Búsqueda evolutiva de vías metabólicas

M. Gerard[1,2], G. Stegmayer[1] and 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

Resumen. Los métodos de búsqueda clásicos permiten encontrar secuencias de operadores para producir determinadas transiciones entre estados. En forma similar, los algoritmos evolutivos realizan la búsqueda usando operadores estadísticos y una función de aptitud que evalúa las características de los individuos, explorando 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 encontrar vías metabólicas entre dos compuestos seleccionados a partir de agrupamientos generados con el modelo IL-SOM. Se describen los operadores empleados, el desarrollo de la función de aptitud y se compara el desempeño del algoritmo propuesto con el de dos métodos clásicos de búsqueda.

Palabras clave: búsqueda, algoritmo evolutivo, vías metabólicas.

1. Introducción

La búsqueda de secuencias de acciones que conducen de un estado inicial a uno final especificado es una tarea habitual en muchos campos de investigación. Los algoritmos de búsqueda clásicos encuentran estas secuencias explorando el espacio de estados generando un árbol de búsqueda explícito. Cada nodo del árbol corresponde a un estado, cada arco es un operador que produce la transición entre dos estados y cada camino es una secuencia de estados conectados por operadores. Cuando sólo se emplea la definición del problema para realizar la búsqueda los algoritmos se denominan de búsqueda no informada. Dentro de este grupo se encuentran los algoritmos de búsqueda en amplitud (BA) y búsqueda en profundidad (BP). Su principal diferencia es el modo en que expanden los nodos durante la búsqueda. Cuando se emplea información específica del problema los algoritmos se denominan de búsqueda informada. Éstos realizan la búsqueda con la ayuda de una heurística, como por ejemplo A^*, para seleccionar el nodo a expandir [1].

Un enfoque alternativo lo constituyen los algoritmos evolutivos (AE), que se basan en el uso de operadores estadísticos y de una función de aptitud que evalúa las soluciones encontradas. Estos algoritmos se caracterizan por: su simplicidad

conceptual, debido a que sólo requieren de la definición de algunos operadores para realizar la búsqueda; pueden utilizarse para diversos problemas en diferentes dominios de aplicación; son fácilmente combinables con otros métodos (por ejemplo gradiente descendente luego de aplicar AE); permiten explorar múltiples puntos del espacio de búsqueda al mismo tiempo y pueden adaptarse fácilmente a cambios del entorno mientras evolucionan [2].

Se han propuesto diferentes estrategias de búsqueda para encontrar vías metabólicas que relacionen compuestos. PathComp [3] 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 [4] asigna a cada operador un costo igual al número de reacciones donde el compuesto participa. PathMiner [5] 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[3]. Recientemente se ha propuesto un modelo denominado IL-SOM [6,7] basado en mapas autoorganizados que permite encontrar agrupamientos a partir de la integración de datos de este tipo. Este modelo aplica el principio denominado "guilt-by-association" [8,9] 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[4], la reconstrucción de éstas a partir de los datos no es sencilla [10].

La motivación de este trabajo es desarrollar un algoritmo evolutivo para encontrar 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. Luego se definen medidas objetivas para cuantificar el desempeño de los algoritmos.

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.

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

[4] Una vía metabólica es una sucesión de reacciones químicas que transforman un sustrato en uno o varios productos a través de una serie de compuestos intermediarios.

2. Algoritmo propuesto

En esta sección se presenta el algoritmo propuesto, que denominaremos algoritmo evolutivo para la búsqueda de vías metabólicas (AEBVM). Primero se define el espacio de estados y los operadores de búsqueda 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 utilizada, se analizan los términos que la componen y se describe el efecto que cada uno produce sobre la búsqueda.

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 que deben excluirse de la búsqueda [11]. Sin embargo definiciones incorrectas pueden excluir compuestos necesarios para producir resultados de interés biológico. Un enfoque diferente fue propuesto en [12] 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[5], exceptuando los polímeros de glucosa, donde los operadores de búsqueda r describen las relaciones binarias permitidas entre compuestos de C. El compuesto sobre el cual se aplica el operador se denominará sustrato s, siendo p el producto o nuevo estado resultante de la transformación r. Los operadores de búsqueda 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 operadores r que transforma $\widehat{s}$ en $\widehat{p}$. 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 transformación.

2.1. Estructura de los cromosomas

La secuencia de operadores de búsqueda que transforman $\widehat{s}$ en $\widehat{p}$ 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 puede variar en el rango $[1, N_{máx}]$, donde $N_{máx}$ es un parámetro que limita el tamaño máximo de reacciones que puede contener la vía metabólica. Cuando el número de reacciones supera esta cota, el cromosoma se trunca para contener sólo las primeras $N_{máx}$ reacciones.

[5] `http://www.genome.jp/kegg/`

2.2. Operadores genéticos

En esta sección se describen los operadores diseñados para el AEBVM. Dados los requerimientos de esta aplicación en particular, ha sido necesario realizar diversas modificaciones a los operadores clásicos, que limitarían la convergencia del algoritmo de ser aplicados directamente. Para facilitar la explicación de estos se definen cuatro conjuntos de operadores de búsqueda. R^* contiene al conjunto completo de los operadores de búsqueda permitidos, $R^1 = \{r_i / r_i = (\widehat{s}, p_i)\} \wedge R^1 \subset R^*$ contiene sólo aquellos operadores que producen transformaciones sobre $\widehat{s}$, $R^N = \{r_i / r_i = (s_i, \widehat{p})\} \wedge R^N \subset R^*$ contiene todos los operadores de búsqueda que transforman algún compuesto en $\widehat{p}$ 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 $N_{inic} \leq N_{máx}$ para cada individuo, que indica el número máximo de genes que puede contener 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 un operador de búsqueda seleccionado al azar del conjunto correspondiente.

Selección. En este algoritmo se utilizó el método tradicional de la ruleta [2]. 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 dos modificaciones respecto del operador clásico. La cruza de dos padres genera un único hijo como descendencia y los puntos de cruza ϕ_1 y ϕ_2 en cada padre son seleccionados para generar en el hijo una reacción válida al empalmar el material genético. La probabilidad de cruza queda determinada por el parámetro p_{cruza}. Dados dos padres $\mathbf{c}_1$ y $\mathbf{c}_2$, los puntos de cruza ϕ_1 y ϕ_2 se elijen donde $\delta(s_i, p_j) = 1$, para $s_i \in \mathbf{c}_1$ y $p_j \in \mathbf{c}_2$. Cuando existe más de un valor de ϕ la elección se realiza aleatoriamente.

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

$$mut(r_i) = \begin{cases} r \in R^+ & \text{si } N = 1, \\ r \in R^1 & \text{si } N > 1 \ \wedge \ i = 1, \\ r \in R^N & \text{si } N > 1 \ \wedge \ i = N, \\ r \in R^* / p = s_{i+1} & \text{si } N > 1 \ \wedge \ u \leq 0{,}5, \\ r \in R^* / s = p_{i-1} & \text{si } N > 1 \ \wedge \ u > 0{,}5, \end{cases} \quad (2)$$

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 $[0, 1]$.

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.

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

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

donde δ es la función delta de Kronecker, que toma valor 1 cuando sus argumentos son iguales. Ésta varía en el rango $[0, 1]$, siendo 1 cuando todos los operadores están concatenados y los compuestos s_1 y p_N son los deseados.

Extremos válidos (E). Este término evalúa los operadores 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 $N_{máx}$.

Tasa de reacciones únicas (Q). Este término penaliza la repetición de operadores de búsqueda 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 operadores 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 $A(\mathbf{c}) = \alpha\,[V(\mathbf{c}) + \beta E(\mathbf{c}) + Q(\mathbf{c}) + I(\mathbf{c})]$, 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 reacciones según la probabilidad de ocurrencia, que está asociada directamente con la abundancia de los compuestos intervinientes.

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 10 y 100 reacciones.

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 [13]. 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 las reacciones químicas almacenan la información de las transformaciones empleando esta 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 $s \rightarrow p$ [14]. Luego del procesamiento de los datos se obtuvieron 5936 compuestos y 14346 relaciones químicas. 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 [15]. 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.

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

3.1. Medidas de evaluación

Para comparar los resultados obtenidos con los distintos algoritmos se midió el tiempo requerido para encontrar una vía metabólica (t)[6], el número de reacciones que contiene la vía (L) y el número de compuestos pertenecientes al agrupamiento del que forman parte los extremos (ψ) que se encuentran participando de la vía. En el caso del AEBVM, también se evaluó el número de generaciones empleado para encontrar una vía (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 cada medida 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 (ver más adelante). Sólo se calcula el valor medio para ψ (indicado como $\overline{\psi}$) debido a que en la mayoría de las vías metabólicas encontradas sólo participan los extremos, 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.

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

Tabla 2. Influencia de la tasas de mutación sobre las búsquedas realizadas por el AEBVM. $AEBVM_{10}$ corresponde a $p_{mut} = 10\,\%$; $AEBVM_{25}$ corresponde a $p_{mut} = 25\,\%$. El tiempo $\widehat{t}$ se expresa en segundos y L en número de reacciones. $|\Psi|$ 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
$G_{máx}$	$AEBVM_{10}$	1883	80520	39287	3000	28385	26495
	$AEBVM_{25}$	1146	791	1261	274	15651	10744
$G_{mín}$	$AEBVM_{10}$	113	19	223	3	255	116
	$AEBVM_{25}$	97	88	28	4	138	330
$\widehat{G}$	$AEBVM_{10}$	375	407	1095	78	1279	1079
	$AEBVM_{25}$	258	265	223	48	418	1009
$L_{máx}$	$AEBVM_{10}$	8	27	30	11	15	25
	$AEBVM_{25}$	11	9	9	7	19	17
$L_{mín}$	$AEBVM_{10}$	4	5	7	3	5	5
	$AEBVM_{25}$	4	5	5	3	5	5
$\widehat{L}$	$AEBVM_{10}$	5	6	9	3.5	8.5	7
	$AEBVM_{25}$	6	7	6	3	7	7
$\psi_{máx}$	$AEBVM_{10}$	3	4	2	4	3	3
	$AEBVM_{25}$	3	3	2	3	3	3
$\overline{\psi}$	$AEBVM_{10}$	2.3	2.4	2.0	2.6	2.1	2.1
	$AEBVM_{25}$	2.1	2.3	2.0	2.1	2.3	2.1
Λ	$AEBVM_{10}$	0.5	0.7	0.3	0.3	0.3	0.3
	$AEBVM_{25}$	0.5	0.5	0.3	0.3	0.3	0.3

3.2. Evaluación del AEBVM empleando dos tasas de mutación

Para evaluar la influencia de la tasas de mutación sobre el resultado de las búsquedas se emplearon probabilidades de 10 y 25 %, y los extremos definidos en la Tabla 1. La elección de estos valores elevados se justifica por el uso de elitismo, dado que garantiza la conservación del individuo más apto. En los experimentos se empleó una población de $P = 1000$, $N_{máx} = 100$, $N_{inic} = 50$, $N_{elite} = 1$ y $p_{cruza} = 80\,\%$. Los resultados obtenidos se presentan en la Tabla 2.

Analizando la tabla se observa que el $AEBVM_{10}$ requiere mayor número de generaciones que el $AEBVM_{25}$ para encontrar vías metabólicas y presenta una dificultad particular en algunos casos, como lo indica $G_{máx}$ para la búsqueda 2. Esta situación no ocurre con $G_{mín}$ donde se observa que en algunos casos el $AEBVM_{25}$ requiere más generaciones que el $AEBVM_{10}$. Estos resultados pueden deberse al modo en que trabaja el operador de mutación. Durante las primeras generaciones, el operador genera bloques de reacciones que aumentan la validez media de la población. Posteriormente, cuando el número de bloques por individuo es bajo, las mutaciones comienzan a modificar los bloques para que el operador de cruza pueda generar una solución. Como consecuencia, valores elevados en la tasa de mutación podrían acelerar ambas etapas del proceso y disminuir el número de generaciones.

En referencia al número de reacciones que contienen las vías encontradas se observa que para ambas tasas de mutación se obtienen resultados similares para

$\widehat{L}$ y $L_{mín}$. Esto último indicaría que el algoritmo tiende a buscar vías con pocas reacciones independientemente de la tasa de mutación.

Con respecto a $\psi_{máx}$ se observa que el algoritmo produce resultados similares para ambas tasas de mutación, resultado reproducido por $\overline{\psi}$ y Λ. Sólo se observan diferencias importantes en la búsqueda 2. Estos resultados son esperables debido a que la incorporación de compuestos del agrupamiento en la vía metabólica buscada no fue modelada en la función de aptitud, por lo que la presencia de más compuestos del agrupamiento puede considerarse un beneficio adicional. Por lo tanto, aunque el AEBVM_{10} encuentra en la mayoría de los casos vías con un mayor número de reacciones, requiere generalmente más generaciones para lograr resultados similares a el AEBVM_{25}. En consecuencia se empleará el AEBVM_{25} para las comparaciones con los algoritmos de búsqueda clásicos.

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

Para evaluar el desempeño del AEBVM se compararon las medidas obtenidas con los algoritmos BA y BP durante la búsqueda de vías metabólicas con los extremos indicados en la Tabla 1. Debido a que resultados preliminares indicaron que para BA $L_{máx} = 6$, la búsqueda se limitó a vías metabólicas de hasta 10 reacciones para hacer comparables los resultados de los tres algoritmos. Para el AEBVM se empleó $N_{inic} = 5$ y los demás parámetros como se definieron anteriormente. Los algoritmos BA y BP fueron modificados para incorporar un control de estados repetidos para eliminar vías metabólicas donde se produjeran bucles, ya que éstas carecen de interés biológico. Además se aleatorizó el orden en que se aplicaron los operadores de búsqueda durante los experimentos debido a que se carecía de información que indicara la mejor forma de aplicarlos. Con ésto se buscó obtener resultados que reflejaran el comportamiento promedio de los algoritmos. En el caso de BA se buscaron 12 vías metabólicas en cada experimento para proveer diversidad en el número de reacciones y generar alternativas que pudieran ser de interés desde el punto de vista biológico. Los resultados obtenidos con los distintos métodos se presentan en la Tabla 3.

Al analizar la tabla se observa que BP encuentra vías metabólicas en tiempos menores al segundo. Éstas poseen el máximo número de reacciones permitido y una dispersión muy baja que se aprecia en la similitud entre $L_{máx}$ y $L_{mín}$. También se observa que los resultados para $\psi_{máx}$, $\overline{\psi}$ y Λ son similares a los encontrados con los otros algoritmos, con excepción de un caso, e indican que no hay preferencia por incorporar en la vía compuestos adicionales del agrupamiento. Por lo tanto, aunque BP supera en velocidad a BA y al AEBVM_{25}, encuentra vías con el máximo número de reacciones fijado y permite descartarlo debido a que este factor condiciona fuertemente la utilidad práctica del resultado obtenido.

En referencia a BA se observa que $\widehat{t}$ presenta los valores más altos, con excepción de un caso. Además, las vías metabólicas encontradas presentan poca variabilidad en el número de reacciones pese a que en cada experimento se buscaron 12 vías para cada par de extremos. Como se aprecia en los valores de Λ,

Tabla 3. Comparación entre BA, BP y AEBVM$_{25}$, limitando la búsqueda a 10 reacciones. El tiempo $\widehat{t}$ se expresa en segundos y L en número de reacciones. $|\Psi|$ 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
$\widehat{t}$	AEBVM$_{25}$	292	300	812	78	810	457
	BA	1297	4694	5907	362	201	819
	BP	0.1	0.2	0.6	0.1	0.1	0.1
$L_{máx}$	AEBVM$_{25}$	7	8	9	6	7	8
	BA	5	6	6	4	6	6
	BP	10	10	10	10	10	10
$L_{mín}$	AEBVM$_{25}$	4	5	5	3	6	5
	BA	4	5	5	3	5	5
	BP	8	9	9	9	9	9
$\widehat{L}$	AEBVM$_{25}$	4	5	6	3	6	6
	BA	5	5	5	4	6	6
	BP	10	10	10	10	10	10
$\psi_{máx}$	AEBVM$_{25}$	3	3	2	3	3	2
	BA	3	3	2	4	3	2
	BP	5	2	2	4	3	2
$\overline{\psi}$	AEBVM$_{25}$	2.1	2.1	2.0	2.1	2.1	2.0
	BA	2.3	2.1	2.0	2.5	2.1	2.0
	BP	2.3	2.0	2.0	2.5	2.3	2.0
Λ	AEBVM$_{25}$	0.5	0.5	0.3	0.3	0.3	0.2
	BA	0.5	0.5	0.3	0.3	0.3	0.2
	BP	0.8	0.3	0.3	0.3	0.3	0.2

habría un fuerte condicionamiento hacia la búsqueda de vías con bajo número de reacciones, que son incapaces de contener todas las relaciones necesarias para explicar el agrupamiento. Respecto a los resultados del AEBVM$_{25}$ se observa que $\widehat{t}$ presenta valores menores que BA en la mayoría de los casos y encuentra vías que poseen un número similar de reacciones a las obtenidas con BA pero con una mayor dispersión de longitudes, lo que se traduce en un aumento en la diversidad de las vías metabólicas encontradas.

Si bien los resultados obtenidos en la comparación entre BA y el AEBVM$_{25}$ presentados en la Tabla 3 indican que el AEBVM encuentra en menor tiempo una mayor diversidad de vías metabólicas, extraer conclusiones a partir de estos datos podría ser válido sólo para el caso particular de $N_{máx} = 10$. Para determinar la influencia de este parámetro sobre el AEBVM se realizó una nueva comparación empleando los mismos parámetros que en la evaluación anterior pero modificando $N_{inic} = 50$ y $N_{máx} = 100$. Aunque vías metabólicas con éste número de reacciones podrían no ser relevantes desde el punto de vista biológico, el elevado valor de $N_{máx}$ puede entenderse como si se le permitiera al AEBVM realizar una búsqueda sin restricción en el número de reacciones. Los resultados obtenidos con ambos algoritmos se presentan en la Tabla 4. No se incluyen los resultados obtenidos para BP debido a que las vías encontradas contenían siempre 100 reacciones.

Tabla 4. Comparación entre BA y AEBVM$_{25}$, limitando la búsqueda a 100 reacciones. El tiempo $\widehat{t}$ se expresa en segundos y L en número de reacciones. $|\Psi|$ 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
$\widehat{t}$	AEBVM$_{25}$	159	170	143	38	229	551
	BA	1297	4694	5907	362	201	819
$L_{máx}$	AEBVM$_{25}$	11	9	9	7	19	17
	BA	5	6	6	4	6	6
$L_{mín}$	AEBVM$_{25}$	4	5	5	3	5	5
	BA	4	5	5	3	5	5
$\widehat{L}$	AEBVM$_{25}$	6	7	6	3	7	7
	BA	5	5	5	4	6	6
$\psi_{máx}$	AEBVM$_{25}$	3	3	2	3	3	3
	BA	3	3	2	4	3	2
$\overline{\psi}$	AEBVM$_{25}$	2.1	2.3	2.0	2.1	2.3	2.1
	BA	2.3	2.1	2.0	2.5	2.1	2.0
Λ	AEBVM$_{25}$	0.5	0.5	0.3	0.3	0.3	0.3
	BA	0.5	0.5	0.3	0.3	0.3	0.2

Al analizar la Tabla 4 se observa que el AEBVM$_{25}$ emplea menor tiempo que BA para encontrar vías similares, aunque con una mayor dispersión de longitudes, como se refleja en los valores $L_{máx}$ y $L_{mín}$. Esto probablemente se debe a que el aumento en $N_{máx}$ posibilita la exploración de un mayor número de relaciones químicas para el mismo P, debido a que permite generar bloques de reacciones de mayor tamaño, proporcionando resultados potencialmente más interesantes, desde el punto de vista biológico, y en un menor tiempo que BA. Por ejemplo, los fragmentos válidos de vías encontrados durante la búsqueda podrían brindar información acerca de nuevas relaciones entre rutas conocidas. Por otra parte las medidas de ψ y Λ presentan valores similares entre ambos métodos indicando que el número de compuestos incorporados en las vías por el AEBVM es independiente del valor $N_{máx}$ empleado. Adicionalmente, al comparar los valores de $\widehat{L}$ y $\overline{\psi}$ para el AEBVM$_{25}$ presentados en las Tablas 3 y 4 se observa que en ambos casos estas medidas resultan similares en todas las búsquedas y permite decir que el valor de $N_{máx}$ empleado sólo modifica el tiempo de búsqueda.

4. Conclusiones y trabajos futuros

En este trabajo se propuso un algoritmo evolutivo para encontrar vías metabólicas entre compuestos seleccionados a partir de dos agrupamientos generados con el modelo IL-SOM, para datos metabólicos y transcripcionales de tomate. Se describieron los operadores empleados, la función de aptitud y las medidas objetivas usadas para evaluar el AEBVM y los algoritmos clásicos BA y BP. Se observó que tasas de mutación elevadas aceleran la convergencia del AEBVM. En la comparación de los tres algoritmos se observó que BP encuentra vías que contienen el máximo número de reacciones permitido, lo que condiciona fuertemente su utilidad. BA empleó el mayor tiempo y encontró vías con el

menor número de reacciones posible y una baja dispersión de longitudes. En cambio, el AEBVM empleó menor tiempo para encontrar las vías y generó una mayor dispersión de longitudes, lo que podría resultar de interés para el posterior análisis biológico. Por otra parte, la similitud en los valores para la tasa de explicación indicó que los tres algoritmos incorporan igual número de compuestos del agrupamiento en las vías. Esto indicaría la necesidad de modificar la función de aptitud para modelar explícitamente la incorporación de compuestos pertenecientes a los agrupamientos en la vía encontrada. Por otro lado, sería importante realizar evaluaciones estadísticas de los resultados obtenidos para sustentar apropiadamente las conclusiones extraídas a partir de las comparaciones realizadas. Finalmente, aunque actualmente el algoritmo emplea una función de aptitud agregativa, resultaría interesante explorar otras estrategias evolutivas multi-objetivo para realizar la búsqueda.

Referencias

1. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach (3 Ed.). Prentice Hall (2009)
2. Sivanandam, S., Deepa, S.: Introduction to Genetic Algorithms. Springer (2008)
3. Ogata, H., et al.: Computation with the KEGG pathway database. BioSystems **47** (1998) 119–128
4. Croes, D., et al.: Metabolic Pathfinding: inferring relevant pathways in biochemical networks. Nucl. Acids Res. **33** (2005) W326–W330
5. McShan, D.C., et al.: PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics **19** (2003) 1692–1698
6. Stegmayer, G., et al: Neural network model for integration and visualization of introgressed genome and metabolite data. In: International Joint Conference on Neural Networks. (2009)
7. Milone, D., et al: 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) (2009) 56–66
8. Saito, K., et al.: Decoding genes with coexpression networks and metabolomics - majority report by precogs. Trends Plant Sci. **13** (2008) 36–43
9. Urbanczyk-Wochniak, E., et al.: Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. EMBO reports **4**(10) (2003) 989–993
10. Kuffner, R., et al.: Pathway analysis in metabolic databases via differential metabolic display (DMD). Bioinformatics **16**(9) (2000) 825–836
11. Kharchenko, P., et al.: Filling gaps in a metabolic network using expression information. Bioinformatics **20** (2004) i178–i185
12. Kotera, M., et al.: Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. J. Am. Chem. Soc. **126** (2004) 16487–16498
13. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucl. Acids Res. **28**(1) (2000) 27–30
14. Faust, K., et al.: Metabolic pathfinding using rpair annotation. J. Mol. Biol. **388** (2009) 390–414
15. Carrari, F., et al.: 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 Physiol. **142** (2006) 1380–1396