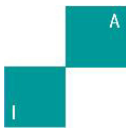
Inteligencia Artificial 0(0), 1-12



INTELIGENCIA ARTIFICIAL

http://erevista.aepia.org/

Algoritmo evolutivo para la búsqueda de vías metabólicas

M. Gerard[1,2], G. Stegmayer[1] y D. Milone[2]
[1]CIDISI-UTN-FRSF, CONICET, Lavaise 610 - Santa Fe (Argentina)
[2]SINC(I)-FICH-UNL, CONICET, Ciudad Universitaria - Santa Fe (Argentina)
mgerard@santafe-conicet.gov.ar

Abstract Los métodos de búsqueda permiten encontrar secuencias que relacionan dos o más estados utilizando un conjunto de transiciones permitidas. Los algoritmos evolutivos realizan la búsqueda guiados por una función de aptitud y emplean operadores estocásticos para explorar múltiples soluciones candidatas a la vez. En bioinformática la búsqueda de vías metabólicas que relacionen dos compuestos es una tarea habitual. En particular esto es de gran interés cuando se quiere descubrir relaciones metabólicas entre compuestos agrupados con técnicas de minería de datos. En este trabajo se propone un algoritmo evolutivo que permite inferir vías metabólicas entre dos compuestos seleccionados a partir de agrupamientos encontrados con un modelo neuronal del tipo mapa auto-organizado. Se describen los operadores y la función de aptitud empleada, se estudia el efecto de la tasa de mutación sobre el algoritmo propuesto y se compara el desempeño de éste con el de dos métodos clásicos de búsqueda.

Keywords: Search Strategies, Evolutive Algorithms, Metabolic Pathways.

1. Introducción

En diversas áreas del conocimiento es habitual el empleo de métodos de búsqueda para resolver diferentes problemas. En muchos casos el uso de métodos clásicos de exploración secuencial del espacio de estados permite encontrar soluciones de manera relativamente rápida. Cuando se exploran exhaustivamente todas las posibles soluciones las estrategias se denominan de búsqueda no informada, y este es el caso de los algoritmos de búsqueda en amplitud (BA) y en profundidad (BP). Sin embargo, cuando la exploración se realiza mediante el uso de una heurística que emplee información del problema, la estrategia se denomina búsqueda informada, como por ejemplo el algoritmo A^* [13].

Es bien conocido el hecho de que existen problemas en los que debe explorarse un número muy elevado de soluciones, lo que hace que los métodos clásicos se vuelvan prácticamente inaplicables. En los últimos años se han propuesto diferentes enfoques para abordar el problema, como por ejemplo el de los algoritmos evolutivos (AE). Éstos utilizan estrategias estocásticas de búsqueda basadas en la evolución de una población de soluciones candidatas, aplicando un conjunto de operadores y una función de aptitud que evalúa la calidad de las soluciones generadas. Algunos aspectos interesantes de esta técnica son la simplicidad de los operadores utilizados, la posibilidad de emplear funciones de aptitud con muy pocos requisitos formales y la capacidad de explorar múltiples puntos del espacio de búsqueda en cada iteración [15].

Un problema de interés en bioinformática es la inferencia de vías metabólicas [8], donde se intenta encontrar secuencias de reacciones bioquímicas que relacionen un conjunto de compuestos de interés.

ISSN: 1988-3064(on-line)

Entender cómo se relacionan los compuestos y cuáles son las reacciones que participan permite formar una imagen más completa de los procesos metabólicos subyacentes en un organismo. Recientemente se han propuesto diferentes estrategias para abordar este problema. PathComp [12] emplea un algoritmo basado en búsqueda en amplitud para construir caminos que conectan los compuestos, tomándolos de a pares y combinándolos a través de relaciones permitidas. Metabolic PathFinding Tool [2] asigna a cada operador un costo igual al número de reacciones donde el compuesto participa. PathMiner [9] utiliza el algoritmo de búsqueda A^*; su función heurística emplea información estructural de los compuestos para generar descriptores característicos y explora el espacio de búsqueda empleando una función de costo basada en la distancia de Manhattan.

Aunque los métodos de búsqueda pueden encontrar secuencias de acciones que relacionen dos estados, es necesario que la relación exista para que la búsqueda produzca un resultado. Una forma de identificar estados potencialmente relacionados es mediante la generación de agrupamientos con técnicas de minería de datos. Sin embargo, éstas ponen de manifiesto la presencia de relaciones pero no las explicitan. Este es un problema habitual en bioinformática, especialmente cuando se trabaja con datos de diferente tipo, como es el caso de perfiles metabólicos y transcripcionales[1]. Recientemente se ha propuesto un modelo denominado IL-SOM [16, 10, 11] basado en mapas auto-organizados que permite encontrar agrupamientos a partir de la integración de datos de este tipo. Este modelo aplica el principio denominado "guilt-by-association" [14, 17] para encontrar genes y metabolitos que varían en forma coordinada. Sin embargo, aunque las relaciones que vinculan estos compuestos en los agrupamientos corresponden a vías metabólicas, la reconstrucción de éstas a partir de los datos no es sencilla [7]. Para realizar esta tarea podría considerarse el siguiente esquema de trabajo: en primer lugar se emplea un algoritmo de agrupamiento para encontrar compuestos metabólicos que se relacionan; luego se selecciona un agrupamiento que contenga compuestos que se intentan relacionar y se emplea un algoritmo de búsqueda para encontrar una vía metabólica que vincule los compuestos.

La motivación de este trabajo es desarrollar un algoritmo evolutivo para la inferencia automática de vías metabólicas que relacionen dos compuestos y comparar su desempeño frente a dos algoritmos de búsqueda clásicos. Para ésto se utiliza el modelo IL-SOM para generar agrupamientos a partir de datos metabólicos y transcripcionales de frutos de tomate y se seleccionan agrupamientos que contienen compuestos de interés entre los cuales buscar vías metabólicas. Luego se definen medidas objetivas para cuantificar el desempeño de los algoritmos y se estudia el efecto de la tasa de mutación sobre el funcionamiento del algoritmo evolutivo.

La organización del trabajo es la siguiente. En la Sección 2 se describe el algoritmo propuesto para la búsqueda evolutiva de rutas metabólicas entre dos compuestos. En la Sección 3 se describen brevemente los datos empleados, las medidas objetivas y los resultados alcanzados. Finalmente se presentan en la Sección 4 las conclusiones del trabajo.

2. Algoritmo propuesto

En esta sección se presenta el algoritmo propuesto, que denominaremos algoritmo evolutivo para la búsqueda de vías metabólicas (AEBVM). Luego de la inicialización y evaluación de esta población inicial, el algoritmo itera los siguientes pasos hasta alcanzar una solución satisfactoria: (1) Selección de los padres a cruzar, (2) Generación de hijos mediante la cruza de los padres seleccionados, (3) Mutación de los hijos, (4) Generación de la nueva población, (5) Evaluación de la aptitud de la población. A continuación se define en primer lugar el espacio de estados y los operadores empleados. Luego se presenta la estructura de los cromosomas y el modo en que se codifica la información. A continuación, se describen los operadores genéticos utilizados y su funcionamiento. Finalmente se presenta la función de aptitud empleada, se analizan los términos que la componen y se describe el efecto que cada uno produce sobre la búsqueda.

2.1. Estructura de los cromosomas

Existen diferentes aproximaciones que permiten reducir el espacio de búsqueda para encontrar vías metabólicas que relacionen dos compuestos. Una propuesta consiste en generar una lista de compuestos

[1] Perfil metabólico: medición de los niveles de concentración de moléculas pequeñas. Perfil transcripcional: medición de los niveles de actividad de un conjunto de genes.

que deben excluirse de la búsqueda [5]. Sin embargo definiciones incorrectas pueden excluir compuestos necesarios para producir resultados de interés biológico. Un enfoque diferente fue propuesto en [6] donde se emplean conjuntos de relaciones binarias "sustrato-producto" para representar las reacciones y se etiqueta cada relación según su función dentro de la reacción. La columna vertebral de las vías se construye empleando sólo las relaciones con información acerca de la transformación de los sustratos.

Siguiendo esta idea se define el espacio de estados como el conjunto C de todos los compuestos metabólicos contenidos en KEGG [4], exceptuando los polímeros de glucosa, donde las transformaciones r describen las relaciones binarias permitidas entre compuestos de C. El compuesto sobre el cual se aplica la transformación se denominará sustrato s, siendo p el producto o nuevo estado resultante de la misma. Las transformaciones se representarán como pares ordenados $r_i = (s_i, p_i)$, con $s_i, p_i \in C$ y $s_i \neq p_i$. Además el sustrato y el producto de r_i se identificarán empleando la notación s_i y p_i respectivamente, siendo $\widehat{s}$ el compuestos inicial y $\widehat{p}$ el compuesto final de la vía metabólica. De este modo una vía metabólica se construye como una secuencia de transformaciones r que producen $\widehat{p}$ a partir de $\widehat{s}$. Finalmente, se define la secuencia de estados posibles $\mathbf{q} = [\widehat{s}, p_1, p_2, \ldots, \widehat{p}]$ como la secuencia de compuestos que intervienen en la secuencia de transformaciones.

La secuencia de transformaciones r que conduce a la producción del compuesto $\widehat{p}$ a partir del compuesto $\widehat{s}$ se codifica en el cromosoma como $\mathbf{c} = [r_1, r_2, \ldots, r_i, \ldots, r_N]$, donde N indica el número de genes y la secuencia se lee de izquierda a derecha. Este valor varía en el rango $[1, N_{\text{máx}}]$, donde $N_{\text{máx}}$ limita el número máximo de transformaciones que puede contener la vía metabólica. Cuando el número de transformaciones supera esta cota, el cromosoma es truncado para contener sólo las primeras $N_{\text{máx}}$. De éste modo, cada gen representa la transformación de un compuesto s en otro compuesto p.

2.2. Operadores genéticos

En esta sección se describen los operadores genéticos diseñados para el AEBVM. Dados los requerimientos de esta aplicación en particular, ha sido necesario realizar diversas modificaciones a los operadores genéticos clásicos, que limitarían la convergencia del algoritmo de ser aplicados directamente. Además, para facilitar la descripción de estos operadores se definen cuatro conjuntos. R^* contiene al conjunto completo de transformaciones permitidas, $R^1 = \{r_i / r_i = (\widehat{s}, p_i)\} \wedge R^1 \subset R^*$ contiene sólo aquellas transformaciones que tienen a $\widehat{s}$ como sustrato, $R^N = \{r_i / r_i = (s_i, \widehat{p})\} \wedge R^N \subset R^*$ contiene todos las transformaciones que producen $\widehat{p}$ y $R^+ = R^1 \cup R^N$ contiene la unión de los dos conjuntos anteriores.

El algoritmo finaliza la ejecución alcanzado un número máximo de generaciones predefinido o encontrando la solución.

Inicialización. El algoritmo se inicializa definiendo el número P de individuos que contiene la población y un valor $N_{inic} \leq N_{\text{máx}}$ para cada individuo, que indica el número de genes que contiene el cromosoma inicialmente. Los cromosomas se construyen según

$$\mathbf{c} = \begin{cases} r_i \in R^+ & \text{si } N = 1, \\ [r_i, r_j], r_i \in R^1, r_j \in R^N & \text{si } N = 2, \\ [r_i, \ldots, r_k, \ldots, r_j], r_i \in R^1, r_k \in R^*, r_j \in R^N & \text{si } N > 2, \end{cases} \quad (1)$$

donde todo gen r es una transformación seleccionada al azar del conjunto correspondiente.

Selección. En este algoritmo se utilizó el método tradicional de la ruleta [15]. Éste se basa en la asignación de un valor f a cada individuo que es proporcional a su contribución a la aptitud media de la población. Además, para garantizar la preservación de los individuos más aptos en cada generación se emplea elitismo; el parámetro N_{elite} determina el número de individuos que serán preservados y pasarán sin modificaciones a la siguiente generación.

Cruza. El operador presenta una modificación importante respecto del operador clásico. El punto de cruza ϕ para cada padre $(\mathbf{c}_1, \mathbf{c}_2)$ es seleccionado aleatoriamente de un conjunto que contiene pares de posiciones (ϕ_1, ϕ_2) que satisfacen $\delta(s_i, p_j) = 1$, donde δ es la función delta de Kronecker que toma valor 1 cuando $s_i \in \mathbf{c}_1$ y $p_j \in \mathbf{c}_2$ son iguales. La Figura 1 presenta un esquema del funcionamiento de este operador para el caso de dos padres que no son completamente válidos (ver definición de *validez* más

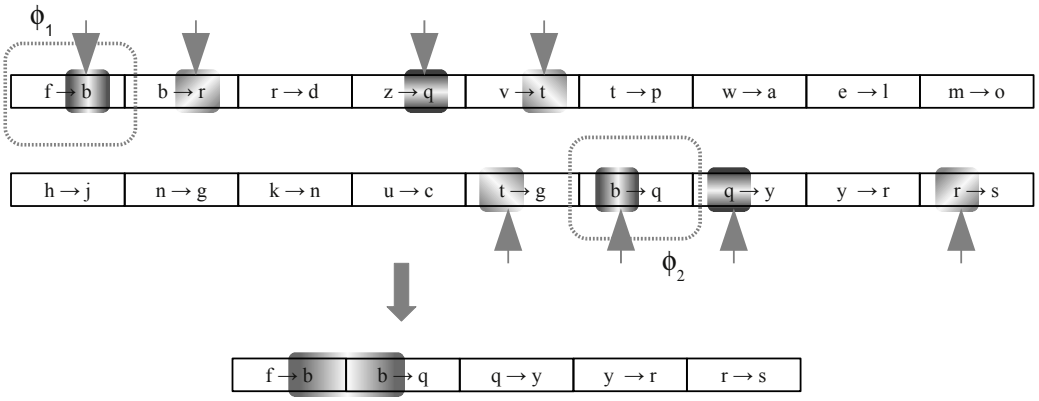


Figura 1: Esquema de funcionamiento del operador de cruza. Cada bloque corresponde a un gen y codifica una transformación. En cada gen, el sustrato y producto están representados por las letras a izquierda y derecha de cada flecha, respectivamente. Los elementos sombreados indican pares de posiciones (ϕ_1, ϕ_2) donde puede realizarse una cruza válida.

adelante). Cada bloque representa un gen y codifica una transformación en donde las letras representan el sustrato y el producto en cada transformación. Puede observarse que si se aplica un método de cruza simple sin tener en cuenta la secuencia de transformaciones, muy probablemente el hijo generado disminuirá su validez (el producto de un gen no será el sustrato del gen siguiente). Sin embargo, si la cruza se realiza en alguno de los pares de posiciones que se encuentran resaltadas con el mismo sombreado (por ejemplo (ϕ_1, ϕ_2)), la validez del hijo se incrementará, o al menos permanecerá constante.

Mutación. Este operador realiza el reemplazo de un gen del cromosoma por otro donde s o p del nuevo gen es p del gen anterior o s del gen siguiente, respectivamente. Cada cromosoma tiene una probabilidad p_{mut} de ser mutado en una única posición seleccionada aleatoriamente. El nuevo gen se obtiene según

$$mut(r_i) = \begin{cases} r \in R^+ & \text{si } N = 1, \\ r \in R^1 & \text{si } N > 1 \ \wedge \ i = 1, \\ r \in R^N & \text{si } N > 1 \ \wedge \ i = N, \\ r \in R^* / p = s_{i+1} & \text{si } N > 1 \ \wedge \ 1 < i < N \ \wedge \ u \leq 0.5, \\ r \in R^* / s = p_{i-1} & \text{si } N > 1 \ \wedge \ 1 < i < N \ \wedge \ u > 0.5, \end{cases} \tag{2}$$

donde s y p son, respectivamente, el sustrato y producto del nuevo gen r; s_{i+1} es el sustrato del gen que se encuentra en la posición siguiente a la del operador r_i mutado y p_{i-1} es el producto del gen que se encuentra en la posición anterior a la del gen mutado. El valor u es seleccionado aleatoriamente en el rango $[0, 1]$. La Figura 2 presenta un esquema del funcionamiento de este operador. En el cromosoma ubicado en la parte superior de la figura se observa que el gen seleccionado para mutar (remarcado en negro) posee una relación válida con el gen anterior. Si se aplicara un operador clásico de mutación sería altamente probable que la validez del cromosoma disminuyera debido a que es baja la probabilidad de insertar un gen que no destruya la relación válida existente. Si por el contrario se aplica una estrategia de mutación válida como la definida en (2), el nuevo cromosoma podría conservar o incrementar su validez, como puede apreciarse en el cromosoma inferior.

2.3. Función de aptitud.

Para cuantificar la calidad de los individuos a lo largo de las generaciones y dirigir la búsqueda de la solución se construyó una función que modela las características que debe reunir la solución del problema.

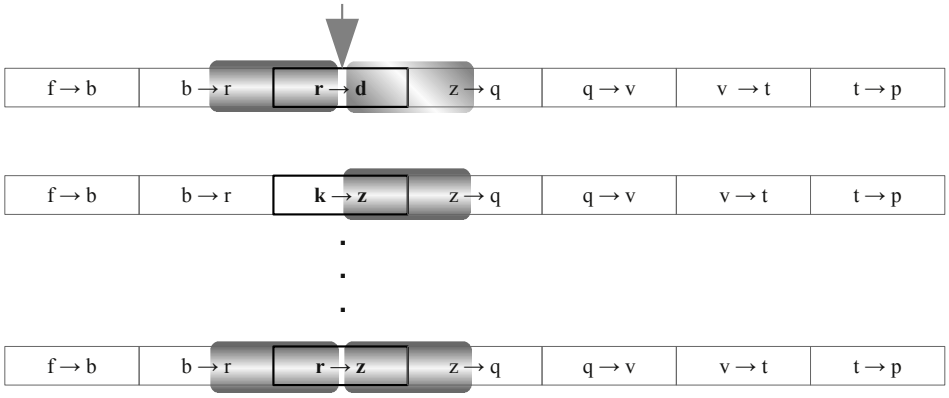


Figura 2: Esquema de funcionamiento del operador de mutación. El gen seleccionado para mutar se encuentra remarcado en negro. Los elementos sombreados indican relaciones válidas entre genes.

Validez (V). Cuantifica el número de concatenaciones válidas presentes en el cromosoma, definiendo éstas como aquellos pares de transformaciones donde el producto p_i de la transformación r_i es el sustrato s_{i+1} de la transformación r_{i+1}. En base a esto, la validez se calcula como

$$V(\mathbf{c}) = \frac{\delta(\widehat{s}, s_1) + \delta(p_N, \widehat{p}) + \sum_{i=1}^{N-1} \delta(s_{i+1}, p_i)}{N+1}. \tag{3}$$

Ésta varía en el rango $[0, 1]$, siendo 1 cuando todas las transformaciones están concatenadas y los compuestos s_1 y p_N son los deseados.

Extremos válidos (E). Este término evalúa las transformaciones r_1 y r_N para verificar que contienen los compuestos $\widehat{s}$ y $\widehat{p}$ deseados. El cálculo se realiza según $E(\mathbf{c}) = \frac{1}{2}\left[\delta(\widehat{s}, s_1) + \delta(p_N, \widehat{p})\right]$. Este término varía en el rango $[0, 1]$ y alcanza su valor máximo cuando los compuestos s_1 y p_N son los deseados. Éste desempeña un papel importante cuando el tamaño de las vías metabólicas supera $N_{\text{máx}}$.

Tasa de reacciones únicas (Q). Este término penaliza la repetición de una transformación en el cromosoma. Para el cálculo se define la función φ que evalúa una secuencia y devuelve el número de elementos únicos en la misma. La tasa se calcula como $Q(\mathbf{c}) = (\varphi(\mathbf{c}) - 1)/(N - 1)$ y se define $Q(\mathbf{c}) = 0$ cuando $N = 1$. Ésta varía en el rango $[0, 1]$ y alcanza su valor mínimo cuando la secuencia contiene un único elemento repetido N veces ($\varphi(\mathbf{c}) = 1$).

Tasa de compuestos únicos (I). Este término penaliza la repetición de compuestos en la vía. La tasa se calcula como $I(\mathbf{c}) = (\varphi(\mathbf{q}) - 2)/(N - 1)$ y se define $I(\mathbf{c}) = 0$ cuando $N = 1$. Ésta varía en el rango $[0, 1]$ y alcanza su valor mínimo cuando el cromosoma contiene transformaciones que conducen solamente a s_1 o p_1. Por ejemplo, si $\mathbf{q} = [a, b, \underline{a}, \underline{b}]$, $N = 3$ y $I(\mathbf{c}) = 0$.

Función de aptitud (A). La función de aptitud A para el cromosoma $\mathbf{c}$ se define como $A(\mathbf{c}) = \alpha\left[V(\mathbf{c}) + \beta E(\mathbf{c}) + Q(\mathbf{c}) + I(\mathbf{c})\right]$, donde $\alpha = 1/(3 + \beta)$ es una constante de normalización que lleva la función al rango $[0, 1]$ y β determina la contribución relativa de E. Esta función toma valor 1 cuando se encuentra una vía metabólica válida y sin bucles, que transforma $\widehat{s}$ en $\widehat{p}$. En caso de contar con información acerca de la abundancia relativa de los compuestos, esta función podría modificarse para ponderar las transformaciones según la probabilidad de ocurrencia, que está asociada directamente con la abundancia de los compuestos intervinientes.

Tabla 1: Agrupamientos seleccionados para buscar vías metabólicas. Solo se presentan los compuestos metabólicos contenidos en el agrupamiento. Los superíndices indican los caminos donde el compuesto fue usado como extremo; la letra indica su participación como extremo inicial ($\widehat{s}$) o final ($\widehat{p}$) de la vía.

Agrupamiento A			Agrupamiento B		
Compuestos	Isómeros		Compuestos	Isómeros	
	I	II		I	II
arginina	C00062$^{1\widehat{s}}$	C00792	asparagina	C00152	C01905
glicerato	C00258$^{2\widehat{s},3\widehat{p}}$		glicina	C00037$^{6\widehat{s}}$	
lisina	C00047$^{1\widehat{p},3\widehat{s}}$		histidina	C00135$^{4\widehat{s},5\widehat{s}}$	
ornitina	C00077$^{2\widehat{p}}$	C00515	isoleucina	C00407	
			serina	C00065$^{4\widehat{p}}$	C00740
			tirosina	C00082$^{5\widehat{p},6\widehat{p}}$	
			treonina	C00188	C00820
			valina	C00183	C06417

3. Resultados y Discusión

En esta sección se presentan los resultados obtenidos en la evaluación del AEBVM y en la comparación con dos métodos de búsqueda clásicos. Primero se describen los datos usados en los experimentos. Luego se presentan las medidas empleadas para comparar los algoritmos. Después se analiza el comportamiento del AEBVM utilizando diferentes tasas de mutación. Finalmente se contrastan las medidas obtenidas con los distintos algoritmos durante la búsqueda de vías metabólicas limitadas a 100 transformaciones.

El conjunto de compuestos válidos para generar vías metabólicas y las reacciones químicas posibles entre estos compuestos se extrajo de la base de datos KEGG. Se empleó esta fuente por ser de acceso libre, encontrarse extensamente citada en la literatura y por contener información de una amplia variedad de organismos. Cada compuesto se encuentra codificado mediante un código único y la información de las reacciones químicas es almacenada con ésta codificación. Además, las reacciones químicas se almacenan como conjuntos de relaciones binarias "sustrato-producto", donde cada una está etiquetada según la función que el par cumple dentro de la reacción química. Para los experimentos se seleccionaron aquellas relaciones etiquetadas como "main" debido a que contienen la información acerca de las transformaciones $s \rightarrow p$ [3]. Luego del procesamiento de los datos se obtuvieron 5936 compuestos y 14346 transformaciones. Los compuestos usados como extremos para las vías metabólicas fueron seleccionados a partir de agrupamientos generados con el modelo IL-SOM y datos de perfiles metabólicos y transcripcionales de frutos de tomate cultivados en condiciones controladas de campo y cosechados en etapa de maduración [1]. Los agrupamientos empleados fueron seleccionados por contener compuestos de interés en el dominio de la aplicación. Los datos transcripcionales sólo se usaron para generar los agrupamientos y no se incluyeron en la construcción de las vías metabólicas. El detalle de los agrupamientos seleccionados se presenta en la Tabla 1. Los isómeros detallados en esta tabla fueron considerados como compuestos diferentes, de manera que el agrupamiento A contiene 6 compuestos y el agrupamiento B contiene 12 compuestos. Para simplificar la notación se usará el código de cada compuesto sin considerar la letra y los ceros que anteceden al número, y se denominará "extremos" a los pares de compuestos entre los que se realizará la búsqueda de la vía metabólica.

3.1. Medidas de evaluación

Para comparar los resultados obtenidos con los distintos algoritmos se mide el tiempo requerido para encontrar una vía metabólica (t)[2], el número de transformaciones que contiene la vía (L) y el número de compuestos del agrupamiento del que forman parte los extremos (ψ) que se encuentran participando de la vía. En el caso del AEBVM, también se evalúa el número de generaciones empleado para encontrar

[2]Los experimentos se realizaron empleando un PC INTEL Pentium IV de 3 GHz con 2 Gb de memoria RAM.

una solución (G).

Cada búsqueda se realiza 12 veces y luego se determinan los valores máximo (indicado por el subíndice "*máx*"), mínimo (indicado por el subíndice "*mín*") y la mediana (indicada por el símbolo "$\widehat{\ }$") para distintas medidas a lo largo de los experimentos. En la mayor parte de las mediciones se emplea la mediana en reemplazo de la media por ser una medida más robusta frente a distribuciones asimétricas, como ocurre en estos casos. Sólo se calcula el valor medio para t y ψ, indicados como $\overline{t}$ y $\overline{\psi}$ respectivamente. Se optó por $\overline{\psi}$ debido a que en la mayoría de las vías metabólicas encontradas sólo los compuestos del agrupamiento a relacionar participan en el camino, haciendo que $\widehat{\psi} = 2$. Otra medida empleada es la tasa de explicación del agrupamiento Λ, calculada como

$$\Lambda = \frac{\max\limits_{k} \{\psi_k\}}{|\Psi|}, \tag{4}$$

donde k indica el número de experimento, ψ_k es el número de compuestos del agrupamiento incluidos en la vía encontrada en el experimento k y $|\Psi|$ es el número total de compuestos del agrupamiento. Esta tasa varía en el rango $[0, 1]$ e indica la proporción de compuestos del agrupamiento presentes en la vía metabólica. Valores de $\Lambda \to 1$ indican que la vía relaciona un gran número de los compuestos del agrupamiento.

3.2. Evaluación del algoritmo evolutivo empleando diferentes tasas de mutación

Para evaluar la influencia de la tasa de mutación sobre las búsquedas se estudia este parámetro entre 1 y 10 %, y los compuestos marcados en la Tabla 1. La mejor solución en cada generación se conserva empleando elitismo ($N_{elite} = 1$). En todos los experimentos se emplea una población de $P = 1000$ individuos y un número inicial máximo de genes por cromosoma $N_{inic} = 50$ para asegurar una buena exploración del espacio de búsqueda. Además se emplea un tamaño máximo de cromosoma $N_{\text{máx}} = 100$ para permitir cruzas que produjeran cromosomas con tamaños superiores al N_{inic} asignado, y una probabilidad de cruza $p_{cruza} = 80\,\%$ que brindó los mejores resultados en experimentos previos.

En la Figura 3 se presenta los tiempos de búsqueda y el número de generaciones requeridas por el AEBVM para encontrar un camino empleando diferentes tasas de mutación. Se puede observar que los valores asociados al número medio de generaciones presentan valores cercanos a 100 generaciones para tasas de mutación de 4, 6 y 7 %, pero toma el valor mínimo en 6 %. Sin embargo, un análisis de la significancia estadística del número medio de generaciones indica que tasas de mutación en el intervalo 3-9 % producen resultados similares ($p = 0.12$). En cuanto al tiempo de búsqueda, se observa un comportamiento similar al que presenta el número medio de generaciones, pero en este caso se emplea el menor tiempo para una tasa de mutación de 4 %. Nuevamente se observa que el efecto de la tasa de mutación sobre el valor medio del tiempo de búsqueda no presenta diferencias estadísticamente significativas ($p = 0.07$) en el intervalo 3-6 %. Con respecto a las medidas de $\psi_{\text{máx}}$, $\overline{\psi}$ y Λ, todas las tasas de mutación conducen a resultados similares.

Si bien los tiempos de búsqueda no presentan diferencias estadísticamente significativas en el rango mencionado anteriormente, para este conjunto de experimentos se observa que con una tasa de mutación de 4 % se consigue el menor tiempo medio y, por lo tanto, se decidió emplear esta tasa de mutación para el AEBVM y su comparación con otros algoritmos.

3.3. Comparación entre búsqueda en amplitud, búsqueda en profundidad y el algoritmo evolutivo propuesto

Para evaluar el desempeño del AEBVM frente a los algoritmos clásicos de BA y BP, se compararon las medidas de desempeño obtenidas para la búsqueda de vías metabólicas. El AEBVM se evaluó empleando los mismos parámetros utilizados durante el estudio de la sección previa. Los algoritmos BA y BP fueron modificados para incorporar un control de estados repetidos que permitiera eliminar vías metabólicas donde se produjeran bucles, debido a que éstas carecen de interés biológico. Para el caso de BP se limitó la profundidad máxima de la búsqueda a 100 transformaciones, con el objetivo de fijar la misma restricción que la establecida para el AEBVM.

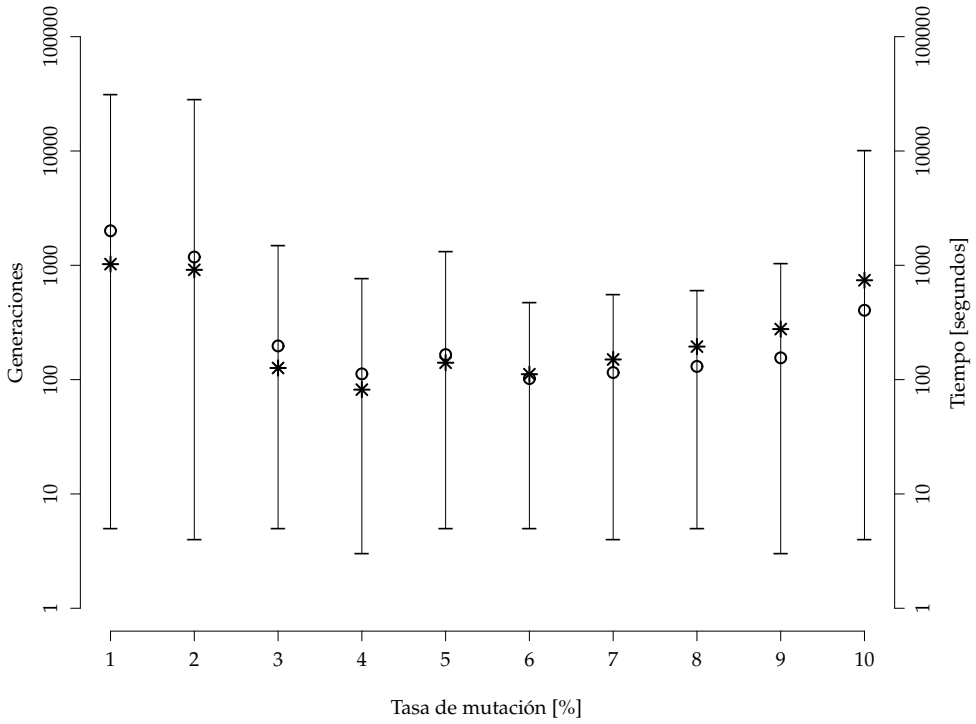


Figura 3: Influencia de la tasa de mutación sobre el tiempo de búsqueda y el número de generaciones requeridas por el AEBVM para encontrar una solución. Los datos están graficados en escala logarítmica. Las estrellas indican el tiempo medio de búsqueda y los círculos corresponden al número medio de generaciones. Las barras verticales indican el número máximo y mínimo de generaciones que el AEBVM empleó para encontrar una solución.

Debido a que el conjunto de soluciones encontradas por los algoritmos de BA y BP está predefinida por el orden en que se aplican los operadores, una única búsqueda puede generar soluciones sesgadas por el ordenamiento inicial de los mismos. Para evitar este efecto se repitieron las búsquedas para diferentes aleatorizaciones de los operadores. En el caso de BP se realizaron 10 búsquedas con cada par de compuestos de la Tabla 1, empleando una aleatorización diferente en cada una. Para BA se realizó la búsqueda de 12 caminos para cada aleatorización y se repitió la búsqueda empleando 10 aleatorizaciones diferentes. Los resultados reportados para este algoritmo corresponden al promedio sobre las 10 aleatorizaciones para cada camino. La búsqueda de 12 caminos se realizó para evaluar la diversidad en el número de transformaciones que contenían los caminos encontrados con cada algoritmo.

Los resultados correspondientes al tiempo de búsqueda y a la diversidad en las longitudes generadas con los tres métodos se presentan en la Figura 4. En la misma se grafican en blanco los diagramas de cajas para los tiempos de búsqueda, y en gris la longitud de las vías encontradas. En cada diagrama el cuerpo de la caja contiene el 50 % central de los datos y el diagrama completo refleja la variabilidad de la medida analizada. Los límites inferior y superior de las cajas corresponden a los cuartiles Q_1 y Q_3 respectivamente, y el segmento que divide cada caja en dos mitades señala la posición del cuartil Q_2 (mediana). La diferencia entre Q_1 y Q_3 se denomina rango intercuartílico (RIC) y los segmentos que se proyectan hacia afuera del cuerpo de la caja se calculan como $L_s = Q_3 + 1{,}5RIC$ y $L_i = Q_1 - 1{,}5RIC$, donde L_s y L_i corresponden a los límites superior e inferior respectivamente. Como se puede observar en la figura, BP encuentra caminos en tiempos cercanos al segundo, pero la longitud de los caminos tiende al valor máximo permitido. Debido a que no resultan de interés biológico vías metabólicas que contengan 100 transformaciones y que sólo relacionen dos de los compuestos del agrupamiento, los resultados alcanzados con este algoritmo se excluyeron de los análisis posteriores.

Aunque los tiempos medios de búsqueda para BA y AEBVM no presentan diferencias estadísticamente

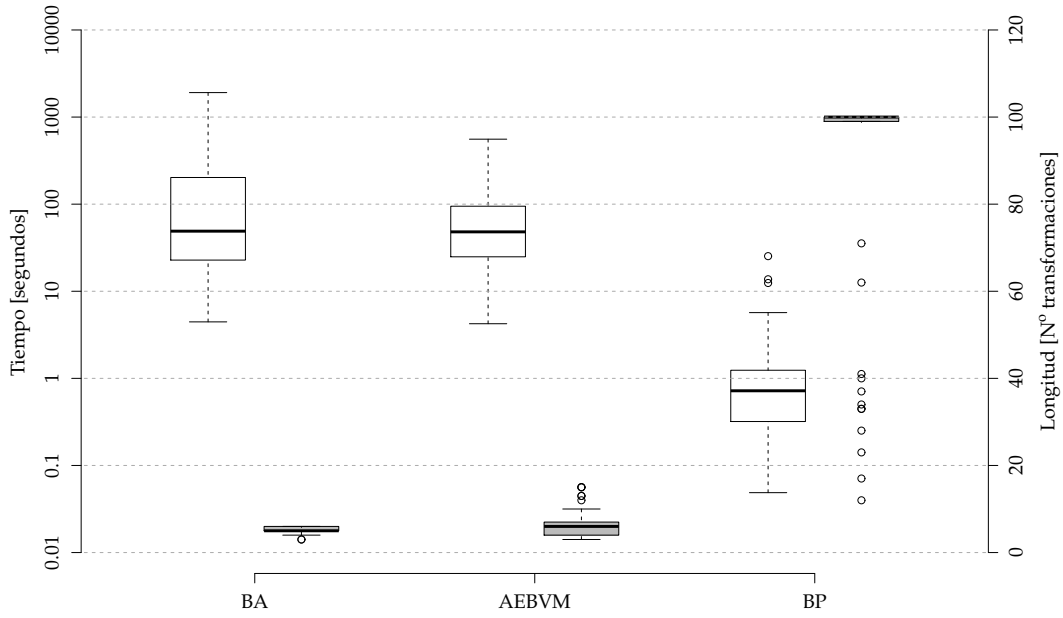


Figura 4: Diagrama de cajas para el tiempo de búsqueda y la longitud de los caminos para los resultados de los algoritmos de AEBVM, BA y BP. En blanco se muestran los tiempos de búsqueda expresados en segundos, y en gris la longitud de los caminos encontrados expresadas en número de transformaciones. Los tiempos se encuentran graficados en escala logarítmica. Cada caja representa el 50 % central de los datos. Los círculos indican mediciones atípicas. Los segmentos que se extienden hacia afuera de las cajas indican los límites máximo y mínimo a partir de los cuales los valores se consideran medidas atípicas.

significativas ($p = 0.17$), BA presenta la mayor dispersión de tiempos de búsqueda y encuentra caminos con el menor número de transformaciones posible, como es de esperar para este algoritmo. En contraposición, el AEBVM genera una menor dispersión en los tiempos de búsqueda, como se refleja principalmente en el Q_3 y en el límite superior del diagrama de cajas (recuérdese que se grafican los tiempos en escala logarítmica), y encuentra caminos con longitudes similares a las encontradas por BA, pero además incluye vías metabólicas con longitudes intermedias a las de BA y BP. Este resultado es interesante debido a que además de encontrar los caminos más cortos entre los compuestos, encuentra caminos con una mayor diversidad de tamaños, lo que brinda alternativas que pueden ser de interés desde el punto de vista de su análisis biológico. Además, es interesante señalar que aunque se empleó un límite de 100 transformaciones para la búsqueda, en ningún caso el AEBVM alcanzó éste límite.

Para analizar con mayor detalle los resultados generados por AEBVM y BA se construyó la Tabla 2, que presenta las medidas obtenidas con cada algoritmo para cada camino. Las filas de la tabla corresponden al tiempo medio de búsqueda, el número de transformaciones máximo, mínimo y la mediana, el número máximo y medio de compuestos del agrupamientos incorporados en la vía y la tasa de explicación del agrupamiento. En las columnas se indican las distintas búsquedas numeradas del 1 al 6, el número de compuestos que contiene cada agrupamiento y los compuestos empleados en cada búsqueda. Analizando esta tabla se observa que BA emplea tiempos 10 veces mayores que AEBVM para realizar las búsquedas 2 y 3, mientras que AEBVM emplea tiempos 5 veces mayores para las búsquedas 5 y 6, y solamente las búsquedas 1 y 4 no presentan diferencias estadísticamente significativas en el tiempo medio de búsqueda ($p > 0.05$). Con respecto a la longitud de los caminos puede apreciarse que la mediana para cada algoritmo es similar, mientras que la diversidad en el número de transformaciones es mayor para el AEBVM, como se refleja en la longitud máxima para cada algoritmo. Por otra parte, se observa que las medidas relacionadas con la presencia de compuestos de los agrupamientos en los caminos es similar para ambos algoritmos. La tasa de explicación de los agrupamientos refleja que ninguno de los algoritmos presenta una preferencia por incorporar compuestos de los agrupamientos en los caminos encontrados.

Tabla 2: Comparación entre BA y AEBVM. El tiempo $\bar{t}$ se expresa en segundos y L en número de transformaciones. $|\Psi|$ indica el número de compuestos del agrupamiento.

Búsqueda		1	2	3	4	5	6
$\|\Psi\|$			6			12	
Extremos		62 - 47	258 - 77	47 - 258	37 - 82	135 - 65	135 - 82
$\bar{t}$	AEBVM	70	40	62	19	121	178
	BA	94	424	458	25	14	41
$L_{\text{máx}}$	AEBVM	9	10	15	5	15	15
	BA	5	6	6	4	6	6
$\widehat{L}$	AEBVM	4	5	6.5	4	6.5	7.5
	BA	5	5	5	4	6	6
$L_{\text{mín}}$	AEBVM	4	5	5	3	5	5
	BA	4	5	5	3	5	5
$\psi_{\text{máx}}$	AEBVM	2	3	2	3	3	2
	BA	3	2	2	3	2	2
$\overline{\psi}$	AEBVM	2.0	2.2	2.0	2.3	2.2	2.0
	BA	2.3	2.1	2.0	2.5	2.1	2.0
Λ	AEBVM	0.3	0.5	0.3	0.3	0.3	0.2
	BA	0.5	0.4	0.3	0.2	0.2	0.2

Se debe destacar que el AEBVM genera una mayor dispersión de longitudes en los caminos, lo que se traduce en un incremento en la variedad de vías encontradas y en la riqueza de posibilidades para su análisis desde un punto de vista biológico.

Para una evaluación biológica preliminar de los resultados producidos por el algoritmo se realizó la búsqueda de una vía alternativa a la glucólisis que relacionara los compuestos C00103 (glucosa) y C00631 (2-fosfoglicerato), ambos característicos de la misma. Para la búsqueda se emplearon los parámetros definidos previamente, fijándose $p_{mut} = 4\,\%$ y $L_{\text{máx}} = 100$. La vía metabólica encontrada presenta la secuencia de transformaciones

$$\text{C00103} \rightarrow [\text{C00031} \rightarrow \text{C00095} \rightarrow \text{C01094} \rightarrow \text{C00577}] \rightarrow \text{C00258} \rightarrow \text{C00631},$$

donde cada flecha corresponde a la transformación del sustrato en producto (izquierda y derecha de cada flecha respectivamente). En la secuencia no se muestran los compuestos adicionales requeridos tales como ATP, NAD^+, etc. La búsqueda proporcionó una nueva vía formada por 6 transformaciones pertenecientes a la glucólisis, a la vía de la manosa y fructosa, a la de los glicerolípidos y a la de las pentosas fosfato. Es interesante destacar que el mayor número de transformaciones ocurre en la segunda vía (indicadas dentro de los corchetes) y que, aunque la secuencia comienza con una transformación perteneciente a la glucólisis, luego emplea transformaciones de diferentes vías para construir una nueva que atraviesa 4 ya conocidas.

4. Conclusiones y trabajos futuros

En este trabajo se propuso un algoritmo evolutivo para la búsqueda de vías metabólicas entre compuestos seleccionados a partir de dos agrupamientos generados con un modelo neuronal del tipo mapa auto-organizado para datos metabólicos y transcripcionales de tomate.

Se estudió el efecto que la tasa de mutación ejerce sobre diferentes medidas de desempeño propuestas para el análisis y se observó que tasas de mutación en el intervalo 3-9 % producen efectos similares en el número de generaciones empleadas en la búsqueda, pero que con una tasa de 4 % se consigue el menor tiempo. En la comparación del AEBVM con los algoritmos de búsqueda en amplitud y en profundidad se observó que éste último emplea tiempos cercanos al segundo, pero encuentra caminos que contienen el número máximo de transformaciones permitido, lo que condiciona fuertemente su utilidad. Se observó también que el algoritmo de búsqueda en amplitud y el AEBVM producen resultados similares para las medidas relacionadas con los agrupamientos, indicando que no poseen ninguna preferencia por

incorporar compuestos adicionales. Este hecho era de esperar dado que en ninguno de los dos algoritmos se modela explícitamente la incorporación de compuestos adicionales de los agrupamientos en los caminos buscados.

Al evaluar la validez biológica de las soluciones encontradas con éste algoritmo se encontró una vía metabólica alternativa a la glucólisis que relacionaba dos de sus compuestos característicos mediante 6 transformaciones que pertenecían a 4 vías diferentes. Sin embargo, aunque las transformaciones encontradas son biológicamente posibles, la falta de restricciones respecto a la reversibilidad de las mismas podría limitar la validez de las soluciones encontradas. El siguiente paso en el modelado consistirá en la incorporación de esta información. Además, aunque actualmente el algoritmo emplea una función de aptitud agregativa, resultaría interesante explorar otras estrategias evolutivas multi-objetivo para realizar la búsqueda, contemplando en el modelado la incorporación de un mayor número de compuestos de los agrupamientos.

Referencias

[1] F. Carrari, C. Baxter, B. Usadel, E. Urbanczyk-Wochniak, M. Zanor, A. Nunes-Nesi, V. Nikiforova, D. Centero, A. Ratzka, M. Pauly, L.J. Sweetlove, and A.R. Fernie. Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiology*, 142:1380–1396, 2006. doi: 10.1104/pp.106.088534.

[2] D. Croes, F. Couche, S.J. Wodak, and J. van Helden. Metabolic Pathfinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Research*, 33:W326–W330, 2005. doi: 10.1093/nar/gki437.

[3] K. Faust, D. Croes, and J. van Helden. Metabolic Pathfinding Using RPAIR Annotation. *Journal of Molecular Biology*, 388:390–414, 2009. doi: 10.1016/j.jmb.2009.03.006.

[4] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27.

[5] P. Kharchenko, D. Vitkup, and G.M. Church. Filling gaps in a metabolic network using expression information. *Bioinformatics*, 20:i178–i185, 2004. doi: 10.1093/bioinformatics/bth930.

[6] M. Kotera, Y. Okuno, M. Hattori, S. Goto, and M. Kanehisa. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society*, 126:16487–16498, 2004. doi: 10.1021/ja0466457.

[7] R. Küffner, R. Zimmer, and T. Lengauer. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16(9):825–836, 2000. doi: 10.1093/bioinformatics/16.9.825.

[8] V. Lacroix, L. Cottret, P. Thébault, and MF. Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5:594–617, 2008. doi: 10.1109/TCBB.2008.79.

[9] D.C. McShan, S. Rao, and I. Shah. PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics*, 19:1692–1698, 2003. doi: 10.1093/bioinformatics/btg217.

[10] D. Milone, G. Stegmayer, M. Gerard, L. Kamenetzky, M. López, and F. Carrari. Métodos de agrupamiento no supervisado para la integración de datos genómicos y metabólicos de múltiples líneas de introgresión. *Revista Iberoamericana de Inteligencia Artificial*, 13(44):56–66, 2009. `http://erevista.aepia.org/index.php/ia/article/viewFile/624/607`.

[11] Diego H. Milone, Georgina S. Stegmayer, Laura Kamenetzky, Mariana López, Je Min Lee, James J. Giovannoni, and Fernando Carrari. *omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. *Bioinformatics*, 11:438–447, 2010. doi: 10.1186/1471-2105-11-438.

[12] H. Ogata, S. Goto, W. Fujibuchi, and M. Kanehisa. Computation with the KEGG pathway database. *BioSystems*, 47:119–128, 1998. doi: 10.1016/S0303-2647(98)00017-3.

[13] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (3 Ed.)*. Prentice Hall, 2009.

[14] Kazuki Saito, Masami Y. Hirai, and Keiko Yonekura-Sakakibara. Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'. *Trends in Plant Science*, 13:36–43, 2008. doi: 10.1016/j.tplants.2007.10.006.

[15] S.N. Sivanandam and S.N. Deepa. *Introduction to Genetic Algorithms*. Springer, 2008.

[16] G. Stegmayer, D. Milone, L. Kamenetzky, M. López, and F. Carrari. Neural Network Model for Integration and Visualization of Introgressed Genome and Metabolite Data. In *Proceedings of the 2009 International Joint Conference on Neural Networks*, 2009. doi: 10.1109/IJCNN.2009.5179039.

[17] E. Urbanczyk-Wochniak, A. Luedemann, J. Kopka, J. Selbig, U. Roessner-Tunali, L. Willmitzer, and A.R. Fernie. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports*, 4(10):989–993, 2003. doi: 10.1038/sj.embor.embor944.