0.1
<i>HIV2_72</i>	10.0	0.1	2.0	0.3	2.0	0.3
<i>membracidae1_81</i>	12.0	0.3	16.0	0.0	2.0	0.8
<i>HIV1_192</i>	7.0	0.2	29.0	0.2	39.0	0.1
<i>RDP II_218</i>	2.0	0.5	2.0	0.5	2.0	0.5
<i>ZILLA_500</i>	5.0	0.1	2.0	0.6	2.0	0.6

Fuente: Elaboración propia, (2017).

4.2 TOMADORES DE DECISIONES

En la literatura existen diferentes tomadores de decisiones que trabajan sobre el espacio objetivo. En este punto de la investigación se ha adaptado estas estrategias para abordar el problema multi-objetivo de inferencia de árboles filogenéticos. El detalle de ellos se encuentra en el Anexo F. Para su evaluación se ha empleado las Fronteras de Pareto obtenidas por MO-MA usando los conjuntos de datos presentados en la Tabla 2.1. Las estrategias implementadas son:

- Método del punto de referencia (RPLIK, RPMD, RPPAR)
- Método de utilidad marginal (MU)
- Métrica L2 (L2).

Además de estos tomadores de decisiones se ha propuesto una nueva estrategia basada en características topológicas.

4.2.1 Descripción del experimento

Los diferentes tomadores de decisiones han sido aplicados y representados sobre el espacio de soluciones. A diferencia de los operadores que trabajan sobre este espacio, se propone un nuevo operador basado en topología. Esta estrategia compara los árboles de las soluciones empleando la métrica de Kendall-Colijn. La matriz de distancia obtenida es usada para construir un agrupamiento empleando el método de PAM con k igual uno. El árbol central es definido como el árbol representativo del conjunto de soluciones (Medoide). Posteriormente, el método PAM es aplicado nuevamente generando un número k de agrupaciones. Los k árboles medoides son estimados y presentados como soluciones representativas secundarias. Los árboles obtenidos para cada conjunto de datos han sido representados visualmente usando MMS (Sección 3.1.3).

4.2.2 Resultados

Las Figuras 4.2 a 4.6 muestran los diferentes árboles seleccionados por los tomadores de decisiones en los diversos conjuntos de datos. Cada uno de ellos ha escogido diferentes árboles desde la Frontera de Pareto, sin evidenciar una relación con su posicionamiento en el espacio de árboles. El valor máximo obtenido para k fue de cinco grupos.

El experimento anterior evidencia una vez más la inexistente relación entre el posicionamiento de los árboles en el espacio objetivo, y su distancia en el espacio de árboles según su topología. Esto implica que el uso de tomadores de decisiones basados en el espacio objetivo y en las características topológicas de árboles, variarán en la elección de una solución. Los árboles representativos seleccionados por las estrategias que emplean el espacio objetivo también resultaron diferentes. Esto significa que cada alternativa para toma de decisiones puede ser empleada indistintamente, no existiendo una alternativa superior a otra.

Los resultados de una investigación más profunda que involucra estos y otros criterios de inferencia han sido publicados en Villalobos-Cid et al. (2017b).

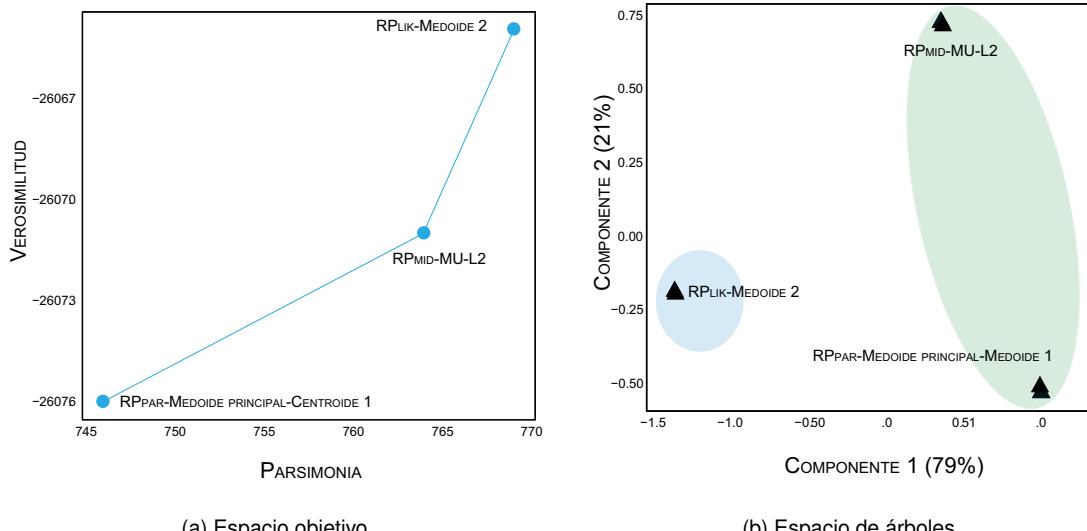


Figura 4.2: Tomadores de decisiones: primates_14.
Fuente: Elaboración propia, (2017).

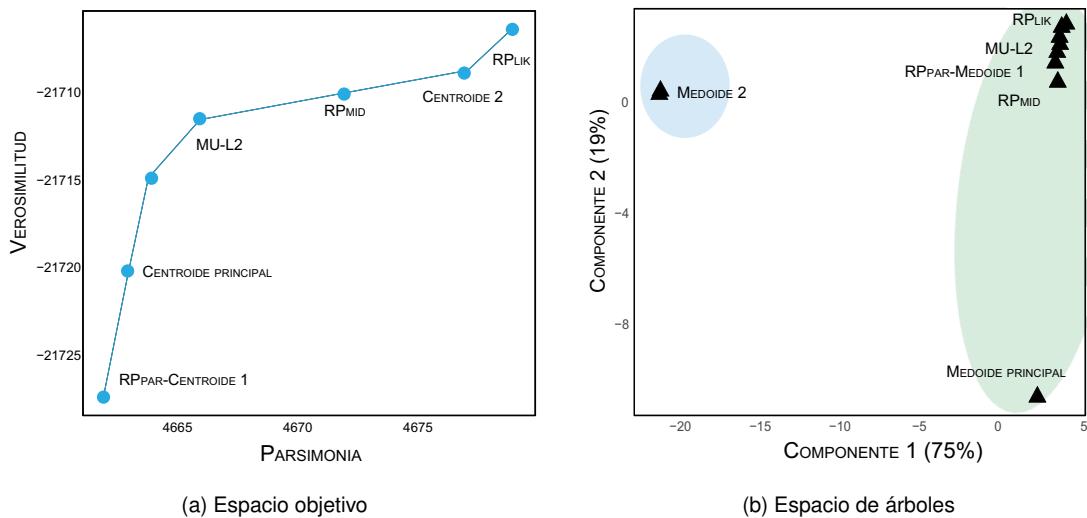


Figura 4.3: Tomadores de decisiones: rbcL_55.
Fuente: Elaboración propia, (2017).

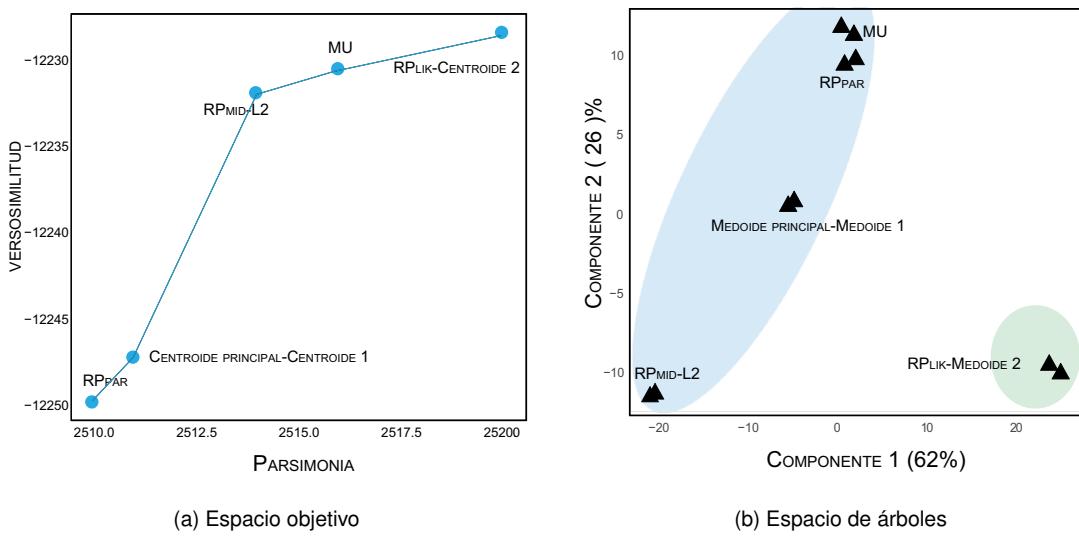


Figura 4.4: Tomadores de decisiones: HIV2_72.
Fuente: Elaboración propia, (2017).

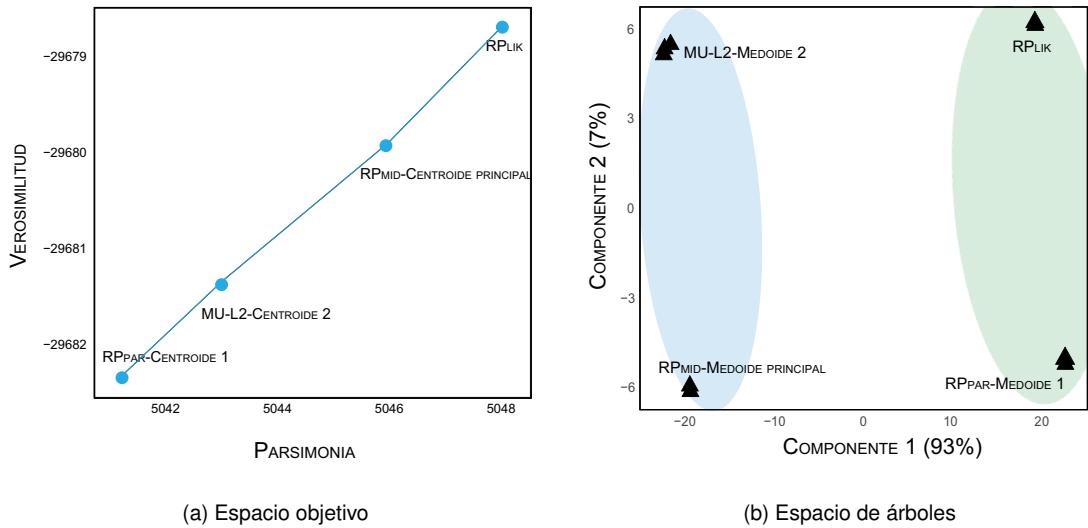


Figura 4.5: Tomadores de decisiones: membracidae1_81.
Fuente: Elaboración propia, (2017).

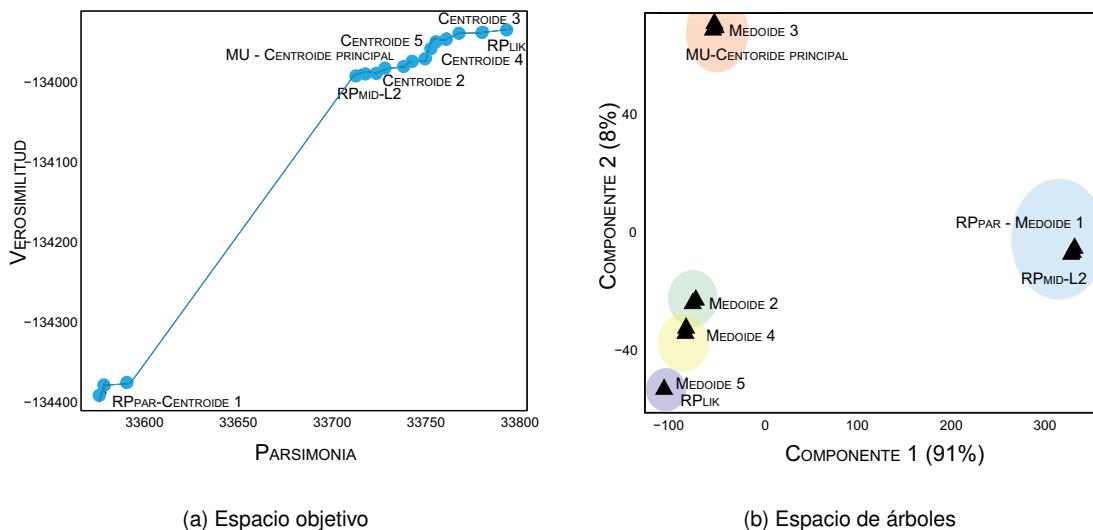


Figura 4.6: Tomadores de decisiones: ZILLA_500.
Fuente: Elaboración propia, (2017).

CAPÍTULO 5. INFERENCIA FILOGENÉTICA MULTI-OBJETIVO BASADO EN REDES

Hasta este punto de la investigación, a modo de resumen, el Capítulo 2 ha demostrado que mediante el modelamiento de inferencia filogenética multi-objetivo se consigue reducir el sesgo asociado a la elección de un criterio de inferencia de árboles filogenéticos, permitiendo obtener un espectro de soluciones que no depende de la elección de uno de ellos en particular. Además, se consiguió comprender el efecto de los diferentes operadores que componen las actuales propuestas, caracterizando sus efectos sobre el modelamiento. Por otro lado, en el Capítulo 3 se propuso una estrategia basada en optimización multi-objectivo que permite la inferencia de árboles filogenéticos empleando diferente evidencia biológica, obteniendo soluciones que consideran dos paradigmas de combinación: TE y TC. Esta estrategia fue capaz de obtener hipótesis evolutivas integrales más exactas en relación a una hipótesis de referencia, comparada con las actuales propuestas de la literatura. Por otro lado, en el Capítulo 4 se reveló que en el modelamiento de árboles no existe relación entre espacio objetivo y espacio de soluciones, lo que condiciona el proceso de inicialización de las actuales propuestas y evidencia el no aseguramiento del significado biológico de las soluciones. Además se demostró que no existe un tomador de decisiones ideal para abordar el problema de inferencia y así reducir el sesgo asociado a la elección de una topología desde una Frontera de Pareto. Este conocimiento debe ser considerado para abordar el objetivo principal de este trabajo, asociado al modelamiento de fenómenos reticulares (Figura 5.1).

En los capítulos anteriores (2, 3 y 4) se ha modelado inferencia filogenética representando las hipótesis evolutivas como árboles filogenéticos. Para ello se asume que la información genética se transfiere exclusivamente de un ancestro a sus descendientes en forma vertical. No obstante, existen otros mecanismos evolutivos que consideran la transferencia de información entre descendientes, transferencia horizontal, y no pueden ser reproducidos por este tipo de representación: paralelismo, convergencia, reversión, hibridación, introgresión, recombinación, transferencia horizontal de genes, fusión de genoma, entre otros (Debevec & Whitfield, 2013; Wheeler, 2015). Para la modelar de estos fenómenos, denominados fenómenos reticulares, es necesario ampliar el concepto de árbol al de red filogenética (Huson et al., 2011).

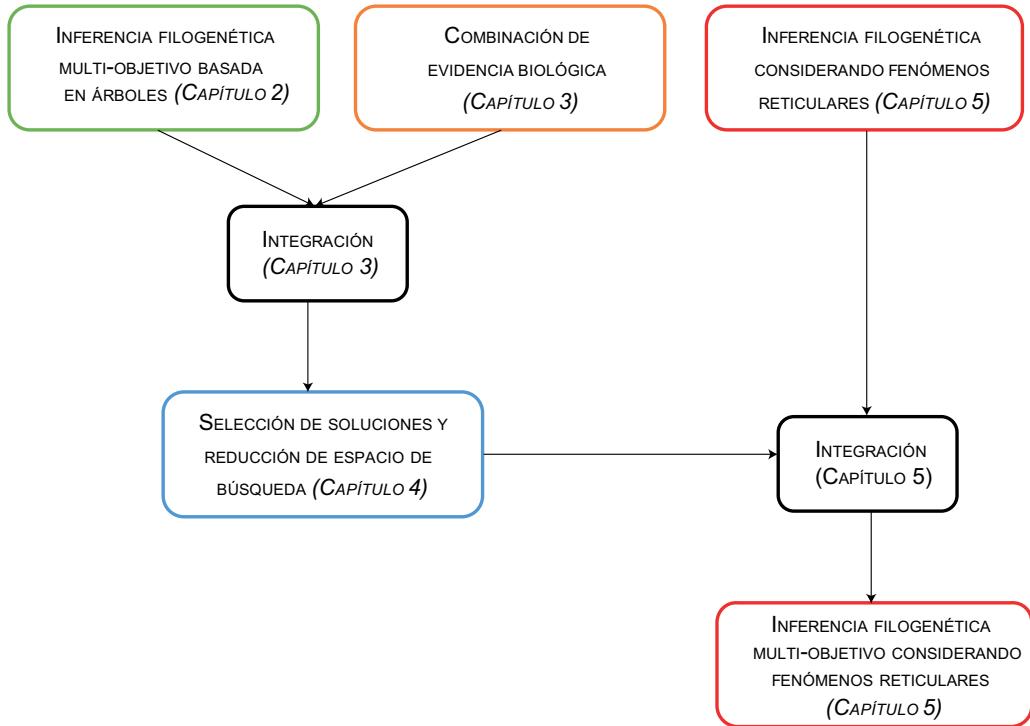


Figura 5.1: Relaciones entre áreas del conocimiento y desarrollo de tesis.

Fuente: Elaboración propia, (2017).

Al igual que un árbol filogenético, una red puede ser enraizada o no enraizada. Se ha diseñado diferentes estrategias para su construcción dependiendo del tipo de red y fenómeno reticular a representar. Algunas de las estrategias se basan en algoritmos exhaustivos, mientras que otras emplean aproximaciones para conseguir topologías que maximicen algún determinado criterio como parsimonia (Wheeler, 2015; Fischer et al., 2015), o verosimilitud (Jin et al., 2006; Yu & Nakhleh, 2015). Se ha demostrado que determinar las topologías que optimizan estos criterios es un problema NP-duro (Yu & Nakhleh, 2015).

Las redes filogenéticas también han sido aplicadas para representar hipótesis evolutivas provenientes de evidencia biológica conflictiva entre sí (Wheeler, 2015) (Capítulo 3). Además de los paradigmas para la combinación de múltiples fuentes (congruencia taxonómica y de evolución total), las redes filogenéticas pueden ser creadas empleando dos sentidos para sus topologías: *softwired* y *hardwired*. Dependiendo del que sea usado, se generan diferentes topologías para un mismo conjunto de especies. Incluso se ha formulado criterios diferenciales para su construcción, como parsimonia con sentido *softwired* y *hardwired*. Lo anterior, junto con la elección de un criterio a optimizar y un modelo evolutivo para el cálculo de verosimilitud, incorpora sesgo en la construcción de una hipótesis evolutiva, repitiendo las dependencias identificadas en el proceso de inferencia de árboles filogenéticos (Capítulo 2).

Como ha sido evidenciado en los capítulos anteriores, una de las ventajas de los

modelos basados en optimización multi-objetivo es que pueden considerar criterios conflictivos entre sí en la obtención de soluciones. En este capítulo se presenta un modelo que buscar resolver la pregunta de investigación planteada en la Sección 1.2, empleando el conocimiento adquirido al estudiar el proceso de inferencia filogenética basado en árboles. Para ello se ha propuesto MO-PhyNet, un algoritmo multi-objetivo derivado del algoritmo NSGA-II capaz de inferir redes filogenéticas enraizadas empleando tres criterios de optimización: (1) búsqueda de la mínima red enraizada (*Minimum rooted phylogenetic network problem*), (2) parsimonia *hardwired*, y (3) *softwired*. Particularmente, el aporte de la investigación descrita en el capítulo corresponde a:

- Una aproximación multi-criterio basada en NSGA-II que permite combinar datos en el contexto del problema de inferencia filogenética empleando representación reticular, considerando:
 - Múltiples entradas (árboles o secuencias alineadas) proveniente de cualquier tipo de evidencia biológica.
 - Tres criterios de optimalidad considerando los sentidos topológicos *hardwired* y *softwired*.
 - Independencia de un modelo evolutivo.
- Estudio de la relación entre espacio de búsqueda y espacio objetivo.
- Una evaluación de la habilidad de diferentes estrategias para reconstruir hipótesis evolutivas reticulares.

5.1 INFERENCIA FILOGENÉTICA BASADA EN REDES

5.1.1 Redes filogenéticas

Una red filogenética es un grafo usado para representar las relaciones evolutivas entre un conjunto de organismos. Corresponde a una generalización de la definición de árboles filogenéticos. Al igual que en estos, los diferentes organismos estudiados son representados por nodos, mientras que las longitudes de las ramas corresponden a tiempo o distancia evolutiva. Una red filogenética también puede ser clasificada como enraizada o no enraizada, dependiendo de todos los organismos comparten un ancestro común.

Una red filogenética puede ser construida empleando como entrada múltiples matrices de distancia, secuencias, o árboles filogenéticos. Cuando una red filogenética contiene

en su topología todos los árboles filogenéticos que se usaron para su generación, se dice que la red tiene un sentido *hardwired*. Por otro lado, si esta no contiene la totalidad de los árboles filogenéticos y, sin embargo, mantiene las mismas agrupaciones de todas las especies que forman estos árboles (*clusters*), se denomina red con sentido *softwired*.

La Figura 5.2 muestra un esquema de ambos tipos de redes filogenéticas, asumiendo como entrada dos árboles filogenéticos T_1 y T_2 . Las agrupaciones que componen a cada uno de estos árboles son señaladas en la Ecuación 5.1, destacándose las comunes entre ambas topologías. La red inferior izquierda tiene un sentido *hardwired* porque contiene en su topología a ambos árboles filogenéticos de entrada o, en otras palabras, contiene las agrupaciones que componen a ambos árboles. En cambio, la red inferior derecha tiene un sentido *softwired* ya que contiene en su topología todas las agrupaciones comunes entre ambas topologías de entrada.

$$T_1 = (\mathbf{A}) (\mathbf{B}) (\mathbf{C}) (\mathbf{D}) (\mathbf{E}) (\mathbf{F})(B, C)(D, E)(A, B, C)(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}) (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{F}) \quad (5.1)$$

$$T_2 = (\mathbf{A}) (\mathbf{B}) (\mathbf{C}) (\mathbf{D}) (\mathbf{E}) (\mathbf{F})(A, B)(C, D)(C, D, E)(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}) (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \mathbf{F})$$

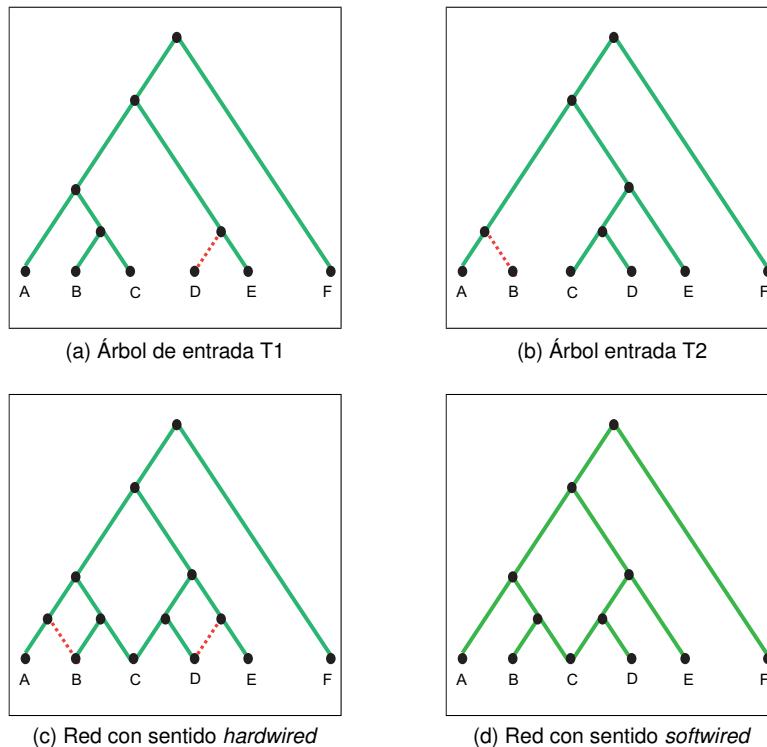


Figura 5.2: Esquema de redes filogenéticas con sentido *hardwired* y *softwired*.
Fuente: (Huson et al., 2011).

5.1.2 Métodos para inferencia de redes filogenéticas

Las estrategias que permiten generar una red filogenética dependen del mecanismo evolutivo a representar y si la red es enraizada o no (Figura 5.3). Una alternativa para representar redes no enraizadas es la aplicación de redes de descomposición (*split network*). Una división (*split*) corresponde a la generación de dos subgrupos complementarios desde un conjunto de organismos a estudiar. Este concepto difiere del de agrupamiento (Sección 5.1.1), ya que es independiente de la topología de la red. Las divisiones de una red pueden ser calculadas empleando el algoritmo *convex hull*, o por medio del algoritmo *circular network* (Huson et al., 2011). A su vez, las estrategias basadas en descomposición dependen del tipo de dato de entrada: matriz de distancia, árboles, o secuencias de caracteres. Entre los métodos basados en distancia se encuentra el método para descomposición de divisiones (*split decomposition method*) (Bandelt & Dress, 1992), y el método de *Neighbor-net* (Bryant & Moulton, 2004). El método de descomposición de divisiones usa una matriz de distancia para generar y consensuar todas las divisiones posibles. Esto último produce pérdida de resolución cuando se estudian más de 100 organismos (Huson et al., 2011). *Neighbor-net* es una generalización de la estrategia *Neighbor joining* (Sección 2.1.2) y corresponde a uno de los métodos más usados en la actualidad. Esto se debe a la fácil visualización de la red resultante, ya que el consenso de divisiones es menos conservativo que el usado por el método para descomposición de divisiones, mostrando un menor número de reticulaciones. Existen otras estrategias que usan como entrada árboles filogenéticos. En este caso, una red filogenética puede ser inferida empleando una generalización de los métodos que permiten construir árboles de consenso (Sección 3.1.3), basándose en la compatibilidad de divisiones. Ejemplos de ello son la regla estricta para consenso de divisiones (*strict consensus split rule*), y la regla de mayoría de consenso (*majority consensus split rule*). También se puede obtener divisiones empleando el algoritmo *convex hull* sobre los caracteres que diferencian un conjunto de secuencias binarias alineadas. En una red mediana (*median network*) (Bandelt et al., 2000) cada nodo representa una secuencia, mientras que las ramas corresponden a la posición del carácter que diferencia a ambos nodos que une. Las redes quasi-medianas (*quasi-median networks*) son una generalización de las redes medianas para considerar secuencias no binarias. Estas pueden ser construidas empleando redes de expansión mínimas (*minimum spanning network*) o el algoritmo *Median-joining* (Bandelt et al., 1999). Además de la representación empleando redes de descomposición, también se ha propuesto el uso de redes de haplotipo y reticulogramas. Para su construcción se puede aplicar criterios basados en parsimonia (Templeton & Sing, 1993) o verosimilitud (Fischer et al., 2015).

Los métodos para construcción de redes filogenéticas enraizadas se basan en el concepto de agrupamiento. En palabras simples, un agrupamiento corresponde al conjunto de

especies que son incluidas por cada subárbol que conforma una red. Existen diferentes tipos de redes enraizadas, por ejemplo, las redes de hibridación (hybridization network) consideran que, además de la existencia de mecanismos de transferencia verticales representados por un árbol filogenético, existen eventos específicos de hibridación génica. Una red de este tipo será capaz de contener un conjunto de árboles filogenéticos de entrada con el mínimo número de eventos reticulados. Hasta hace poco tiempo, solo se podía aplicar este método para consensuar dos topologías de árboles. Sin embargo, recientemente Albrecht (2015) ha propuesto un algoritmo que permite encontrar todas las redes de hibridación para múltiples árboles binarios. Otras estrategias han sido propuestas para representar otros mecanismos de transferencia génica, como redes de recombinación (*recombination networks*) (Hein, 1993; Gusfield et al., 2003), y redes DLT (*duplication-loss-transfer*) (Tofigh et al., 2011). Al igual que para redes no enraizadas, se han propuesto estrategias basadas en criterios de optimización que maximizan parsimonia (Wheeler, 2015), o verosimilitud (Yu & Nakhleh, 2015).

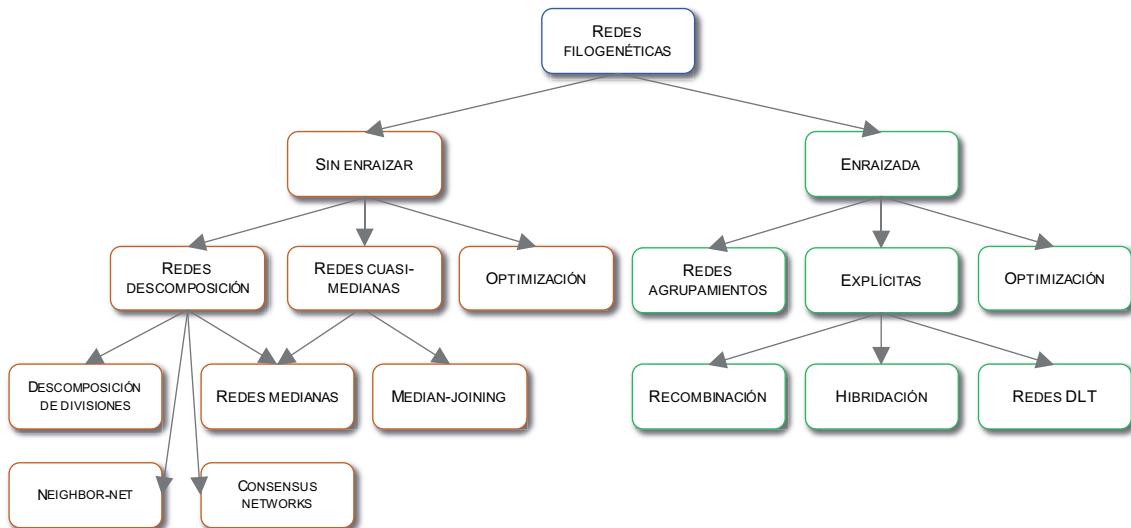


Figura 5.3: Esquema con métodos para inferencia de redes filogenéticas.

Fuente: (Huson et al., 2011).

5.1.3 El problema de inferencia multi-objetivo de redes filogenéticas

Ya se discutió en capítulos anteriores que los métodos para inferencia de árboles filogenéticos poseen sesgo asociado a: (1) la elección de un criterio a optimizar para establecer una determinada topología, (2) el modelo evolutivo definido para el cálculo de verosimilitud, (3) la fuente biología empleada para el proceso de inferencia, y (4) el paradigma asociado en la combinación de datos. Además de estos puntos, al proceso de inferencia filogenética basado en

redes se debe sumar otra fuente más de sesgo: el sentido softwired o hardwired de la red.

Los capítulos previos también han demostrado que los modelos basados en optimización multi-objetivo han conseguido reducir el sesgo asociado al proceso de inferencia de árboles. Considerando el conocimiento adquirido en estas áreas de la investigación, es posible efectuar una definición para el problema de inferencia filogenética multi-objetivo basado en redes. Para esto se puede emplear una derivación de la Ecuación 2.3:

$$\text{maximizar } \vec{z} = \vec{f}(x) = (f_1(x), f_2(x), f_3(x)), x \in X \quad (5.2)$$

En este caso, x es una solución correspondiente a una red filogenética enraizada en el conjunto de todas las posibles topologías o soluciones X , y $z = \vec{f}(x)$ es el vector de objetivos. Específicamente, f_1 equivale al negativo de la función que cuenta el número de eventos reticulados (*minimum rooted phylogenetic network problem*) (Huson et al., 2011), f_2 corresponde a la función de parsimonia en sentido *hardwired*, y f_3 a la función de parsimonia en sentido *softwired*. Perfectamente se puede incluir otras funciones en $\vec{f}(x)$ relacionados a verosimilitud, sin embargo, este depende de la estimación de un modelo evolutivo.

La mínima red filogenética reticulada que representa un conjunto de árboles filogenéticos de entrada en sentido *softwired*, es aquella que posee el número mínimo de nodos con reticulaciones. Por otro lado, la función de parsimonia en un sentido *hardwired* de una red filogenética (N) corresponde a una generalización del concepto de parsimonia en árboles. Es decir, la búsqueda de la red filogenética que minimiza el número de cambios necesarios ($w(e)$) entre caracteres (C) de una secuencia, para explicar los datos de entrada (Ecuación 5.3). El problema de este criterio es que su valor para una red filogenética puede ser muy superior al obtenido para los diferentes árboles filogenéticos que la componen (t en $T(N)$). Esto no ocurre con criterio de parsimonia en un sentido *softwired*, ya que este corresponde a la sumatoria de la parsimonia mínima obtenida entre los diferentes árboles que forman parte de la red para cada una de las secuencias de entrada (Ecuación 5.4).

$$Par_{hard}(N, C) = \sum_{c \in C} \sum_{e \in N} w_c(e) \quad (5.3)$$

$$Par_{soft}(N, C) = \sum_{c \in C} \min_{(t \in T(N))} t^c \quad (5.4)$$

En este trabajo se ha propuesto un algoritmo basado en NSGA-II que optimiza estos criterios, siendo capaz de inferir redes filogenéticas enraizadas desde cualquier tipo de fuente biológica, sin requerir la especificación de un modelo evolutivo.

5.2 APROXIMACIÓN BASADA EN OPTIMIZACIÓN MULTI-OBJETIVO

Esta sección describe un algoritmo multi-objetivo propuesto para abordar el problema de inferencia de redes filogenéticas, MO-PhyNet, y como se ha integrado el conocimiento adquirido en los capítulos anteriores para su construcción. También se detalla diferentes experimentos efectuados para su evaluación.

5.2.1 Descripción de algoritmo

En el Algoritmo 5.1 se ha adaptado el algoritmo NSGA-II para abordar el problema de inferencia filogenética basado en redes. La estructura principal de la propuesta es la misma empleada por MO-CS, donde D corresponde a los conjuntos de entrada, ps es el tamaño de la población, cr y mu son la tasa de cruce y mutación respectivamente. Las siguientes subsecciones describen la estructura de MO-PhyNet.

Algoritmo 5.1: Algoritmo genético MO para inferencia de redes filogenéticas - MO-PhyNet

Input: D, ps, cr, mr, nt

Output: Población P de redes filogenéticas (Frontera de Pareto)

```
/* Inicialización de población */  
1  $P \leftarrow INCIALIZACION\_POBLACION(D, ps)$   
2 while condición de término no es alcanzada do  
3   for each  $p$  en  $P$  do  
4     /* Operaciones genéticas */  
5      $[R_1, R_2] \leftarrow SELECCION\_TORNEO\_BINARIO(P);$   
6      $Q[p] \leftarrow CRUZAMIENTO(R_1, R_2, cr);$   
7      $Q[p] \leftarrow MUTACION(Q[p], mr);$   
8   end  
9   /* Actualización de la Frontera de Pareto */  
10   $P \leftarrow ORDENAMIENTO\_NO\_DOMINADO(P, Q, ps);$   
11 end
```

5.2.1.1 Conjunto de datos de entrada

Al igual que MO-CS, MO-PhyNet fue diseñado para trabajar usando como entradas múltiples secuencias alineadas de caracteres o árboles filogenéticos inferidos previamente. Las secuencias de caracteres deben estar almacenadas en diferentes archivos usando el formato PHYLIP, mientras que los árboles filogenéticos deben usar el formato NEWICK.

5.2.1.2 Inicialización de la población

El concepto de población varía en relación a MO-MA y MO-CS. En esta nueva aproximación, la función *INCIALIZAR_POBLACION* está asociada a una población formada por ps individuos, donde cada uno corresponde a una red filogenética. Cada individuo representa al consenso entre un conjunto de nt árboles filogenéticos. Estos corresponden a los genes en el contexto de algoritmos evolutivos. El valor de nt fue definido como 30, según los resultados de experimentos parciales previamente desarrollados.

Los conjuntos de entrada son combinados igual que MO-CS (Sección 3.2.1.2). Los árboles filogenéticos de cada individuo son generados aplicando nt NNI operaciones sobre el árbol inicial T_i . Posteriormente un operador de consenso es aplicado para generar las ps topologías de red que componen la población P . Se ha empleado este operador de mutación, ya que el proceso de investigación relacionado a MO-MA concluyó que era el operador más eficiente (Capítulo 2).

5.2.1.3 Criterios de optimalidad

MO-PhyNet considera tres criterios de optimalidad: (1) mínima red enraizada, (2) parsimonia *softwired*, y (3) parsimonia *hardwired*. Todos estos criterios han sido implementados en R como una nueva función siguiendo las estrategias desarrolladas por Wheeler (2015). Se ha seleccionado estos criterios con objetivo de reducir dos fuentes de sesgo: (1) la elección de un modelo evolutivo y (2) la dependencia de un sentido de la red.

5.2.1.4 Operación de cruzamiento

La operación de cruzamiento comienza usando un torneo binario para seleccionar dos padres R_1 y R_2 desde la población P . La operación se efectúa por medio de un operador uniforme que se aplica según probabilidad cr . Este selecciona aleatoriamente e intercambia los árboles que componen R_1 y R_2 para generar dos nuevos individuos, R'_1 y R'_2 , que representan sus redes de consenso. Padres y descendientes son comparados según dominancia, dejando en la población Q las dos soluciones dominantes o con mayor distancia de aglomeración.

5.2.1.5 Operación de mutación

Los árboles filogenéticos que componen cada individuo de la población Q son modificados usando el operador NNI considerando una probabilidad mr . Las topologías de las redes Q cuyos árboles han sido modificados son recalculadas. La aplicación de estos operadores en redes se fundamenta en la estructura basal de los individuos, compuestos por árboles filogenéticos.

5.2.1.6 Ordenamiento no dominado

Las soluciones de P y Q son comparadas usando un algoritmo de ordenamiento no dominado que incluye distancia de aglomeración. Las soluciones son ordenadas considerando este criterio y los primeros ps individuos son seleccionados para actualizar la población P .

5.2.2 Parametrización del algoritmo

MO-PhyNet fue parametrizado con los mismos valores obtenidos para el componente genético de MO-MA (Capítulo 2). La personalización de los parámetros no fue efectuada, debido al elevado tiempo requerido para efectuar las ejecuciones necesarias para esta tarea (Anexo G).

5.2.3 Evaluación de reconstrucción de hipótesis evolutivas reticuladas

MO-PhyNet fue evaluado considerando 31 ejecuciones, empleando diferentes conjuntos de datos que incluyen árboles filogenéticos y secuencias moleculares (Tabla 5.1 y 5.2). Con objetivo de caracterizar su funcionamiento, se ha calculado la métrica de hipervolumen normalizando el valor de las soluciones entre 0 y 1. El punto de referencia empleado fue de (2,2).

5.2.4 Espacio de soluciones y espacio objetivo

Se seleccionaron como soluciones representativas de cada conjunto de datos, aquellas que componen la Frontera de Pareto cuyo hipervolumen corresponde a la mediana de todas las ejecuciones. Con objetivo de comprender las ventajas de emplear un modelo basado en optimización multi-objetivo en inferencia de redes filogenéticas se ha efectuado dos experimentos. En una primera instancia, se ha evaluado la diferencia entre las topologías de las redes que componen las Fronteras de Pareto empleando una adaptación de la métrica Robinson-Foulds. Si no existe diferencia entre las topologías obtenidas, el uso de un modelo basado en optimización multi-objetivo carece de sentido y puede ser reemplazado por la optimización de un único objetivo (o la combinación lineal de estos). También se ha contrastado la diferencia de topologías (espacio de soluciones) con los criterios de optimización empleados (espacio objetivo). Para ello se ha efectuado el mismo procedimiento descrito en el Capítulo 4 que compara los agrupamientos obtenidos sobre ambos espacios por medio del IJ.

5.2.5 Comparación con otras propuestas

A nivel funcional no existe una amplia gama de estrategias para la construcción de redes filogenéticas enraizadas, por lo que se ha decidido comparar MO-PhyNet con una estrategia basada en consenso empleando una regla estricta (ConsensusNET). Se efectuó una ejecución para cada conjunto de datos, debido a que este último emplea un algoritmo exacto para la construcción de una topología (Holland et al., 2004).

5.3 RESULTADOS

Los algoritmos, experimentos y análisis estadísticos desarrollados en esta investigación fueron efectuados usando el entorno R 3.3.2 y RStudio 0.99.491. Particularmente se emplearon las bibliotecas *phangorn* (Schliep, 2011), *phytools* (Revell, 2012), *emoa* (Mersmann, 2011), e *igraph* (Csardi & Nepusz, 2006). Los experimentos fueron realizados usando un procesador HP ProLiant SL230s Gen8 con 10 núcleos Intel Xeon E5-2660, 48 GRam y 128 GB de disco duro. Los conjuntos de datos fueron obtenidos desde la literatura relacionada. Las Tablas 5.1 y 5.2 muestran los detalles de las secuencias utilizadas y sus correspondientes referencias.

Tabla 5.1: Conjuntos de datos empleados en experimentos

Datos	Tipo	#Species	#Número
<i>dengue_17</i>	Árbol	17	500
<i>fungi_20</i>	Secuencias DNA	20	1927
<i>haemoglobin_20</i>	Árbol	20	2
<i>ureasas_126</i>	Secuencias AA-DNA	126	609
<i>ureasas_126</i>	Árbol	126	2
<i>flu_165</i>	Árbol	165	200

Fuente: Elaboración propia, (2017).

Tabla 5.2: Referencia conjuntos de datos empleados en experimentos

Datos	Fuente
<i>dengue_17</i>	(Jombart et al., 2017)
<i>fungi_20</i>	(Mahe et al., 2012)
<i>haemoglobin_20</i>	(Berman et al., 2000)
<i>ureasas_126</i>	(Carlini & Ligabue-Braun, 2016)
<i>flu_165</i>	(Jombart et al., 2017)

Fuente: Elaboración propia, (2017).

5.3.1 Evaluación de reconstrucción de hipótesis evolutivas reticuladas

Las Tablas 5.3, 5.4 y 5.5 muestran los valores de cada objetivo para las soluciones extremas, pertenecientes a la Frontera de Pareto de la ejecución que consiguió la mediana de hipervolumen en cada conjunto de datos. Esta métrica apenas tuvo una variación del 5% considerando todas las ejecuciones.

Por definición se podría esperar la existencia de una fuerte relación entre el número de reticulaciones y el criterio de parsimonia *hardwired*. Esto ocurrió al estudiar los conjuntos más pequeños, *haemoglobin_20* y *dengue_17*, ya que la solución con menor número de reticulaciones obtuvo al mismo tiempo el mejor puntaje de parsimonia *hardwired*. Sin embargo, este fenómeno no ocurrió en los conjuntos de datos de mayor tamaño. También se puede observar esta relación al graficar las Fronteras de Pareto obtenida para cada conjunto de datos (Figuras 5.4 y 5.5). En este caso las Fronteras de Pareto de los conjuntos de datos más pequeños, *fungi_20*, *haemoglobin_20* y *dengue_17*, aparecen representadas como una diagonal dentro del espacio objetivo, mientras que en el resto de los conjuntos esta genera una superficie.

Tabla 5.3: Evaluación de MO-PhyNet: Solución de menor número de reticulaciones

Nombre	Conjunto de datos Detalle	Hiper.	Reticulación		
			#Ret	Hard	Soft
fungi_20	DNA-DNA	7.46 (0.24)	1	5822	2911
flu_165	Árbol	7.89 (0.03)	21	125572	62389
haemoglobin_20	Árbol	7.18 (0.33)	4	158	78
ureasas_126	AA-DNA	7.43 (0.41)	1	73326	36663
ureasas_126	Árbol	7.34 (0.33)	6	2296	1146
dengue_17	Árbol	7.58 (0.14)	5	18777	9271

Fuente: Elaboración propia,(2017).

Tabla 5.4: Evaluación de MO-PhyNet: Solución de mejor puntaje de parsimonia con sentido *hardwired*

Nombre	Conjunto de datos Detalle	Hiper.	hardwired		
			#Ret	Hard	Soft
fungi_20	DNA-DNA	7.46 (0.24)	4	5821	2901
flu_165	Árbol	7.89 (0.03)	23	125483	62329
haemoglobin_20	Árbol	7.18 (0.33)	4	158	78
ureasas_126	AA-DNA	7.43 (0.41)	14	73233	36580
ureasas_126	Árbol	7.34 (0.33)	206	2292	1116
dengue_17	Árbol	7.58 (0.14)	5	18777	9271

Fuente: Elaboración propia,(2017).

Tabla 5.5: Evaluación de MO-PhyNet: Solución de mejor puntaje de parsimonia con sentido *softwired*

Nombre	Conjunto de datos Detalle	Hiper.	softwired		
			#Ret	Hard	Soft
fungi_20	DNA-DNA	7.46 (0.24)	10	5831	2886
flu_165	Árbol	7.89 (0.03)	152	125584	61950
haemoglobin_20	Árbol	7.18 (0.33)	17	159	70
ureasas_126	AA-DNA	7.43 (0.41)	13	73257	36579
ureasas_126	Árbol	7.34 (0.33)	205	2297	1108
dengue_17	Árbol	7.58 (0.14)	22	18776	8583

Fuente: Elaboración propia, (2017).

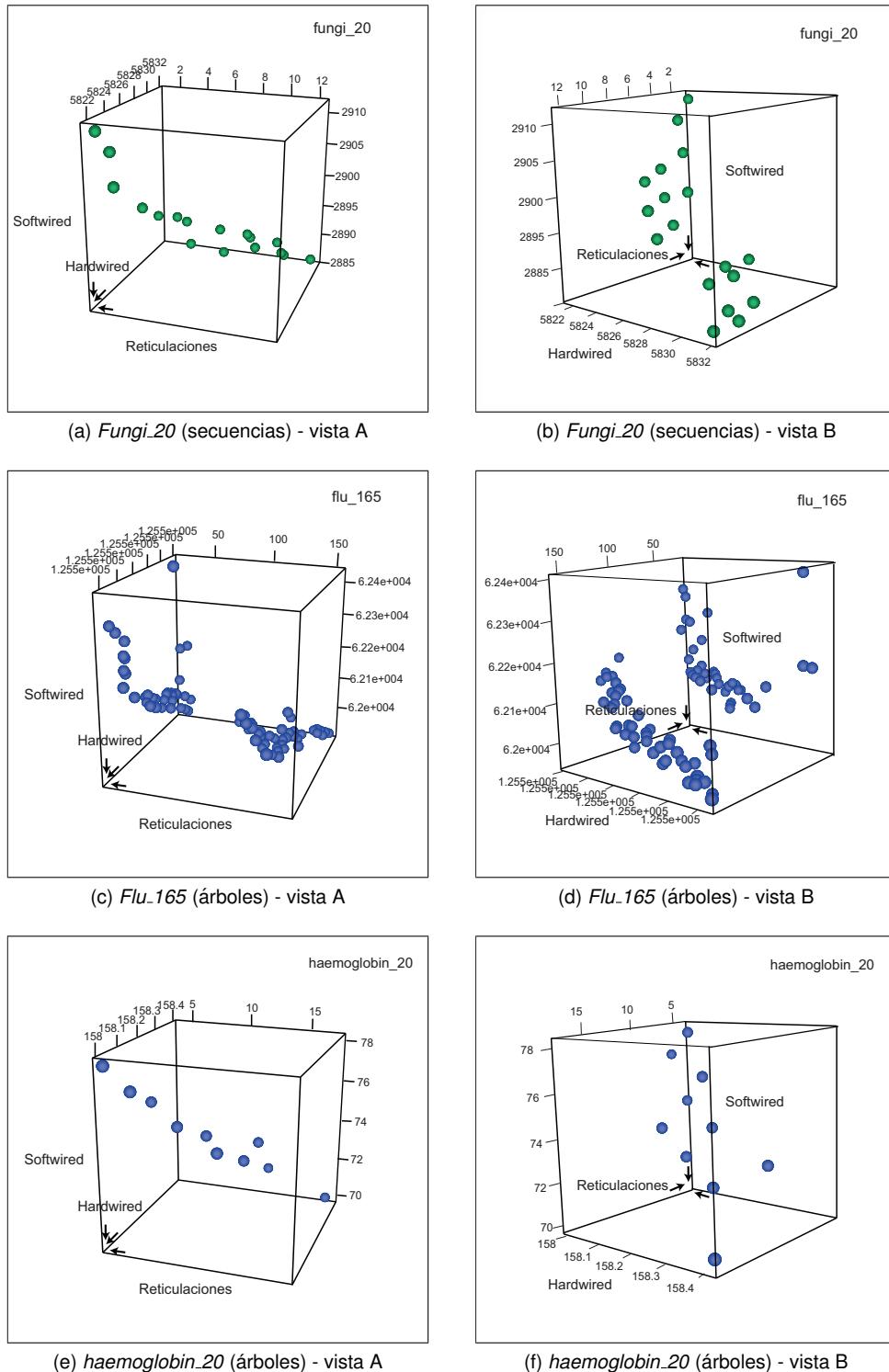


Figura 5.4: Fronteras de pareto obtenidas por MO-PhyNet para diferentes conjuntos de datos.
Fuente: Elaboración propia, (2017).

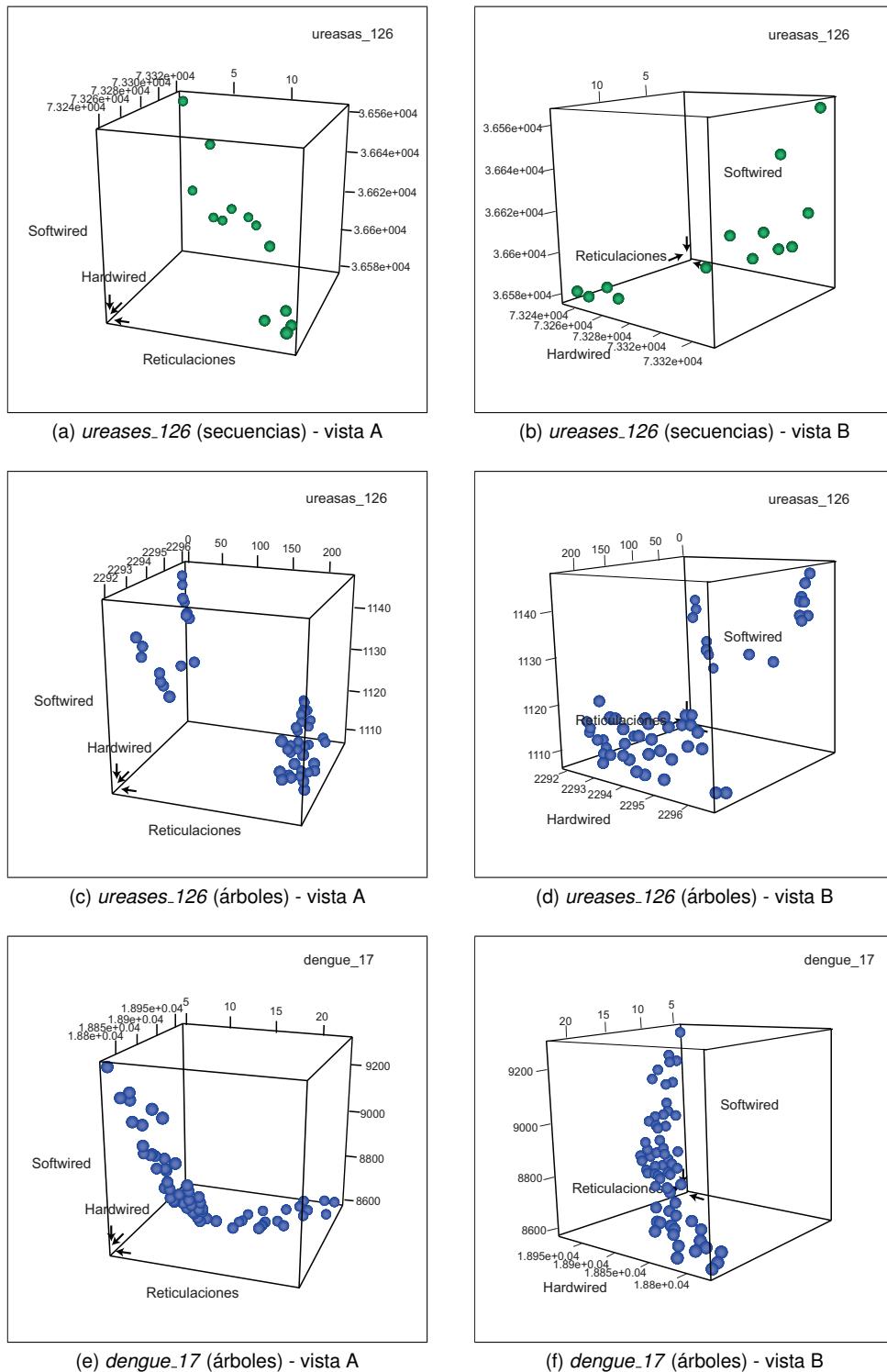


Figura 5.5: Fronteras de pareto obtenidas por MO-PhyNet para diferentes conjuntos de datos.
Fuente: Elaboración propia, (2017).

5.3.2 Espacio de soluciones y espacio objetivo

La Figura 5.6 muestra las matrices de distancia obtenidas luego de la aplicación de la métrica de Robinson-Foulds, sobre las diferentes redes filogenéticas que componen las Fronteras de Pareto representativas pertenecientes a cada conjunto de datos. En todas ellas las soluciones extremas difieren entre sí y del resto de las topologías. Esto confirma la existencia de sesgo asociado a la elección de un criterio para inferir una red filogenética, y justifica la aplicación de un modelo basado en optimización multi-objetivo. Sin embargo, también se puede observar que existen soluciones no extremas en la población final con topologías repetidas o similares.

La comparación entre espacio de soluciones y espacio objetivo empleando el IJ (Tabla 5.6) dan cuenta de una relación entre la minimización de reticulaciones y la métrica de Robinson-Foulds, alcanzando un valor de 1 para dos conjuntos de datos. Sin embargo, al integrar el resto de los objetivos, la relación entre espacio de soluciones (topologías) y espacio objetivo (criterios) disminuye hasta un promedio de 0.5, con máximo de 0.9 para el conjunto de datos *flu_165* y un mínimo de 0.3 para *dengue_17*.

Tabla 5.6: Comparación entre espacio de búsqueda de soluciones y espacio objetivo

Nombre	Detalle	ret		hard		soft		todos	
		N	JC	N	JC	N	JC	N	JC
fungi_20	DNA-DNA	5.0	0.7	7.0	0.7	6.0	0.7	7.0	0.7
flu_165	Árbol	2.0	1.0	2.0	0.5	2.0	0.5	2.0	0.9
haemo_20	Árbol	12.0	0.4	9.0	0.5	12.0	0.5	16.0	0.4
ureasas_126	AA-DNA	2.0	0.5	2.0	0.5	2.0	0.5	2.0	0.5
ureasas_126	Árbol	2.0	1.0	14.0	0.3	14.0	0.4	14.0	0.4
dengue_17	Árbol	17.0	0.2	42.0	0.2	42.0	0.2	42.0	0.3

Fuente: Elaboración propia, (2017).

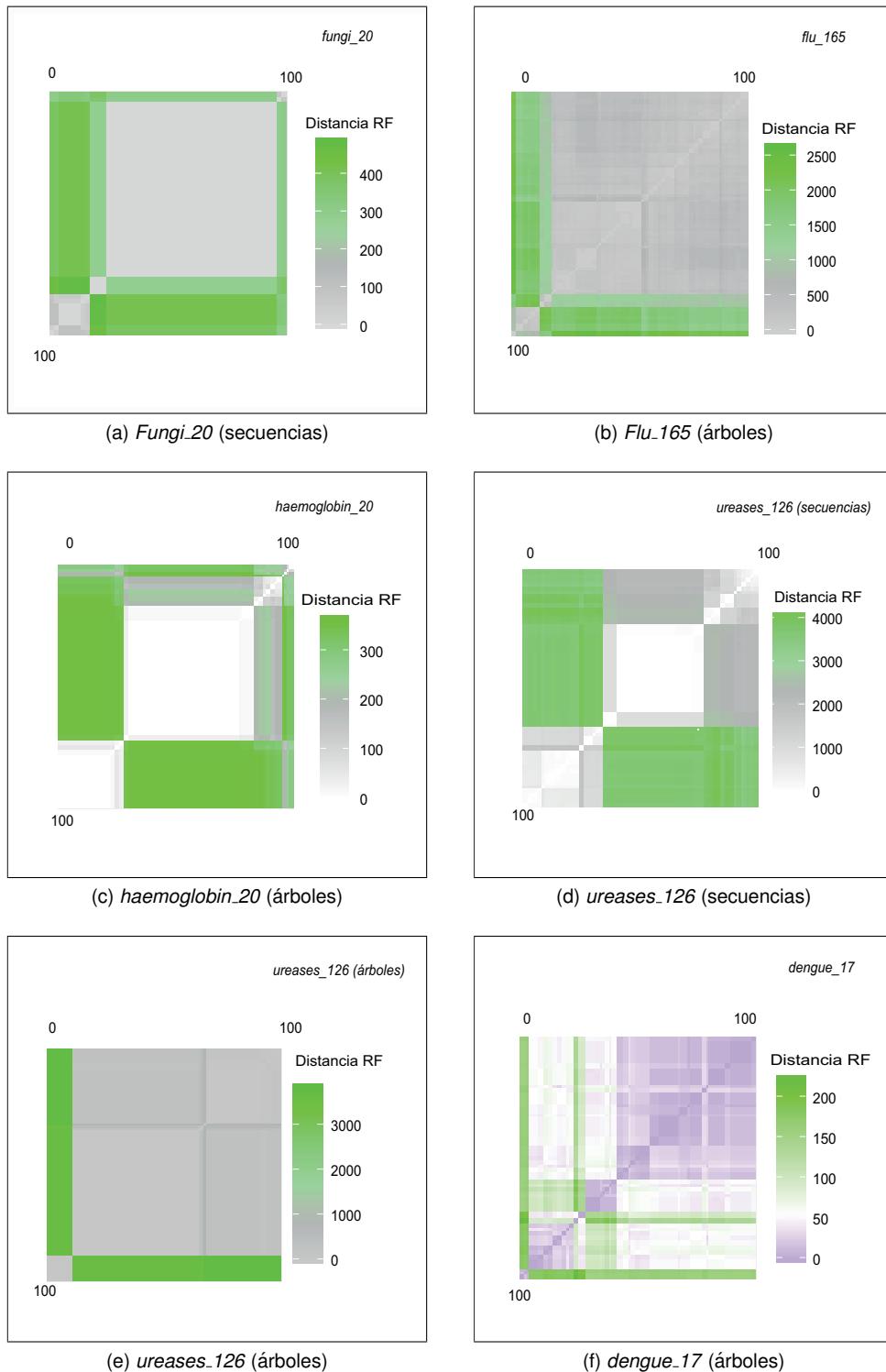


Figura 5.6: Distancia RF entre redes filogenéticas para diferentes conjuntos de datos estudiados.
Fuente: Elaboración propia, (2017).

5.3.3 Comparación con otras propuestas

Al comparar las soluciones obtenidas por MO-PhyNet (Tablas 5.3, 5.4 y 5.5) y una herramienta basada en reglas de consenso (ConsensusNET), es posible apreciar que MO-PhyNet domina en todos los conjuntos de datos las soluciones obtenidas por esta última (Tabla 5.7). A excepción de los conjuntos de datos *fungi_20* y *haemoglobin_20*, el gran número de reticulaciones de las redes obtenidas por ConsensusNet hace imposible su representación gráfica sin perder el sentido de la hipótesis evolutiva que caracteriza.

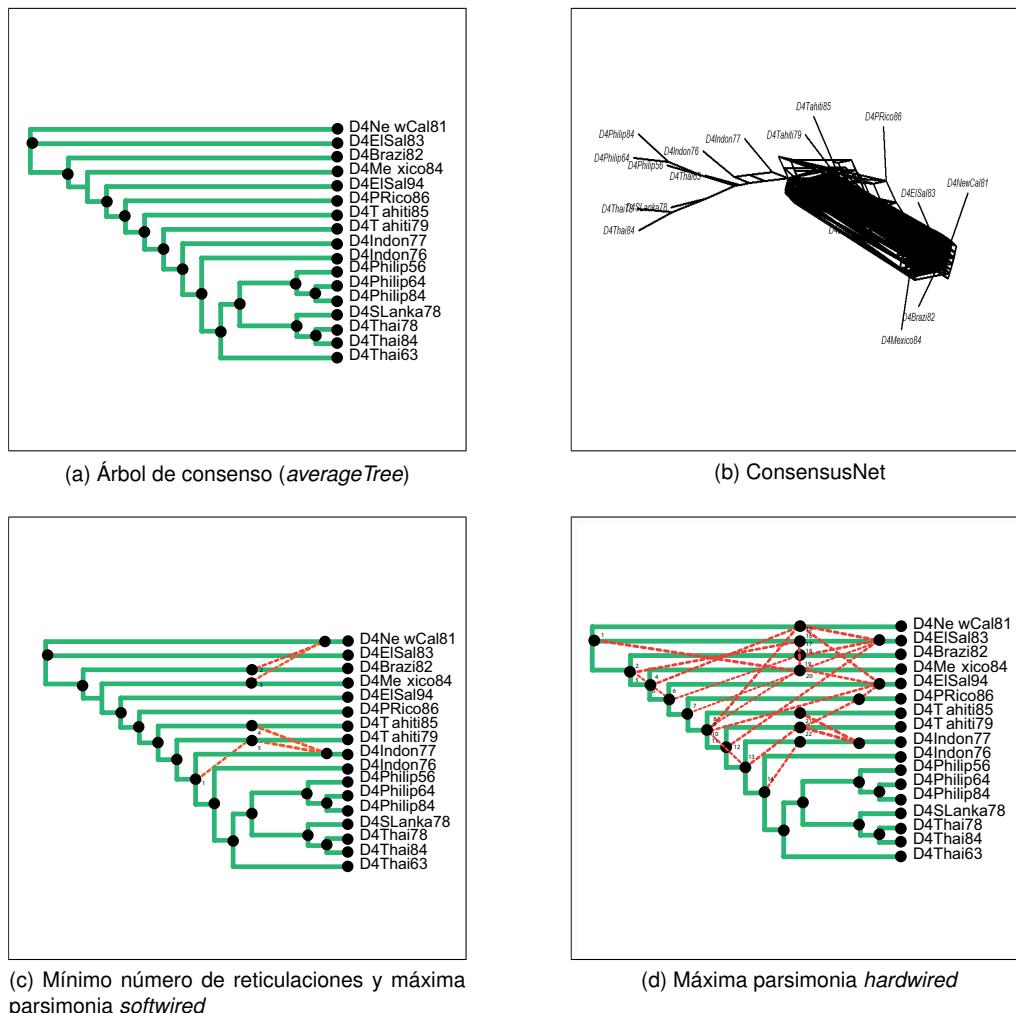


Figura 5.7: Comparación entre MO-PhyNet y ConsensusNET (*dengue_17*)
Fuente: Elaboración propia, (2017).

Tabla 5.7: Resultados ConsensusNET

Nombre	Conjunto de datos objet	ConsensusNET		
		#Ret	Hard	Soft
fungi_20	DNA-DNA	24	6007	2888
flu_165	Árbol	17468	150356	132600
haemoglobin_20	Árbol	48	174	91
ureasas_126	AA-DNA	239	77043	36772
ureasas_126	Árbol	445	2814	1369
dengue_17	Árbol	271	19998	10500

Fuente: elaboración propia, 2017.

La Figura 5.7 muestra cuatro hipótesis evolutivas alternativas para el conjunto de datos *dengue_17*. La primera de ellas (Figura 5.7a) corresponde al árbol filogenético obtenido al combinar 500 topologías mediante el método del árbol consenso medio (Sección 3.1.3). La segunda topología (Figura 5.7b) representa a la red obtenida mediante ConsensusNET, este contiene 271 eventos reticulados que complican su visualización. La tercera topología (Figura 5.7c) corresponde a una topología obtenida por MO-PhyNet, con el mínimo número de eventos reticulados y el máximo puntaje de parsimonia *hardwired*. Por último, la Figura 5.7d representa la solución de mejor puntaje de parsimonia *softwired* obtenida por MO-PhyNet. Es interesante observar que la topología del árbol de consenso coincide con la estructura principal de las redes generadas por MO-PhyNet.

5.4 CONCLUSIONES

El desarrollo de nuevas estrategias para construir hipótesis evolutivas con topologías reticulares es un actual foco de investigación en inferencia filogenética. Este se debe a su capacidad de representar mecanismos de transferencia de información más complejos en relación a los representados por árboles filogenéticos. Estos últimos asumen exclusivamente que la información hereditaria se transfiere entre un parente y dos hijos.

En esta etapa de investigación se propone la primera aproximación para enfrentar el problema multi-objetivo de inferencia filogenética basada en redes, empleando técnicas evolutivas: MO-PhyNet. Mediante esta estrategia se intenta reducir el sesgo asociado a: (1) criterio de selección para inferencia de topologías, (2) dependencia de una fuente biológica estudiada, (3) selección de un modelo evolutivo particular asociado al cálculo de verosimilitud, (4) vinculación al paradigma para combinación de datos, y (5) sentido de la topología de la red filogenética.

Las soluciones obtenidas por MO-PhyNet corresponden a una Frontera de Pareto

en tres dimensiones, mostrando la existencia de una relación entre el número de reticulaciones y la parsimonia *hirewired* en conjuntos de datos pequeños. Sin embargo, al estudiar conjuntos de datos mayores, las topologías asociadas a los tres criterios empleados resultan diferentes entre sí según la métrica Robinson-Foulds. Esto demuestra que si se selecciona inicialmente solo uno de estos criterios en el proceso de inferencia filogenética, se obtiene una hipótesis evolutiva sesgada.

A nivel algorítmico, MO-PhyNet resulta en una población final que posee un número importante de soluciones repetidas o muy similares dentro de la Frontera de Pareto. Esto se debe a las características elitistas del algoritmo NSGA-II, que prioriza el ordenamiento basado en dominancia y luego en aglomeración. Además, en los experimentos, no se consigue determinar una relación clara entre el espacio objetivo y espacio de soluciones. Esto indica que no es posible determinar *a priori* si conocimiento previo respecto a las topologías de un determinado conjunto de especies, puede aportar en el proceso multi-objetivo de inferencia filogenética a nivel de convergencia u optimización de objetivos. Esto excluye topologías previas que hayan sido obtenidas empleando alguno de los criterios considerados por MO-PhyNet.

Las topologías obtenidas por la estrategia multi-objetivo dominan a las inferidas por ConsensusNET en todos los conjuntos de datos. Este último tiene un alto número de eventos reticulados, haciendo difícil su representación visual e interpretación. Es interesante que para conjuntos de datos pequeños, la estructura principal de la red filogenética es coincidente con el árbol filogenético de consenso. Esto plantea una nueva arista de investigación para evaluar si por medio de estas estrategias de consenso se lograría acotar el espacio de búsqueda de soluciones.

A pesar de que los resultados de este trabajo son prometedores, sobretodo considerando que se usa una estrategia basada en un algoritmo clásico como NSGA-II, aún existen aspectos importantes a mejorar en la diversificación de soluciones. Así como demostró MO-MA, esta podría ser aumentada con la incorporación de una estrategia de búsqueda local. Para ello se debe evaluar diferentes estrategias: operadores, tiempo, condiciones de término, definición de vecindad, entre otros. La búsqueda de soluciones representativas empleando tomadores de decisiones no fue implementada, ya que se demostró en el capítulo anterior que todas las estrategias pueden derivar en la selección de distintas topologías.

CAPÍTULO 6. CONCLUSIONES Y TRABAJO FUTURO

6.1 CONCLUSIONES

A lo largo de esta investigación se aborda uno de los problemas más recientes dentro del área de inferencia filogenética: la reducción del sesgo asociado a la construcción de modelos que permiten representar mecanismos evolutivos complejos mediante topologías reticulares. Debido a que este problema pertenece a un área del conocimiento poco explorada, se opta por comprobar la hipótesis de esta investigación estudiando diferentes subproblemas para los cuales la literatura da cuenta de un mayor entendimiento: inferencia de árboles filogenéticos, combinación de evidencia biológica y modelamiento de fenómenos reticulares.

Con el estudio del problema de inferencia filogenética basado en árboles desarrollado en el Capítulo 2, se consigue comprobar que los recientes modelos propuestos basados en optimización multi-objetivo logran disminuir el sesgo asociado a la elección de un criterio de optimización particular. Sin embargo, no han alcanzado a resolver la dependencia de la evidencia biológica seleccionada para construir una hipótesis evolutiva, ni del modelo evolutivo seleccionado para el cálculo de verosimilitud. También se ha determinado que, debido a la gran diversidad de heurísticas propuestas en el último tiempo para el proceso de optimización multi-objetivo y estrategias de reordenamiento diseñadas para efectuar operaciones genéticas, se hace muy difícil establecer criterios para evaluar el desempeño de las diferentes propuestas. Debido a ello se propone un modelo compuesto por un algoritmo genérico como NSGA-II para caracterizar diferentes estrategias de reordenamiento como operadores genéticos y de búsqueda local. Estudiando las ventajas y desventajas de cada operador se ha propuesto MO-MA, un algoritmo memético multi-objetivo capaz de hallar hipótesis evolutivas según los criterios de parsimonia y verosimilitud, que ha demostrado tener un desempeño superior a las estrategias propuestas en la literatura. Con este desarrollo se consigue cumplir con el segundo objetivo específico planteado en esta investigación (Sección 1.3.3).

Paralelamente, con objetivo de independizar el proceso de inferencia filogenética del uso de un marcador o evidencia biológica en particular, y así cumplir con el tercer objetivo específico de esta investigación (Sección 1.3.3), se estudia el problema de combinación de evidencia filogenética (Capítulo 3). En la literatura se dispone de diferentes métodos para combinación de datos que dependen de dos paradigmas de combinación conflictivos entre sí, lo que da origen a diferentes hipótesis evolutivas para un mismo conjunto de datos. A raíz de ello se propone MO-CS, un algoritmo genético multi-objetivo capaz de obtener hipótesis evolutivas integrales según los diferentes paradigmas de combinación, cuyos resultados poseen una mayor exactitud que las actuales propuestas de la literatura. Este es capaz de combinar datos

provenientes desde cualquier evidencia biológica sin depender de la definición de un modelo evolutivo en particular, estableciendo la posibilidad de aplicar métricas de soporte directamente sobre la elección de una topología.

Las estrategias basadas en optimización multi-objetivo minimizan la dependencia de varias fuentes de sesgo asociada al propio modelamiento de inferencia filogenética basado en árboles. Sin embargo, debido a que la solución presentada por este tipo de estrategias consiste en múltiples topologías que conforman una Frontera de Pareto, implícitamente se genera una nueva fuente de sesgo asociada a la elección de una hipótesis evolutiva representativa. Para resolver esto, en el Capítulo 4 se estudia relación entre espacio objetivo y espacio de soluciones, así como la aplicación de estrategias para toma de decisión, cumpliendo con el cuarto objetivo específico planteado (Sección 1.3.3). Los resultados no muestran una relación clara entre ambos espacios, estableciendo que dos topologías diferentes pueden ser representadas por el mismo punto en el espacio objetivo, o que, por otro lado, dos soluciones muy lejanas en este último pueden corresponder a topología similares. Esto quiere decir que estrategias elitistas como NSGA-II, que dependen de un ordenamiento no dominado y la medición de distancia de aglomeración acorde al espacio objetivo, pueden descartar soluciones con potencial significancia biológica. También se demuestra que el uso de topologías previamente definidas no aportan en la convergencia de los algoritmos evolutivos, a menos que su inferencia se haya efectuado considerando los mismos criterios aplicados por las estrategias multi-objetivo. Esto cobra relevancia no tan sólo a nivel algorítmico, sino que también plantea un parámetro a ser considerado en la generación de bases de datos para almacenar hipótesis evolutivas y como estos datos pueden ser utilizados para el proceso de inferencia. En cuanto a los tomadores de decisiones estudiados, estos seleccionan diferentes soluciones representativas sobre el espacio objetivo, por lo que no se consigue establecer un consenso entre ellos, lo que conlleva a que todos pueden ser empleados indistintamente para abordar el problema de inferencia filogenética multi-objetivo. Además, esta investigación propone un nuevo tomador de decisiones cuyo proceso de selección es efectuado sobre el espacio de soluciones.

Finalmente, con el conocimiento adquirido en las etapas previas, se aborda el problema de inferencia de redes filogenéticas desde un punto de vista multi-objetivo. Dado que redes filogenéticas pueden ser estudiadas considerando diferentes sentidos para sus topologías (*hardwired* y *softwired*), se adiciona una nueva condicionante para la obtención de una hipótesis evolutiva. Con objetivo de reducir el sesgo asociado a ella, se propone MO-PhyNet, una estrategia basada en el NSGA-II que permite inferir redes filogenéticas considerando tres criterios de optimización bajo diferentes sentidos topológicos, integrando los diferentes paradigmas de combinación en la integración de evidencia biológica, e independencia de un modelo evolutivo. La evaluación de las topologías generadas demuestra la existencia de fuentes de sesgo, y

la inexistencia de una relación entre espacio objetivo y espacio de soluciones, dejando en evidencia el problema planteado en esta investigación. Además, se valida el uso de estrategia basada en optimización multi-objetivo que contempla todas estas relaciones de conflicto en sus soluciones. Por otro lado, MO-PhyNet muestra un mejor desempeño en relación a otra estrategia diseñada para representar fenómenos reticulares. Con esto se logra cumplir con el primer objetivo específico planteado en esta investigación (Sección 1.3.3).

Finalmente, se puede dar respuesta a la pregunta de investigación propuesta al inicio de este trabajo: *¿Es posible el modelamiento de inferencia filogenética considerando fenómenos reticulares, reduciendo las fuentes de sesgo descritas en la Sección 1.2, e integrando diferentes criterios de optimalidad?* La respuesta a esta pregunta es: Sí, si se puede. Mediante la experimentación desarrollada a lo largo de este trabajo se confirma cuantitativamente la existencia de sesgo en el proceso de inferencia filogenética, proponiendo diferentes estrategias basadas en optimización multi-objetivo para su reducción, tanto a nivel de inferencia basada en árboles, combinación de evidencia biológica y generación de topologías reticulares. La reducción del sesgo se basa en la generación de soluciones que combinan en su construcción diferentes criterios y paradigmas que, al ser usados en forma individual, presentan topologías conflictivas. Dado que se ha conseguido cumplir con los objetivos específicos propuestos al inicio de esta investigación, se logra cumplir con el objetivo general, validando la hipótesis presentada en la Sección 1.3.

6.2 TRABAJO FUTURO

La exploración de esta área de conocimiento ha dejado en evidencia varios puntos por desarrollar con posterioridad a esta investigación, como, por ejemplo:

- **Mejoramiento de modelos.** Los actuales modelos aplicados a inferencia basados en árboles, redes y combinación de evidencia biológica pueden ser mejorados considerando diversos aspectos, entre ellos:
 - **Algoritmos multi-objetivo y estrategias de reordenamiento.** A nivel algorítmico se debe explorar nuevas estrategias que consigan mejorar el desempeño de las actuales propuestas, o sean capaces de hallar nuevas soluciones no contempladas por estas últimas. Para ello se puede explorar otros algoritmos multi-objetivo bioinspirados que sean capaces de integrar más de tres criterios en sus funciones objetivo, o que consideren nuevas estrategias de reordenamiento en sus operaciones genéticas y de búsqueda local. En este último punto aún quedan bastantes aspectos por abordar: modificación de operadores, tipo de búsqueda local, tiempo, condiciones de término,

definición de vecindad, entre otros. Debido a la naturaleza del problema de inferencia filogenética, se requiere de operadores que consideren información del espacio de soluciones y objetivo.

- **Relación entre espacio objetivo y espacio de soluciones.** En este trabajo se ha explorado la relación entre espacio de soluciones y espacio objetivo, considerando los criterios de verosimilitud y parsimonia. Sin embargo, existen otros criterios empleados para inferir filogenia cuya relación entre estos espacios es desconocida. Un trabajo interesante a desarrollar corresponde al estudio de las relaciones entre los diferentes criterios y como estos establecen una correspondencia entre los diferentes espacios. Conocer esta relación es clave en la creación de estrategias para reducir el tiempo de convergencia de los algoritmos basados en uno o múltiples objetivos, aumentar la diversidad de las soluciones, y para la interpretación de las topologías resultantes.
- **Combinación de evidencia biológica y robustez.** En este trabajo de investigación se ha evaluado la capacidad de diferentes estrategias en integrar evidencia biológica, y así generar una hipótesis evolutiva unificada. Sin embargo, aún es un desafío en el área determinar si combinar diferente evidencia biológica favorece la obtención de hipótesis evolutivas más robustas, en comparación al uso exclusivo de evidencia. Por ahora es una tarea difícil de realizar, ya que las métricas actualmente disponibles para evaluar soporte también son dependientes del tipo de dato empleado para el proceso de inferencia. Probablemente se requiera de diferentes experimentos no in silico para complementar este análisis.
- **Estudio de criterios para modelamiento de fenómenos reticulares.** Si bien este corresponde a un área actual de investigación a nivel de modelamiento matemático, ningún trabajo ha evaluado si las topologías obtenidas con los criterios de parsimonia *softwired* o *hardwired* cumplen con el sentido de las mismas. Es decir que las topologías resultantes efectivamente contengan total o parcialmente las características de las hipótesis evolutivas empleadas como entrada. También corresponde a un nuevo desafío el diseño de nuevos modelos evolutivos o estrategias para su combinación, ya que los actuales algoritmos basados en verosimilitud combinan las diferentes hipótesis de entrada asumiendo un modelo uniforme. Esto último es una aproximación grosera, ya que marcadores genéticos pueden perfectamente tener tasas evolutivas diferentes.
- **Diseño de herramientas para visualización de redes filogenéticas.** En la actualidad se dispone de pocas herramientas funcionales específicas para inferencia filogenética que permitan representar fenómenos reticulares, convirtiéndola en un área de desarrollo.
- **Creación de herramientas funcionales para inferencia filogenética basados en optimización multi-objetivo.** Hasta el momento la totalidad de las herramientas destinadas

a inferir filogenia se basan en la optimización de un objetivo único. Esto se debe a que la mayoría de las propuestas basadas en optimización multi-objetivo se han limitado al desarrollo de bibliotecas para *python*, *java* y *R*, no existiendo herramientas funcionales que permitan a un usuario no familiarizado con el área computacional, construir fácilmente hipótesis evolutivas.

LISTADO DE ACRÓNIMOS

- AA: aminoácidos.
- ACCTRAN: del inglés, Accelerates the evolutionary transformation of a character.
- BE: del inglés, Branch Exchange.
- BioNJ: del inglés, Bio-neighbor joining.
- CS: operador de consenso.
- CSave: árbol promedio.
- CSmaj: método de consenso basado en la regla de consenso mayoría.
- CSstr: método de consenso basado en la regla de consenso estricta.
- DNA: del inglés, Deoxyribonucleic acid.
- DNAPARS: del inglés, DNA parsimony program.
- IJ: índice de Jaccard.
- KC: métrica de Kendall-Colijn.
- L2: métrica L2.
- MEGA: del inglés, Molecular evolutionary genetics analysis.
- MMAX: Método multi-modal basado en el máximo.
- MMEAN: Método multi-modal basado en el promedio.
- MMIN: Método multi-modal basado en el mínimo.
- MMS: del inglés, Metric multidimensional scaling method.
- MO: del inglés, multi-objective.
- MOO: del inglés, multi-objective optimisation.
- MO-ABC: del inglés, Multi-objective ant bee colony.
- MO-CS: algoritmo multi-objetivo para combinación de evidencia biológica en inferencia filogenética.
- MO-MA: del inglés, Multi-objective memetic algorithm.
- MO-FA: del inglés, Multi-objective Firefly Algorithm.

- MO-PhyNe: Modelo multi-objetivo para inferencia de redes filogenéticas.
- MPROD: Método multi-modal basado en producto.
- MRP: del inglés, Matrix representation with parsimony method.
- MU: Método de utilidad marginal.
- NEWICK: del inglés, New Hampshire tree format.
- NJ: del inglés, Neighbor joining.
- NNI: del inglés, Nearest-neighbor interchange.
- NSGA-II: del inglés, Non-dominated sorting genetic algorithm II.
- NSGA-II EM: Non-dominated sorting genetic algorithm II que incluye operador de cruzamiento para modelos evolutivos.
- PhyloMOEA: algoritmo evolutivo multi-objetivo para inferencia filogenética.
- PHYLIP: del inglés, Phylogeny Inference Package computer programs for inferring phylogenies.
- PHYML: del inglés, Phylogenetic estimation using (Maximum) Likelihood.
- PDG: del inglés, Prune-Delete-Graft algorithm.
- PDGm: Prune-Delete-Graft modificado.
- PLS: del inglés, Pareto local search.
- RAxML: del inglés, Randomized Axelerated Maximum Likelihood.
- RF: del inglés, Robinson-Foulds distance.
- RPlik: Método del punto de referencia basado en verosimilitud.
- RPmid: Método del punto de referencia basado en el promedio.
- RPpar: Método del punto de referencia basado en parsimonia.
- SA: del inglés, Simulated Annealing.
- SPR: del inglés, Sub-tree Pruning and Re-grafting.
- TBR: del inglés, Tree Bisection and Reconnection.
- TC: del inglés, Taxonomic congruence.

- TCG: del inglés, Triangulated Colored Graphs.
- TE: del inglés, Total evidence.
- UPGMA: del inglés, Unweighted Pair Group Method with Arithmetic Mean.
- WPGMA: del inglés, Weighted Pair Group Method with Arithmetic Mean.

REFERENCIAS BIBLIOGRÁFICAS

- Abascal, F., Aguirre, E., & León, A. (2014). *Bioinformática con N*. Publiciones CreateSpace.
- Albrecht, B. (2015). Computing all hybridization networks for multiple binary phylogenetic input trees. *BMC Bioinformatics*, 16(1), 439–441.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410.
- Bandelt, H.-J., & Dress, A. W. (1992). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92(1), 47–105.
- Bandelt, H. J., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1), 37–48.
- Bandelt, H. J., Macaulay, V., & Richards, M. (2000). Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Molecular Phylogenetics and Evolution*, 16(1), 8–28.
- Bandyopadhyay, S., Saha, S., Maulik, U., & Deb, K. (2008). A simulated annealing-based multiobjective optimization algorithm: Amosa. *IEEE Transactions on Evolutionary Computation*, 12(3), 269–283.
- Barry, D., & Hartigan, J. (1987). Asynchronous distance between homologous DNA sequences. *Biometrics*, 43(1), 261–276.
- Baser, P., & Saini, J. R. (2015). Agent based stock clustering for efficient portfolio management. *International Journal of Computer Applications*, 116(3), 1.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28(1), 235–242.
- Beyer, A., Stein, L., Smith, F., & Ulam, M. (1974). A molecular sequence metric and evolutionary trees. *Mathematical Biosciences*, 19(1), 9–25.
- Bicego, M., Dellaglio, F., & Felis, G. E. (2007). Multimodal phylogeny for taxonomy: Integrating information from nucleotide and amino acid sequences. *Journal Bioinformatics and Computational Biology*, 5(5), 1069–1085.
- Billera, L. J., Holmes, S. P., & Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4), 733–767.
- Borges, L. M. S., Hollatz, C., Lobo, J., Cunha, A. M., Vilela, A. P., Calado, G., Coelho, R., Costa, A. C., Ferreira, M. S. G., Costa, M. H., & Costa, F. O. (2016). With a little help from DNA barcoding: investigating the diversity of gastropoda from the portuguese coast. *Scientific Reports*, 6(1).
- Borguesan, B., Barbachan, M., Grisci, B., Inostroza-Ponta, M., & Dorn, M. (2015). APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. *Computational Biology and Chemistry*, 59, Part A(1), 142–157.
- Bryant, D. (2003). A classification of consensus methods for phylogenetics. In *Bioconsensus (Piscataway, NJ, 2000/2001)*, vol. 61 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, (pp. 163–183). American Mathematical Society.
- Bryant, D., & Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2), 255–265.

- Cancino, W., & Delbem, A. C. B. (2007). A multi-objective evolutionary approach for phylogenetic inference. In S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, & T. Murata (Eds.) *Evolutionary Multi-Criterion Optimization: 4th International Conference, EMO, Matsushima, Japan. Proceedings*, 1, (pp. 428–442). Springer Berlin Heidelberg.
- Cantú-Paz, E., & Goldberg, D. E. (2003). *Are Multiple Runs of Genetic Algorithms Better than One?*, (pp. 801–812). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Carlini, C. R., & Ligabue-Braun, R. (2016). Ureases as multifunctional toxic proteins: A review. *Journal of The International Society on Toxicology*, 110(1), 90–109.
- Cavalli-Sforza, L., & Edwards, A. (1967). Phylogenetic analysis. Models and estimation procedures. *The American Journal of Human Genetics*, 19(1), 233–257.
- Chaudhary, R., Burleigh, J. G., & Fernandez-Baca, D. (2012). Fast local search for unrooted robinson-foulds supertrees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4), 1004–1013.
- Chor, B., & Tuller, T. (2005). Maximum likelihood of evolutionary trees is hard. In *Recomb*, vol. 3500 of *Lecture Notes in Computer Science*, (pp. 296–310). Springer.
- Clark, C., & Kalita, J. (2015). A multiobjective memetic algorithm for PPI network alignment. *Bioinformatics*, 31(12), 1988–1998.
- Coelho, G. P., da Silva, A., & Von Zuben, F. (2010). An immune-inspired multi-objective approach to the reconstruction of phylogenetic trees. *Neural Computing and Applications*, 19(8), 1103–1132.
- Coelho, G. P., & Zuben, F. J. V. (2007). A Multiobjective Approach to Phylogenetic Trees: Selecting the Most Promising Solutions from the Pareto Front. In *Seventh International Conference on Intelligent Systems Design and Applications*, 1, (pp. 837–842). Institute of Electrical and Electronics Engineers (IEEE).
- Congdon, C. B. (2002). Gaphyl: An evolutionary algorithms approach for the study of natural evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 1, (pp. 1057–1064). Morgan Kaufmann Publishers Inc.
- Cotta, C., & Moscato, P. (2002). Inferring phylogenetic trees using evolutionary algorithms. In *Proceedings of the 7th International Conference on Parallel Problem Solving from Nature, PPSN VII*, (pp. 720–729). London, UK: Springer-Verlag.
- Cotta, C., & Moscato, P. (2003). A memetic-aided approach to hierarchical clustering from distance matrices: application to gene expression clustering and phylogeny. *BioSystems*, 2439(1), 75–97.
- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*(1), 1695.
- Day, W. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(1), 461–467.
- de Bruyn, A., Martin, D. P., & Lefevre, P. (2014). Phylogenetic reconstruction methods: An overview. In P. Besse (Ed.) *Molecular Plant Taxonomy: Methods and Protocols*, 1, (pp. 257–277). Humana Press.
- Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197.
- Debevec, A. H., & Whitfield, J. B. (2013). Introduction to phylogenetic networks. *Systematic Biology*, 62(1), 177.

- Desper, R., & Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9(1), 687–705.
- Dikow, T. (2009). A phylogenetic hypothesis for asilidae based on a total evidence analysis of morphological and DNA sequence data (insecta: Diptera: Brachycera: Asiloidea). *Organisms Diversity & Evolution*, 9(3), 165–188.
- Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Inter-coder reliability and validity of webplotdigitizer in extracting graphed data. *Behaviour Modification*, 41(2), 323–339.
- Drugan, M. M., & Thierens, D. (2012). Stochastic Pareto local search: Pareto neighbourhood exploration and perturbation strategies. *Journal of Heuristics*, 18(5), 727–766.
- Dubois-Lacoste, J., López-Ibáñez, M., & Stützle, T. (2012). Pareto local search algorithms for anytime bi-objective optimization. In J. K. Hao, & M. Middendorf (Eds.) *Evolutionary Computation in Combinatorial Optimization: 12th European Conference, EvoCOP, Málaga, Spain. Proceedings*, (pp. 206–217). Springer Berlin Heidelberg.
- Eernisse, D. J., & Kluge, A. G. (1993). Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Molecular Biology and Evolution*, 10(6), 1170–1195.
- Eguiarte, L. (2007). *Ecología molecular*. Secretaría de Medio Ambiente y Recursos Naturales, Instituto Nacional de Ecología.
- Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *The American Journal of Human Genetics*, 25(1), 471–492.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(1), 368–376.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Macmillan Education.
- Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6.
- Fischer, M., van Iersel, L., Kelk, S., & Scornavacca, C. (2015). On computing the maximum parsimony score of a phylogenetic network. *SIAM Journal on Discrete Mathematics*, 29(1), 559–585.
- Fitch, W. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(1), 406–416.
- Fitch, W., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(1), 279–284.
- Fleischauer, M., & Böcker, S. (2016). Collecting reliable clades using the greedy strict consensus merger. *PeerJ*, 4(1), e2172.
- Gallardo, J. E., Cotta, C., & Fernández, A. J. (2007). Reconstructing Phylogenies with Memetic Algorithms and Branch-and-bound. In *Analysis of Biological Data: A Soft Computing Approach*, (pp. 59–84). World Scientific.
- Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(1), 685–695.
- Gascuel, O. (2000). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Molecular Biology and Evolution*, 17(3), 401.
- Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., & Drummond, A. J. (2017). Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, 66(1), 57.

- Giribet, G. (2007). Efficient tree searches with available algorithms. *Evolutionary Bioinformatics*, 3(3), 341–356.
- Goloboff, P., Farris, J., & Nixon, K. (2008). Tnt, a free program for phylogenetic analysis. *Cladistics*, 24, 774–786.
- Grechko, V. V. (2002). Molecular dna markers in phylogeny and systematics. *Russian Journal of Genetics*, 38(8), 851–868.
- Guermeur, Y., Geourjon, C., Gallinari, P., & Deleage, G. (1999). Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15(5), 413–421.
- Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *System Biology*, 52(5), 696–704.
- Gusfield, D., Eddhu, S., & Langley, C. (2003). Efficient reconstruction of phylogenetic networks with constrained recombination. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, 1, (pp. 363–374).
- Handl, J., Kell, D., & Knowles, J. (2007). Multiobjective optimization in bioinformatics and computational biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(1), 279–292.
- Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the Human-Ape splitting by a molecular clock of mitochondrial-DNA. *Journal of Molecular Evolution*, 22(1), 160–174.
- Hasnat, A., & Molla, A. U. (2016). Feature selection in cancer microarray data using multi-objective genetic algorithm combined with correlation coefficient. In *2016 International Conference on Emerging Technological Trends (ICETT)*. IEEE.
- Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 1(1), 396–405.
- Hinchliff, C., Smith, S., Allman, J., Burleigh, J., Chaudhary, R., Coghill, L., Crandall, K., Deng, J., Drew, B., Gazis, R., Gude, K., Hibbett, D., Katz, L., Laughinghouse, H., McTavish, E., Midford, P., Owen, C., Ree, R., Rees, J., Soltis, D., Williams, T., & Cranston, K. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112, 12764–12769.
- Holland, B. R., Huber, K. T., Moulton, V., & Lockhart, P. J. (2004). Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution*, 21(7), 1459–1461.
- Horiike, T., Miyata, D., Tateno, Y., & Minai, R. (2011). HGT-Gen: a tool for generating a phylogenetic tree with horizontal gene transfer. *Bioinformation*, 7(1), 211–213.
- Huelsenbeck, J. P., Bull, J., & Cunningham, C. W. (1996). Combining data in phylogenetic analysis. *Trends in Ecology & Evolution*, 11(4), 152–158.
- Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(1), 754–755.
- Huson, D. H., Rupp, R., & Scornavacca, C. (2011). *Phylogenetic Networks: Concepts, Algorithms and Applications*. New York, NY, USA: Cambridge University Press.
- Ishibuchi, H., Yoshida, T., & Murata, T. (2003). Balance between genetic search and local search in memetic algorithms for multiobjective permutation flow-shop scheduling. *IEEE Transactions on Evolutionary Computation*, 7(2), 204–223.

- Jayaswal, V., Poladian, L., & Jermiin, L. (2007). Single- and multi-objective phylogenetic analysis of primate evolution using a genetic algorithm. In *Proceedings of the IEEE Congress on Evolutionary Computation*, (pp. 4146–4153).
- Jiang, S., Ong, Y., Zhang, J., & Feng, L. (2014). Consistencies and contradictions of performance metrics in multiobjective optimization. *IEEE Transactions on Cybernetics*, 44(12), 2391–2404.
- Jin, G., Nakhleh, L., Snir, S., & Tuller, T. (2006). Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22(21), 2604.
- Jombart, T., Kendall, M., Almagro-Garcia, J., & Colijn, C. (2017). treespace: statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology Resources*, -(), n/a–n/a.
- Jukes, T., & Cantor, C. (1969). Evolution of protein molecules: Mammalian Protein Metabolism.
- Katoh, K., Kuma, K., & Miyata, T. (2001). Genetic algorithm-based maximum-likelihood analysis for molecular phylogeny. *Journal of Molecular Evolution*, 53(1), 477–484.
- Kendall, M., & Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular Biology and Evolution*, 33(10), 2735–2743.
- Kidd, K., & Sgaramella-Zonta, L. (1971). Phylogenetic analysis: concepts and methods. *The American Journal of Human Genetics*, 23, 235–252.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(1), 111–120.
- King, R. D., & Sternberg, M. J. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*, 5(11), 2298–2310.
- Kluge, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among epicrates (boidae, serpentes). *Systematic Biology*, 38(1), 7.
- Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3), 459–468.
- Kumar, S., & Gadagkar, S. R. (2001). Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics*, 158(3), 1321–1327.
- Leigh, J. W., Susko, E., Baumgartner, M., & Roger, A. J. (2008). Testing congruence in phylogenomic analysis. *Systematic Biology*, 57(1), 104.
- Lemmon, A. R., & Milinkovitch, M. C. (2002). The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proceedings of the National Academy of Sciences*, 99(16), 10516–10521.
- Levasseur, C., & Lapointe, F.-J. (2006). Total evidence, average consensus and matrix representation with parsimony: What a difference distances make. *Evolutionary Bioinformatics*, 2(1).
- Lewis, P. (1998). A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution*, 15(3), 277–283.
- Li, B., & Lecointre, G. (2009). Formalizing reliability in the taxonomic congruence approach. *Zoologica Scripta*, 38(1), 101–112.
- Li, Y. F., Pedroni, N., & Zio, E. (2013). A memetic evolutionary multi-objective optimization method for environmental power unit commitment. *IEEE Transactions on Power Systems*, 28(3), 2660–2669.

- Ligabue-Braun, R., Andreis, F., Verli, H., & Carlini, C. (2013). 3-to-1: unraveling structural transitions in ureases. *Naturwissenschaften*, 100(5), 459–467.
- Lin, Y., Fang, S., & Thorne, J. (2007). A tabu search algorithm for maximum parsimony phylogeny inference. *European Journal of Operational Research*, 176(1), 1908–1917.
- Lobo, J., Teixeira, M. A. L., Borges, L. M. S., Ferreira, M. S. G., Hollatz, C., Gomes, P. T., Sousa, R., Ravara, A., Costa, M. H., & Costa, F. O. (2016). Starting a dna barcode reference library for shallow water polychaetes from the southern european atlantic coast. *Molecular Ecology Resources*, 16(1), 298–313.
- Mahe, S., Duhamel, M., Le Calvez, T., Guillot, L., Sarbu, L., Bretaudeau, A., Collin, O., Dufresne, A., Kiers, E. T., & Vandenkoornhuyse, P. (2012). PHMYCO-DB: a curated database for analyses of fungal diversity and evolution. *PLoS ONE*, 7(9), e43117.
- Makarenkov, V., & Legendre, P. (2004). From a phylogenetic tree to a reticulated network. *Journal of Computational Biology*, 11(1), 195–212.
- Mariano, F., Olivera, A., & Tohmé, F. (2010). A memetic algorithm based on a NSGA-II scheme for the flexible job-shop scheduling problem. *Annals of Operations Research*, 181(1), 745–765.
- Matsuda, H. (1995). Construction of phylogenetic trees from amino acid sequences using a genetic algorithm. *Proceedings of Genome Informatics Workshop*, 6(6), 19–28.
- Matsuda, H. (1996). Protein phylogenetic inference using maximum likelihood with a genetic algorithm. *Pacific Symposium on Biocomputing*, 1(1), 512–523.
- Mersmann, O. (2011). Emoa: Evolutionary multiobjective optimization algorithms. <http://CRAN.R-project.org/package=emoa>. (R package version 0.4-8).
- Moilanen, A. (1999). Searching for most parsimonious trees with simulated evolutionary optimization. *Cladistics*, 15(1), 39–50.
- Mora, C., Tittensor, D., Adl, S., Simpson, A., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biol*, 9, 100–127.
- Ochoa, G., Verel, S., & Tomassini, M. (2010). First-improvement vs. best-improvement local optima networks of nk landscapes. In R. Schaefer, C. Cotta, J. Kołodziej, & G. Rudolph (Eds.) *Parallel Problem Solving from Nature, PPSN XI: 11th International Conference, Kraków, Poland, Proceedings, Part I*, (pp. 104–113). Springer Berlin Heidelberg.
- Ortuno, F., Florido, J. P., Urquiza, J. M., Pomares, H., Prieto, A., & Rojas, I. (2012). Optimization of multiple sequence alignment methodologies using a multiobjective evolutionary algorithm based on NSGA-II. In *2012 IEEE Congress on Evolutionary Computation*, (pp. 1–8). IEEE.
- Padhye, N., & Deb, K. (2011). Multi-objective optimisation and multi-criteria decision making for fdm using evolutionary approaches. In L. Wang, A. H. C. Ng, & K. Deb (Eds.) *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing*, (pp. 219–247). Springer.
- Parraga-Alava, J., & Inostroza-Ponta, M. (2016). A bi-objective model for gene clustering combining expression data and external biological knowledge. In *2016 XLII Latin American Computing Conference (CLEI)*, (pp. 1–12). IEEE.
- Phillips, E. (1992). The PhD: assessing quality at different stages of its development. *Starting Research: Supervision and Training. Brisbane, Queensland: Tertiary Education Institute, University of Queensland*, 1(6), 1.
- Pirkwieser, S., & Raidl, G. (2008). Finding consensus trees by evolutionary, variable neighbourhood search, and hybrid algorithms. In *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, (pp. 323–330). New York, NY, USA: ACM.

- Plonski, P., & Radomski, J. (2013). Neighbor Joining Plus-algorithm for phylogenetic tree reconstruction with proper nodes assignment. *Populations and Evolution.*, 1(1), 1–18.
- Poladian, L., & Jermiin, L. (2006). Multi-objective evolutionary algorithms and phylogenetic inference with multiple data sets. *Soft Computing*, 10(4), 359–368.
- Pólya, G. (2004). *How to Solve It: a New Aspect of Mathematical Method*. Princeton Science Library. Princeton University Press.
- Pyron, R. A. (2017). Novel approaches for phylogenetic inference from morphological data and total-evidence dating in squamate reptiles (lizards, snakes, and amphisbaenians). *Systematic Biology*, 66(1), 38.
- Queiroz, A. D., & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in Ecology & Evolution*, 22(1), 34–41.
- Real, R., & Vargas, J. M. (1996). The probabilistic basis of jaccard's index of similarity. *Systematic Biology*, 45(3), 380.
- Reijmers, T., Wehrens, R., Daeyaert, F., Lewi, P., & Buydens, L. (1999). Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences. *Biosystems*, 49(1), 31–43.
- Revell, L. J. (2012). phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2), 217–223.
- Riquelme, N., Lucken, C. V., & Baran, B. (2015). Performance metrics in multi-objective optimization. In *2015 Latin American Computing Conference (CLEI)*. IEEE.
- Robinson, D., & Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1), 131–147.
- Rodríguez, M., Vargas, M., Antúnez, K., Gerding, M., Ovidio-Castro, F., & Zapata, N. (2014). Prevalence and phylogenetic analysis of honey bee viruses in the Biobío Region of Chile and their association with other honey bee pathogens. *Chilean journal of agricultural research*, 74(1), 170–177.
- Rohatgi, A., & ZlatanStanojevic (2017). Webplotdigitizer: Version 3.11 of webplotdigitizer. URL <https://doi.org/10.5281/zenodo.32375>
- Rokas, A., Williams, B., King, N., & Carroll, S. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(1), 798–804.
- Rost, B., Sander, C., & Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, 235(1), 13–26.
- Rota-Stabelli, O., & Telford, M. (2008). A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Molecular Phylogenetics and Evolution*, 48, 103–111.
- Rubio-Largo, A., Vega-Rodríguez, M. A., & González-Álvarez, D. L. (2016). A Hybrid Multiobjective Memetic Metaheuristic for Multiple Sequence Alignment. *IEEE Transactions on Evolutionary Computation*, 20(4), 499–514.
- Rzhetsky, A., & Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5), 1073–1095.
- Sadava, D., Hillis, D., & Heller, H. (2011). *Life: The Science of Biology*. Life: The Science of Biology. W. H. Freeman.

- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(1), 406–425.
- Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 1(5), 35–42.
- Santander-Jiménez, S., & Vega-Rodríguez, M. (2013a). Applying a multiobjective metaheuristic inspired by honey bees to phylogenetic inference. *Biosystems*, 114(1), 39–55.
- Santander-Jiménez, S., & Vega-Rodríguez, M. (2013b). A comparative study on distance methods applied to a multiobjective firefly algorithm for phylogenetic inference. In *Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '13 Companion, (pp. 1587–1594).
- Santander-Jiménez, S., & Vega-Rodríguez, M. (2013c). A multiobjective proposal based on the firefly algorithm for inferring phylogenies. In *Proceedings of the 11th European Conference EvoBIO 2013*, 1, (pp. 141–152). Springer Berlin Heidelberg.
- Santander-Jiménez, S., & Vega-Rodríguez, M. (2014). Inferring multiobjective phylogenetic hypotheses by using a parallel indicator-based evolutionary algorithm. In A.-H. Dediu, M. Lozano, & C. Martín-Vide (Eds.) *Theory and Practice of Natural Computing: Third International Conference*, (pp. 205–217). Springer International Publishing.
- Santander-Jiménez, S., & Vega-Rodríguez, M. (2015a). A hybrid approach to parallelize a fast non-dominated sorting genetic algorithm for phylogenetic inference. *Concurrency and Computation: Practice and Experience*, 27(3), 702–734.
- Santander-Jiménez, S., & Vega-Rodríguez, M. (2015b). On the design of shared memory approaches to parallelize a multiobjective bee-inspired proposal for phylogenetic reconstruction. *Journal of Information Science*, 324(-), 163–185.
- Santander-Jiménez, S., & Vega-Rodríguez, M. (2016). Performance evaluation of dominance-based and indicator-based multiobjective approaches for phylogenetic inference. *Information Sciences*, 330(C), 293–314.
- Santander-Jiménez, S., Vega-Rodríguez, M., Pulido, J. G., & Sánchez-Pérez, J. (2012). Comparing different operators and models to improve a multiobjective artificial bee colony algorithm for inferring phylogenies. In *Theory and Practice of Natural Computing*, vol. 7505, (pp. 187–200). Springer.
- Schliep, K. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4), 592–593.
- Sengupta, S., & Bandyopadhyay, S. (2012). De novo design of potential reca inhibitors using multiobjective optimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4), 1139–1154.
- Sheneman, L., & Foster, J. (2006). Estimating the destructiveness of crossover on binary tree representations. In M. Keijzer, & M. Cattolico (Eds.) *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, vol. 2, (pp. 1427–1428). ACM Press.
- Sievers, F., & Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. In D. J. Russell (Ed.) *Multiple Sequence Alignment Methods*, (pp. 105–116). Totowa, NJ: Humana Press.
- Skourikhine, A. (2000). Phylogenetic tree reconstruction using self-adaptive genetic algorithm. In *BIBE*, (pp. 129–134). IEEE Computer Society.
- Smith, S., Moore, M., Brown, J., & Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, 15(1), 150–160.

- Srinivas, N., & Deb, K. (1994). Multiobjective optimization using non dominated sorting in genetic algorithms. *Evolutionary Computation*, 2(2), 221–248.
- Stamatakis, A. (2004). *Distributed and parallel algorithms and systems for inference of huge phylogenetic trees based on the maximum likelihood method*. Ph.D. thesis, Technical University Munich, Germany.
- Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(1), 2688–2690.
- Steel, M. A., & Penny, D. (1993). Distributions of tree comparison metrics—some new results. *Systematic Biology*, 42(2), 126.
- Strachan, T., & Andrew, R. (2011). *Human Molecular Genetics, Fourth Edition*, vol. 1. Garland Science, 4 ed.
- Sudhir, K., Glen, S., & Koichiro, T. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger data-sets. *Molecular Biology and Evolution*, 33(7), 1870–1874.
- Sumner, J., Jarvis, P., Fernandez-Sanchez, J., Kaine, B., Woodhams, M., & Holland, B. (2012). Is the general time-reversible model bad for molecular phylogenetics? *System Biology*, 61(1), 1069–1074.
- Swofford, D., & Maddison, W. (1987). Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, 87(2), 199–229.
- Swofford, D., Waddell, P., Helsenbeck, J., Foster, P., Lewis, P., & Rogers, J. (2001). Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology*, 50(4), 525–539.
- Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, 9, 678–687.
- Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3), 512–526.
- Tapia, D., Eissler, Y., Torres, P., Jorquera, E., Espinoza, J., & Kuznar, J. (2015). Detection and phylogenetic analysis of infectious pancreatic necrosis virus in Chile. *Diseases of Aquatic Organisms*, 116(1), 173–184.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17(1), 57–86.
- Templeton, A. R., & Sing, C. F. (1993). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics*, 134(2), 659–669.
- Thiel, V., Tank, M., Neulinger, L., S.and Gehrmann, Dorador, C., & Imhoff, J. F. (2010). Unique communities of anoxygenic phototrophic bacteria in saline lakes of Salar de Atacama (chile): evidence for a new phylogenetic lineage of phototrophic gammaproteobacteria from puflm gene analyses. *FEMS Microbiology Ecology*, 74(1), 510–522.
- Tofigh, A., Hallett, M., & Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2), 517–535.
- Torres, P., Eissler, Y., Tapia, D., Espinoza, J. C., & Kuznar, J. (2016). Genotipificación y relación hospedador-específica del virus de la necrosis pancreática infecciosa en Chile. *Latin american journal of aquatic research*, 1(1), 860–868.

- Ulloa, P. M., Hernández, C. E., Rivera, R. J., & Ibáñez, C. M. (2017). Biogeografía histórica de los calamares de la familia Loliginidae (Teuthoidea: Myopsida). *Latin american journal of aquatic research*, 45(1), 113–129.
- Venegas, M., Alvarado-Mora, M., Villanueva, R., Rebello, J., Carrilho, F., Locarnini, S., Yuen, L., & Brahm, J. (2011). Phylogenetic analysis of hepatitis b virus genotype f complete genome sequences from chilean patients with chronic infection. *Journal of Medical Virology*, 83(1), 1530–1536.
- Vila, I., Morales, P., Scott, S., Poulin, E., Veliz, D., Harrod, C., & Mendez, M. (2013). Phylogenetic and phylogeographic analysis of the genus Orestias (Teleostei: Cyprinodontidae) in the southern Chilean Altiplano: the relevance of ancient and recent divergence processes in speciation. *Journal of Fish Biology*, 82(1), 927–943.
- Villalobos-Cid, M., Dorn, M., Ligabue-Braun, R., & Inostroza-Ponta, M. (2017a). A memetic algorithm based on an NSGA-II scheme for phylogenetic tree inference. *IEEE Transactions on Evolutionary Computation*.
- Villalobos-Cid, M., Vega-Araya, D., & Inostroza-Ponta, M. (2017b). Application of different multi-objective decision making techniques in the phylogenetic inference problem. In *36th International Conference of the Chilean Computer Science Society, SCCC 2017, Arica, Chile, October 16-20, 2017*.
- von Haeseler, A. (2012). Do we still need supertrees? *BMC Biology*, 10(1), 13.
- Wang, F., & Zhu, Z. (2013). Global path planning of wheeled robots using a multi-objective memetic algorithm. In *14th International Conference, IDEAL, IDEAL 2013*, (pp. 437–444).
- Wernicke, S. (2003). *On the Algorithmic Tractability of Single Nucleotide Polymorphism (SNP) Analysis and Related Problems*. Ph.D. thesis, University Tübingen.
- Wheeler, W. (2015). Phylogenetic network analysis as a parsimony optimization problem. *BMC Bioinformatics*, 16, 296–305.
- Wilgenbusch, J. C., Huang, W., & Gallivan, K. A. (2017). Visualizing phylogenetic tree landscapes. *BMC Bioinformatics*, 18(1), 85:1–85:12.
- Wróbel, B., Bogdanowicz, D., & Giaro, K. (2012). Treecmp: Comparison of trees in polynomial time. *Evolutionary Bioinformatics*, 8, 475–487.
- Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13, 303–314.
- Yassin, A., Lienau, E. K., Narechania, A., & DeSalle, R. (2010). Catching the phylogenetic history through the ontogenetic hourglass: a phylogenomic analysis of drosophila body segmentation genes. *Evolution & Development*, 12(3), 288–295.
- Yijie, S., & Gongzhang, S. (2008). Improved NSGA-II multi-objective genetic algorithm based on hybridization-encouraged mechanism. *Chinese Journal of Aeronautics*, 21(6), 540–549.
- Yu, Y., & Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(1), 10–28.
- Zambrano-Vega, C., Nebro, A., & Aldana-Montes, J. (2016). Mo-phylogenetics: a phylogenetic inference software tool with multi-objective evolutionary metaheuristics. *Methods in Ecology and Evolution*, 7(7), 800–805.
- Zararsiz, G., & Coşgun, E. (2014). Introduction to Statistical Methods for MicroRNA Analysis. In M. Yousef, & J. Allmer (Eds.) *miRNomeics: MicroRNA Biology and Computational Analysis*, vol. 1107 of *Methods in Molecular Biology*, (pp. 129–155). Humana Press.

- Zhang, C., Stadler, T., Klopstein, S., Heath, T. A., & Ronquist, F. (2015). Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65(2), 228–249.
- Zhihua, D., Feng, L., & Usman, R. (2005). Reconstruction of large phylogenetic trees: A parallel approach. *Computational Biology and Chemistry*, 29(4), 273–280.
- Zrzavý, J., Říha, P., Piálek, L., & Janouškovec, J. (2009). Phylogeny of annelida (lophotrochozoa): total-evidence analysis of morphology and six genes. *BMC Evolutionary Biology*, 9(1), 189.
- Zwickl, D. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Ph.D. thesis, The University of Texas at Austin.

ANEXO A. NÚMERO DE PUBLICACIONES POR AÑO PERTINENTES AL ÁREA DE INVESTIGACIÓN

Las siguientes tablas muestran el número de publicaciones con tópicos relacionados a esta investigación registradas durante el último quinquenio en las bases de datos **PubMed** (Tabla A.1), **ScienceDirect** (Tabla A.2) y **ProQuest** (Tabla A.3). Los datos fueron obtenidos durante el mes de mayo 2017.

Tabla A.1: Número de publicaciones relacionada en Pubmed: 2007-2017.

Palabras claves	Pubmed				
	2013	2014	2015	2016	2017
Phylogeny	13162	13112	13421	8948	2258
Phylogenetic inference	287	272	323	381	162
Phylogeny analysis	5684	5835	6203	5980	2534
Phylogenetic tree	1470	1528	1603	1852	833
Phylogenetic network	230	268	287	297	155
Optimization (Optimisation)	8113	8888	9433	10372	5438
Muti-objetive optimization	69	65	85	104	52
Muti-objetive optimization & phylogeny	0	0	0	0	0

Fuente: Elaboración propia, (2017). Datos extraídos de *PubMed*, mayo 2017.

Tabla A.2: Número de publicaciones relacionada en ScienceDirect: 2007-2017.

Palabras claves	ScienceDirect				
	2013	2014	2015	2016	2017
Phylogeny	2849	2790	2995	3421	1966
Phylogenetic inference	1031	1063	1218	1495	908
Phylogeny analysis	7650	7818	8595	9531	5734
Phylogenetic tree	5010	5105	5614	6422	3820
Phylogenetic network	1734	1777	2052	2287	1363
Optimization (Optimisation)	56855	62823	69370	75747	49658
Muti-objetive optimization	5846	6929	7535	8718	5816
Muti-objetive optimization & phylogeny	26	17	18	16	18

Fuente: Elaboración propia, (2017). Datos extraídos de *ScienceDirect*, mayo 2017.

Tabla A.3: Número de publicaciones relacionada en ProQuest: 2007-2017.

Palabras claves	ProQuest				
	2013	2014	2015	2016	2017
Phylogeny	13258	12817	12044	9943	2787
Phylogenetic inference	4728	4688	4543	3688	984
Phylogeny analysis	22630	22612	22352	19184	5655
Phylogenetic tree	12643	12856	12653	10332	2747
Phylogenetic network	6237	6409	6167	5035	1257
Optimization (Optimisation)	144015	162308	166189	162483	62826
Muti-objetive optimization	387	480	492	364	89
Muti-objetive optimization & phylogeny	1	0	0	2	3

Fuente: Elaboración propia, (2017). Datos extraídos de *ProQuest*, mayo 2017.

El número de publicaciones relacionadas al concepto de *multi-objetive optimization* considera el número de publicaciones de asociadas al la búsqueda de *multi-criteria optimization*.

ANEXO B. DEFINICIONES DE ORIGINALIDAD PARA UN TRABAJO DE INVESTIGACIÓN

Usualmente se exige a los doctorandos generar una **contribución original** al conocimiento. Esto produce estrés y desgaste a la hora de identificar problemas y proponer soluciones, ya que como parte de la formación no se explica lo que califica como un **trabajo original**. Ante ello, surge la idea que el trabajo desarrollado debe describir un problema no estudiado previamente o aplicar una solución inédita a un problema particular. No obstante, según Phillips (1992) la investigación de un doctorando puede ser original de diferentes maneras, como por ejemplo:

1. Realizar un trabajo empírico no hecho anteriormente.
2. Efectuar una síntesis no realizado hasta el momento.
3. Usar material ya conocido para dar una nueva interpretación.
4. Probar a nivel local algo ya realizado en otro lugar.
5. Aplicar una técnica concreta a una nueva área.
6. Proporcionar nuevas pruebas para fundamentar una cuestión de larga data.
7. Ser interdisciplinario y emplear diferentes metodologías.
8. Considerar áreas que una determinada disciplina no abordaba antes.
9. Aportar conocimientos de una manera inédita.

En el desarrollo de esta investigación se trata un problema que combina diferentes áreas de conocimiento, entre ellas: **modelamiento computacional**, mediante la aplicación de algoritmos evolutivos multi-objetivo, e **inferencia filogenética**. Considerando que: (1) la mayoría de los algoritmos propuestos en este trabajo han sido aplicados exitosamente en otros campos de investigación, y que (2) el problema de inferencia filogenética para modelamiento de fenómenos reticulares no ha sido tratado previamente en la literatura con un enfoque multi-objetivo, la investigación desarrollada cumple con las definiciones 1, 5, 7 y 8.

ANEXO C. MÉTODOS PARA RECONSTRUCCIÓN DE ÁRBOLES FILOGENÉTICOS

C.1 MÉTODOS DE RECONSTRUCCIÓN FILOGENÉTICA BASADOS EN DISTANCIA

Las aproximaciones basadas en distancias cuentan con algoritmos muy rápidos que resultan útiles para efectuar análisis preliminares. Los resultados obtenidos pueden ser similares o iguales a árboles estimados con modelos probabilistas si se utilizan datos con poco ruido, pero tienen un desempeño bastante pobre cuando las asunciones del modelo evolutivo no corresponden a los datos (Abascal et al., 2014).

C.1.1 Algoritmos para generación de árboles ultramétricos

Formalmente un árbol enraizado es ultramétrico si la distancia desde su raíz es la misma para todos los nodos terminales, y cumple con la condición métrica de los tres puntos ($d_{AC} \leq \max(d_{AB}, d_{BC})$). Esto implica que los descendientes han evolucionado desde el ancestro principal a una tasa constante (hipótesis del reloj molecular). Sin embargo, si los datos o el modelo estudiado presentan una tasa de evolución desigual, estos algoritmos presentarán una topología incorrecta.

Los algoritmos trabajan considerando como entrada una matriz D de distancia (d_{ij}) construida al comparar n_i especies. Las etapas para construir un árbol son las siguientes:

1. Encontrar la distancia d_{ij} menor en la matriz D .
2. Combinar los elementos i y j en un grupo k .
3. Recalcular las distancias d_{km} empleando:

$$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma |d_{im} - d_{jm}| \quad (\text{C.1})$$

Donde m representa cualquier elemento o grupo distinto de k . Estas nuevas distancias reemplazan a d_{im} y d_{jm} en D . Los valores de α_i, α_j , β y γ dependerán del algoritmo específico utilizado.

4. Repetir los pasos 1 y 2 hasta que se consiga un grupo que englobe a todos los elementos.

Los algoritmos específicos a utilizar se describen a continuación y sus resultados se muestran en la Figura C.1

- Vinculación individual: también conocido como *Nearest Neighbor Clustering* (NNC). En este caso la distancia entre dos grupos está determinada por los elementos más cercanos. Los valores de las constantes son: $\alpha_i = \alpha_j = 0,5$, $\beta = 0$ y $\gamma = -0,5$.
- Vinculación completa: también conocido como *Furthest Neighbo* (FN). En este caso la distancia entre dos grupos está determinada por los elementos más lejanos. Los valores de las constantes son: $\alpha_i = \alpha_j = 0,5$, $\beta = 0$ y $\gamma = 0,5$.
- Promedio simple: también se conoce como Grupos ponderados usando media aritmética (*Weighted Pair Group Method with Arithmetic Mean*, WPGMA). En este caso la distancia entre dos grupos está determinada por la distancia promedio entre sus elementos. Los valores de las constantes son: $\alpha_i = \alpha_j = 0,5$, $\beta = 0$ y $\gamma = 0$.

- Centroide: también conocido *Unweighted pair-group centroid method* (UPGMC). En este caso la distancia entre dos grupos está determinada por el centroide o centro de gravedad de sus elementos. Los valores de las constantes son: $\alpha_i = \frac{n_i}{n_k}$, $\alpha_j = \frac{n_j}{n_k}$, $\beta = -\alpha_i\alpha_j$ y $\gamma = 0$.
- Mediana: también se conoce como *Weighted pair-group centroid method* (WPGMC). En este caso la distancia entre dos grupos está determinada por la mediana entre sus elementos. Los valores de las constantes son: $\alpha_i = \alpha_j = 0,5$, $\beta = -0,25$ y $\gamma = 0$.
- Promedio de grupo: también se conoce como Unweighted Pair Group Method with Arithmetic Mean (UPGMA) y es el algoritmo más empleado de este tipo. Los valores de las constantes son: $\alpha_i = \frac{n_i}{n_k}$, $\alpha_j = \frac{n_j}{n_k}$, $\beta = 0$ y $\gamma = 0$.
- Mínima varianza de Ward: Este método agrupa minimizando la suma cuadrados de la varianza dentro de cada grupo. Los valores de las constantes son: $\alpha_i = \frac{n_j+n_m}{n_k+n_m}$, $\alpha_j = \frac{n_j+n_m}{n_k+n_m}$, $\beta = \frac{-n_m}{n_k+n_m}$ y $\gamma = 0$.

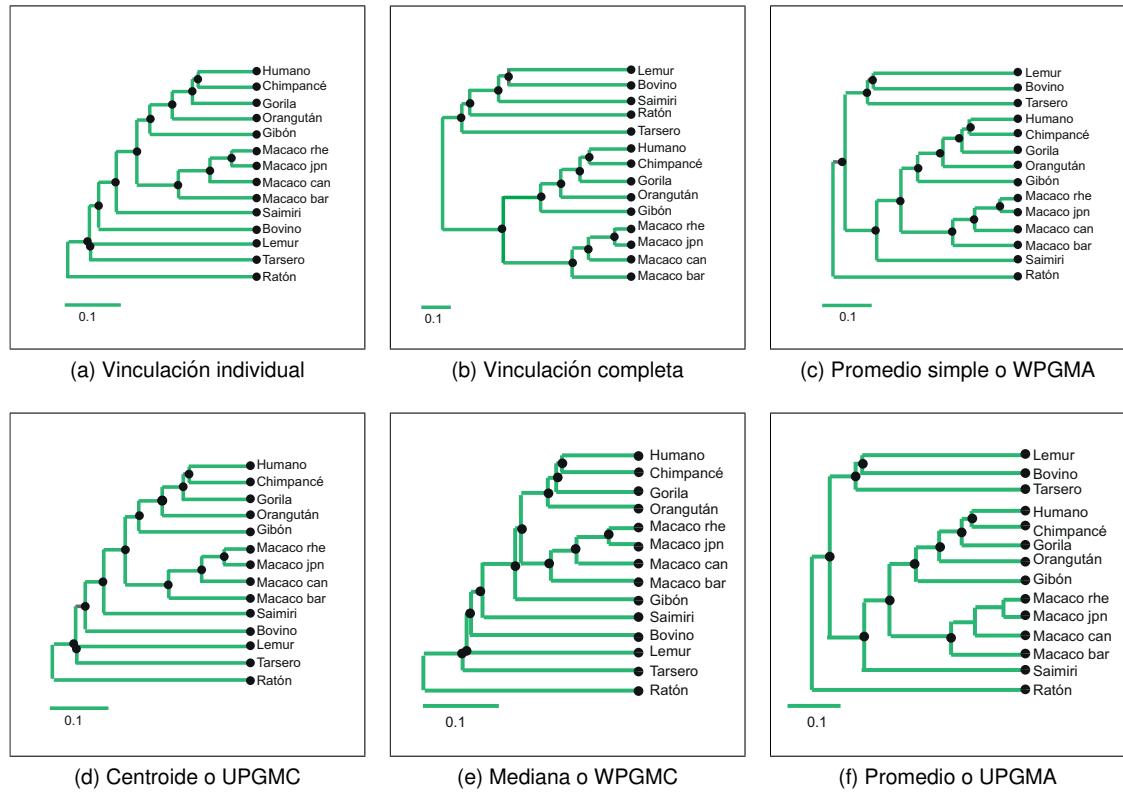


Figura C.1: Resultados obtenidos al emplear diferentes algoritmos aglomerativos con el conjunto de datos *primates_14*

Fuente: Elaboración propia, (2017).

C.1.2 Algoritmos para generación de árboles additivos

Formalmente un árbol es aditivo si la distancia evolutiva entre cada par de especies resulta igual a la suma de las longitudes de las ramas que las unen, y cumple la condición métrica de los cuatro puntos ($d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$). Bajo aditividad no se asume una tasa de evolución constante ni un ancestro común como raíz. Ultrametricidad implica aditividad, pero no a la inversa.

Saitou & Nei (1987) propusieron un algoritmo que permite generar este tipo de árboles rápidamente: *Neighbor joining* (NJ). Éste utiliza como entrada una matriz de distancia D y aplica el siguiente proceso:

1. Para cada especie de un conjunto de largo n , calcular $u_i = \sum_{j:j \neq i}^n \frac{D_{ij}}{(n-2)}$.
2. Seleccionar i y j en que $D_{ij} - u_i - u_j$ es menor.
3. Unir en un nuevo grupo a i y j . Calcular el tamaño de la rama desde i a un nuevo nodo v_i y desde j a un nuevo nodo v_j empleando la siguiente Ecuación:

$$v_i = \frac{1}{2}D_{ij} + \frac{1}{2}(u_i - u_j) \quad v_j = \frac{1}{2}D_{ij} + \frac{1}{2}(u_j - u_i) \quad (\text{C.2})$$

4. Calcular la distancia $D_{(ij)} = \frac{(D_{ik} + D_{jk} - D_{ij})}{2}$ entre el nuevo nodo (ij) y cada uno de los nodos restantes.
5. Borrar las especies i y j y reemplazarlos por un nuevo nodo ij .
6. Si existen más de dos nodos se debe regresar al primer paso. Caso contrario, conectar los dos nodos restantes l y m por una rama de distancia D_{lm} .

El algoritmo NJ se ha caracterizado por presentar soluciones rápidas. Sin embargo, estas resultan menos precisas que la entregada por algoritmos basados en probabilidad y a menudo presenta ramas negativas cuando se utilizan datos ruidosos. Se han diseñado otros algoritmos que buscan mejorar la precisión del algoritmo NJ como BIONJ (Gascuel, 1997) o NJ+ (Plonski & Radomski, 2013). La Figura C.2 muestra un ejemplo de aplicación de NJ y BIONJ en un mismo conjunto de datos, presentando ambos resultados equivalentes (Distancia Robinson-Foulds igual a cero).

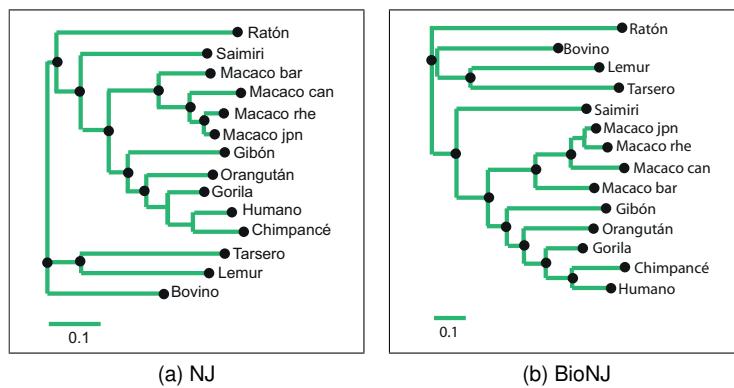


Figura C.2: Resultados obtenidos al emplear NJ y BIONJ con el conjunto de datos *primates_14*.
Fuente: Elaboración propia, (2017).

C.1.3 Métodos basados en optimización

C.1.3.1 Mínimos cuadrados

Los algoritmos basados en distancia utilizan una matriz de distancias observadas D_{ij} , para efectuar una aproximación a un modelo de árbol compuestos por pares d_{ij} . Los algoritmos basados en mínimos cuadrados (*Least-squares method, LS*) minimizan la discrepancia Q entre estas distancias observadas y esperadas mediante la siguiente ecuación:

$$Q = \sum_{i=1}^n \sum_{j=1}^n w_{ij} (D_{ij} - d_{ij})^2 \quad (\text{C.3})$$

Donde w_{ij} son pesos que difieren entre propuestas: $w_{ij} = 1$ (Cavalli-Sforza & Edwards, 1967), $w_{ij} = 1/D_{ij}^2$ (Fitch & Margoliash, 1967) y $w_{ij} = 1/D_{ij}$ (Beyer et al., 1974). El criterio de los mínimos cuadrados ha resultado ser más exacto que NJ pero menos eficiente, ya que se ha demostrado que encontrar el árbol óptimo mediante este criterio con cualquier factor de corrección es un problema NP-completo. La Figura C.3 muestra ejemplos de aplicación.

C.1.3.2 Evolución mínima

Método propuesto por Kidd & Sgaramella-Zonta (1971) emplea el criterio de los mínimos cuadrados no ponderados para ajustar el tamaño de las ramas, evaluando y comparando los árboles según su largo total L . La Ecuación C.4 es empleada para calcular L , donde e_i son los $2n - 3$ largos de ramas calculadas a partir de las distancias pares entre las secuencias.

$$L = \sum_{i=1}^{2n-3} |e_i| \quad (\text{C.4})$$

Usualmente se emplea NJ para construir una topología inicial. Sin embargo, Desper & Gascuel (2002) propusieron una nueva alternativa mas eficiente que utiliza un algoritmo goloso para construir las topologías iniciales: *Fast minimum evolution* (FastME). La Figura C.3 muestra ejemplos de aplicación.

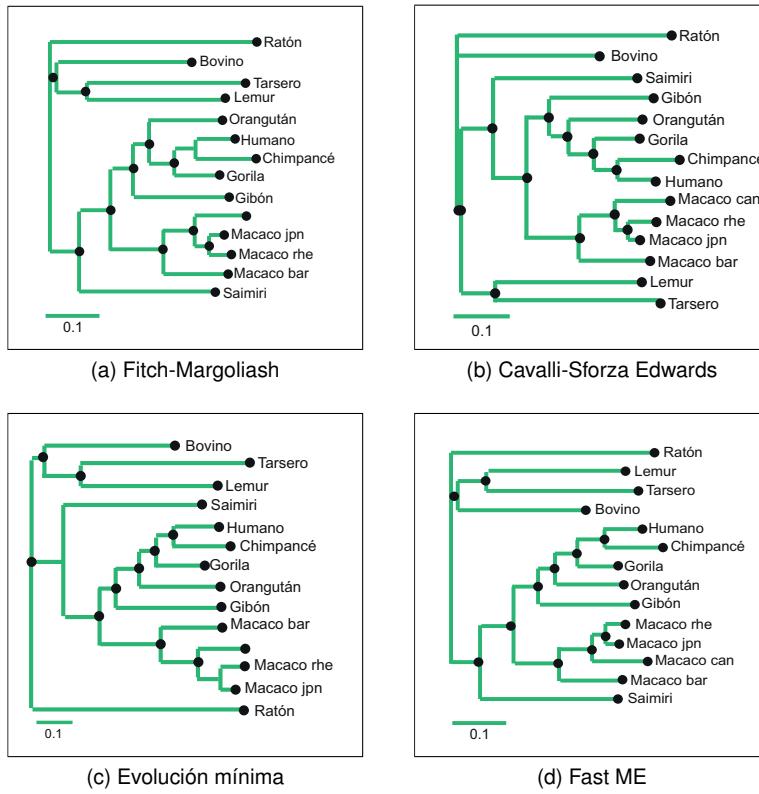


Figura C.3: Resultados obtenidos al emplear el criterio de mínimos cuadrados y evolución mínima con el conjunto de datos *primates_14*.

Fuente: Elaboración propia, (2017).

C.2 MÉTODOS DE RECONSTRUCCIÓN FILOGENÉTICA BASADOS EN CARACTERES

C.2.1 Máxima parsimonia

El problema de máxima parsimonia se fundamenta en el principio de Ockham, el que señala que si existen diferentes hipótesis para explicar un fenómeno, se debe seleccionar aquella que presenta una menor complejidad. Esto se traduce en la selección del árbol que contiene el menor número de cambios evolutivos durante el estudio de la diversificación de un conjunto de secuencias.

El problema de encontrar el árbol más parsimonioso pertenece a la categoría de problemas NP-duro (Felsenstein, 2004). Un algoritmo diseñado para abordar este problema debe considerar dos aspectos: (1) la capacidad de proponer una filogenia determinando y minimizando el número de eventos evolutivos, y (2) recorrer todo el espacio de configuraciones hasta hallar la solución con el menor número de cambios.

A diferencia de los algoritmos de optimización que actúan sobre matrices de distancias, los algoritmos diseñados para maximizar parsimonia actúan directamente sobre los caracteres de una secuencia (Figura C.4). Algoritmos clásicos de este tipo son los algoritmos de Fitch (Fitch, 1971) y de Sankoff (Sankoff, 1975). Como resultado de un algoritmo basado

en maximización de parsimonia es posible tener varios árboles que cumplen con el criterio de optimalidad, por lo que es común la estimación de árboles de consenso.

C.2.2 Máxima verosimilitud

El concepto de máxima verosimilitud se puede explicar mediante el cociente de probabilidad. Dado dos hipótesis, H_1 y H_2 , para un conjunto de datos D la $Prob(H|D) = Prob(H \cap D)/Prob(B) = Prob(D|H)Prob(H)/Prob(D)$. A raíz de ello el cociente de probabilidad es:

$$\frac{Prob(H_1|D)}{Prob(H_2|D)} = \frac{Prob(D|H_1)}{Prob(D|H_2)} \frac{Prob(H_1)}{Prob(H_2)} \quad (C.5)$$

La componente $Prob(D|H)$ se denomina verosimilitud de la hipótesis H . Por lo que ante un conjunto de diferentes datos u observaciones, la verosimilitud será: $Prob(D|H) = Prob(D^1|H_1)Prob(D^2|H_1)\dotsProb(D^n|H_1)$ y la Ecuación C.5 se convertirá en:

$$\frac{Prob(H_1|D)}{Prob(H_2|D)} = \left(\prod_{i=1}^n \frac{Prob(D|H_i)}{Prob(D|H_2)} \right) \frac{Prob(H_1)}{Prob(H_2)} \quad (C.6)$$

El concepto aplicado a filogenia fue establecido por Felsenstein (1973). Consiste en encontrar la topología de árbol más probable que explica los datos observados empleando un modelo evolutivo. Para ello es necesario hacer dos suposiciones: (1) la evolución en diferentes sitios en un árbol determinado es independiente, y (2) la evolución por linajes también es independiente (Felsenstein, 2004).

LA Figura C.4 muestra un ejemplo de aplicación. Si se quiere efectuar inferencia filogenética empleando un carácter perteneciente a una secuencia de ADN de 5 especies (A,C,C,C,G) en un conjunto de datos T , la probabilidad de hipótesis estará determinada por la Ecuación C.7. Donde x, t, z y w son los estados internos.

$$Prob(D^{(i)}|T) = \sum_x \sum_y \sum_z \sum_w Prob(A, C, C, C, G, x, y, z, w|T) \quad (C.7)$$

Se ha demostrado que el problema de encontrar el árbol mediante máxima verosimilitud es NP-duro (Chor & Tuller, 2005). Un árbol con n número de especies tendrá $n - 1$ nodos internos, con 4 estados (nucleótidos), por lo que existirán 4^{n-1} configuraciones. Por ejemplo, para estudiar 20 especies se requerirá revisar 274×10^9 topologías.

C.2.3 Aproximación bayesiana

La inferencia Bayesiana en filogenia se basa en el cálculo de la probabilidad posterior de un árbol, suponiendo que este sea correcto para un conjunto de datos y un modelo evolutivo.

La probabilidad posterior se calcula mediante el Teorema de Bayes, en que la probabilidad posterior de un árbol $P(T, \theta|D)$ es proporcional a su probabilidad previa $P(T, \theta)$ multiplicada por su verosimilitud, según lo especificado en la Ecuación C.8, donde $P(D)$ es la probabilidad marginal de los datos (Abascal et al., 2014).

$$P(T, \theta) = \frac{P(T, \theta)P(D|T, \theta)}{P(D)} \quad (C.8)$$

El cálculo de la probabilidad posterior implica evaluar todos los posibles árboles, y para cada uno de ellos determinar las combinaciones de longitudes de ramas y parámetros de modelos evolutivos. Debido a la impracticabilidad desde el punto de vista computacional, se han empleado Cadenas de Markov Monte Carlo (*Markov Chain Monte Carlo, MCMC*) para estimar la distribución de probabilidad. Este tipo de algoritmos suele quedar estancado en zonas sin gradientes o máximos locales, por lo que herramientas como MrBayes (Huelsenbeck & Ronquist, 2001), emplean varios modelos a la vez: una cadena fría y tres calientes.

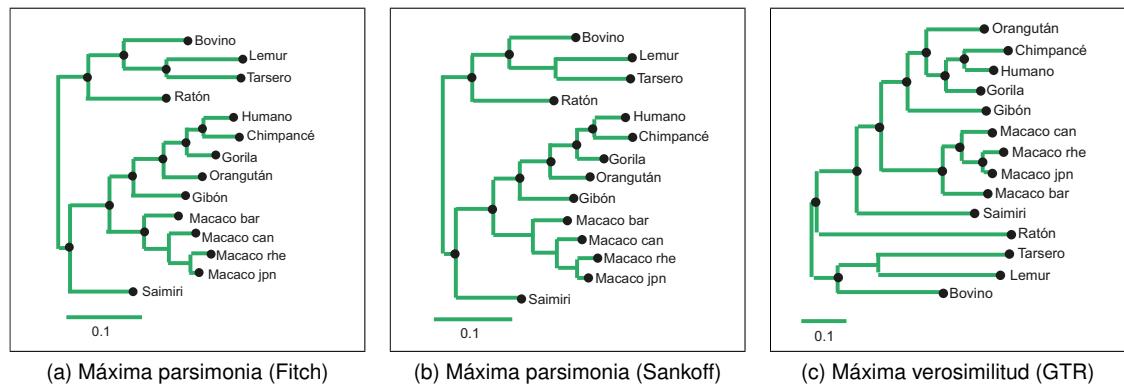


Figura C.4: Resultados obtenidos al emplear los criterios de máxima parsimonia y verosimilitud con el conjunto de datos *primates_14*.

Fuente: Elaboración propia, (2017).

C.3 MODELOS EVOLUTIVOS

Por medio de las distancias de edición es posible evaluar cuánto difieren dos secuencias según el número de caracteres que deben ser modificados para convertir una en la otra. Sin embargo, un nucleótido o aminoácido puede haber pasado por estados intermedios antes de convertirse en la molécula de una secuencia final. A raíz de ello se han desarrollado modelos evolutivos que incorporan probabilidades de las tasas de cambios (Figura C.5).

El modelo evolutivo más simple para secuencias de ADN es el diseñado por Jukes-Cantor en 1969 (*JC69*). Este asume que todas las bases a lo largo de una secuencia tienen igual frecuencia ($\pi_T = \pi_C = \pi_G = \pi_A = \frac{1}{4}$) y probabilidad de cambio. El valor de la distancia (D_{jc}) se puede calcular mediante la Ecuación C.9, donde d es la proporción de sitios con nucleótidos diferentes (Jukes & Cantor, 1969).

$$D_{jc} = -\frac{3}{4} \ln(1 - \frac{3}{4}d) \quad (\text{C.9})$$

En la realidad las probabilidades de una transición (base purina por purina o pirimidina por pirimidina) es mayor a la transversión (purina por pirimidina o viceversa). Kimura (1980) propuso un modelo (*K80*) basado en estos dos parámetros: α y β respectivamente. Luego de combinar estas probabilidades en el modelo de *JC69*, la distancia resultante (D_{k80}) se puede calcular empleando la Ecuación C.10, donde P y Q son las frecuencias de los sitios con transición y transversión.

$$D_{k80} = \frac{-\ln(1 - 2P - Q)}{-\ln(1 - 2Q)} - \frac{1}{2} \quad (\text{C.10})$$

Felsenstein (1981) (*F81*) generalizó el modelo *JC69* modificando la frecuencia de

aparición de cada base ($\pi_T \neq \pi_C \neq \pi_G \neq \pi_A$) . La Ecuación de distancia D_{f81} está dada por:

$$D_{f81} = b \ln(1 - \frac{p}{b}) \quad b = \frac{1}{2}(1 - \sum_{i=1}^4 g_i^2 + \frac{p^2}{c}) \quad c = \sum_{i=1}^3 \sum_{j=i+1}^4 \frac{x_{ij}^2}{2g_i g_j} \quad (\text{C.11})$$

Donde x_{ij} es la proporción de pares de nucleótidos i y j entre los dos secuencias de ADN, g_i y g_j son las frecuencias de equilibrio de los $i-th$ nucleótidos, y p la proporción de sitios con nucleótidos diferentes. Hasegawa et al. (1985) (HYK85) combinaron los modelos F81 y K80 distinguiendo entre transversiones y transiciones empleando un parámetro k . Barry & Hartigan (1987) desarrollaron una distancia basada en proporciones observadas experimentalmente entre cambios de bases (BH+I). Sin embargo, esta medida de distancia no resultó simétrica. Tamura (1992) (T92) extendió el modelo K80 considerando sesgo producido por el contenido GC (θ). La distancia (D_{t82}) se puede calcular mediante la Ecuación C.12. Al año siguiente Tamura & Nei (1993) propusieron un modelo que integra diferentes probabilidades según el tipo transversión (TN93).

$$D_{t82} = h \ln(1 - \frac{p}{h} - q) - \frac{1}{2}(1 - h) \ln(1 - 2q) \quad \text{Donde } h = 2\theta(1 - \theta) \quad \text{y } \theta \in (0, 1) \quad (\text{C.12})$$

El modelo mas empleado en la actualidad (Sumner et al., 2012) es el GTR propuesto por Tavaré (1986). Éste permite integrar 6 tasas diferentes de sustituciones reversibles y diferenciar la frecuencia de cada una de las bases. Bajo el mismo principio se han construido modelos matriciales basados en probabilidad para evaluar sustituciones en aminoácido: WAG, JTT, LG, Dayhoff, cpREV, mtmam, mtArt, MtZoa, mtREV24, VT, RtREV, HIVw, HIVb, FLU, Blossum62, Dayhoff-DCMut, JTT-DCMut, entre otros. Es posible hallar una descripción de cada uno de ellos en Schliep (2011).

Existen distintos criterios que pueden utilizarse para elegir el modelo que mejor se ajusta a los datos. Tradicionalmente se utilizan contrastes basados en la tasa de verosimilitud (*Likelihood ratio test, LRT*). Sin embargo, en la actualidad se ha extendido ampliamente el uso del criterio de información de Akaike (Abascal et al., 2014).

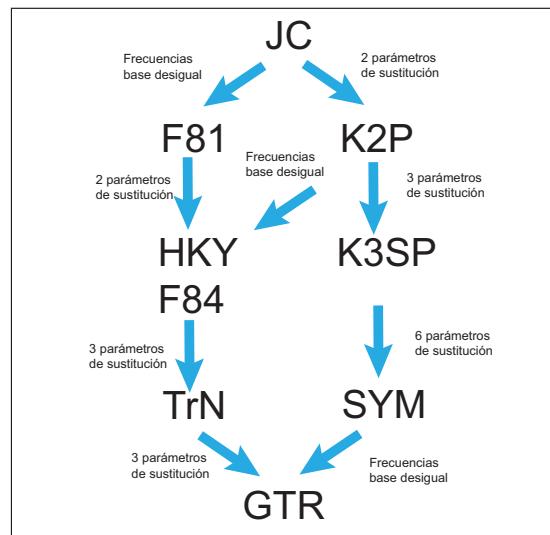


Figura C.5: Relaciones entre modelos evolutivos para secuencias de nucleótidos.
Fuente: Huson et al. (2011).

ANEXO D. ALGORITMO MEMÉTICO MULTI-OBJETIVO PARA INFERENCIA DE ÁRBOLES FILOGENÉTICOS

D.1 PARÁMETROS PARA ESTRATEGIAS DE BÚSQUEDA LOCAL

D.1.1 Número de iteraciones para algoritmo de búsqueda local

Con objetivo de definir el número de iteraciones para las estrategias de búsqueda local (parámetro ls), se efectuaron 31 ejecuciones de MO-MA empleando la configuración NJ-PDG-NNI-G sobre tres conjuntos de datos. Se evaluó:

- La relación entre el número de iteraciones y el tiempo de ejecución.
- La relación entre el número de iteraciones y la métrica de hipervolumen.

La Figura D.1 (superior izquierda) muestra la relación entre el parámetro ls (rango [1-100]) y el tiempo de ejecución normalizado por el valor promedio obtenido para NSGA-II EM. La Figura D.1 (superior derecha) muestra la relación entre el parámetro ls y la métrica de hipervolumen. Esta métrica incrementa en dos rangos de iteraciones para los tres conjuntos de datos: [10-15] y [75-100]. Se ha decidido usar el valor de ls que maximiza el valor del hipervolumen y minimiza el tiempo de ejecución: 10 iteraciones.

Básicamente el algoritmo emplea ls iteraciones, buscando un vecino p' para cada uno de los p individuos que componen una población de tamaño P . Si p' domina a p , este último es incorporado a la población P . Finalmente, una estrategia de ordenamiento no dominado controla el tamaño de la población P .

Algoritmo D.1: Algoritmo Pareto local search

Input: Un conjunto de P soluciones.

Output: Un nuevo conjunto de P soluciones.

```
1 for  $i \in 1 : ls$  do
2    $F[p] \leftarrow \text{FALSE } \forall p \in P;$ 
3   while ( $\exists f \in F : f = \text{FALSE}$ ) do
4      $p \leftarrow P[f] : f = \text{FALSE}, f \in F;$ 
5     for each  $p' \in \text{Vecindario}(p)$  do
6       if ( $p' \not\sim p$ ) then
7          $F[p'] \leftarrow \text{FALSE} ;$ 
8          $P \leftarrow P \cup p' ;$ 
9       end
10       $F[p] \leftarrow \text{TRUE} ;$ 
11    end
12     $P = \text{ORDENAMIENTO\_NO\_DOMINADO}(P);$ 
13  end
14 end
15 return ( $P$ );
```

D.1.2 Parámetros para algoritmo SA

Con objetivo de definir el coeficiente de temperatura (α) y el valor de temperatura inicial se aplicó el mismo método empleado en la sección anterior, reemplazando la estrategia golosa por el algoritmo SA. Se comparó:

- La relación entre la temperatura inicial y la métrica de hipervolumen (rango [1-1000])
- La relación entre el coeficiente de temperatura y la métrica de hipervolumen (rango [0.1-0.9])

Se ha seleccionado el valor inicial de temperatura (Figura D.1, inferior izquierda) y el valor de α (Figura D.1, inferior derecha) que maximiza la métrica de hipervolumen. Estos valores son 10 y 0.9 respectivamente.

Esta estrategia funciona similar a PLS, sin embargo posee un mayor grado de flexibilidad minimizando el riesgo de estancamiento en mínimos locales. En este caso, una función de temperatura que es modificada a medida que avanza el número de iteraciones l_s , controla el nivel de aceptancia de una solución p' en la población P (Algoritmo D.2), dependiendo de una probabilidad ($Prob$).

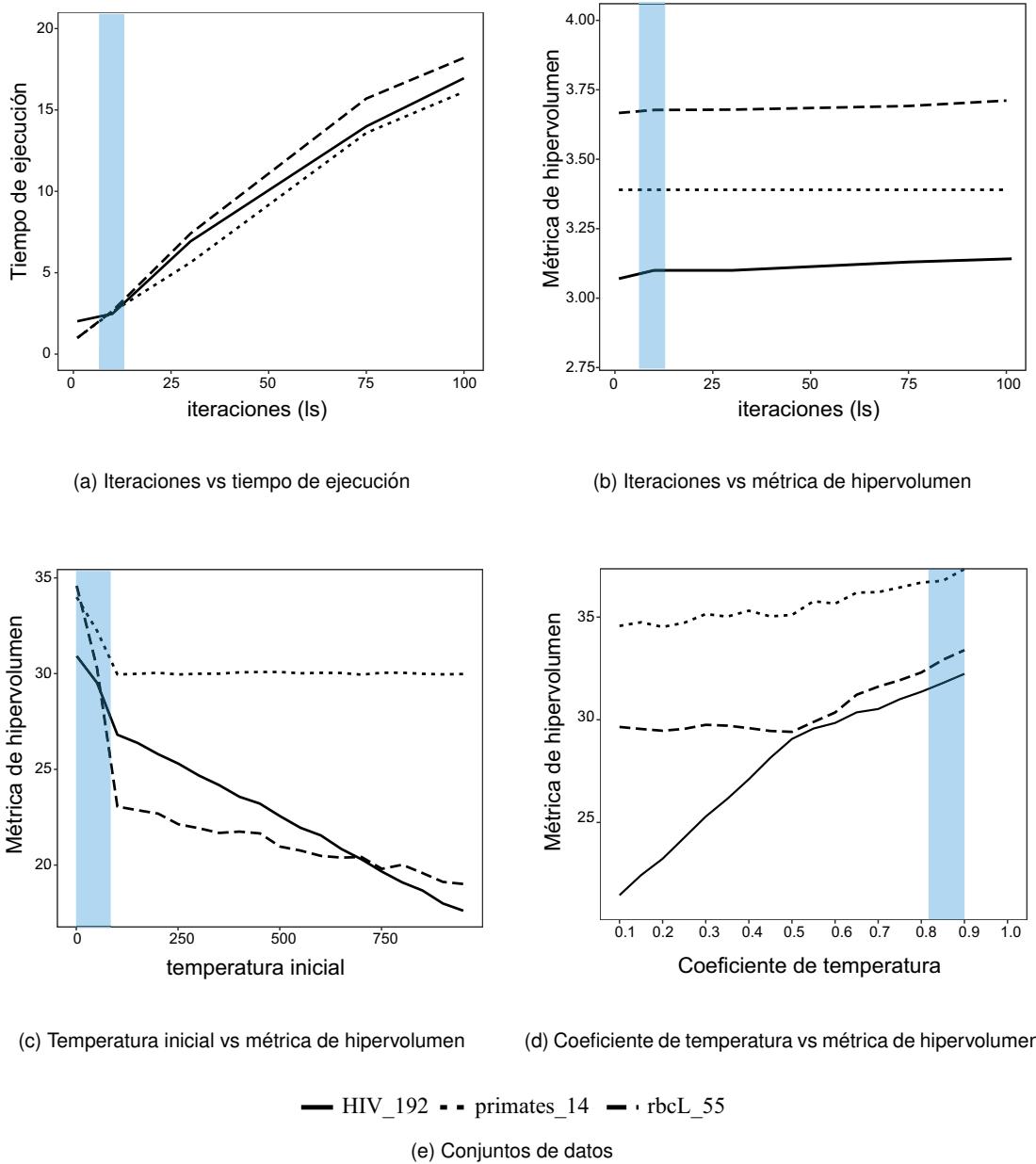


Figura D.1: Evaluación de diferentes parámetros para el proceso de búsqueda local: relación entre el número de iteraciones y tiempo de ejecución (superior izquierda), relación entre el número de iteraciones y la métrica de hipervolumen (superior derecha), y la relación entre la temperatura inicial (inferior izquierda) y el coeficiente de temperatura alfa (inferior derecha) con la métrica de hipervolumen. Los rangos con los mejores parámetros han sido coloreados.

Fuente: Elaboración propia, (2017).

Algoritmo D.2: Algoritmo Simulated Annealing

Input: Un conjunto de P soluciones, T, α .

Output: Un nuevo conjunto de P soluciones.

```
1 for  $i \in 1 : ls$  do
2    $F[p] \leftarrow \text{FALSE } \forall p \in P;$ 
3    $T = T * \alpha^i$ 
4   while ( $\exists f \in F : f = \text{FALSE}$ ) do
5      $p \leftarrow P[f] : f = \text{FALSE}, f \in F;$ 
6     for each  $p' \in \text{Vecindario}(p)$  do
7       if ( $p' \neq p$ ) then
8          $F[p'] \leftarrow \text{FALSE};$ 
9          $P \leftarrow P \cup p';$ 
10      end
11    else
12       $Prob = \exp^{-(\Delta f_{par,ver}(p,p'))/T};$ 
13      if ( $Prob > \text{Aleatorio}(0, 1)$ ) then
14         $P \leftarrow P \cup p';$ 
15      end
16    end
17     $F[p] \leftarrow \text{TRUE};$ 
18  end
19   $P = \text{ORDENAMIENTO\_NO\_DOMINADO}(P);$ 
20 end
21 end
22 return ( $P$ );
```

D.2 HERRAMIENTAS BASADAS EN OPTIMIZACIÓN DE OBJETIVO ÚNICO

Dado que el tiempo requerido para calcular parsimonia y verosimilitud difiere, y que algunas herramientas no pueden ser parametrizadas para reducir su tiempo de aplicación, se ha decidido emplear los parámetros recomendados para herramienta basada en objetivo único (Tabla D.1). En el caso de NSGA-II EM, los parámetros especificados fueron obtenidos desde Santander-Jiménez & Vega-Rodríguez (2013a). Con el fin de realizar una comparación justa, se ha ajustado el número de generaciones de MO-MA usando el mismo tiempo requerido por NSGA-II EM para realizar 100 generaciones (Tabla D.2).

Tabla D.1: Parámetros usados en MO-MA para comparación con herramientas basadas en optimización de objetivo único.

Herramientas	Parámetros	Ejecuciones
DNAPARS	Number of trees to save: 1000 Threshold parsimony: no Transversion parsimony: no	30
MEGApars	Substitution type: nucleotides Gaps: Use all sites Codon positions: all MP search method: SPR 5 levels -N initial trees: 100 -MP search level: 2 -Max N trees to retain:1000	30
MEGAlik	Data: DNA Substitution model: GTR Gamma distributed: 5 Gaps: Use all sites ML heuristic: SPR 5 levels Initial tree: True	30
PHYML	Data: DNA Substitution model: GTR Proportion of invariable sites: fixed Number of relative substitutions rate: 4 Gamma distribution parameter: estimated Starting tree: BioNJ Tree topology: NNI	30
RAXML	Data: DNA Substitution model: GTRCAT Algorithm: High-climbing Runs: 100	30
MO-MA	Selection of parents: binary tournament popSize 100 (ps) crossover 80 % (cr) mutation 5 % (mu) maxGenerations (P) - primates_14: 58 - rbcL_55 - HIV2_72 - membracidae1_81 - mtDNA_186 - HIV1_192 - RDPII_218 - ZILLA_500	30

Fuente: Elaboración propia, (2017).

Tabla D.2: Tiempo promedio requerido por NSGA-II EM para realizar 100 generaciones en cada conjunto de datos.

Datos	Tiempo[s]	Datos	Tiempo[s]
<i>primates_14</i>	132	<i>mtDNA_186</i>	1088
<i>rbcL_55</i>	350	<i>HIV1_192</i>	764
<i>HIV2_72</i>	410	<i>RPDII_218</i>	2360
<i>membracidae1_81</i>	1000	<i>ZILLA_500</i>	2600

Fuente: Elaboración propia, (2017).

D.3 HERRAMIENTAS BASADAS EN OPTIMIZACIÓN DE MÚLTIPLES OBJETIVOS

PhyloMOEA (Cancino & Delbem, 2007), MO-Phyl (Santander-Jiménez & Vega-Rodríguez, 2015a), y NSGA-II EM son algoritmos basados en población con una estructura similar a NSGA-II (Santander-Jiménez & Vega-Rodríguez, 2013a) que no incluye búsqueda local. En estas aproximaciones el número de evaluaciones de parsimonia y verosimilitud depende del tamaño de la población (ps) y el número de generaciones. Los parámetros para estos algoritmos han sido publicados en la literatura relacionada (Tabla D.3). Por otro lado, trabajos previos han adaptado los parámetros de MO-FA (Santander-Jiménez & Vega-Rodríguez, 2013a) y MO-ABC (Santander-Jiménez & Vega-Rodríguez, 2013a) para efectuar una comparación con NSGA-II limitando el tiempo de ejecución. Siguiendo la misma estrategia, con el fin de efectuar una comparación justa, se ha limitado el número de generaciones de MO-MA acorde al tiempo requerido por NSGA-II EM para desarrollar 100 generaciones (Tabla D.2).

También se han efectuado experimentos para comparar el desempeño entre MO-MA y NSGA-II EM. La métrica de hipervolumen fue calculada considerando diez ejecuciones de 20000 segundos. Se han probado tres conjuntos de datos *primates_14*, *rbcL_55*, y *HIV1_192*. Según los resultados mostrados en la Figura D.2, MO-MA tiene el mayor hipervolumen en todos los conjuntos de datos estudiados.

Tabla D.3: Parámetros usados en MO-MA para comparación con herramientas basadas en optimización multi-objetivo.

Aproximación	Parámetros	Ejecuciones
MO-ABC	maxGenerations 100 swarmSize 100 mutation 5 % limit 15	31
MO-FA	maxGenerations 100 swarmSize 100 beta 1 gama 0.5 alfa 0.05	30
NSGA-II	maxGenerations 100 popSize 100 crossover 80 % mutation 5 % Selection of parents: binary tournament	31
NSGA-II EM	maxGenerations 100 popSize 100 crossover 80 % mutation 5 % Selection of parents: binary tournament	30
MO-Phyl	maxGenerations 100 popSize 100 crossoverProb 70 % mutationProb 5 %	30
PhyloMOEA	maxGenerations 100 popSize 100 crossover 80 % mutation 5 % Selection of parents: binary tournament	31
MO-Phylogenetics	max evaluations 6000 % rearrange search o PPN 0.5 % rearrange search by node: 0.3 max transversiones 20	20
NSGA-II EM	maxGenerations: 100 popSize: 100 crossover: 80 % mutation: 5 %	30
MO-MA	popSize 100 (ps) crossover 80 % (cr) mutation 5 % (mu) maxGenerations (P) - primates_14: 20 - rbcL_55: 15 - HIV2_72: 20 - membracidae1_81: 17 - mtDNA_186: 16 - HIV1_192: 20 - RDPII_218: 17 - ZILLA_500: 18	30

Fuente: Elaboración propia, (2017).

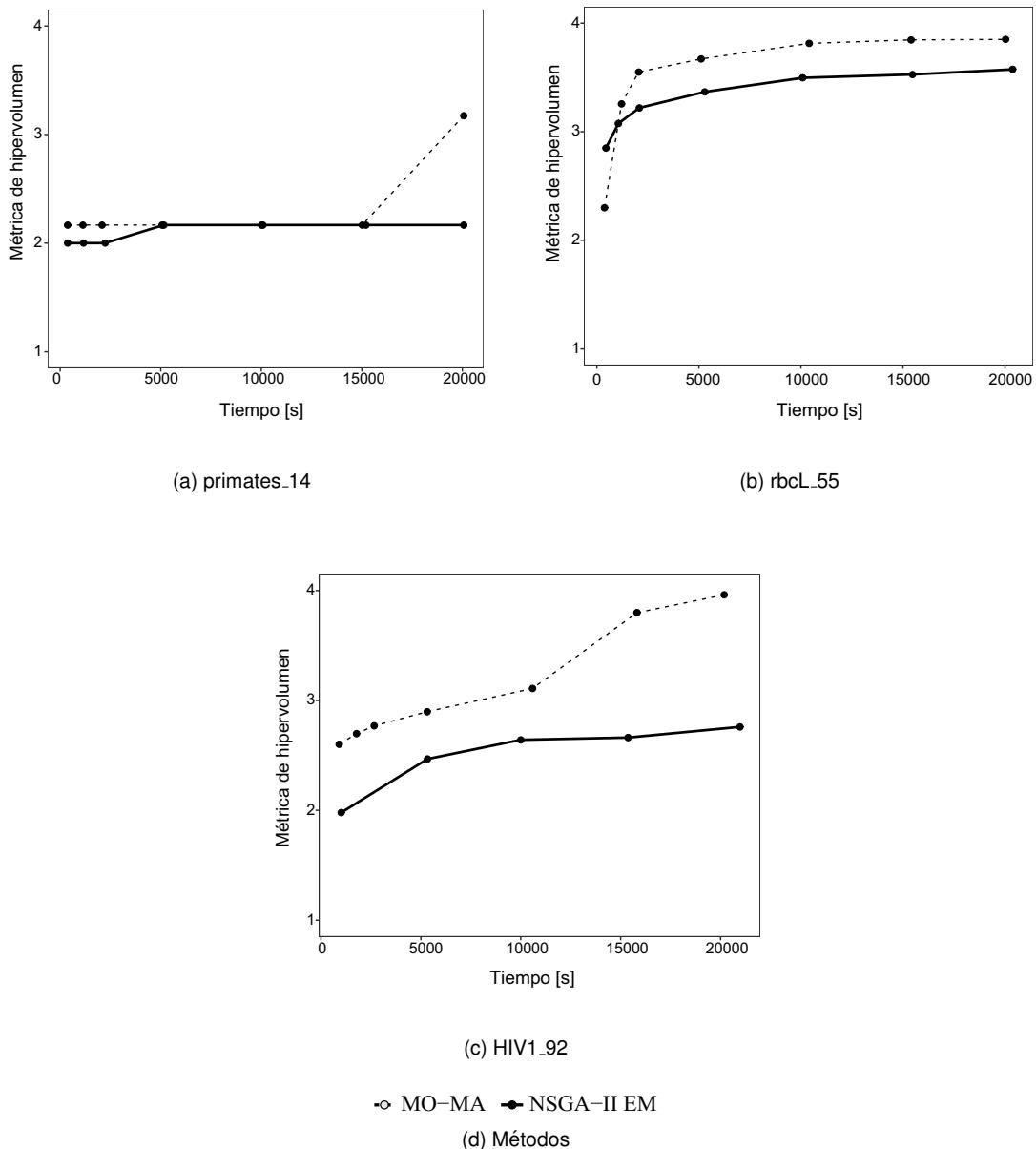


Figura D.2: Comparación entre el desempeño de MO-MA y NSGA-II EM.
Fuente: Elaboración propia, (2017).

D.4 CONDICIÓN DE BALANCE EN OPERADORES DE CRUZAMIENTO

La diferencia en la métrica de Robinson-Foulds entre hijos y cada padre para los diferentes operadores de cruzamiento son mostrados en la Figura D.3. Exceptuando por el método de consenso, todos los operadores tuvieron una diferencia estadística significante con cada padre. En estos casos, la distancia entre el padre podado fue considerablemente mayor que el padre reinsertado (condición de desbalance).

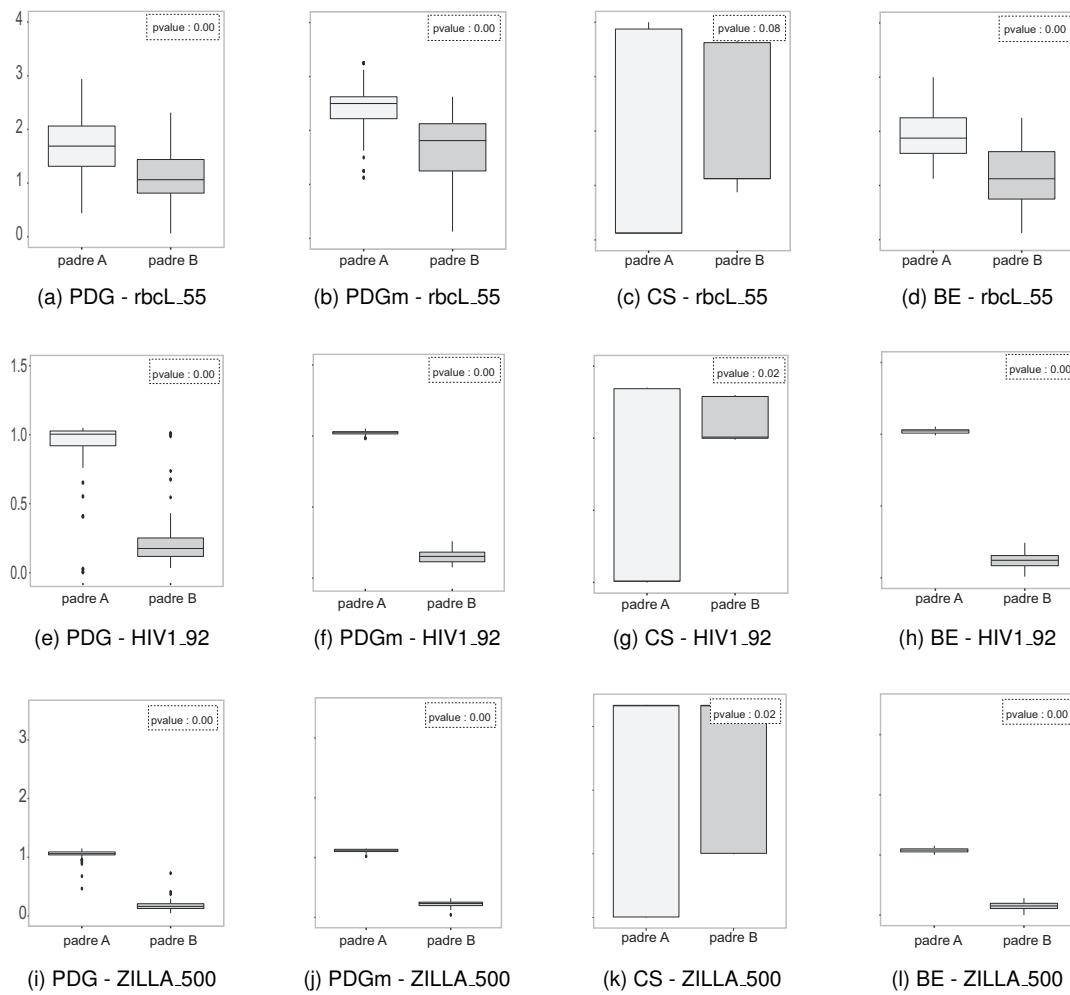


Figura D.3: Métrica de Robinson-Foulds normalizada entre hijos y padres para diferentes estrategias de cruzamiento y tres conjuntos de datos.

Fuente: Elaboración propia, (2017).

D.5 COMPARACIÓN DE OPERADORES DE CRUZAMIENTO Y MUTACIÓN

La Figura D.4 muestra la métrica de Robinson-Foulds después de la aplicación de diferentes operadores de cruzamiento (izquierda) y mutación (derecha). Las diferencias significativas han sido destacadas en negrita. En el caso de los operadores de cruzamiento, el

método CS tiene la mayor diferencia significativa entre padres y descendientes. En tanto entre los operadores de mutación, los descendientes de NNI tienen una menor distancia en relación a SPR y TBR. No se encontró diferencia significativa entre SPR y TBR.

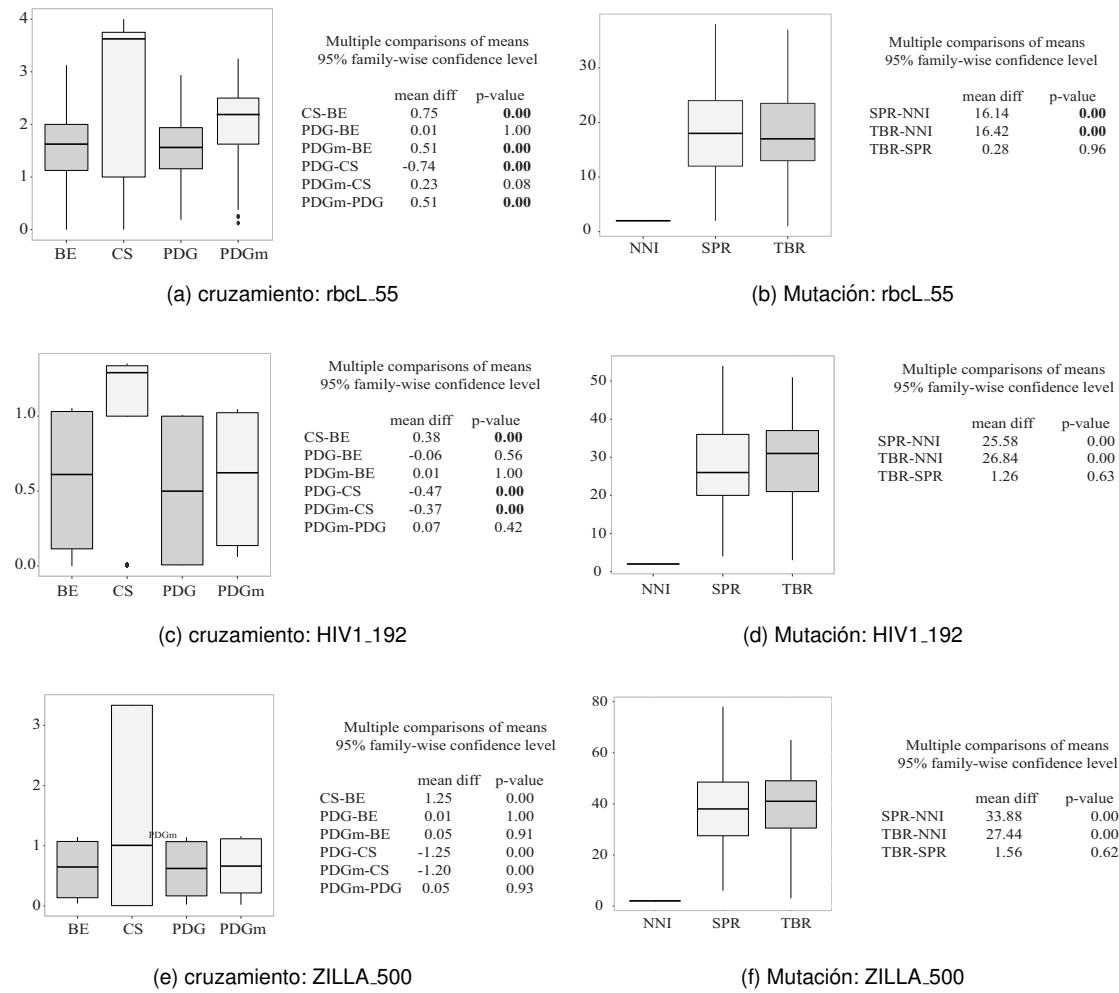


Figura D.4: Métrica de Robinson-Foulds normalizada entre hijos y padres para diferentes operadores genéticos y tres conjuntos de datos.

Fuente: Elaboración propia, (2017).

ANEXO E. MÉTRICAS DE RENDIMIENTO PARA ESTRATEGIAS MULTI-OBJETIVO

En la literatura se han propuesto múltiples estrategias evolutivas y bio-inspiradas basadas en optimización multi-objetivo, diseñadas para abordar diferentes problemas en múltiples áreas del conocimiento. Esto ha llevado a la necesidad de contar con métricas de calidad, que permitan comparar el desempeño de estas propuestas. Estas métricas evalúan diferentes aspectos de las soluciones: (1) **convergencia**, (2) **diversidad** y (3) **número**. En el transcurso de esta tesis se emplean tres métricas de calidad: (1) Hipervolumen, (2) Cobertura, y (3) Representatividad de las soluciones en la Frontera de Pareto.

E.1 HIPERVOLUMEN

El hipervolumen es la métrica de calidad más utilizada por la literatura (Riquelme et al., 2015), ya que combina los tres aspectos de las soluciones (convergencia, diversidad y número). Esta se calcula sumando el aporte en volumen que es cubierto por cada una de las soluciones que componen una Frontera de Pareto en el espacio objetivo (Figura E.1). Cuando se comparan múltiples estrategias, los valores objetivos deben estar normalizados entre 0 y 1. Valores altos de hipervolumen implican mejor calidad de soluciones.

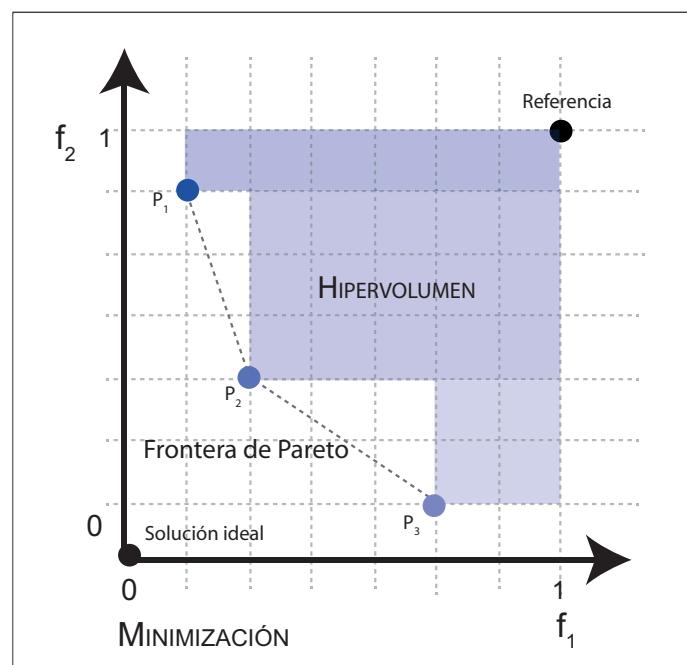


Figura E.1: Esquema métrica de hipervolumen.
Fuente: Elaboración propia, (2017).

E.2 REPRESENTATIVIDAD DE LAS SOLUCIONES EN LA FRONTERA DE PARETO

Esta métrica, también llamada Cobertura de frontera, tiene como objetivo representar el porcentaje de participación de las soluciones de un método particular, en relación a una Frontera Pareto de referencia. Para ello se cuantifica el número de soluciones comunes entre ellas (Figura E.2). Formalmente, sea A_{rep} una aproximación a la Frontera Pareto-óptima y B un conjunto de soluciones entregado por un método algorítmico, la cobertura de la frontera se formula como:

$$Part(B) = \frac{|\{b \in B / b \in A_{rep}\}|}{|A_{rep}|} \quad (\text{E.1})$$

Cuando $Part(B) = 1$, todas las soluciones del conjunto B pertenecen a la frontera Pareto-representativa. Si $Part(B) = 0$, ninguna solución de B es parte de la frontera Pareto-representativa.

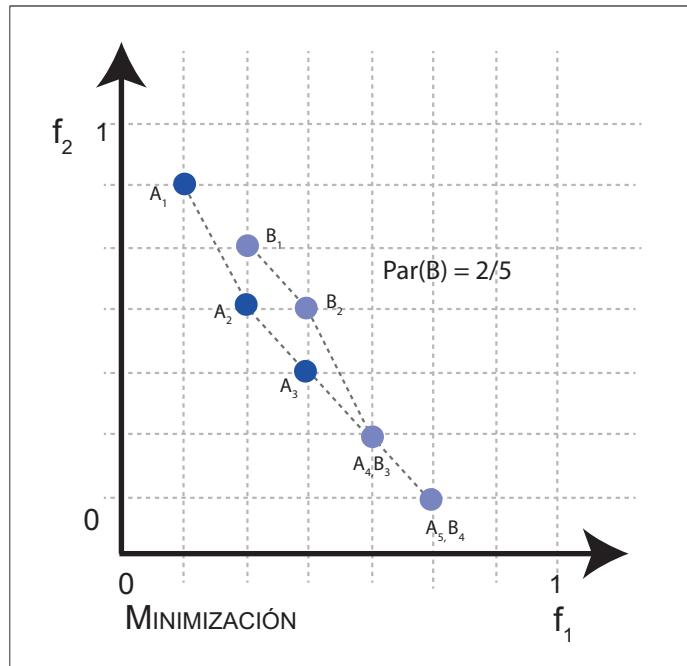


Figura E.2: Métrica de Representatividad de las soluciones en Frontera de Pareto.
Fuente: Elaboración propia, (2017).

E.3 COBERTURA

Esta métrica de cobertura compara dos Fronteras de Pareto, cuantificando el número de soluciones que son dominadas entre ellas (Figura E.3). Dado dos conjuntos de soluciones A y B . La métrica de cobertura puede ser formulada como:

$$Cob(A, B) = \frac{|\{b \in B / \exists a \in A : a \succ b\}|}{|B|} \quad (\text{E.2})$$

Cuando $Cob(A, B) = 1$, significa que todas las soluciones de B son dominadas por las soluciones de A . Por otro lado $Cob(A, B) = 0$ indica que ninguna de las soluciones de B son dominadas por alguna solución de A .

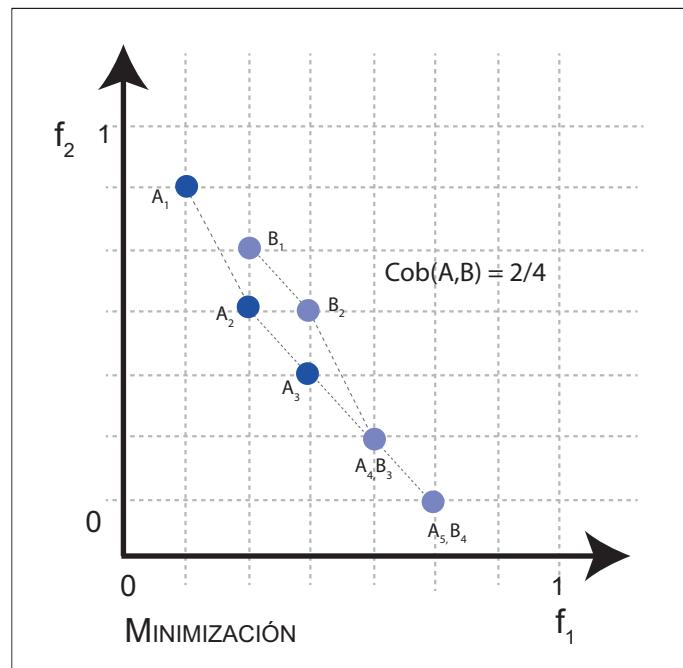


Figura E.3: Métrica de Cobertura.
Fuente: Elaboración propia, (2017).

ANEXO F. TOMADORES DE DECISIONES MULTI-OBJETIVO

Cuando se obtiene un conjunto de soluciones luego de aplicar alguna estrategia basada en optimización multi-objetivo, es necesario definir cuál de ellas resulta ser la más representativa de la Frontera de Pareto. Esta selección puede ser realizada empleando conocimiento experto, o mediante alguna estrategia para toma decisiones: Método del punto de referencia (*Reference Point Method*), Método de utilidad marginal (*Marginal Utility Method*) y métrica L2 (*L2-metric*) (Padhye & Deb, 2011).

- **Método del punto de referencia.** Esta estrategia selecciona aquella solución que se acerca en mayor medida a algún punto de referencia definido previamente en el espacio de soluciones (punto aspiracional). En esta investigación se ha empleado distancia Euclíadiana para medir la diferencia entre cada solución y los siguientes puntos aspiracionales:

$$RPmid = \left[\frac{\min(par) + \max(par)}{2}, \frac{\min(ver) + \max(ver)}{2} \right] \quad (F.1)$$

$$RPar = \left[\min(par), \frac{\min(ver) + \max(ver)}{2} \right] \quad (F.2)$$

$$RPlik = \left[\frac{\min(par) + \max(par)}{2}, \min(ver) \right] \quad (F.3)$$

- **Método de utilidad marginal.** Esta estrategia no requiere de información previa para efectuar la elección de una solución, ya que se basa en el concepto de afinidad (*AF*) entre soluciones en el espacio objetivo. El cálculo de afinidad para una vecindad considera tres puntos de la Frontera de Pareto: P_1 , P_0 y P_2 , en que ($\text{Par_}P_1 \leqslant \text{Par_}P_0 \leqslant \text{Par_}P_2$) y ($\text{Lik_}P_1 \geqslant \text{Lik_}P_0 \geqslant \text{Lik_}P_2$). Ejemplificare el cálculo de la afinidad del punto P_0 : Para efectuar la selección de P_1 y P_2 , se deben evaluar k vecinos hacia cada uno de los extremos de la Frontera de Pareto desde P_0 . Las soluciones más cercanas al centroide de cada una de estas dos regiones corresponderán a P_1 y P_2 . Para esta investigación se empleó arbitrariamente un valor de k igual a uno. Una vez que los punto P_1 y P_2 han sido definidos, se debe calcular la función de afinidad de cada uno de los puntos no extremos de la Frontera de Pareto:

$$AF = \max(w1, w2); w1 = \frac{\text{Ver_}P_0 - \text{Ver_}P_1}{\text{Par_}P_1 - \text{Ver_}P_0}; w2 = \frac{\text{Ver_}P_2 - \text{Ver_}P_0}{\text{Par_}P_0 - \text{Ver_}P_2} \quad (F.4)$$

La solución representativa corresponderá a la que minimize este parámetro.

- **Métrica L_2 .** Esta estrategia selecciona la solución que minimiza la distancia Euclíadiana (L_2) en relación a una solución hipotéticamente ideal. En esta investigación, debido a las características de los criterios de parsimonia y verosimilitud, el punto ideal fue definido como (0,0).

ANEXO G. TIEMPO DE EJECUCIÓN MO-PHYNET

La siguiente tabla detalla el tiempo de ejecución empleado por MO-PhyNet para inferir hipótesis considerando los diferentes conjunto de datos. Además se incluyen los días requeridos para estimar los cuatro parámetros del algoritmo considerando 31 ejecuciones.

Tabla G.1: Tiempo de ejecución MO-PhyNet.

Conjunto de datos	Tiempo por ejecución[seg]	Tiempo para parametrización [días])
<i>dengue_17</i>	93024	801
<i>fungi_20</i>	3134	27
<i>hamoglobin_20</i>	2661	23
<i>ureasas_126 (sec)</i>	16117	139
<i>ureasas_126 (árbol)</i>	4610	40
<i>flu_165</i>	178770	1539

Fuente: Elaboración propia, (2017).