

Un algoritmo transgénico vs algoritmos de optimización

CORREA-VILLALÓN, Christian José†, LUNA-RAMÍREZ, Enrique*, SORIA-CRUZ, Jorge y CRUZ-VALENZUELA, Roberto

Instituto de Educación de Aguascalientes. San Luis Potosí 601, Ojocaliente. Aguascalientes, Ags. C.P. 20256
Instituto Tecnológico El Llano Aguascalientes. Km. 18 Carr. Ags.-S.L.P. El Llano Aguascalientes. C.P. 20330

Recibido Octubre 29, 2015; Aceptado Marzo 12, 2016

Resumen

En este artículo se presenta un algoritmo transgénico diseñado para resolver problemas de optimización basándose en el tratamiento de funciones engañosas. Para ello, se incorporan elementos novedosos al algoritmo transgénico: homología de genes, alineamiento de cromosomas, resolución positiva-negativa y replicación de cultivos. De esta manera, es posible encontrar los individuos más aptos de una población, lo cual es discutido mediante la presentación de los resultados obtenidos de la aplicación del algoritmo a un grupo de funciones engañosas, cuya finalidad es precisamente demeritar la eficacia de las técnicas heurísticas. Tales resultados fueron comparados con resultados de otros algoritmos de optimización en el estado del arte, haciéndose evidente la eficiencia de nuestro algoritmo y concluyéndose que éste no requiere de poblaciones grandes, ni de un gran número de generaciones para poder obtener buenos resultados.

Algoritmos transgénicos, Problemas de optimización

Abstract

In this paper, a transgenic algorithm designed to solve optimization problems, based on the treatment of deceptive functions, is presented. To deal with it, innovative elements are incorporated to the transgenic algorithm: homology of genes, alignment of chromosomes, positive-negative resolution and crop replication. In this way, it is possible to find the best individuals of a population, as discussed through the presentation of the results obtained from the algorithm application to a group of deceptive functions, which purpose is just to break the effectiveness of heuristic techniques. Such results were compared with the results of other optimization algorithms in the state of art, which allowed to show the efficiency of our algorithm, concluding that it does not require big populations, neither a big number of generations in order to obtain good results.

Transgenic algorithms, Optimization problems

Citación: CORREA-VILLALÓN, Christian José, LUNA-RAMÍREZ, Enrique, SORIA-CRUZ, Jorge y CRUZ-VALENZUELA, Robert. Un algoritmo transgénico vs algoritmos de optimización. Revista de Análisis Cuantitativo y Estadístico. 2016. 3-7: 53-61

*Correspondencia al Autor (Correo Electrónico: elunaram@hotmail.com)

† Investigador contribuyendo como primer autor.

Introducción

La inspiración para la creación de la heurística propuesta en este artículo es la Biotecnología, sobre la que existen diversas definiciones, siendo una de las más relevantes la definición mencionada en la Convention on Biological Diversity (1993), abordada ampliamente por Celesti et al. (2016) en su tratado sobre sistemas computacionales aplicados al procesamiento masivo de secuencias de DNA. Esta definición enuncia que “La Biotecnología es cualquier aplicación tecnológica que haga uso de sistemas biológicos, organismos vivos o derivados, para crear o modificar productos o procesos para un uso específico”. Un ejemplo típico de esto es la manipulación genética de organismos vivos, razón por la cual se les conoce como organismos transgénicos (Villalobos, 2007). Es importante mencionar que los cambios aplicados a un organismo transgénico se realizan buscando un objetivo específico, relacionado con la mejora de algunas de sus características.

En la actualidad existen diversos ejemplos de organismos transgénicos, que pueden ir desde modificaciones genéticas en cosas tan comunes como los cereales que consumimos a diario, hasta la modificación genética de una planta de tabaco, a la cual se le puede agregar el gen que permite generar luz a las luciérnagas, de manera que al entrar en contacto con una sustancia específica, la planta pueda emitir una radiación de luz (Ow et al., 1986). Así, el primer algoritmo transgénico (Correa et al., 2007) fue diseñado con base en el proceso de mejoramiento genético en las plantas. De esta manera, en un primer intento por emular este proceso, sólo fueron usadas algunas bases teóricas. No obstante, en la siguiente sección, se describe una versión mejorada del algoritmo transgénico, que considera factores significativos en la producción de transgénicos, que no habían sido considerados en trabajos anteriores.

Algoritmo Transgénico con Homología

Apoyándose en el uso de técnicas para la creación de individuos modificados genéticamente (Ignacimuthu, 2008), se diseñó una nueva versión del algoritmo transgénico, mostrada en la figura 1. Como se puede observar en la figura, este algoritmo incorpora nuevas características tales como la selección elitista inversa, que, a diferencia de la selección elitista, permite determinar los individuos menos aptos en una población, cuyos genes necesariamente desaparecerán en generaciones posteriores. De esta manera, las subpoblaciones conformadas por porcentajes de las poblaciones de los individuos más aptos y menos aptos, conocidas como población positiva y población negativa, respectivamente, son sometidas a la técnica de homología de genes, que consiste en un alineamiento de cromosomas mediante el cual es posible identificar el dominio de un gen sobre otros. Esta técnica es precisamente la que le da el nombre a nuestro algoritmo, Algoritmo Transgénico con Homología (ATH).

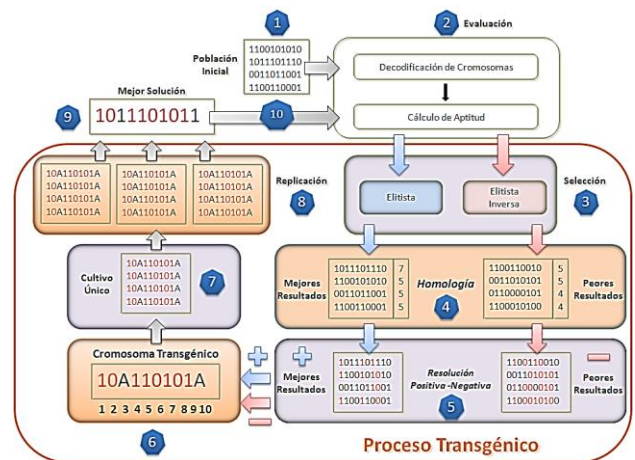


Figura 1 Algoritmo Transgénico con Homología

Antes de proceder a la descripción del ATH, es preciso resaltar otras características importantes del mismo, como es el caso de la combinación de las llamadas resolución positiva y resolución negativa, que al operar juntas, se generan mejores resultados, dado que en una misma generación, los genes que producen buenos resultados son preservados, mientras que los genes que producen malos resultados son rechazados (Correa et al., 2010).

La replicación de cultivos en una misma generación es otra característica importante del ATH, ya que permite llegar más rápidamente a la solución óptima, como se discutirá en la presentación de resultados.

Creación de la población inicial. La población inicial para comenzar a operar el ATH se crea en forma totalmente aleatoria, como sucede con otras técnicas heurísticas, por ejemplo, los algoritmos genéticos (Goldberg, 1989).

Así, en el paso 1 de la figura 2, se pueden observar soluciones iniciales conformadas por ceros y unos.

Evaluación. En esta etapa se evalúa la población inicial, obteniéndose para cada individuo el nivel de satisfacción de la función objetivo (la función que será optimizada). En nuestro caso, se optó por utilizar la función conocida como OneMax (Schaffer & Eshelman, 1991), dado que se trata de una función típica para medir la eficiencia de heurísticas, la cual básicamente mide la máxima cantidad de elementos "1" en cada individuo. Esto implica que los datos contenidos en cada individuo deben ser decodificados, de manera que sea posible determinar dicha cantidad. Hasta este punto, el ATH opera de manera similar a otras técnicas heurísticas; no obstante, en lo sucesivo se mostrará su parte original.

Selección. En el mundo de las técnicas heurísticas, existen diferentes tipos de selección de individuos. Para el caso del ATH, se optó por una selección elitista, cuyo objetivo es identificar a los individuos más aptos de una población. Así, si se tratara de un algoritmo genético, los individuos más aptos serían aquellos que mejor satisfacen la función objetivo, por lo que serían usados para realizar una cruce entre ellos y poder generar nuevos individuos con mejores características. En nuestro algoritmo, la selección elitista juega el mismo papel, excepto que los individuos no son cruzados entre sí, sino más bien son sometidos a un alineamiento de cromosomas, como se verá más adelante. Sin embargo, la principal diferencia respecto a otras heurísticas radica en lo que hemos denominado selección elitista inversa, cuyo objetivo es identificar a los individuos menos aptos, entendiéndose éstos como aquellos que se alejan de una posible solución.

Homología. Una vez identificadas las poblaciones de individuos más aptos e individuos menos aptos, se derivan de ellas subpoblaciones tomando un porcentaje de cada una, comúnmente entre el 30% y 40%, dependiendo del tamaño del problema (Correa et al., 2010). Estas subpoblaciones, conocidas como población positiva y población negativa, son sometidas a la técnica de homología de genes. En un plano epistemológico, homología se refiere a la correspondencia que existe entre dos o más rasgos de diferentes organismos, es decir, a la similitud que existe entre ellos. No es una relación directamente observable, sino más bien es una relación que se infiere a través de algún método de análisis.

Así, la homología es usada en el proceso de creación de organismos transgénicos para detectar patrones existentes en los genes, tanto de organismos receptores de nuevos genes, como de organismos donadores.

Con base en esto, es posible determinar qué genes conviene modificar y qué genes no deben de ser modificados en un organismo. En el contexto computacional, la homología también tiene el propósito de detectar patrones en los cromosomas de los individuos de una población, lo cual es realizado mediante un alineamiento de cromosomas, cuya idea básica es representar y comparar dos o más secuencias/cadenas de ADN u otras estructuras proteicas primarias con el objeto de resaltar sus zonas de similitud, mismas que podrían indicar relaciones funcionales o evolutivas entre genes (De Luna & Mishler, 1996). Bajo esta idea, los cromosomas (secuencias de ceros y unos) son almacenados como filas de una matriz, en los cuales se insertan espacios para que las zonas con estructuras idénticas o similares se alineen.

En la figura 2 se presenta un típico alineamiento de cromosomas de la población positiva, en la cual se realiza la búsqueda de patrones mediante el recorrido de cada una de las columnas que conforman esta población. Como se puede observar en la figura, solo fueron seleccionados cuatro de los diez individuos de la población, resaltándose sus genes dominantes con color rojo y encerrados en círculo, identificados con base en el mayor número de presencias (ceros o unos) en cada columna. Por ejemplo, la primera columna contiene un gen dominante correspondiente al número “1”, por lo que éste prevalecerá en el siguiente cultivo en esa misma posición. Por el contrario, un ejemplo donde no se observa un gen dominante es en la segunda columna, en la cual existe la misma cantidad de ceros y unos, por lo que se tendrá un valor aleatorio en esta posición en el siguiente cultivo. Este criterio se aplica por igual a las otras columnas donde no existe un gen dominante (columnas 3, 4, 6, 9 y 10).

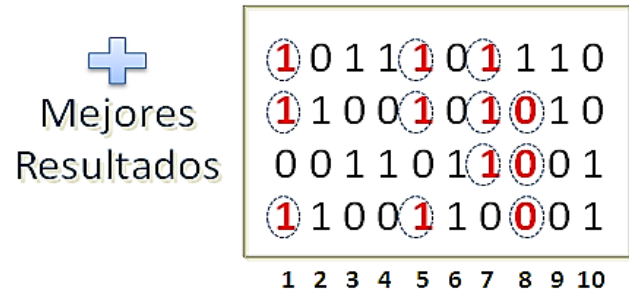


Figura 2 Alineamiento positivo de cromosomas

Al finalizar el alineamiento de cromosomas en la población positiva, se obtiene un cromosoma transgénico positivo o vector K, utilizado para crear el siguiente cultivo. En la figura 3 se muestra dicho cromosoma, el cual se mantiene el gen dominante “1” en las posiciones 1, 5, 7 y el gen dominante “0” en la posición 8.

Para las otras posiciones, como se mencionó, se generan valores aleatorios, indicados en la figura con la letra “A”.

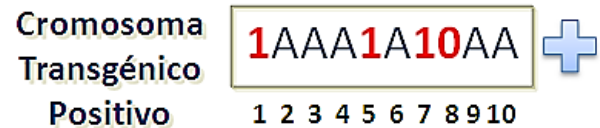


Figura 3 Cromosoma resultante del alineamiento

El alineamiento de cromosomas sobre una población negativa sigue el mismo procedimiento descrito anteriormente, por lo que el resultado también es un cromosoma transgénico, llamado en este caso cromosoma transgénico negativo o vector K' (véase figura 4). Sin embargo, contrario al caso del vector k, los genes dominantes de este cromosoma serán rechazados en la siguiente generación, debido a que éstos representan resultados no deseados al haber sido derivados de una población negativa.

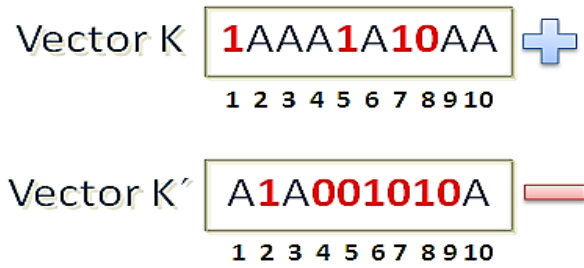


Figura 4 Cromosomas positivo y negativo juntos

Replicación. Una vez creado el cromosoma transgénico, se procede a generar el siguiente cultivo con base en este cromosoma. En esta etapa, el ATH echa mano del principio usado en Biotecnología para generar una población transgénica: para obtener mejores resultados a lo largo de nuevas generaciones, se puede utilizar un único cultivo o se pueden utilizar réplicas de un mismo cultivo. De esta manera, la mejor solución encontrada en una generación es incorporada a la siguiente generación, repitiéndose el algoritmo a partir del segundo paso hasta llegar a la mejor solución posible o, en su defecto, a la condición especificada para detener el algoritmo.

Resultados

En esta sección se discuten los resultados obtenidos de someter el ATH al tratamiento de funciones engañosas, cuya finalidad es precisamente demeritar la eficacia de técnicas heurísticas. De manera concreta, estas funciones permiten indagar la forma en qué los algoritmos orientados a la optimización fallan en la determinación de óptimos globales.

Las funciones engañosas están asociadas a lo se conoce como decepción, que se refiere al estado de reducción en la aptitud de un individuo generado por una combinación de buenos elementos constructores, cuando la lógica indicaría lo contrario, un aumento en la aptitud.

Este fenómeno fue sugerido originalmente por Goldberg (1989) para explicar el mal desempeño de los algoritmos genéticos en algunos problemas de optimización. Para comprender mejor esto, considérese el siguiente ejemplo relativo a una función de aptitud (función engañosa), cuyo dominio está compuesto por cadenas binarias de longitud 3 y su contradominio son propiamente los valores de aptitud:

Cadena	Aptitud	Cadena	Aptitud
000	70	100	30
001	50	101	2
010	49	110	3
011	1	111	80

Tabla 1

En este ejemplo se puede observar que la tendencia es a tener una mayor aptitud, a medida que una cadena contiene más ceros; no obstante, el óptimo global se localiza en la cadena que contiene sólo unos. En este caso, un algoritmo genético tendería típicamente a ir seleccionando cadenas con más ceros en la búsqueda de una solución óptima, por lo que en la cadena “111”, que es propiamente el óptimo global en este ejemplo, nunca sería determinada como tal. Con base en estas consideraciones, el ATH fue sometido a diversas funciones engañosas y los resultados obtenidos fueron comparados con los resultados de dos algoritmos trascendentales en el estado del arte, a saber, un algoritmo evolutivo que utiliza mapas de Markov para modelar soluciones prometedoras y un algoritmo genético con varios tipos de cruzamiento. En el Evolutionary Algorithm Based on a Markov Graphical Model Selection of Promising Solutions (Ponce et al., 2004), al cual se hará referencia como algoritmo evolutivo, los autores proponen una variante del algoritmo genético, que básicamente consiste en realizar la etapa de selección usando mapas de Markov.

Este algoritmo fue sometido a diversas funciones engañosas, en particular a la función conocida como FC₂ (Mühlenbein et al., 1998), considerada en la actualidad como una de las funciones con mayor grado de complejidad, razón por la cual fue seleccionada para realizar nuestro análisis comparativo entre algoritmos. La función FC₂ se define de la siguiente manera:

$$FC_2 = \sum_{j=1}^m F^5_{muhl}(s_j)$$

donde $s_j = (x_{5j-4}, x_{5j-3}, x_{5j-2}, x_{5j-1}, x_{5j})$ (1)

$$F^5_{muhl}(x) = \begin{cases} 3.0 & \text{for } x = (0,0,0,0,1) \\ 2.0 & \text{for } x = (0,0,0,1,1) \\ 1.0 & \text{for } x = (0,0,1,1,1) \\ 3.5 & \text{for } x = (1,1,1,1,1) \\ 4.0 & \text{for } x = (0,0,0,0,0) \\ 0.0 & \text{otherwise} \end{cases}$$

En la siguiente tabla se presentan los parámetros de experimentación, habiéndose utilizado los mismos valores para el tamaño de cadena y el tamaño de población para ambos algoritmos, esto con la finalidad de que operen en condiciones similares:

Algoritmo Evolutivo	
Problema	Valores para FC ₂
Tamaño de Cadena	10
Tamaño de Población	100,300
Porcentaje de Selección	70%
ATH	
Problema	Valores para FC ₂
Tamaño de Cadena	10
Tamaño de Población	100,300
% sel. positiva, negativa	2%, 2%
Porcentaje de dureza	90%

Tabla 2

Nótese que en la tabla anterior aparece el concepto de dureza, el cual se refiere al criterio (porcentaje) utilizado para determinar si un gen es dominante. Así, un 90% de dureza se refiere a que el número de veces de aparición de un elemento (“0” o “1”) en una columna debe ser al menos el 90% del total de elementos en esa columna para que pueda ser considerado como un gen dominante. Los parámetros anteriores produjeron los siguientes resultados:

Algoritmo Evolutivo		
Problema	FC ₂	FC ₂
Tamaño de Cadena	10	10
Tamaño de Población	100	300
Medida de generaciones	24	18
ATH		
Problema	FC ₂	FC ₂
Tamaño de Cadena	10	10
Tamaño de Población	100	300
Medida de generaciones	6	2

Tabla 3

En esta tabla, la media de generaciones es el resultado que permite medir la eficiencia de los algoritmos, la cual se refiere al número de generaciones que se requirieron para alcanzar el óptimo de la función FC₂. Así, los resultados muestran que el algoritmo ATH es significativamente más eficiente que el algoritmo evolutivo, ya que para ambos tamaños de población (100 y 300), se requirieron significativamente menos generaciones para alcanzar el óptimo.

Heywood (2006) muestra el comportamiento de un algoritmo genético ante un problema engañoso usando diferentes tipos de cruzamiento. El problema engañoso usado por Heywood, relativo a la teoría de bloques constructores, se muestra en la siguiente tabla:

$F(0^{**}) > f(1^{**})$	$F(00^*) > f(11^*), f(01^*), f(10^*)$
$F(*0^*) > f(*1^*)$	$F(0^*0) > f(1^*1), f(0^*1), f(1^*0)$
$F(**0) > f(**1)$	$F(*00) > f(*11), f(*01), f(*10)$

Tabla 4

Para el caso particular de orden 3, es decir, cuando se utilizan cadenas de 3 bits en el dominio de la función f , el problema asume los valores mostrados en la siguiente tabla:

$f(000)=28$	$f(001)=26$
$f(010)=22$	$f(100)=14$
$f(110)=0$	$f(011)=0$
$f(101)=0$	$f(111)=30$

Tabla 5

En la siguiente tabla se presentan los parámetros de experimentación, habiéndose incrementado a 30 el número de bits en las cadenas (Bloques Constructores con 10 subcadenas de Orden 3 – $BCO^{3(10)}$), aumentándose por tanto de manera significativa la complejidad del problema:

Algoritmo Genético	
Problema	$BCO^{3(10)}$
Tamaño de cadena	30
Tamaño de población	10,200,2000
Cruzamiento	1 Punto, Uniforme, “Tagged”
Evaluaciones	300,10000,50000
ATH	
Problema	Valores para FC_2
Tamaño de la cadena	30
Tamaño de la población	10,200,2000
%sel. Positiva, negativa	2%, 2%
Evaluaciones	300,10000,50000
Número de cultivos	2
Porcentaje de dureza	90%

Tabla 6

Los resultados obtenidos para ambos algoritmos, el genético y el ATH, fueron los siguientes:

Algoritmo Genético			
Problema	$BCO^{3(10)}$	$BCO^{3(10)}$	$BCO^{3(10)}$
Tamaño de la población	10	200	2000
Evaluaciones	300	10000	50000
Cruzamiento		1 Punto	1 Punto
Porc Resolución		27%	38%
Cruzamiento	No reporta	Uniforme	Uniforme
Porc Resolución		27%	35%
Cruzamiento		“Tagged”	“Tagged”
Porc Resolución		53%	64%
ATH			
Problema	$BCO^{3(10)}$	$BCO^{3(10)}$	$BCO^{3(10)}$
Tamaño de población	10	200	2000
Evaluaciones	300	10000	50000
Porc Resolución	60%	70%	100%

Tabla 7

Como lo muestran los resultados, también en este caso el ATH es más eficiente en la determinación de la solución óptima, al haber logrado siempre un mayor porcentaje de resolución del problema, llegando incluso a una resolución del 100% con un tamaño de población y un número de evaluaciones suficientes (2000 y 50000, respectivamente).

Conclusiones y trabajo futuro

Se presentó un algoritmo transgénico (ATH) diseñado para resolver problemas de optimización tomando como base el tratamiento de funciones engañosas, las cuales permiten indagar la forma en que los algoritmos orientados a la optimización fallan en la determinación de óptimos globales. El diseño de un primer algoritmo transgénico estuvo basado en el proceso de mejoramiento genético en las plantas, de manera que en un primer intento por emular este proceso, sólo fueron usadas algunas bases teóricas. No obstante, en este trabajo, se describió una versión ampliada y mejorada del algoritmo, que considera otros factores importantes en la producción de transgénicos, tales como homología de genes, alineamiento de cromosomas, resolución positiva-negativa y replicación de cultivos.

CORREA-VILLALÓN, Christian José, LUNA-RAMÍREZ, Enrique, SORIA-CRUZ, Jorge y CRUZ-VALENZUELA, Robert. Un algoritmo transgénico vs algoritmos de optimización. Revista de Análisis Cuantitativo y Estadístico. 2016

Nuestro algoritmo fue sometido a diversas funciones engañosas y los resultados obtenidos fueron comparados con los resultados de dos algoritmos recientes y de trascendencia en el estado del arte: un algoritmo evolutivo que utiliza mapas de Markov para modelar soluciones prometedoras y un algoritmo genético con varios tipos de cruzamiento. Un análisis comparativo mostró que el ATH es significativamente más eficiente que ambos algoritmos, ya que en la comparación con el algoritmo evolutivo, se observó que se requieren significativamente menos generaciones para llegar a la solución óptima, además de que la razón de eficiencia del ATH aumentó al aumentar el tamaño de la población inicial. Respecto a la comparación con el algoritmo genético, el ATH logró siempre un mayor porcentaje de resolución, llegando incluso a una resolución del 100% utilizando un tamaño de población y un número de evaluaciones suficientes, lo que al algoritmo genético nunca logró (su resolución máxima fue del 64%). Además, los resultados mostraron que el ATH no requiere de un tamaño de población grande para lograr buenos resultados. Como trabajo futuro, se tiene considerado realizar más análisis comparativos con algoritmos de optimización que operan bajo la filosofía de colonia de hormigas (Dorigo & Stützle, 2004), que cuenta con una aceptación importante en la actualidad debido a desempeño en problemas de optimización combinatoria.

Referencias

- Convention on Biological Diversity. Use of Terms. United Nations Treaty Series, Vol. 1760 I-30619. 1993.
- Celesti, A., Fazio, M., Celesti, F., Sannino, G., Campo, S., Villari, M. New Trends in Biotechnology: The Point on NGS Cloud Computing Solutions. En: IEEE Workshop on ICT solutions for eHealth 2016.
- Correa, C. J., Padilla, F., Pérez, E. & Quezada, F. Algoritmo Transgénico para resolver el problema del TSP. En: Tercer Congreso Mexicano de Computación Evolutiva COMCEV' 07. Aguascalientes, México, 2007, pp. 75-78.
- Correa, C. J., Padilla, F., Pérez, E., Ponce de León, E., Quezada, F., Ponce, J. & Padilla, A. New Algorithm Transgenic with Homology to Solve the Problem of OneMax. Revista Computación y Sistemas del CIC-IPN, 13(4): 3-8. Abril-Junio 2010.
- De Luna, E. & Mishler, B.D. El concepto de homología filogenética y la selección de caracteres taxonómicos. Boletín de la Sociedad Botánica de México, 59(1): 131-146. Mayo 1996.
- Dorigo, M. & Stützle, T. Ant Colony Optimization. MIT Press. 2004.
- Goldberg, D.E. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley. 1989.
- Heywood, M.I. Deception in Genetic Search. <http://web.cs.dal.ca/~mheywood/CSCI6506/Handouts/N04-Deception.pdf>. 2006. Sitio visitado el 31 de agosto de 2016.
- Ignacimuthu, S. Biotechnology: An Introduction. Alpha Science International, Ltd. 2008.
- Mühlenbein, Mahnig, and Rodriguez O. Schemata, distributions and graphical models in evolutionary optimization. Journal of Heuristics, 5: 215-247. 1998.

Ow, D. W., de Wet, J. R., Helinski, D. R., Howell, S. H., Wood, K. V. & de Luca, M. Transient and stable expression of the firefly luciferase gene in plant cells and transgenic plants. *Science*, 234(4778): 856-859. November 1986.

Ponce, E., Díaz E., & Padilla, F. Evolutionary Algorithm based on a Markov graphical model selection of promising solutions. En: *International Conference on Computing*. Mexico City, Mexico, 2004, pp. 3-9.

Schaffer, J. D. & Eshelman, L. J. On crossover as an evolutionary viable strategy. En: *Proceedings of the 4th International Conference on Genetic Algorithms*. Morgan Kaufmann, 1991, pp. 61-68.

Villalobos, V. *Los Transgénicos: oportunidades y amenazas*. Mundi-Prensa (Madrid, España). 2007.