Hibridación Jerárquica ACO-Firefly para la Selección de Características Genómicas en Cáncer de Mama

Jaime Lugo
Maestría en Ciencias de la Computación, Mención Inteligencia Artificial / Universidad Centroccidental Lisandro Alvarado
Barquisimeto, Venezuela
lugotorrealba@gmail.com
Zaidibeth Ramos
Maestría en Ciencias de la Computación, Mención Inteligencia Artificial / Universidad Centroccidental Lisandro Alvarado
Barquisimeto, Venezuela
zaidibethramos@gmail.com
Gustavo Rivero
Maestría en Ciencias de la Computación, Mención Inteligencia Artificial / Universidad Centroccidental Lisandro Alvarado
Cabudare, Venezuela
gustavoerivero12@gmail.com

Resumen (Abstract)

Se propone un enfoque computacional jerárquico para la selección de características en datos genómicos oncológicos de alta dimensionalidad, integrando Optimización por Colonia de Hormigas (ACO) y el Algoritmo de Luciérnagas (Firefly). Para mitigar la maldición de la dimensionalidad, el sistema opera mediante un pre-filtrado estadístico por Información Mutua seguido de una evaluación Wrapper. Dentro de esta arquitectura híbrida, ACO realiza una exploración global identificando subespacios prometedores, mientras Firefly ejecuta una explotación local para podar redundancias. Para neutralizar el sobreajuste evolutivo inherente a los microarreglos, se acotó matemáticamente el espacio de búsqueda mediante una penalización estricta sobre la cardinalidad del subconjunto y un clasificador evaluador LinearSVC fuertemente regularizado. La emergencia estigmérgica del enjambre fue validada cuantitativamente mediante el decaimiento de la Entropía de Shannon. Aplicado a la clasificación de subtipos de cáncer de mama, el modelo híbrido alcanzó una alta precisión predictiva sobre un conjunto de prueba aislado (F1-Macro = 0.9440), descubriendo un subconjunto óptimo de apenas 13 biomarcadores frente a los 200 del genoma pre-filtrado. Esto representa una compresión dimensional masiva del genoma, demostrando matemáticamente la superioridad de esta arquitectura jerárquica frente a las metaheurísticas aisladas y priorizando la viabilidad clínica mediante el Principio de Parsimonia.

Palabras clave

Inteligencia de Enjambre, Selección de Características, Estigmergia, Cáncer de Mama, Máquinas de Vectores de Soporte.

1. Introducción

El diagnóstico temprano y preciso en oncología constituye uno de los mayores desafíos de la biomedicina moderna. Con el avance de las tecnologías de secuenciación de próxima generación y los microarreglos de ADN, es posible analizar simultáneamente miles de expresiones génicas en una sola muestra tisular. Sin embargo, esto genera el problema algorítmico conocido como “maldición de la dimensionalidad”: conjuntos de datos con miles de características (genes) pero un número reducido de muestras (pacientes) [1]. Esta desproporción provoca sobreajuste computacional, ruido estadístico y modelos de difícil interpretación clínica.

Para mitigar este problema, la selección de características (feature selection) se vuelve un paso de preprocesamiento indispensable, cuyo objetivo es identificar el subconjunto mínimo de genes con máximo poder predictivo. Los métodos clásicos de filtrado estadístico o búsqueda exhaustiva resultan computacionalmente inviables pues el espacio de subconjuntos posibles crece de forma exponencial con la dimensionalidad: dada una matriz de m características, el número de subconjuntos candidatos es 2^m, lo que define un problema de complejidad NP-Difícil por reducción desde el problema de cobertura de conjuntos [2]. Por ello, las metaheurísticas basadas en inteligencia colectiva han emergido como alternativas robustas [3].

Investigaciones recientes han demostrado la viabilidad de algoritmos aislados como la Optimización por Colonia de Hormigas (ACO) [4], [5], [6] o el Algoritmo de la Luciérnaga (Firefly Algorithm, FA) [7], [8], para reducir la dimensionalidad en conjuntos de datos oncológicos, demostrando que la cooperación descentralizada entre agentes simples puede encontrar subconjuntos de características altamente predictivos. No obstante, la mayoría de los enfoques actuales aplican estos algoritmos de forma plana, estancándose prematuramente en óptimos locales o siendo ineficientes ante la extrema dimensionalidad.

El presente trabajo propone una arquitectura híbrida jerárquica que integra la exploración global de la Optimización por Colonia de Hormigas (ACO) con la explotación local del Algoritmo de Luciérnagas (Firefly), optimizada para datos genómicos de alta dimensionalidad. El resto de este artículo se organiza de la siguiente manera: la Sección 2 presenta el marco teórico y los trabajos relacionados; la Sección 3 detalla la formulación del problema y la función de aptitud; la Sección 4 describe la metodología del sistema propuesto; la Sección 5 expone el diseño experimental; y finalmente, las Secciones 6 y 7 presentan los resultados y la discusión académica, respectivamente.

2. Estado del Arte

2.1 Fundamentos de ACO y sus variantes

El paradigma de Optimización por Colonia de Hormigas fue introducido por Dorigo, Maniezzo y Colorni [5] mediante el algoritmo Ant System (AS), en el que múltiples agentes construyen soluciones incrementales guiados por feromonas artificiales y heurística local. Dorigo y Gambardella [6] propusieron posteriormente el Ant Colony System (ACS), que incorpora una regla de transición pseudo-aleatoria proporcional y una actualización de feromona online que recompensa exclusivamente a la mejor hormiga, mejorando significativamente la velocidad de convergencia. Una revisión comprensiva de las variantes de ACO y sus fundamentos teóricos puede encontrarse en Dorigo y Stützle [2]. En el contexto de selección de características, ACO destaca por su capacidad para explorar espacios discretos y de alta dimensionalidad mediante el uso de grafos y estigmergia, lo que lo hace especialmente adecuado para problemas de minimización de atributos en conjuntos de datos tabulares de alta dimensionalidad.

2.2 Metaheurísticas para selección de características en datos biomédicos de alta dimensionalidad

La aplicación de metaheurísticas para la selección de características en biomedicina y oncología ha experimentado un crecimiento sostenido, motivado por la incapacidad de los métodos estadísticos clásicos para escalar a espacios de miles de atributos. Revisiones sistemáticas recientes refuerzan esta tendencia. Piri et al. [1] analizaron 35 trabajos publicados entre 2009 y 2022 sobre híbridos evolutivos para selección de características, destacando que las combinaciones de ACO, FA, PSO y Algoritmos Genéticos (GA) dominan

por su capacidad de equilibrar exploración y explotación. Asimismo, Zendehbad et al. [3] revisaron aplicaciones de inteligencia de enjambre en ingeniería biomédica, resaltando la robustez de ACO y FFA en entornos ruidosos y de alta dimensionalidad.

Ali et al. [10], confirman que los enfoques híbridos (filtro + wrapper) basados en metaheurísticas bio-inspiradas logran reducciones de dimensionalidad extremas (a menudo superiores al 99%) manteniendo precisiones de clasificación de entre el 90 % y el 100 % en datos de expresión génica de cáncer de mama de alta dimensionalidad. La eficacia de estas metaheurísticas también se ha comprobado en bases de datos clínicas clásicas: Ashokkumar et al. [4] aplicaron ACO sobre el Wisconsin Breast Cancer Dataset (WBCD) y el Wisconsin Diagnostic Breast Cancer (WDBC), conjuntos de baja dimensionalidad basados en características citológicas y nucleares digitalizadas, respectivamente, alcanzando precisiones del 99.79% y 99.71% mediante votación suave de cinco clasificadores (kNN, SVM, LR, DT, RF). Estas cifras demuestran el potencial de ACO en espacios de búsqueda reducidos, si bien su extrapolación directa al dominio transcriptómico de extrema dimensionalidad requiere las precauciones metodológicas discutidas en la sección 2.4. Por su parte, Marovac et al. [11] compararon métodos estadísticos de filtrado versus wrappers basados en inteligencia de enjambre en datos biomédicos, concluyendo que los wrappers superan a los filtros tanto en precisión (0.97 vs. 0.91) como en reducción drástica de características.

A nivel de algoritmos individuales, el algoritmo Firefly (FA) ha demostrado superioridad frente a enfoques clásicos como PSO y algoritmos genéticos en tareas de selección de características, gracias a su capacidad para equilibrar exploración y explotación mientras minimiza el número de atributos, maximizando la capacidad de generalización del modelo al penalizar matemáticamente los subconjuntos extensos [7]. Esta ventaja se potencia mediante hibridación: Ibrahim et al. [8] propusieron el híbrido Firefly + Simulated Annealing (FASA), alcanzando una precisión promedio del 93,51 % en 14 datasets de cáncer, al mejorar significativamente la explotación local y superar al FFA aislado (68,63%).

2.3 ACO en clasificación oncológica y su hibridación con FFA

La aplicación de ACO sobre datasets de baja dimensionalidad con atributos clínicos o de imagen (como WBCD y WDBC [4]) muestra que la eficacia estigmérgica no se transfiere de forma directa al dominio transcriptómico, donde el espacio de búsqueda supera los 50.000 atributos y la relación muestra-característica ($n \ll m$) exacerba la presión sobre la diversidad del enjambre..

La hibridación directa ACO-FFA ha sido explorada de forma preliminar por Jona [12], quien propuso un esquema donde FFA actúa como motor de búsqueda local de ACO y alcanza 93,0% de precisión en la base de datos mini-MIAS (imágenes de mamografía). Aunque mejora la convergencia, este enfoque presenta dos limitaciones que nuestra investigación supera: (i) opera sobre características de imagen limitadas y no sobre perfiles de expresión génica, y (ii) emplea una arquitectura plana, donde ambos algoritmos comparten simultáneamente el mismo espacio de búsqueda completo sin división explícita entre exploración global y explotación local. En contraste, nuestra propuesta introduce una estructura estrictamente jerárquica en la que Firefly actúa como un operador de poda subordinado a la selección global de ACO, permitiendo una explotación local agresiva exclusivamente sobre los subespacios prometedores.

La arquitectura base de delegar la exploración global a ACO y el refinamiento local a FFA tiene sus raíces en la optimización matemática continua [13], y su eficacia ha sido validada recientemente en la resolución de problemas de ultra-complejidad en redes eléctricas a gran escala [14]. La presente investigación transpone esta arquitectura hacia un dominio de optimización combinatoria binaria NP-Difícil, adaptando la mecánica continua del enjambre a la selección discreta de biomarcadores oncológicos.

2.4 Brecha científica identificada

El análisis de la literatura revela que, a pesar de la eficacia demostrada por ACO y FFA de forma individual en datos biomédicos [1], [3], [4], [8], [9], [10], [11] y de un antecedente de hibridación ACO-FFA aplicado a imágenes [12], no se ha reportado una arquitectura jerárquica diseñada específicamente para la selección de características en datos transcriptómicos de alta dimensionalidad.

Esto se confirma en trabajos que utilizan el dataset GSE45827 (CuMiDa): Grisci et al. (2019) [9] aplicaron neuroevolución detectando 177 genes relevantes (82 validados como asociados a cáncer), logrando accuracies superiores al 95% pero sin reducción drástica de dimensionalidad.

Esta ausencia genera dos problemas fundamentales en el estado del arte: (i) los enfoques planos tratan cada gen como una entidad independiente sin división explícita entre exploración global y explotación local, limitando la capacidad de escapar de óptimos locales en espacios de búsqueda inmensos $\left(2^m \, con \, m > 50000\right)$; y (ii) los modelos de evaluación Wrapper actuales sufren de sobreajuste evolutivo al carecer de mecanismos explícitos que limiten la cardinalidad del subconjunto, seleccionando firmas genéticas extensas que memorizan el ruido estadístico de los microarreglos.

En consecuencia, se identifica una oportunidad científica significativa: el desarrollo de un enfoque de hibridación jerárquica ACO-Firefly que integre un control estricto sobre el tamaño del subconjunto en su función de aptitud. El presente trabajo aborda esta brecha proponiendo una arquitectura donde ACO realiza la exploración global, Firefly refina localmente los candidatos, y el entorno de evaluación penaliza matemáticamente la redundancia. El objetivo es minimizar drásticamente el número de biomarcadores seleccionados garantizando al mismo tiempo la máxima capacidad de generalización en la clasificación de subtipos moleculares de cáncer de mama.

3. Formulación del problema

El problema de selección de características se define formalmente como un problema de optimización combinatoria. Sea un conjunto de datos estructurados (microarreglo) representado por una matriz $D \in \Re^{nxm}$, donde n es el número de muestras (pacientes) y m es el número total de características (genes). Dada la alta dimensionalidad del dominio, se cumple la condición de $m \gg n$ [9]. Cada paciente está asociado a una etiqueta de clase $y_i \in \{1, 2, ..., C\}$, donde C es el número de clases o subtipos moleculares.

Dado que m resulta computacionalmente intratable para una búsqueda metaheurística directa (como se establece en la sección 1), se asume una etapa de pre-filtrado que reduce el espacio de m a K características ($K \ll m$). El criterio y la justificación de este pre-filtrado se detallan en la Sección 4.

Sobre el espacio reducido, el problema se modela siguiendo los fundamentos topológicos propuestos por Dorigo y Stützle [2], mediante un grafo de construcción completamente conectado $G = (V, E)$, donde cada nodo $v_i \in V$ representa una de las K características retenidas y $E = V \times V$ define las transiciones posibles entre ellas. Cada nodo posee dos atributos: una feromona τ_i que refleja la calidad histórica acumulada del gen en subconjuntos previos, y una información heurística η_i que cuantifica su relevancia predictiva individual respecto a la variable de clase.

El objetivo algorítmico es encontrar un vector binario de decisión $X = [x_1, x_2, \ldots, x_k]$, tal que $x_i \in \{0, 1\}$, donde $x_i = 1$ indica que la característica i es seleccionada y $x_i = 0$ indica su descarte.

Para evaluar la calidad de cada vector candidato X (o subconjunto seleccionado S), se emplea un enfoque *Wrapper*. Siguiendo el estándar propuesto por Emary et al. [7], la función objetivo (*Fitness*) se modela mediante una escalarización lineal que incorpora una penalización por cardinalidad condicionada por un límite máximo estricto::

$$Fitness(S) = \begin{cases} \alpha \times F1_{macro}(S) - \beta \times \dfrac{|S|}{K}, & si\ |S| \le L_{max} \\ 0, & en\ caso\ contrario \end{cases}$$

Donde:

- $F1_{macro}(S)$ es la precisión de clasificación macro-promediada, estimada por un clasificador wrapper LinearSVC fuertemente regularizado ($C = 0.05$) con validación cruzada estratificada.
- $|S|$ denota la cardinalidad del subconjunto (cantidad de características con $x_i = 1$).
- $\frac{|S|}{K}$ es el factor de penalización por redundancia dimensional.
- L_{max} es el límite máximo de dimensionalidad permitida.

Los hiperparámetros de la función de evaluación están sujetos a la restricción heurística α + β = 1, calibrando el equilibrio entre poder predictivo y parsimonia [7].

4. Metodología Propuesta

El enfoque propuesto se define como una arquitectura híbrida jerárquica de dos niveles diseñada para la selección de características en datos de expresión génica de ultra-alta dimensionalidad. La jerarquía se establece mediante un Nivel 1 de Exploración Global, donde la Optimización por Colonia de Hormigas (ACO) identifica regiones prometedoras del transcriptoma, y un Nivel 2 de Explotación Local, donde el Algoritmo de Luciérnagas (Firefly) realiza una poda estocástica para minimizar la redundancia.

Dado el régimen de extrema alta dimensionalidad del dominio ($n \ll m$, con m = 54.676 sondas, caracterizado en las Secciones 1 y 3), la arquitectura incorpora una Fase 0 de Pre-filtrado Supervisado por Información Mutua (MI) con estrategia One-vs-Rest (OvR). Esta elección no supervisada captura dependencias no lineales entre expresión génica y subtipo molecular comprimiendo el espacio de búsqueda a los K biomarcadores más informativos previo al inicio del enjambre.

4.A. Nivel 1: Exploración Global mediante ACO

Se emplea el grafo de construcción G = (V, E) definido en la sección anterior, donde cada nodo representa un gen del espacio pre-filtrado. Cada hormiga k construye un subconjunto candidato S^k seleccionando genes de forma incremental. Partiendo de un gen inicial aleatorio, la hormiga añade iterativamente genes al subconjunto hasta alcanzar un criterio de parada (número máximo de genes o convergencia de fitness). La probabilidad de que la hormiga k seleccione el gen j como siguiente característica a añadir se rige por:

$$p_j^k = \frac{[\tau_j]^\alpha [\eta_j]^\beta}{\Sigma_{l \in C^k} [\tau_l]^\alpha [\eta_l]^\beta}$$

Donde τ_j representa la cantidad de feromona acumulada en el gen j (reflejando su participación histórica en subconjuntos exitosos), η_j es la información heurística (relevancia predictiva de la variable), C^k es el conjunto de genes candidatos aún no seleccionados por la hormiga k, y los parámetros α y β controlan el peso relativo del rastro histórico frente a la heurística estadística y tasa de evaporación $\rho \in (0, 1]$. Tras cada iteración del enjambre, los rastros de feromona se actualizan dinámicamente incorporando una tasa de evaporación $\rho \in (0, 1]$ para evitar la convergencia prematura en óptimos locales.

Algoritmo 1. Descripción de alto nivel de la Fase de Exploración Global (ACO)

```
/* Inicialización de la iteración */
S_iterBest ← 0
F_iterBest ← 0
/* Construcción de soluciones por el enjambre */
Para k = 1 hasta N_A (número de hormigas) hacer
    S_k ← 0
    /* La hormiga k construye el subconjunto respetando el espacio delimitado*/
    Mientras |S_k| < L_max hacer
        Seleccionar el siguiente gen j probabilísticamente usando la ecuación de transición p_j^k
        S_k ← S_k ∪ {j}
    Fin Mientras
    /* Evaluación Wrapper de la hormiga k */
    F_k ← EvaluateWrapper(S_k, LinearSVC)
    /* Actualización del élite local */
    Si F_k > F_iterBest entonces
        S_iterBest ← S_k
        F_iterBest ← F_k
    Fin Si
Fin Para
Retornar S_iterBest /* Subconjunto dominante que será transferido a Firefly */
```

4.B. Nivel 2: Explotación Local mediante Firefly

A diferencia de los enfoques híbridos tradicionales, esta propuesta adopta una arquitectura memética de tipo lamarckiana para garantizar escalabilidad. Una vez que el enjambre ACO finaliza una iteración completa creando múltiples subconjuntos candidatos, únicamente el genotipo globalmente dominante de esa iteración es transferido al nivel 2. A partir de esa semilla, el algoritmo Firefly refina la selección a nivel local. Esta delegación de roles se fundamenta en la arquitectura conceptual propuesta por Rizk-Allah et al. [14], quienes demostraron matemáticamente que FA opera de manera óptima como un motor de búsqueda local para sintonizar y refinar las posiciones globales previamente descubiertas por las hormigas. Cada luciérnaga se inicializa como un vector binario sobre las características base, donde cada posición indica si el gen correspondiente está incluido (1) o excluido (0). La atracción entre las luciérnagas está regida por la función Fitness descrita en la sección anterior. El movimiento de una luciérnaga i hacia una solución más brillante j se define adaptando la ecuación de atracción estándar al dominio discreto:

$$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}\left(x_j - x_i\right) + \alpha(rand - 0.5)$$

Donde $\beta_0$es el coeficiente de atracción inicial (atracción máxima a distancia cero), γ es el coeficiente de absorción que controla la decaída de la atracción con la distancia, $r_{ij} = ||x_i - x_j||$es la distancia euclidiana entre las soluciones i y j en el espacio continuo, y $\alpha \in [0, 1]$ es el parámetro de aleatoriedad que controla la magnitud de la exploración estocástica.

Dado que el problema de selección de características requiere vectores binarios de alta dimensionalidad espacial, el resultado continuo de esta ecuación se discretiza mediante una función de transferencia sigmoide sesgada (biased sigmoid) orientada a maximizar la parsimonia:

$S(x) = \frac{1}{1+e^{-(x-\theta)}}$, donde $\theta = 1.5$ es el umbral heurístico o bias.

Esta formulación evita que la probabilidad de retención ronde el 50% de forma aleatoria, forzando inherentemente el vector hacia un estado escaso (sparse) y mitigando el incremento indiscriminado de la dimensionalidad en la firma genética.

Algoritmo 2. Descripción de alto nivel de la Fase de Explotación Local (Firefly)

```
/* Inicialización Lamarckiana a partir de la semilla de ACO */
Entrada: S_iterBest (Semilla élite de ACO)
Para i = 1 hasta N_F (número de luciérnagas) hacer
    x_i ← S_iterBest
    Aplicar mutación aleatoria estocástica leve a x_i para generar diversidad local
    I_i ← Fitness(x_i) /* Intensidad de luz inicial */
Fin Para
/* Ciclo de refinamiento */
Para t = 1 hasta maxIterF hacer
    Para i = 1 hasta N_F hacer
        Para j = 1 hasta N_F hacer
            Si I_j > I_i entonces
                /* Atracción en espacio continuo */
                x_temp ← x_i + β_o e^(−γ r_ij^2)(x_j − x_i) + α_t(rand − 0.5)
                /* Discretación forzando Parsimonia. Usando Sigmoide Sesgada S(x) */
                x_tempBin ← Discretizar(x_temp)
                /* Actualización individual si hay mejora */
                F_temp ← Fitness(x_tempBin)
                Si F_temp > I_i entonces
                    x_i ← x_tempBin
                    I_i ← F_temp
                Fin Si
            Fin Si
        Fin Para
    Fin Para
Fin Para
Retornar x_best /* Luciérnaga con mayor intensidad de luz final */
```

4.C. Mecanismo de Cooperación

El valor de aptitud (*fitness* o brillo) de la mejor luciérnaga se retroalimenta al nivel ACO como refuerzo de feromona. Específicamente, tras cada ciclo de optimización local de Firefly sobre el subconjunto seleccionado por una hormiga k, se incrementa la feromona depositada en cada gen i que pertenece al mejor subconjunto encontrado, proporcionalmente al fitness de dicha solución:

$$\Delta\tau_i^k = Fitness\,(S_{best}^k)$$

Donde S_{best}^k es el subconjunto de características con mayor F*itness* obtenido por la fase Firefly. Esta estigmergia cierra el bucle jerárquico: los genes que participan en soluciones de alta calidad reciben más feromonas, aumentando su probabilidad de ser seleccionados por futuras hormigas. De esta forma, la exploración global de ACO se dirige progresivamente hacia regiones del espacio de características que contienen genes predictivamente superiores.

Evaluación de la propiedad emergente del sistema

Además del desempeño predictivo del clasificador, se evaluará explícitamente la propiedad emergente del enjambre (autoorganización estigmérgica) mediante el decaimiento de la Entropía de Shannon aplicada sobre la distribución normalizada de feromonas. En cada iteración t se calcula:

$$H(t) = -\sum_{i=1}^{K} p_i Log_2 p_i\,(t)$$

Donde $p_i = \tau_i / \sum \tau$ es la probabilidad normalizada de la feromona en el gen i. Se espera un decaimiento progresivo y monotónico de $H(t)$ a lo largo de las iteraciones, indicando que el sistema pasa de un estado de alta incertidumbre (exploración uniforme) a un estado de baja entropía (concentración emergente de feromonas en un subconjunto reducido de biomarcadores). Este análisis cuantitativo se realiza independientemente del F1-Macro del clasificador y sirve como evidencia directa de la emergencia de orden colectivo sin control centralizado.

4.D. Evaluación y Wrapper

La función de aptitud definida en la sección 3 requiere un oráculo de generalización que estime la calidad predictiva de cada subconjunto candidato generado por el enjambre. Este rol lo cumple un clasificador LinearSVC con parámetro de regularización C = 0.05. La elección de un clasificador lineal fuertemente regularizado responde a dos criterios complementarios: primero, construye hiperplanos de decisión simples que capturan relaciones lineales en el espacio génico reducido sin sobreajustar al ruido estadístico de los microarreglos; segundo, su costo computacional es significativamente inferior al de kernels no lineales, lo que lo hace viable para el volumen de evaluaciones anidadas que requiere la arquitectura jerárquica.

Dado el desbalance inherente del dataset GSE45827, donde ciertos subtipos moleculares presentan una representación significativamente menor, el uso de la exactitud (accuracy) tradicional resultaría en un sesgo hacia la clase mayoritaria. Por ello, el evaluador emplea el F1-macro como métrica interna, que garantiza una penalización equitativa por los errores de Tipo I y Tipo II en todas las categorías de igual modo, independientemente de su frecuencia relativa.

Algoritmo 3. Hibridación Jerárquica ACO-Firefly para transcriptómica

Entrada: Matriz D (genoma), $maxIter$, límite esparzo L_{max}
Salida: Firma genómica óptima $S_{globalBest}$

/* Fase 0: Pre-filtrado supervisado */
$S_{MI} \leftarrow$ Extraer K genes usando Información Mutua OvR sobre D
Para cada gen $i \in S_{MI}$ hacer
    $\tau_i(0) \leftarrow \tau_0$ /* Inicializar matriz de feromonas */
Fin Para

/* Fase metaheurística híbrida */
$S_{globalBest} \leftarrow 0,\ F_{globalBest} \leftarrow 0$
Para $t = 1$ hasta $maxIter$ hacer
    /* Nivel 1: Exploración Global */
    $S_{iterBest} \leftarrow$ Ejecutar **Algoritmo 1** (ACO)

    /* Nivel 2: Explotación Local */
    $S_{refinado} \leftarrow$ Ejecutar **Algoritmo 2** (Firefly usando $S_{iterBest}$ como entrada)
    $F_{refinado} \leftarrow Fitness(S_{refinado})$

    /* Mecanismo de cooperación (Estigmergia discreta) */
    Para cada gen $i \in S_{MI}$ hacer
        $\tau_i(t) \leftarrow (1 - \rho) \times \tau_i(t)$ /* Evaporación general */
    Fin Para
    Para cada gen $i \in S_{refinado}$ hacer
        $\tau_i(t) \leftarrow \tau_i(t) + F_{refinado}$ /* Depósito de feromona elitista */
    Fin Para

    /* Actualización global */
    Si $F_{refinado} > F_{globalBest}$ entonces
        $S_{globalBest} \leftarrow S_{refinado}$
        $F_{globalBest} \leftarrow F_{refinado}$
    Fin Si
Fin Para
Devolver $S_{globalBest}$

5. Diseño experimental

5.A. Dataset

Se empleó el conjunto de datos CuMiDa (*Curated Microarray Database*) correspondiente al perfil transcriptómico de cáncer de mama (GEO: GSE45827) [9]. Este conjunto de datos, disponible también en el repositorio Kaggle, contiene 151 muestras biológicas descritas por 54.676 sondas de expresión (plataforma Affymetrix HG-U133 Plus 2.0). La distribución de subtipos moleculares presenta un evidente desbalance: Basal (41), HER2+ (30), Luminal B (30), Luminal A (29), Línea celular (14) y Normal (7). Es necesario señalar que no se evaluó la matriz de intensidades en crudo; el dataset proviene de un proceso de curación

estandarizado donde las señales ópticas del chip fueron corregidas y transformadas logarítmicamente mediante el algoritmo RMA (*Robust Multi-array Average*). Esta pre-normalización mitiga el ruido de fondo del hardware y asegura que la varianza modelada corresponda a la expresión biológica real.

5.B. Pre-filtrado

Dada la intratabilidad computacional de operar directamente sobre 54.676 dimensiones con un método Wrapper, se aplicó un pre-filtrado estadístico basado en Información Mutua (IM) con estrategia One-vs-Rest (OvR). Para cada una de las seis clases, se calculó la IM entre cada gen y una etiqueta binaria (clase actual vs. resto), obteniéndose una matriz de puntajes. El puntaje final de cada gen correspondió al máximo a lo largo de las seis comparaciones OvR. Se retuvieron los $K = 200$ genes con mayor puntaje de IM, constituyendo así el espacio de búsqueda para las metaheurísticas. Cabe señalar que este cálculo se realizó exclusivamente sobre la partición de entrenamiento post-divisióncon el conjunto de prueba ya aislado. Esta separación garantiza que ninguna información estadística del conjunto de evaluación final influya sobre la selección de características, asegurando que la métrica F1-macro reportada sobre prueba sea una estimación honesta de la capacidad de generalización del modelo.

5.C. Partición de datos

El conjunto de datos se dividió inicialmente en conjuntos de entrenamiento (80%) y prueba (20%) mediante muestreo estratificado, garantizando que cada partición preservase la proporción original de los seis subtipos. Se utilizó una semilla aleatoria fija ($randomState = 0$) para asegurar la reproducibilidad. Dado que la clase "Normal" cuenta con solo 7 muestras, se requiere un mecanismo de estabilización para permitir la validación cruzada. Este proceso de duplicación estocástica de clases minoritarias (hasta alcanzar 6 ejemplos por clase) se aplicó exclusiva y estrictamente sobre el conjunto de entrenamiento, con el conjunto de prueba completamente aislado desde el momento de la división inicial (véase la sección 5.B).

5.D. Hiperparámetros del modelo híbrido

La parametrización de algoritmos bio-inspirados en espacios de alta dimensionalidad es un desafío crítico. Por ello, la sintonización de los hiper parámetros no se realiza de manera arbitraria, sino mediante un enfoque de calibración fundamentado en la literatura base de Inteligencia de Enjambres [2], [5], adaptando empíricamente a las restricciones de parsimonia del dominio genómico oncológico. La Tabla 1 resume los hiperparámetros utilizados en la ejecución experimental, organizados por componente algorítmico. Los valores corresponden a la configuración default.yaml del repositorio.

1. **Parámetros de la Colonia de Hormigas (ACO):** Siguiendo las directrices teóricas de Dorigo y Stützle [2] para problemas NP-Difíciles, se establece un equilibrio simétrico entre la influencia del rastro histórico y la heurística estadística $\left(\alpha_{ACO} = 1.0,\ \beta_{ACO} = 1.0\right)$, permitiendo que la Información Mutua guíe el inicio de la búsqueda mientras el aprendizaje estigmérgico domina las iteraciones tardías. La tasa de evaporación se fijó en un valor conservador ($\rho = 0.15$) para evitar la pérdida prematura de memoria colectiva en el vasto espacio de 200 genes, asegurando una exploración sostenida.
2. **Parámetro de Luciérnagas (Firefly):** Basado en las adaptaciones discretas de FA para selección de características propuestas por Emary et al. [7], la absorción lumínica ($\gamma = 1.0$) y la atracción base $\left(\beta_0 = 1.0\right)$ se configuraron para forzar una

convergencia local rápida. Dado que FA opera aquí como un refinador quirúrgico de la hormiga élite y no como explorador global, el componente estocástico se inicializó bajo $\left(\alpha_{rand} = 0.05\right)$ con un decaimiento severo $\left(\alpha_{decay} = 0.95\right)$, garantizando que la luciérnaga pode características redundantes sin destruir la topología del subespacio prometedor entregado por ACO.

3. **Restricciones Dimensionales y de Aptitud:** Los límites de construcción heurística $\left(g_{min} = 5,\ g_{max} = 18\right)$ y los pesos de la función de aptitud ($\alpha = 0.90,\ \beta = 0.10$) fueron determinados mediante experimentación empírica y respaldados por literatura biomédica [11]. Una ponderación de $\beta = 0.10$ castiga matemáticamente la hipertrofia del genotipo, asegurando que el enjambre descarte firmas genómicas clínicamente inviables y favorezca agrupaciones extremadamente parsimoniosas, acordes con la realidad económica de las pruebas PCR múltiplex.

Tabla 1. Hiperparámetros experimentales del modelo híbrido ACO-Firefly.

Componente	Parámetro	Símbolo	Valor	Rol
Datos	Genes pre-filtrados	K	200	Dimensión del espacio de búsqueda
ACO	Número de hormigas	n_ants	10	Soluciones construidas por iteración
ACO	Iteraciones máximas	T_{ACO}	25	Ciclos del bucle externo ACO
ACO	Peso feromona	α_{ACO}	1.0	Influencia de la feromona en la transición probabilística
ACO	Peso heurística	β_{ACO}	1.0	Influencia de la información heurística (MI)
ACO	Tasa de evaporación	ρ	0.15	Fracción de feromona evaporada por iteración
ACO	Feromona inicial	τ_{init}	1.0	Valor inicial uniforme del vector de feromona
ACO	Feromona mínima	τ_{min}	0.05	Cota inferior (evita estancamiento)
ACO	Feromona máxima	τ_{max}	10.0	Cota superior (evita dominancia prematura)
ACO	Genes mínimos por subconjunto	g_{min}	5	Tamaño mínimo de construcción de la hormiga
ACO	Genes máximos por subconjunto	g_{max}	18	Tamaño máximo de construcción de la hormiga
Firefly	Número de luciérnagas	n_ff	5	Tamaño de la población del FA
Firefly	Iteraciones FA	T_{FA}	3	Ciclos de refinamiento local por cada solución ACO
Firefly	Atracción base	β_0	1.0	Intensidad de atracción a distancia cero
Firefly	Absorción lumínica	γ	1.0	Tasa de decaimiento de la atracción con la distancia
Firefly	Escala de aleatoriedad	α_{rand}	0.05	Magnitud del componente estocástico
Firefly	Decay de aleatoriedad	α_{decay}	0.95	Factor multiplicativo de reducción por iteración
Fitness	Peso de clasificación	α	0.90	Ponderación del F1-Macro en la función objetivo
Fitness	Peso de parsimonia	β	0.10	Penalización proporcional al tamaño del subconjunto
General	Semilla aleatoria	–	0	Reproducibilidad de todos los procesos

estocásticos

Se fijó un máximo de 25 iteraciones ACO como criterio de parada, ya que corridas más largas producen sobreajuste evolutivo: el fitness interno (validación cruzada en entrenamiento) continúa aumentando, pero el F1-Macro en el conjunto de prueba aislado comienza a degradarse.

5.E. Métodos de comparación (baselines)

Para contextualizar el rendimiento del enfoque híbrido, se implementaron cuatro métodos de referencia:

1. **Todos los genes (All Genes):** Utiliza los $K = 200$ genes pre-filtrados sin ningún proceso de selección adicional. El clasificador LinearSVC se entrena con la totalidad de las características disponibles.
2. **Selección Aleatoria (Random):** Genera subconjuntos aleatorios cuyo tamaño se muestrea uniformemente y evalúa cada uno con la función de aptitud Wrapper. Este baseline cuantifica el beneficio marginal de la búsqueda guiada respecto al azar.
3. **ACO sin refinamiento (ACO-Only):** Ejecuta el ciclo completo de construcción de subconjuntos y actualización de feromona descrito en la Sección 4, pero omite la optimización local con Firefly.
4. **Firefly sin guía ACO (Firefly-Only):** Inicializa cada corrida del Algoritmo de Luciérnagas con un subconjunto aleatorio. De este modo se evalúa la capacidad de refinamiento local del FA en ausencia de la señal estigmérgica global.

5.F. Esquema de validación

Para garantizar la integridad de los resultados, la evaluación se ejecuta en dos etapas disjuntas. Durante la búsqueda metaheurística, la función de aptitud definida en la seccion 3 se estima mediante el promedio de una validación cruzada estratificada de cinco pliegues aplicada exclusivamente sobre la partición de entrenamiento. Este procedimiento asegura que la selección de características sea robusta frente a variaciones en la muestra y evita que el enjambre sobreajuste a una partición particular.

Una vez obtenida la firma genómica óptima al finalizar la optimización jerárquica ACO-Firefly, la evaluación final se realiza mediante un pase predictivo singular (single-pass): el modelo se entrena con la totalidad de la partición de entrenamiento y se valida sobre el conjunto de prueba aislado, que representa el 20% de las muestras originales y no ha intervenido en ninguna fase de optimización ni de ajuste de hiperparámetros. Esta metodología proporciona una medida objetiva de la capacidad de generalización del modelo en un escenario clínico simulado.

5.G. Entorno computacional

El desarrollo, la validación y la ejecución de los experimentos se llevaron a cabo en un entorno basado en Python 3, utilizando sistemas operativos Windows 11 y macOS para verificar la portabilidad de la arquitectura propuesta. La gestión de las dependencias de software se realizó mediante entornos virtuales aislados, garantizando la consistencia de las versiones de las librerías empleadas. Para la manipulación y vectorización de la matriz transcriptómica se utilizaron las librerías NumPy y Pandas, mientras que el pre-filtrado por Información Mutua y el modelado del evaluador LinearSVC se fundamentaron en el ecosistema de Scikit-learn y SciPy. La representación y gestión de la topología de grafos requerida para el proceso de construcción de soluciones en la fase de colonia de hormigas se llevó a cabo mediante NetworkX.

Con el objetivo de optimizar la eficiencia temporal frente a la complejidad del modelo jerárquico, la evaluación de las soluciones se ejecutó de forma asíncrona mediante la librería Joblib. Esta implementación permitió la paralelización de la validación cruzada a través de la

totalidad de los núcleos de procesamiento lógico disponibles en el hardware, mitigando de forma efectiva el costo computacional de las evaluaciones anidadas. Asimismo, la ingesta del conjunto de datos desde el repositorio fuente se automatizó mediante KaggleHub, asegurando un acceso reproducible a las muestras originales. Finalmente, la orquestación de los hiperparámetros y el análisis cuantitativo de la convergencia y la entropía se realizaron mediante PyYAML, Matplotlib y Seaborn, conformando un flujo de trabajo documentado bajo los estándares de la computación científica.

6. Resultados

6.A. Desempeño predictivo y reducción dimensional

La Tabla 2 y la Figura 1 presentan el rendimiento de los cinco métodos evaluados sobre el conjunto de datos de prueba aislada (Test Set).

Tabla 2. Comparación de métodos sobre el conjunto de prueba (K = 200 genes pre-filtrados)

Método	F1-Macro (Test)	Genes retenidos	Tiempo de búsqueda (s)
Todos los genes (All Genes)	0.9720	200	0.0
Selección aleatoria (Random)	0.9111	12	8.6
ACO puro (Sin refinamiento)	0.8552	15	3.1
Firefly puro (Sin guía ACO)	0.8892	22	0.8
Híbrido ACO-Firefly	0.9440	13	4.9

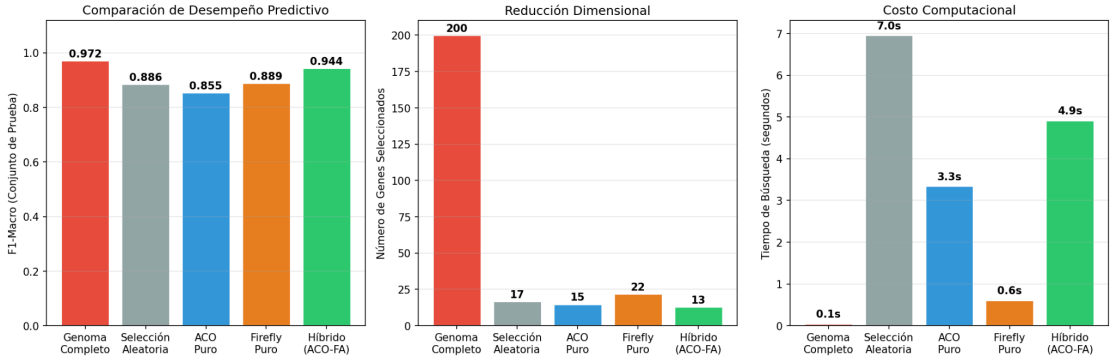

Figura 1. Panel comparativo de desempeño predictivo, reducción dimensional y costo computacional entre el modelo propuesto y las líneas base.

La Tabla 2 y la Figura 1 sintetizan el rendimiento comparativo de los cinco métodos sobre el conjunto de prueba aislado; el análisis de las implicaciones algorítmicas y clínicas de estos valores se desarrolla en la Sección 7.

6.B. Dinámica de convergencia y emergencia estigmérgica

Como se observa en la Figura 2, aunque el sistema identificó óptimos locales fuertes en etapas tempranas, la exploración sostenida permitió descubrir el óptimo global definitivo en la iteración 25, donde la integración de la hormiga élite con el refinamiento de Firefly alcanzó una aptitud interna máxima ($f(S) = 0.8386$).

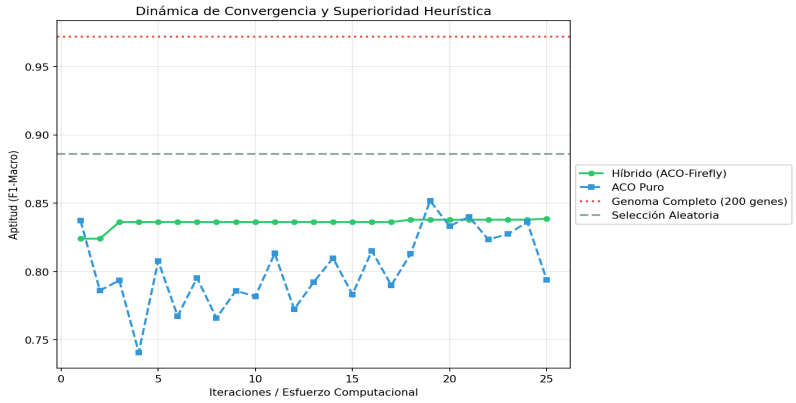

Figura 2. Dinámica de convergencia del modelo híbrido.

El decaimiento de la Entropía de Shannon sobre la distribución normalizada de feromonas, registrado durante las 25 iteraciones (Figura 3), muestra un descenso de 7.5942 bits a 6.6026 bits, una reducción del 13% en la incertidumbre colectiva del enjambre.

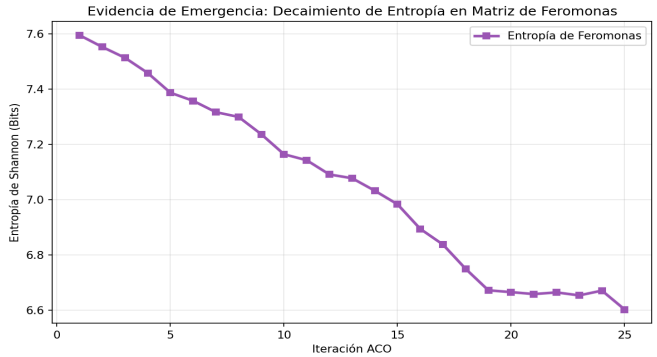

Figura 3. Decaimiento de la Entropía de Shannon.

6.C. Validación clínica y separabilidad espacial

Para validar la pertinencia biológica de los 13 genes seleccionados, se evalúa el comportamiento diagnóstico detallado del clasificador SVM final. La Matriz de Confusión (Figura 4) demuestra una alta tasa de verdaderos positivos en la discriminación de los 6 subtipos moleculares.

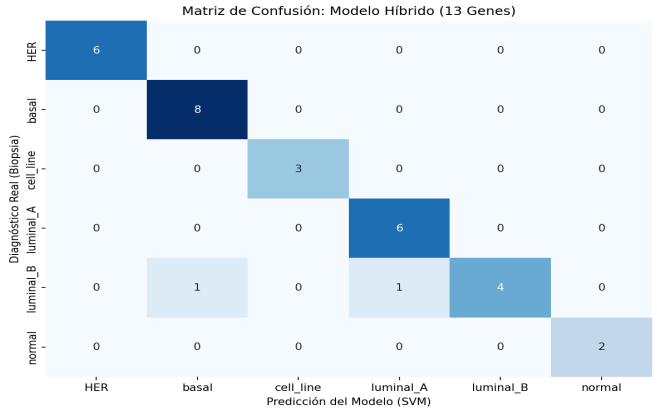

Figura 4. Matriz de confusión del modelo SVM lineal.

Esta asertividad clínica se corrobora espacialmente mediante un análisis de componentes principales (PCA). Al proyectar las muestras del conjunto de datos de prueba en un plano

bidimensional empleando únicamente los 13 genes de la firma híbrida (Figura 5), se observa una clara separabilidad topológica de los subtipos de cáncer.

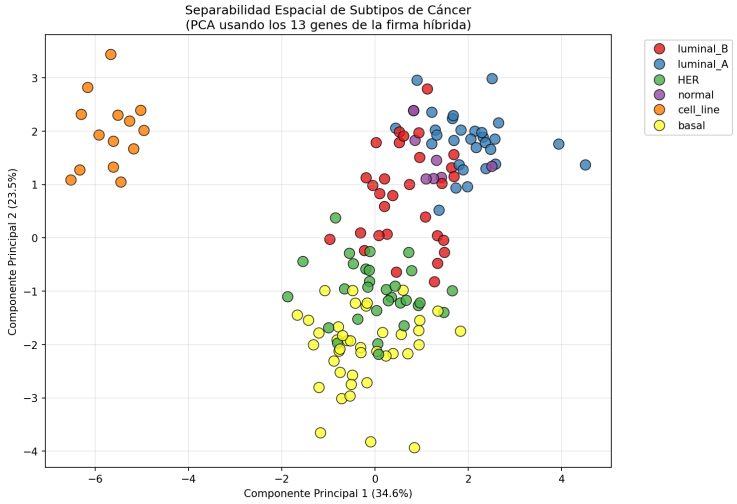

Figura 5. Proyección del Análisis de Componentes Principales (2D).

Finalmente, la Figura 6 detalla la relevancia biológica individual de los 13 genes retenidos, clasificados según su Puntuación de Información Mutua (OvR-MI). La distribución de puntuaciones MI-OvR sobre los 13 genes retenidos, representada en la Figura 6, muestra una concentración de puntajes por encima del umbral mediano del espacio pre-filtrado, con ERBB2 (HER2) y KDM5A registrando las mayores relevancias discriminativas individuales.

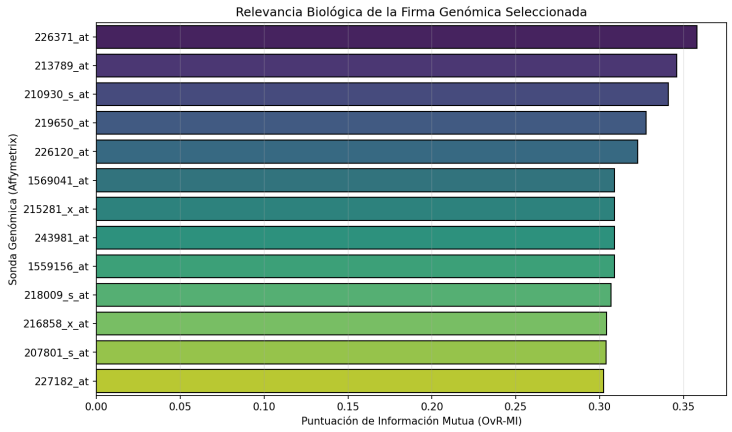

Figura 6. Relevancia biológica de la firma genómica seleccionada.

6.D. Firma genómica seleccionada

La ejecución de la hibridación jerárquica culminó con la selección de un subconjunto estable compuesto exactamente por 13 sondas correspondientes a la plataforma Affymetrix HG-U133 Plus 2.0. A continuación, en la Tabla 3, se detallan los identificadores técnicos extraídos por el modelo y su naturaleza de hibridación.

Tabla 3. Sondas seleccionadas por el modelo híbrido ACO-Firefly y sus símbolos génicos correspondientes

ID Sonda (Affymetrix)	Símbolo del Gen	ID Sonda (Affymetrix)	Símbolo del Gen
226371_at	KDM5A	243981_at	STK4

213789_at	Desconocido (No codificante)		1569041_at	Desconocido (No codificante)	
210930_s_at	ERBB2 (HER2)		218009_s_at	PRC1	
219650_at	ERCC6L		216858_x_at	Desconocido (No codificante)	
226120_at	TTC8		207801_s_at	RNF10	
215281_x_at	POGZ		227182_at	SUSD3	
1559156_at	Desconocido (No codificante)				

7. Discusión

7.A. Superioridad del enfoque jerárquico

La ventaja del híbrido sobre las metaheurísticas aisladas no es un accidente numérico: es el producto directo de la división de roles que define la arquitectura. ACO puro, sin refinamiento local, construye subconjuntos guiados por la señal estigmérgica acumulada, pero carece de un mecanismo que elimine genes redundantes dentro del subconjunto ya seleccionado: su exploración es global pero su explotación intrasubconjunto es nula. El resultado observado, un F1-Macro inferior con un subconjunto de 15 genes, refleja que la colonia converge hacia regiones prometedoras del espacio génico pero no depura la redundancia dentro de ellas.

Firefly puro, por el contrario, es un refinador eficiente pero ciego: inicializado desde subconjuntos aleatorios sin señal histórica, no tiene información sobre qué regiones del espacio génico concentran genes predictivamente coherentes. Su exploración estocástica termina reteniendo más genes (22 en promedio) para compensar la falta de orientación global, lo que penaliza su fitness según el término de parsimonia definido en la sección 3.

La arquitectura jerárquica resuelve ambas limitaciones simultáneamente. ACO provee la orientación global mediante la convergencia de feromonas hacia genes de alta calidad histórica, validada cuantitativamente por el decaimiento de la Entropía de Shannon documentado en la sección 6.B. Firefly recibe esa orientación como semilla lamarckiana y la depura quirúrgicamente mediante la función de transferencia sigmoide sesgada ($\theta = 1.5$), eliminando los genes cuya contribución marginal no supera el costo dimensional impuesto por el parámetro β. La interacción entre ambos niveles produce una firma de 13 genes que es, simultáneamente, más pequeña que la de Firefly puro y más precisa que la de ACO puro.

En contraste, el modelo de referencia sin selección («Todos los genes») opera sobre 200 características con un clasificador LinealSVC en un régimen donde la dimensionalidad es alta relativa al número de muestras de entrenamiento (~120). El diferencial del 5.6% en F1-Macro respecto al híbrido debe interpretarse con cautela: a diferencia del modelo híbrido, el baseline no ha sido sometido a ninguna presión de parsimonia y su ventaja predictiva viene acompañada de un panel de biomarcadores clínicamente inviable por costo de secuenciación.

7.B. Comparación con la literatura

A diferencia de los estudios citados [4] [12], que reportan altas precisiones (>93%) operando sobre atributos clínicos o características extraídas de imágenes médicas con dimensionalidad moderada, este trabajo resuelve un espacio transcriptómico masivo de 54.676 características. Al realizar una comparación directa sobre este mismo dominio (GSE45827), Grisci et al. [9] alcanzaron precisiones superiores al 95% mediante neuroevolución, pero requirieron una firma extensa de 177 características. En contraste, el modelo híbrido propuesto logra un

desempeño altamente competitivo, con un F1-Macro de 0.9440, comprimiendo el panel a tan solo 13 características. Esta reducción extrema no destruyó los límites de decisión entre clases, lo que confirma que el algoritmo logra extraer características con dependencias no lineales profundas. Estos resultados validan empíricamente que la poda estocástica lamarckiana del nivel Firefly erradica la redundancia transcriptómica sin comprometer la capacidad de discriminación clínica del clasificador final, aun cuando el panel se comprime a un orden de magnitud inferior al reportado por el único trabajo comparable en el mismo dominio [9].

7.C. Rol del pre-filtrado

El pre-filtrado por Información Mutua OvR es un paso ineludible para la viabilidad computacional, reduciendo el espacio de búsqueda en un 99.63%, comprimiendo el transcriptoma de 54 676 a 200 genes candidatos y haciendo posible que el modelo converja sin caer en explosión combinatoria.

Desde el punto de vista estadístico, la IM con estrategia OvR captura dependencias no lineales entre los genes y los subtipos moleculares que los métodos univariados de varianza o t-test no pueden detectar, ya que la información mutua no asume distribución paramétrica alguna. Este pre-filtrado actúa como un primer nivel de criba biológica que preserva genes con relación genuina con la variable de clase y descarta expresiones génicas constitutivas sin valor discriminativo. Los procedimientos que garantizan la integridad estadística de este paso se describen en la sección 5.B.

7.D. Análisis del trade-off en la función de aptitud

La función de aptitud definida en la sección 3 establece un equilibrio crítico entre el poder discriminativo (ponderado por $\alpha = 0.90$) y el Principio de Parsimonia (penalizado por $\beta = 0.10$). Los resultados confirman que este equilibrio es clínicamente adecuado: la presión selectiva impuesta por el término β fue suficiente para reducir el espacio post-filtro al 9.5% (13 de 200 genes), sin destruir la separabilidad de los subtipos tumorales evidenciada en la proyección PCA de la sección 6.C.

El valor β = 0.10 fue determinado empíricamente, pero su efecto puede interpretarse en términos de umbral de relevancia marginal: un gen que contribuye a F1_macro con menos del 10% de lo que cuesta incorporar una característica adicional (medido como la fracción |S|/K) es descartado. Este criterio opera de forma análoga al Principio de Parsimonia en modelado estadístico: privilegia la hipótesis más simple que explique los datos observados con suficiente fidelidad. Omitir esta restricción equivaldría a permitir al enjambre maximizar la métrica de clasificación sin costo dimensional, lo que en el dominio de los microarreglos conduce invariablemente a firmas de decenas de genes que memorizan el ruido estadístico de las 120 muestras de entrenamiento.

Una dirección de trabajo futuro de alto valor sería reformular este balance como un problema de optimización multi-objetivo explícito (véase §7.G), generando una frontera de Pareto que permita al clínico seleccionar el punto de operación deseado según las restricciones económicas de la prueba PCR disponible.

7.E. Análisis biológico de la firma genómica (nueva sección)

La firma de 13 sondas seleccionada por el modelo opera sobre identificadores técnicos de hibridación (Probe Set IDs) de la plataforma Affymetrix HG-U133 Plus 2.0, no directamente sobre genes conceptuales. Esta distinción es metodológicamente relevante: una sonda no es necesariamente un gen, y su interpretación biológica requiere anotación bioinformática y verificación de especificidad de hibridación antes de cualquier implementación clínica.

La presencia de ERBB2 (HER2) entre los 13 biomarcadores seleccionados constituye el hallazgo de validación biológica más robusto de la firma: ERBB2 es el marcador canónico del subtipo HER2+ y su sobreexpresión define diagnósticamente este fenotipo molecular. Que el enjambre lo seleccione de forma autónoma, sin conocimiento a priori de su relevancia clínica, es evidencia de que la señal estigmérgica converge hacia genes con genuino poder discriminativo.

Dos sondas de la firma presentan sufijos de hibridación cruzada (_x_at: 215281_x_at, 216858_x_at), lo que indica que pueden detectar múltiples transcritos relacionados. Esto constituye una fuente de ruido biológico potencial: en un contexto de implementación clínica real, estas sondas deben ser reemplazadas por primers de PCR con especificidad génica verificada. De igual forma, las cuatro sondas sin anotación conocida (213789_at, 1559156_at, 1569041_at, 216858_x_at) representan un área de incertidumbre que requiere validación experimental independiente, posiblemente mediante RNA-seq, antes de ser incluidas en un panel diagnóstico.

7.F. Limitaciones

A diferencia de los enfoques planos tradicionales, el modelo jerárquico propuesto presenta un posicionamiento algorítmico claro: prioriza la eficacia en la selección de características mediante la división explícita entre exploración global (ACO) y explotación local (Firefly). Sin embargo, este diseño no está exento de limitaciones. Es necesario delimitar las restricciones metodológicas del modelo en su estado actual:

1. **Carga computacional por evaluaciones anidadas:** La principal restricción radica en el volumen total de evaluaciones de fitness requeridas por el acoplamiento de dos enjambres bio-inspirados. Aunque el uso de un clasificador lineal regularizado y la restricción memética (refinar únicamente la solución élite de cada ciclo) mitigan el riesgo de una explosión combinatoria, el sistema sigue demandando un costo algorítmico inherente a las arquitecturas anidadas.
2. **Sensibilidad estocástica y paramétrica:** La eficacia de la metaheurística depende intrínsecamente de hiperparámetros fijados de forma empírica (tasa de evaporación ρ, decaimiento de aleatoriedad α_{decay}, límites de construcción g_{min}, g_{max}). Se desconoce si existe una configuración matemática superior.
3. **Dependencia del evaluador lineal:** LinearSVC condiciona topológicamente los genes para hiperplanos lineales; clasificadores no lineales podrían requerir combinaciones diferentes.
4. **Validación monocéntrica:** Los experimentos se ejecutaron sobre un único origen transcriptómico (GSE45827).
5. **Naturaleza técnica de las sondas y colinealidad biológica:** El modelo seleccionó sondas con riesgo de hibridación cruzada (sufijos _x_at) y no anotadas (UNKNOWN). Lo que exige que para fases de despliegue clínico, la firma sea depurada antes de validarse por PCR.

7.G. Trabajo futuro

A partir de las limitaciones expuestas, se proyectan las siguientes extensiones investigativas:

1. **Frontera de Pareto explícita:** Reformular la función de aptitud hacia una optimización multi-objetivo formal, permitiendo al investigador médico seleccionar visualmente el punto de equilibrio deseado entre la precisión del panel de diagnóstico y el número total de biomarcadores requeridos.
2. **Validación multicéntrica y multi-ómica:** Desplegar la arquitectura jerárquica sobre bases de datos oncológicas internacionales, verificando la consistencia algorítmica ante perfiles de expresión derivados de secuenciación masiva paralela.

3. **Calibración Bayesiana:** Acoplar un motor de inferencia bayesiana (como Optuna) que gobierne y auto-sintonice los pesos estigmérgicos del enjambre en tiempo real, erradicando la dependencia de la calibración manual humana.
4. **Integración de conocimiento a priori (biológicos):** Modificar la función de aptitud del algoritmo Wrapper para penalizar no solo la dimensionalidad ($|S|/K$), sino también la redundancia topológica (multicolinealidad) basándose en redes de interacción génica (Gene Ontology), forzando al enjambre a seleccionar genes de rutas metabólicas estrictamente independientes.

8. Conclusiones

La presente investigación formalizó matemáticamente y validó empíricamente una arquitectura metaheurística jerárquica (ACO-Firefly) diseñada para enfrentar la maldición de la dimensionalidad en microarreglos oncológicos. La evidencia acumulada sugiere que la separación explícita de exploración y explotación en arquitecturas metaheurísticas anidadas constituye un principio de diseño robusto para problemas de selección de características en espacios de dimensionalidad extrema, con implicaciones directas para el diseño de paneles de diagnóstico molecular en oncología de precisión.

De la ejecución rigurosa y libre de fuga de datos se derivan las siguientes conclusiones fundamentales:

1. **Eficacia del enfoque jerárquico:** El modelo propuesto alcanzó un óptimo de Pareto al diagnosticar correctamente los seis subtipos tumorales con un F1-Macro de 0.9440 sobre un conjunto de prueba estrictamente aislado. Este resultado superó la capacidad predictiva y dimensional de los algoritmos de referencia aislados (ACO y Firefly), validando la superioridad de dividir la exploración y la explotación.
2. **Reducción dimensional extrema:** El pipeline logró comprimir el espacio original de 54.676 genes a un panel accionable de apenas 13 biomarcadores. Esta reducción masiva (reteniendo solo el 9.5% de los atributos pre-filtrados) se consiguió con una concesión marginal del 5.7% en rendimiento frente a la evaluación del genoma completo, lo que garantiza la viabilidad clínica y económica de implementar la firma en ensayos PCR rutinarios.
3. **Validación información-teórica de la emergencia estigmérgica:** El decaimiento cuantificado de la Entropía de Shannon (de 7.59 a 6.60 bits en 25 iteraciones) sobre la distribución de feromonas constituye evidencia directa de que la convergencia del enjambre no es producto del azar, sino de un proceso de autoorganización colectiva. El sistema redujo su incertidumbre exploratoria de forma progresiva y monotónica, produciendo una concentración estructurada de memoria feromonal en el subconjunto de biomarcadores con mayor coherencia predictiva.

Referencias

[1] J. Piri et al., "Literature review on hybrid evolutionary approaches for feature selection" *Algorithms*, vol. 16, no. 4, p. 167, 2023.

[2] M. Dorigo, T. Stützle, Ant Colony Optimization. Cambridge, MA: MIT Press, 2004

[3] S. A. Zendehbad, E. M. Rad, y S. S. Bajestani, "Swarm intelligence in biomedical engineering," *Intelligence-Based Medicine*, 2025.

[4] P. Ashokkumar, T. V. S. Kumar, M. Khan, M. M. Su'ud, M. M. Alam, y S. Mallik, "Ant Colony Optimization for feature selection in breast cancer classification," *Egyptian Informatics Journal*, 2025.
[5] M. Dorigo, V. Maniezzo y A. Colorni “Ant system: optimization by a colony of cooperating agents” *IEEE Trans. Syst., Man, Cybernetics B*, vol. 26 no. 1, pp. 29-41, 1996.
[6] M. Dorigo y L. M. Gambardella, “Ant colony system: A cooperative learning approach to the traveling salesman problem” *IEEE Trans. Evolutionary Computation*, vol. 1, no. 1, pp 53-66, 1997.
[7] E. Emary, H. M. Zawbaa, K. K. A. Ghany, A. E. Hassanien, y B. Parv, "Firefly optimization algorithm for feature selection," en BCI '15: *Proceedings of the 7th Balkan Conference on Informatics*, ACM, 2015.
[8] H. T. Ibrahim, W. J. Mazher, y Z. F. Yaseen, "A hybrid feature selection approach based on Firefly algorithm and Simulated Annealing for cancer datasets," *University of Thi-Qar Journal for Engineering Sciences*, vol. 14, no. 1, 2024.Xemar
[9] B. R. G. Grisci, G. E. B. Feltes, M. Dorn, "Neuroevolution as a tool for microarray gene expression pattern identification in cancer research," *Journal of Biomedical Informatics*, vol. 89, pp. 122-133, 2019. Dataset CuMiDa (GSE45827) disponible en: https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida
[10] N. M. Ali, R. Besar, N. A. A. Aziz, "Hybrid feature selection of breast cancer gene expression microarray data based on metaheuristic methods: A comprehensive review," Symmetry, vol. 14, no. 10, p. 1955, 2022
[11] U. Marovac, *et al.*, "Feature selection for biomedical data classification: Statistical vs. swarm intelligence methods," *Journal of Scientific and Industrial Research*, vol. 84, no. 6, pp. 672-680, 2025.
[12] J. B. Jona, "A hybrid of ACO-FFA algorithm for feature selection in digital mammogram," *Int. J. Multidisciplinary Res.*, vol. 7, no. 4, 2025.
[13] R. M. Rizk-Allah, E. M. Zaki, y A. A. El-Sawy, "Hybridizing ant colony optimization with firefly algorithm for unconstrained optimization problems," *Applied Mathematics and Computation*, vol. 224, pp. 473-483, 2013.
[14] P. G. A. Kumar, P. A. Jeyanthy, y D. Devaraj, "Hybrid multi-objective method based on ant colony optimization and firefly algorithm for renewable energy sources" *Sustainable Computing: Informatics and Systems*, vol. 36, p. 100810, 2022.