



universidad
de león



FACULTAD DE CIENCIAS BIOLÓGICAS Y AMBIENTALES

**OPTIMIZACIÓN DE MEDIOS DE CULTIVO
MEDIANTE ALGORITMO HÍBRIDO GENÉTICO
CON REGRESIÓN MÚLTIPLE**

**OPTIMIZATION OF CULTURE MEDIA USING
MULTIPLE REGRESSION AND GENETIC HYBRID
ALGORITHM**

Autor: Miriam Velasco Díez

GRADO EN BIOTECNOLOGÍA

Junio, 2020

Índice

1. Introducción	1
2. Objetivo y metodología	4
3. Desarrollo del planteamiento	6
3.1. Función de ajuste a los datos conocidos	6
3.1.1. Regresión por mínimos cuadrados (wxMaxima)	6
3.1.2. Búsqueda aleatoria (Perl)	9
3.2. Optimización mediante Algoritmos Genéticos	10
3.2.1. Operador de selección	11
3.2.2. Operador de sobrecruzamiento/recombinación	13
3.2.3. Operador de mutación	13
3.2.4. Algoritmo Híbrido Genético con Regresión Múltiple (Perl)	15
4. Resultados	17
4.1. Función de ajuste por mínimos cuadrados	17
4.1.1. Medio de cultivo de <i>Nannochloropsis gaditana</i>	18
4.1.2. Hidrólisis enzimática de plasma bovino	19
4.2. Función de ajuste por búsqueda aleatoria	20
4.2.1. Medio de cultivo de <i>Nannochloropsis gaditana</i>	20
4.2.2. Hidrólisis enzimática de plasma bovino	22
4.3. Algoritmo Híbrido Genético con Regresión Múltiple	22
4.3.1. Medio de cultivo de <i>Nannochloropsis gaditana</i>	23
4.3.2. Hidrólisis enzimática de plasma bovino	25
4.4. Análisis estadístico	26
5. Conclusión y perspectivas futuras	28
6. Referencias	29
Anexo I. Simulaciones para medio de cultivo de <i>Nannochloropsis gaditana</i>	31
Anexo II. Simulaciones para hidrólisis enzimática de plasma bovino	39
Anexo III. Especificaciones de los programas desarrollados	41

RESUMEN

Se ha implementado un Algoritmo Híbrido Genético para optimizar dos procesos en paralelo. En el primero, se ha buscado lograr la máxima productividad, al menor coste, de *Nannochloropsis gaditana*, en base a la elección de un medio de cultivo cuyos componentes se encontrasen en la concentración adecuada. En el segundo, se ha realizado la optimización de las condiciones (pH, temperatura y relación enzima/sustrato) de la reacción de hidrólisis enzimática a partir de proteínas de plasma bovino. Para ello, se han empleado los datos experimentales de estudios previos para hallar mediante regresión múltiple la función de aproximación de datos, necesaria para el funcionamiento del algoritmo. También se ha desarrollado un programa para encontrar dicha función mediante búsqueda aleatoria, minimizando el error entre el resultado real y el obtenido por la función. Se ha logrado obtener una productividad de $0,86 \text{ g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$ para *N. gaditana* a un precio de 0,13 €/L, del mismo modo que se ha conseguido un 23,56 % de hidrólisis enzimática a partir de plasma bovino. La regresión múltiple ha resultado ser un método adecuado para problemas a optimizar con bajo número de variables y elevado número de muestras, mientras que el algoritmo de búsqueda aleatoria ha sido apto para múltiples variables con menor cantidad de muestras. El Algoritmo Genético ha mostrado gran capacidad de optimización en ambas soluciones.

ABSTRACT

A Genetic Hybrid Algorithm has been implemented to optimize two processes at the same time. On the one hand, we have looked for to achieve *Nannochloropsis gaditana* maximum productivity, at the lowest cost, based on the suitable components concentration choice of the culture medium. On the other hand, we have optimized the enzymatic hydrolysis conditions (pH, temperature and enzyme/substrate ratio) from bovine plasma proteins. Fitness function, required for the algorithm, has been found by multiple regression from experimental data from previous studies. We have also developed a program to find a fitting function through random search, by error minimization between the real result and the function result. We have obtained a productivity of $0.86 \text{ g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$ for *N. gaditana* at a price of € 0.13 per litre, in the same way that 23.56% of enzymatic hydrolysis has been achieved from bovine plasma. Multiple regression has turned out a suitable method for optimization problems with low number of variables and high number of samples, while we have been able to use our program for multiple variables problems with fewer samples. The Genetic Algorithm has shown great optimization capacity in both solutions.

Palabras clave: Algoritmo de búsqueda aleatoria, Algoritmo genético, Función de aptitud, Medios de cultivo, Optimización, Regresión múltiple.

Keywords: Culture media, Fitness function, Genetic algorithm, Multiple regression, Optimization, Random search algorithm.

1. INTRODUCCIÓN

En términos generales, se denomina optimización al proceso basado en el descubrimiento del conjunto de condiciones requeridas para lograr la mejor solución en una situación dada (Venkateswarlu y Jujjavarapu, 2020b).

Desde el punto de vista de la química, los medios de cultivo celular son una compleja mezcla de moléculas utilizadas para el crecimiento y el metabolismo celular. Para su elaboración, es importante decidir su composición (y la concentración óptima de cada molécula) en base a los requisitos metabólicos de cada línea celular, lo que conlleva una experimentación iterativa que consume una gran cantidad de tiempo, persiguiendo la optimización de la formulación del medio. Asimismo, se plantean otros desafíos como las diferencias de concentración, en cuanto a órdenes de magnitud, entre los componentes de los medios que dificultan su precisa medición y la caracterización de los medios (Galbraith *et al.*, 2018).

Los medios definidos son extremadamente complejos de realizar, debido a la incertidumbre en torno a los requisitos de las diferentes líneas celulares y al efecto de los componentes sobre dichas líneas, así como por la gran cantidad de componentes en los medios. Por ejemplo, el agotamiento de un aminoácido esencial puede conllevar el uso de un aminoácido diferente en su lugar. De igual modo, los cofactores bioquímicos (como los iones) son necesarios para las funciones celulares, pues pueden ocurrir anomalías debido al agotamiento o exceso de estos. La optimización de la formulación de medios debe realizarse para cada proceso individual, ya que los medios óptimos para una línea celular no son siempre adecuados para otra. La decisión sobre qué estrategia utilizar para la optimización de los medios de cultivo depende de muchos factores, como el tipo de producto generado, el tipo de célula y el tipo de proceso (continuo, discontinuo o semicontinuo) (Chen *et al.*, 2011; Galbraith *et al.*, 2018). Las tecnologías ómicas colaboran en la comprensión del comportamiento celular (distinta expresión génica en diferentes condiciones) y, por tanto, ayudan a la predicción de formulación de medios mediante el conocimiento de las rutas celulares y metabólicas (Galbraith *et al.*, 2018; Mahajan *et al.*, 2019).

Existen varias prácticas heurísticas llevadas a cabo en la industria mas no se ha desarrollado aún un enfoque universal para la formulación y optimización de medios (figura 1). Por un lado, se establecen las estrategias cerradas con un número fijo de componentes del medio cuyas concentraciones óptimas hay que averiguar. En estos casos, se optimiza rápidamente pero no se encuentran aquellos componentes que pueden ser beneficiosos. En cambio, las estrategias abiertas no limitan la cantidad de componentes por lo que el proceso de

optimización es tan tedioso que resulta inviable. El enfoque tradicional se basa en utilizar un medio inicial fijo y partir de estudios anteriores de otros investigadores (relacionados con la línea celular de estudio propia), tras lo cual se pueden seguir diferentes procedimientos (Galbraith *et al.*, 2018; Venkateswarlu y Jujjavarapu, 2020a).

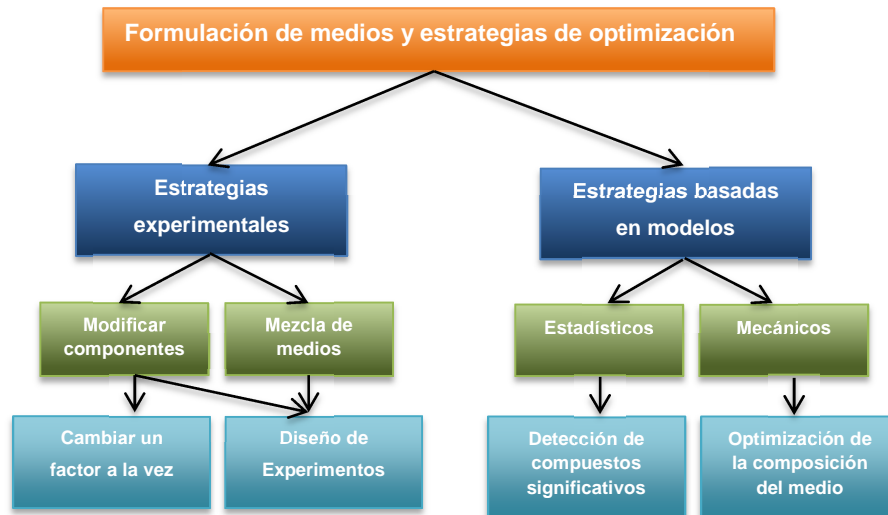


Figura 1. Enfoques para la optimización y formulación de medios de cultivo, que engloban las estrategias a nivel experimental y aquellas basadas en modelos (Modificado de Galbraith *et al.*, 2018).

Cambiar un Factor a la vez (One Factor at a Time) es una técnica de optimización clásica, en la que solo se varía la concentración de uno de los componentes del medio durante un período de tiempo, mientras se mantienen constantes los demás factores. Aunque es un método relativamente simple, pues se observan los efectos sin necesidad de análisis estadísticos, no estudia las interacciones entre las diferentes variables y requiere de un gran número de experimentos y un elevado coste (Galbraith *et al.*, 2018; Mahajan *et al.*, 2019).

La mezcla de medios sirve para optimizar muchos ingredientes de medios simultáneamente. La composición de los medios de partida debe ser conocida y significativamente diferente para que las mezclas resultantes sean únicas. Dado que los componentes se cambian simultáneamente, es imposible sacar conclusiones sobre las relaciones causales entre ellos, pero la combinación de medios reduce el tiempo y el coste de la optimización (Galbraith *et al.*, 2018).

El Diseño de Experimentos es un enfoque tradicional que permite explorar los efectos de interacción de los componentes, ya que demuestra que cambiar varios factores del medio a la vez puede ser más eficiente que alterar uno sólo. Los principales diseños experimentales utilizados son el Diseño Factorial (completo o parcial), el Diseño Central Compuesto, el diseño de Plackett-Burman y el diseño de Box–Behnken, entre otros. Tanto para originar medios para bacterias, hongos y células de mamíferos como para la formulación de productos

farmacéuticos, se ha utilizado la Metodología de Superficie de Respuesta (MSR) con un Diseño Central Compuesto, el cual es muy útil para experimentos secuenciales y optimización de múltiples variables (Kennedy y Krouse, 1999; Galbraith *et al.*, 2018; Mahajan *et al.*, 2019). La MSR implica el diseño estadístico de experimentos, el establecimiento de relaciones matemáticas (entre los factores causales y las variables de respuesta) y la optimización de los factores de formulación que satisfacen las características deseadas del producto (Venkateswarlu y Jujjavarapu, 2020a). Se considera uno de los métodos de optimización más utilizados pero existen muchos otros como Análisis Canónico, Regresión Lineal Múltiple (RLM), método de Gauss–Seidel, Redes Neuronales Artificiales y Algoritmos Genéticos (AG) (Kennedy y Krouse, 1999).

Los modelos matemáticos poseen un alto potencial para reemplazar la experimentación en laboratorios, pues ya se ha logrado reducir significativamente el número de experimentos necesarios en muchos problemas. El objetivo de la optimización matemática es encontrar los valores de las variables de decisión (parámetros del modelo) que producen los mejores valores (máximos o mínimos) de la función objetivo. Si las entradas del modelo son componentes del medio, la formulación del mismo puede optimizarse para producir la salida deseada. Aunque aún no existe un enfoque formal basado en modelos, se han desarrollado algunos modelos para biofabricación que podrían adaptarse a la formulación de medios. Cabe destacar que no se puede exagerar la capacidad del modelo de hacer predicciones sobre el rendimiento del cultivo en diferentes formulaciones de medios sin datos experimentales (o con un conjunto de datos limitado) (Galbraith *et al.*, 2018; Venkateswarlu y Jujjavarapu, 2020b).

Como ya hemos comentado, uno de los usos más frecuentes de las técnicas de optimización hace referencia a la formulación de medios de cultivo de distintos microorganismos. Algunas microalgas, como *Nannochloropsis gaditana*, se consideran la base de la cadena alimentaria de la acuicultura, ya que poseen un alto valor nutricional por su elevado contenido en proteínas y ácidos grasos poliinsaturados (Camacho-Rodríguez *et al.*, 2013, 2015). También se han realizado estudios que indican que determinados compuestos como el ácido eicosanopentaenoico (EPA) son metabolitos primarios por lo que, en general, incrementan a medida que aumenta la cantidad del microorganismo (Camacho-Rodríguez *et al.*, 2014).

En el año 2015, J. Camacho-Rodríguez, M.C. Cerón-García, J.M. Fernández-Sevilla y E. Molina-Grima realizaron un estudio que pretendía optimizar la formulación del medio de cultivo para el crecimiento de *N. gaditana*, maximizando la productividad de biomasa y de EPA y minimizando el coste de producción (Camacho-Rodríguez *et al.*, 2015). Para proceder

con el presente trabajo de investigación, únicamente recabamos los datos experimentales de este estudio con el fin de utilizarlos para la búsqueda de la función objetivo necesaria para el AG. Para ello, empleamos la composición de los medios de cultivo (normalizando los datos en el intervalo [0,1]) y la productividad de la biomasa resultante.

Estrechamente relacionada con la formulación de los nutrientes del medio de cultivo, se encuentra la optimización de las condiciones de reacción. A modo de ejemplo se puede citar la hidrólisis de proteínas, que se utiliza en procesos alimentarios y en la mejora de funciones biológicas; concretamente, los hidrolizados de plasma sanguíneo bovino y porcino contienen péptidos inhibidores de la Enzima Convertidora de la Angiotensina (ACE) y antígenotóxicos, así como también presentan propiedades antioxidantes y antimicrobianas (Figueroa *et al.*, 2016).

En 2016 se llevó a cabo un estudio, realizado por Omar A. Figueroa, José E. Zapata y Claudia P. Sánchez (Figueroa *et al.*, 2016), que perseguía la maximización del grado de hidrólisis enzimática de proteínas de plasma bovino. A través de MSR y AG, optimizaron las condiciones de pH, temperatura y la proporción enzima/sustrato, variables cuyos datos de entrada normalizaron entre los valores -1,68 y 1,68. De igual modo, en el presente trabajo de investigación realizado, utilizamos los datos de las condiciones de este segundo estudio para la búsqueda de resultados mediante aproximación de datos por RLM (o búsqueda aleatoria) y optimización mediante AG.

2. OBJETIVO Y METODOLOGÍA

El objetivo del estudio que hemos desarrollado consistió en la optimización de las variables del problema mediante AG, dados unos datos experimentales previos. Dicho algoritmo requiere de una función de aptitud, que hallamos a través de RLM o mediante un algoritmo de búsqueda aleatoria (dando lugar a un algoritmo híbrido). Por un lado, tratamos de optimizar las condiciones del medio de cultivo de un determinado microorganismo (*Nannochloropsis gaditana*) para lograr la máxima productividad de biomasa del mismo. A su vez, perseguimos la minimización de los costes de producción de manera que se pudiese establecer un proceso industrial idóneo para su realización. Por otro lado, también optimizamos las condiciones de la reacción de hidrólisis enzimática de plasma bovino para, finalmente, poder aplicar los AG a problemas de optimización de manera general.

La metodología seguida para el desarrollo del presente trabajo, se fundamentó en una serie de pasos para alcanzar tal objetivo:

En primer lugar, buscamos la información necesaria (datos previos de otros experimentos, artículos con objetivos similares y antecedentes teóricos sobre la materia) para la realización de la propia investigación, mediante una serie de motores de búsqueda y bases de datos.

Seguidamente, tras profundizar en las referencias que tratamos, recopilamos los datos experimentales de dos artículos científicos: un estudio de optimización de medios de *N. gaditana* realizado por J. Camacho-Rodríguez, M.C. Cerón-García, J.M. Fernández-Sevilla y E. Molina-Grima (Camacho-Rodríguez *et al.*, 2015) y un estudio de optimización de la hidrólisis enzimática de plasma bovino por Omar A. Figueroa, José E. Zapata y Claudia P. Sánchez (Figueroa *et al.*, 2016). Adaptamos la estructura de datos de estas muestras (ver anexo III) para poder ser leídas desde los siguientes programas:

- a) wxMaxima (Vodopivec *et al.*, 2004): empleamos esta interfaz gráfica para el sistema de Álgebra Computacional Maxima (Frederick Schelter, 1982), con el fin de buscar la función objetivo mediante RLM. Esta función es necesaria para la optimización de las variables mediante el uso de AG, denominándose función de aptitud.
- b) ActivePerl (ActiveState, 2017): utilizamos esta herramienta como intérprete del lenguaje de programación Perl para la elaboración de un programa que buscara de manera aleatoria la función objetivo. El código de dicho programa, desarrollado mediante el editor de texto NotePad++ (Ho, 2003), también era capaz de ejecutar el AG una vez obtenida la función.

Posteriormente, efectuamos numerosas pruebas con los programas realizados para la búsqueda de la función, hasta lograr unos resultados aptos (función de ajuste de los datos con error mínimo). De este modo, ejecutamos el AG con las funciones obtenidas para hallar los valores óptimos de los factores estudiados que permitieron mejorar la respuesta a evaluar.

A continuación, con los resultados obtenidos tras la optimización de la respuesta, establecimos las conclusiones en base al objetivo de la investigación realizada, determinando el beneficio de la misma sobre distintos problemas en diversos campos de trabajo.

Por último, debido a que en un principio trabajamos exclusivamente sobre los artículos anteriormente mencionados, generalizamos los programas desarrollados en wxMaxima y ActivePerl de manera que se pudiesen utilizar para diversos problemas de optimización, dados unos datos experimentales con independencia del número de variables a evaluar.

3. DESARROLLO DEL PLANTEAMIENTO

Para empezar, necesitamos determinar la función objetivo que mejor se ajustase a cada uno de los artículos utilizados (ver páginas 3 y 4, sección 1), que estableceremos como función de aptitud del AG que optimizará los problemas propuestos.

3.1. Función de ajuste a los datos conocidos

Para encontrar una función de aproximación de datos aplicamos dos procedimientos que pretendían minimizar el error de la misma respecto a los resultados experimentales.

3.1.1. Regresión por mínimos cuadrados (wxMaxima)

La regresión lineal es un modelo que se basa en la determinación de la ecuación que mejor ajusta la relación entre la variable independiente o respuesta y las variables dependientes que la modifican. Para realizar la regresión, se utiliza el ajuste por mínimos cuadrados que nos permite obtener la ecuación (ver figura 3) que minimiza la suma de los errores cuadráticos (diferencia entre el punto real de la nube y el punto esperado por la función) (Carrasquilla-Batista *et al.*, 2016). Existen otros modelos de regresión como el logarítmico y el exponencial, pero la regresión lineal es el más utilizado por su mayor facilidad de uso (Raposo, 2016).

Este método se usa para realizar predicciones, por lo que en la actualidad está estrechamente relacionado con campos tales como el aprendizaje de redes neuronales artificiales. Anteriormente, también se utilizó para métodos analíticos en los cuales era necesario obtener la relación de calibración entre los analitos y la respuesta del instrumento. Se busca obtener una función de regresión que aproxime de manera aceptable los datos conocidos de las muestras mediante el criterio de mínimos cuadrados. Una de las utilidades más importantes, aún sin explorar completamente, es la optimización de datos. En esta, se determina la ecuación necesaria para hallar una combinación de valores de las variables \underline{x} de entrada que maximice o minimice la variable y de salida (Carrasquilla-Batista *et al.*, 2016; Raposo, 2016). Para demostrar la eficacia de este método, sabemos que dada una colección de k puntos, $(x_1, y_1), \dots, (x_k, y_k)$, el objetivo de la regresión múltiple por mínimos cuadrados consiste en conseguir el mejor ajuste de la nube de puntos. Para ello, se pretende encontrar una función lineal en a_i , como las que se muestran en las funciones 1 y 2 (ver figuras 2 y 3).

$$g(x) = a_0 \cdot f_0(x) + a_1 \cdot f_1(x) + \dots + a_n \cdot f_n(x) \quad (1)$$

Figura 2. Muestra la función de aproximación de datos (función 1) siendo $f_i(x)$ funciones establecidas *a priori*, a_i los parámetros que deseamos encontrar, x la variable correspondiente de los datos de entrada, i el índice del sumatorio y $n + 1$ el número de coeficientes a determinar. En este caso, se muestra la función para $m = 1$ (siendo m el número de variables del conjunto de entrada), es decir, cuando el conjunto de entrada de datos es \mathbb{R} , ya que solo existe la variable de entrada x (Modificado de García-Tejeda y Barrera-Figueroa, 2019).

$$g(x_1, x_2, \dots, x_m) = a_0 \cdot f_0(x_1, x_2, \dots, x_m) + a_1 \cdot f_1(x_1, x_2, \dots, x_m) + \dots + a_n \cdot f_n(x_1, x_2, \dots, x_m) \quad (2)$$

Figura 3. Representa la función de aproximación de datos (función 2) expresada de forma genérica para $m > 1$. Conlleva el mismo procedimiento que la función 1, a excepción del número de variables utilizadas (m), por lo que se denomina regresión lineal múltiple (Modificado de García-Tejeda y Barrera-Figueroa, 2019).

Cada muestra $\underline{x} = (x_1, x_2, \dots, x_m)$ de la nube de puntos conlleva un error, cometido por la función de ajuste $g(\underline{x})$, el cual se define como la diferencia entre el valor y esperado, y el valor aproximado obtenido por la función $g(\underline{x})$. Se obtiene el error cuadrático (ver función 3, figura 4) para evitar errores por los cambios de signo y la suma de los mismos en las diferentes muestras (García-Tejeda y Barrera-Figueroa, 2019).

$$E(a_0, a_1, \dots, a_n) = \sum_{i=1}^k (g(\underline{x}_i) - y_i)^2 \quad (3)$$

Figura 4. Corresponde a la función 3 que indica el error cuadrático total cometido al aproximar mediante la función de ajuste $g(\underline{x})$, obtenido de la suma de los errores cuadráticos de la nube de puntos para esa función (Modificado de García-Tejeda y Barrera-Figueroa, 2019).

Se pretende por tanto, encontrar los valores de $a_i \in \mathbb{R}$, de tal forma que se minimice el error de aproximación en las k muestras conocidas. La minimización se logra realizando la derivada $\frac{\partial E}{\partial a_j}$ respecto a a_j e igualándola a 0, dando lugar a la ecuación 4 (ver figura 5) (García-Tejeda y Barrera-Figueroa, 2019).

$$a_0 \cdot \sum_{i=1}^k f_0(\underline{x}_i) \cdot f_j(\underline{x}_i) + a_1 \cdot \sum_{i=1}^k f_1(\underline{x}_i) \cdot f_j(\underline{x}_i) + \dots + a_n \cdot \sum_{i=1}^k f_n(\underline{x}_i) \cdot f_j(\underline{x}_i) = \sum_{i=1}^k y_i \cdot f_j(\underline{x}_i), \forall j \quad (4)$$

$$B^t \cdot B \cdot A = B^t \cdot Y$$

$$\underbrace{\begin{pmatrix} f_0(\underline{x}_1) & f_0(\underline{x}_2) & \dots & f_0(\underline{x}_k) \\ f_1(\underline{x}_1) & f_1(\underline{x}_2) & \dots & f_1(\underline{x}_k) \\ \dots & \dots & \dots & \dots \\ f_n(\underline{x}_1) & f_n(\underline{x}_2) & \dots & f_n(\underline{x}_k) \end{pmatrix}}_{B^t}, \underbrace{\begin{pmatrix} f_0(\underline{x}_1) & f_1(\underline{x}_1) & \dots & f_n(\underline{x}_1) \\ f_0(\underline{x}_2) & f_1(\underline{x}_2) & \dots & f_n(\underline{x}_2) \\ \dots & \dots & \dots & \dots \\ f_0(\underline{x}_k) & f_1(\underline{x}_k) & \dots & f_n(\underline{x}_k) \end{pmatrix}}_B, \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_n \end{pmatrix}}_A, \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_k \end{pmatrix}}_Y$$

Figura 5. Muestra las matrices del sistema de ecuaciones (ecuación 4) procedente de la derivación de la función de error cuadrático, igualándolo a 0 y expresado en forma explícita y matricial. Se muestran también las matrices que conforman el sistema, del cual se pretende despejar la matriz A , correspondiente a la solución de la regresión (Modificado de García-Tejeda y Barrera-Figueroa, 2019).

A partir de la ecuación 4, podemos despejar la solución, es decir, los valores de a_i (o la matriz A) que dan lugar al error mínimo mediante el proceso de regresión múltiple explicado.

Para desarrollar el modelo de regresión necesario, usamos la interfaz gráfica wxMaxima de la versión 5.38.1 del programa Maxima, cuyos cálculos se basaron en la forma matricial de la expresión matemática explicada en el apartado anterior. Por tanto, mediante la ayuda de un soporte informático de álgebra computacional logramos obtener los valores a_i de forma rápida y eficiente, de tal manera que obtuvimos una función que se ajusta a los datos conocidos. Esta se utilizó posteriormente como función de aptitud para el algoritmo genético, imprescindible para la optimización.

En distintos estudios (Desai *et al.*, 2008; Figueroa *et al.*, 2016; Shirodkar y Muraleedharan, 2017; Sirohi *et al.*, 2018), se observó que para determinar la función de optimización, los métodos más usados son la metodología de superficie de respuesta y la regresión, siendo la función 5 (ver figura 6) la fórmula más utilizada en los mismos.

$$g(\underline{x}) = a_0 + \sum_{i=1}^n a_i \cdot x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^n a_{ii} \cdot x_i^2 \quad (5)$$

Figura 6. Corresponde a la expresión polinómica de segundo grado (función 5), utilizada para los diseños de MSR, que utiliza la regresión lineal múltiple, donde se establece la influencia de los factores (variables de entrada) y sus interacciones (entre factores tomados de 2 en 2), así como los efectos cuadráticos de los mismos (Desai *et al.*, 2008; Figueroa *et al.*, 2016; Shirodkar y Muraleedharan, 2017; Sirohi *et al.*, 2018).

Esta fórmula permite obtener un ajuste preciso pero, ya que las muestras conocidas del primer artículo en el que nos basamos (ver página 3, sección 1) contienen un número elevado de variables a evaluar, el programa Maxima v.5.38.1 no soportó determinados cálculos matriciales, por lo que procedimos a realizar nuevas búsquedas de funciones en base a un menor número $n + 1$ de parámetros (tanto a_i como $f_i(\underline{x}_i)$).

En el presente trabajo, probamos diversas funciones $f_i(\underline{x})$ en función de los resultados que íbamos obteniendo. A su vez, aplicamos diferentes funciones de transferencia (las cuales denominamos $h(\underline{x})$) para la reducción al caso lineal de modelos no lineales. Entre las funciones $g(\underline{x})$ utilizadas destacaron la suma de variables, la suma de sus cuadrados o la suma de sus productos, y dentro de las funciones de transferencia, predominaron la función cuadrática o la función sigmoide. Asimismo, efectuamos la normalización de los datos y la comparamos con los resultados sin normalizar, escogiendo el mejor ajuste de todas las pruebas realizadas tal como se puede observar en el apartado de resultados.

Finalmente, introducimos la función elegida en el AG que, tras los procesos de selección, sobrecruzamiento y mutación, optimizará los datos de las generaciones obteniendo los mejores valores para el máximo de respuesta.

A pesar de utilizarse para los datos de los estudios mencionados (optimización del medio de *N. gaditana* y de la reacción de hidrólisis de plasma bovino; ver páginas 3 y 4 en sección 1), en realidad, este proceso de búsqueda de función de aptitud mediante regresión, sería válido para múltiples problemas de optimización con unos datos de entrada y salida conocidos (ver figura 43 en anexo III). Por tanto, el usuario puede introducir el tipo de muestras que prefiera en base al problema estudiado y obtener la función de aproximación de datos.

3.1.2. Búsqueda aleatoria (Perl)

Por otro lado, con la ayuda del intérprete ActivePerl 5.24.3, realizamos un programa capaz de buscar la función objetivo $g(\underline{x})$ necesaria para el AG (ver función 2, figura 3).

$$f_i(\underline{x}) = x_i^m \quad (6) \qquad f_i(\underline{x}) = \frac{1}{x_i} \quad (7)$$

$$f_i(\underline{x}) = e^{x_i} \quad (8) \qquad f_i(\underline{x}) = b(x_i) \quad (9)$$

Figura 7. Se observan las funciones 6, 7, 8 y 9 elegidas al azar para su disposición en la función 2, donde x_i son las variables de entrada a optimizar, m son los exponentes de dichas variables y b son funciones trigonométricas (seno o coseno) elegidas al azar. Tanto los parámetros m y b como los coeficientes a_i (de la función 2) se fijaron al azar entre los valores de un intervalo cuyo rango definimos previamente.

Para ello, buscamos una función $g(\underline{x})$ (función 2) eligiendo aleatoriamente sus funciones $f_i(\underline{x})$ entre las funciones de la figura 7. Realizamos múltiples iteraciones, y en cada una de ellas, obtuvimos una función $g(\underline{x})$ y la aplicamos a los datos experimentales de los estudios (ver páginas 3 y 4, sección 1), dando lugar a los valores de salida o respuesta ($g(\underline{x})$). Dichos valores se compararon con la respuesta y conocida, permitiéndonos calcular el error en valor absoluto cometido por la función. Llevamos a cabo numerosas ejecuciones para encontrar multitud de funciones aleatorias diferentes, guardando en un archivo la mejor función obtenida teniendo en cuenta la diferencia entre el valor de salida conocido y el obtenido por la función (en valor absoluto). Tras obtener la mejor función, la introducimos directamente en el AG para obtener la optimización de las variables del medio.

$$h(\underline{x}) = \frac{1}{1 + e^{g(\underline{x})}} \quad (10)$$

Figura 8. Presenta la función de transferencia de tipo sigmoide (función 10) en la cual $g(\underline{x})$ se dispuso en base a las funciones $f_i(\underline{x})$ de la figura 7, elegidas aleatoriamente y halladas por el programa desarrollado en Perl.

De igual modo, seguimos el mismo procedimiento para aplicar una función de transferencia de tipo sigmoide (ver función 10, figura 8), pues al realizar la regresión por mínimos cuadrados con wxMaxima, observamos un mejor ajuste. Esta función también se introducirá posteriormente en el AG como función objetivo.

Por otro lado, una vez obtenidas estas funciones, implementamos una mejora del sistema de búsqueda. Una vez obtenida la mejor función con el proceso efectuado, procedimos a realizar nuevas búsquedas aleatorias cambiando un único parámetro (a , m , b) de la función, también elegido al azar, para apreciar algún resultado mejor. Además, el programa incluye la opción de generar una nueva función aleatoria eliminando todos los resultados previamente guardados, lo que nos permite volver al inicio del proceso.

En un principio, ejecutamos directamente el programa con los ajustes necesarios para obtener los resultados esperados, mientras que, al final, realizamos un menú (ver anexo III, figura 45) para que se mostrara visualmente por pantalla al usuario, con las posibilidades a ejecutar. De esta manera, el programa es válido para aquellos usuarios que dispongan de un problema cuyos datos pretenden optimizar. Finalmente, en el menú incorporamos la búsqueda aleatoria de funciones y el procesamiento del AG con esas funciones, o con la obtenida por RLM.

3.2. Optimización mediante Algoritmos Genéticos

El Algoritmo Genético, desarrollado por John Henry Holland entre 1960 y 1970, pertenece al campo de los Algoritmos Evolutivos y está inspirado en la teoría de la selección natural de Darwin que explica el origen de las especies (Konak *et al.*, 2006; Sirohi *et al.*, 2018). En la naturaleza, las especies con características menos propensas a la supervivencia, se acaban extinguiendo mientras que las especies con mejores aptitudes, tienen una mayor oportunidad de transmitir sus genes a su descendencia (Konak *et al.*, 2006). Por tanto, el AG consiste en una búsqueda evolutiva heurística que imita el proceso de selección natural y utiliza operadores inspirados en la naturaleza para identificar buenas soluciones de trabajo (Kar, 2016).

Dentro del AG, un vector de solución, denominado individuo o cromosoma, está compuesto por unidades discretas llamadas genes, que controlan características del cromosoma, manteniendo el paralelismo con los conceptos de la teoría evolutiva. Originalmente, los genes se expresaban como una combinación de dígitos binarios pero, en la actualidad, se han introducido tipos de genes más variados. Se parte de una colección de cromosomas, llamada población, que se inicializa al azar. Análogamente a lo que ocurre en la naturaleza, el AG combina las variables independientes, es decir, los genes, mediante tres operadores (selección,

cruce y mutación) para producir mejores descendientes. A medida que la búsqueda del algoritmo avanza, la población incluye soluciones más adecuadas hasta concluir en una única solución, en función de la aptitud de los individuos (Konak *et al.*, 2006; Sirohi *et al.*, 2018).

Hay diversos tipos de codificaciones utilizadas en los AG pero la más habitual es la codificación binaria, que consiste en convertir cada posible individuo o solución al problema en una subcadena de longitud predefinida que contiene como valores o bien 0, o bien 1. En general, cualquier valor real establecido dentro de un intervalo se puede codificar en binario con la precisión decimal que el problema requiera como se observa en el ejemplo de la figura 9. La codificación de número real a cadena binaria es un paso opcional previo a los operadores de selección, recombinación y mutación (Brinc y Belič, 2019; Passos y Barrenechea, 2019).

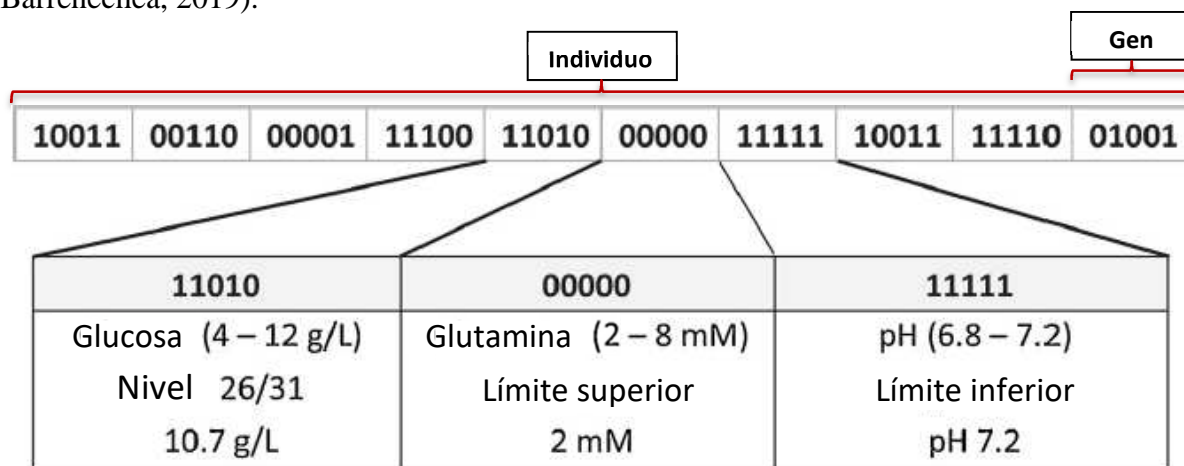


Figura 9. En este ejemplo de codificación binaria, se muestra un cromosoma de 10 genes o parámetros a tener en cuenta para cada individuo. Cada gen se encuentra codificado en binario por 5 bits, por lo que puede codificar 32 niveles (2^5 niveles). Los genes en la posición 5, 6 y 7 corresponden a los niveles de glucosa, glutamina y pH, respectivamente, cada uno de los cuales se encuentra dentro de un rango real (determinado experimentalmente). La mínima concentración de glutamina (2 mM) corresponde con 00000, la máxima concentración de pH (7,2) se codifica en binario como 11111, y la concentración de glucosa (10,7 g/L) corresponde a 11010 (Modificado de Brinc y Belič, 2019).

En nuestro caso, no implementamos codificación binaria por no considerarse necesaria al obtener buenos resultados sin dicha codificación. Por otra parte, en dichas pruebas sí se observó que es necesaria la normalización de los datos para conseguir una buena solución al problema.

3.2.1. Operador de selección

Inicialmente, los individuos se originan aleatoriamente, y el proceso de selección determina aquellos que pasarán a las siguientes generaciones en función del valor de aptitud que presenten, siendo la aptitud una función que permite calificar a los individuos para establecer su probabilidad de supervivencia. Se calcula la función de aptitud a partir de un conjunto de

números enteros, vectores, matrices, listas vinculadas u otra estructura de datos, en base a la relación entre las variables del problema a tratar y el objetivo que se pretende alcanzar (maximizar o minimizar el valor de respuesta de la función) (Konak *et al.*, 2006; Figueroa *et al.*, 2016; Kar, 2016; Sirohi *et al.*, 2018).

Existen varios procedimientos de selección utilizados (Passos y Barrenechea, 2019):

a) Método de la ruleta

También conocido como selección proporcional de aptitud, consiste en establecer unos valores o pesos para los cromosomas proporcionales a su aptitud, es decir, repartir a cada individuo, un fragmento de ruleta correspondiente al valor de su aptitud, enmarcándolo en un intervalo de la ruleta. Se simulan tantas tiradas de la ruleta como individuos existan y, después de cada giro, se selecciona un individuo para formar parte de la población de los padres. Para llevar a cabo esta simulación, en el programa se saca un número aleatorio entre 0 y la suma de aptitudes proporcionales de los cromosomas, tras lo cual se selecciona para la siguiente generación el individuo a cuyo intervalo pertenezca el número aleatorio. Los individuos con mayor aptitud tomarán un intervalo de longitud mayor, por lo que tendrán más probabilidad de que el número se encuentre en este, como podemos observar en la figura 10 (Figueroa *et al.*, 2016; Passos y Barrenechea, 2019; Shojaedini *et al.*, 2019).

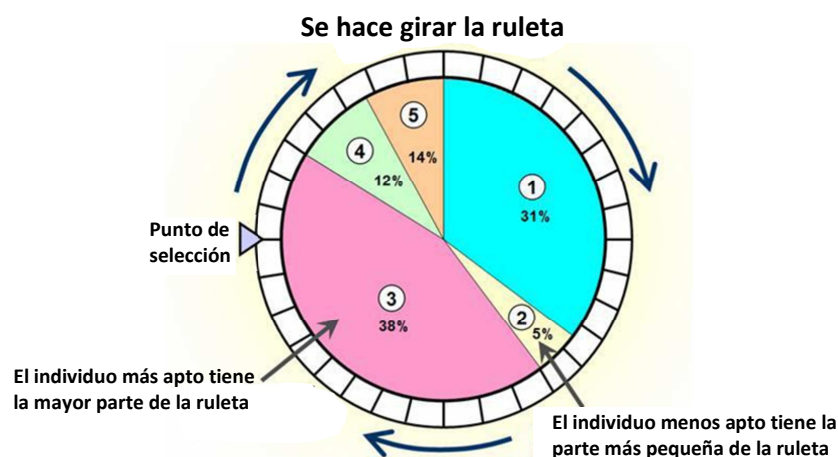


Figura 10. En este ejemplo de método de selección de la ruleta, se observan 5 individuos en la ruleta con fracciones proporcionales a la aptitud de cada uno. Al girar la ruleta, el individuo 2 será el que menos probabilidades tenga de ser seleccionado mientras que el 3 será el mejor individuo y el que permanecerá con más frecuencia en la siguiente generación (Modificado de El-Desoky *et al.*, 2016).

b) Elitismo

Se plantean dos tipos de elitismo, uno de los cuales se basa en mantener uno o más cromosomas en la próxima generación, garantizando que sus genes no se perderán después

del cruce y la mutación. De esta manera, el mejor individuo de cada generación siempre se mantendrá en la siguiente. Para ello, uno de los procedimientos llevado a cabo consiste en guardar los mejores cromosomas al salir de la mutación y añadirlos después de la selección de la generación siguiente, siendo necesario desechar el mismo número de cromosomas con peor aptitud. Este proceso de selección se utiliza para AG simples pero no para AG multiobjetivo. Otro tipo de elitismo consiste en asignar las casillas de la ruleta restantes al mejor individuo; existen casillas adicionales restantes en la ruleta porque se añaden tantas como individuos haya, para que al menos cada individuo tenga una casilla asignada en el caso de que todas (100% de la ruleta) fuesen a un solo individuo (Konak *et al.*, 2006; Passos y Barrenechea, 2019).

Para generar la población de la próxima generación usamos dos operadores genéticos: cruce y mutación. Estos dan lugar a una solución “secundaria” que retiene muchas de las características positivas de los parentales a la vez que disminuye las características menos útiles (Kar, 2016; Shojaedini *et al.*, 2019).

3.2.2. Operador de sobrecruzamiento/recombinación

El objetivo del sobrecruzamiento, al igual que en la evolución biológica, consiste en la combinación de dos cromosomas (parentales) para formar dos nuevos cromosomas (descendientes) que generalmente, tengan una mejor aptitud que los padres (Shojaedini *et al.*, 2019; Manzoni *et al.*, 2020). En primer lugar, entre los cromosomas existentes en la población, se escogen los parentales y se selecciona la zona de cruce (sitio de corte) de manera aleatoria. Tras esto, los fragmentos posteriores al sitio de corte de ambas cadenas se intercambian, dando lugar a dos nuevas cadenas o individuos descendientes, como se muestra en la figura 11. En el AG, se establece previamente la tasa de sobrecruzamiento o probabilidad de cruce (P_c), que determina la frecuencia con la que ocurre. Al aplicar iterativamente el operador de cruce, se espera que los genes de los mejores individuos aparezcan con mayor frecuencia en la población, lo que acaba dando lugar a una buena solución general. Existen más tipos de cruce como el cruce uniforme o cruce de dos puntos (Konak *et al.*, 2006; Figueroa *et al.*, 2016; Sirohi *et al.*, 2018; Passos y Barrenechea, 2019; Shojaedini *et al.*, 2019).

3.2.3. Operador de mutación

El operador de mutación introduce cambios aleatorios en los genes de los cromosomas. En los tipos de AG más conocidos, la tasa de mutación P_m (probabilidad de cambiar las propiedades de un gen), es muy pequeña y depende de la longitud del cromosoma, por lo que el nuevo

cromosoma mutado no será muy diferente del original. Sin embargo, se puede elegir una elevada tasa para mejorar significativamente el ajuste de la solución (la descendencia) para evitar la optimización local y los desafíos asociados con los niveles intermedios (Konak *et al.*, 2006; Kar, 2016; Sirohi *et al.*, 2018).

Si seguimos la terminología conocida en la rama biológica, la mutación consta de alteraciones de los genotipos de los individuos, es decir, los descendientes presentan genes que no se encontraban en sus antepasados. Por tanto, los nuevos cromosomas poseen nuevas características que pueden transmitirse a la descendencia futura del individuo, modificando a largo plazo los atributos de la especie. De esta manera, se concluye que este operador genético es esencial para la variabilidad de individuos y, en consecuencia, evita alcanzar de manera prematura una solución no optimizada. Sin embargo, si la tasa de mutación es muy elevada, la intensa variación impide la convergencia en la mejor solución. En una población cuyos individuos (cromosomas) están codificados en una cadena de valores 0 o 1, este operador provocaría que si un elemento o gen sufriese mutación, se modificase al valor contrario, es decir, pasaría de ser un 0 a un 1 y viceversa, como se observa en la figura 11. También se pueden implementar otro tipo de mutaciones como multiplicar un gen por un determinado valor o sumarle cierta cantidad numérica (Passos y Barrenechea, 2019).

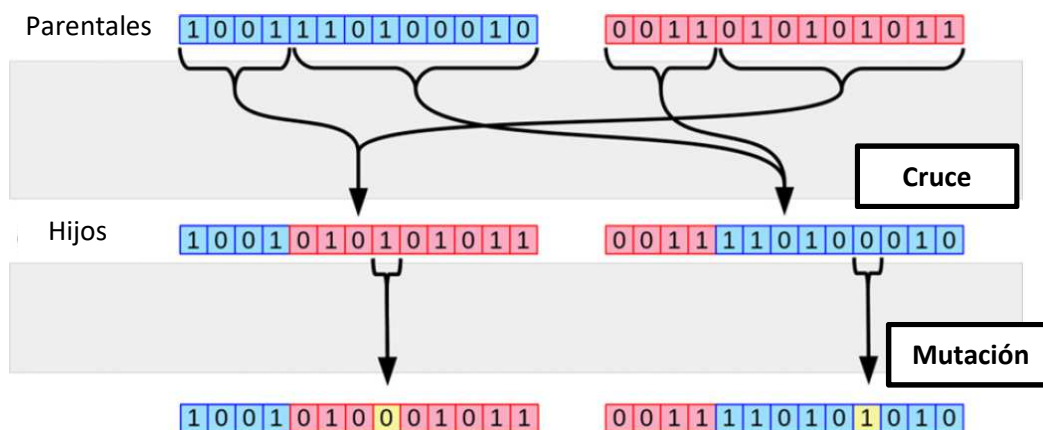


Figura 11. Se observa el proceso de recombinación y mutación en algoritmos genéticos en cromosomas cuyos genes se establecen como un 0 o un 1. Tras el proceso de selección, se procede a intercambiar fragmentos de las cadenas que conforman los cromosomas a partir de la zona de cruce, lo cual tiene lugar con una frecuencia de recombinación (P_c) establecida por el usuario en el algoritmo. Después, tiene lugar la mutación de uno o más genes de los cromosomas de la descendencia, lo que aumenta la variabilidad y las probabilidades de encontrar una solución óptima con éxito, en función de la frecuencia de mutación (P_m) utilizada (Modificado de Urso *et al.*, 2019).

Por último, se conforma una nueva generación cuyos individuos iniciales serán los mejores individuos obtenidos por el algoritmo en la generación anterior. Se realizarán los mismos procesos de selección, recombinación y mutación sobre la nueva generación y se repetirá el

procedimiento hasta alcanzar el número máximo de generaciones (para obtener un buen resultado en la última generación) (El-Desoky *et al.*, 2016; Passos y Barrenechea, 2019).

El algoritmo genético se implementó para asemejar la forma en la que trabaja la naturaleza, usando técnicas computacionales para obtener soluciones de trabajo adecuadas mientras se crean generaciones futuras en organismos biológicos. Se ha utilizado significativamente para resolver una variedad de problemas de objetivos únicos o múltiples que son de naturaleza combinatoria y no determinista en los campos de la ciencia, la ingeniería, los negocios y las ciencias sociales (Kar, 2016; Venkateswarlu y Jujjavarapu, 2020c).

Uno de los campos en los que se utilizan los algoritmos genéticos es la optimización de problemas de gestión de operaciones tanto para diseño de procesos y productos, como planificación, control y mejora de operaciones. Todas las organizaciones necesitan operaciones para producir una combinación de servicios y productos, mediante una serie de funciones (marketing, desarrollo y operaciones) (Lee, 2018). Sin embargo, los AG se han utilizado principalmente para buscar respuesta a problemas de maximización/minimización como el problema del vendedor ambulante, problemas de clasificación y toma de decisiones con objetivos o criterios múltiples. Por esto, también se han aplicado a la ingeniería industrial para análisis de redes (rutas), robótica, aprendizaje automático, etc., así como en muchos otros ámbitos (Kar, 2016; Choi *et al.*, 2017; Venkateswarlu y Jujjavarapu, 2020c).

Los AG conllevan una serie de ventajas como métodos de optimización, pues son capaces de proporcionar una lista de soluciones óptimas (en lugar de una única), son útiles frente a los modelos matemáticos o numéricos, optimizan tanto funciones continuas como discretas y son capaces de trabajar con grandes problemas de optimización de múltiples parámetros con funciones estocásticas no lineales y objetivas. En cambio, también ocasionan una serie de inconvenientes. Una incorrecta elección de la función aptitud puede conducir a soluciones inferiores, del mismo modo que la convergencia prematura puede suponer un problema en su implementación. Al tratarse de funciones estocásticas, no hay garantía de calidad en la solución y el cálculo repetido del valor de aptitud puede generar una carga computacional y disminuir su rendimiento (Kar, 2016; Choi *et al.*, 2017; Venkateswarlu y Jujjavarapu, 2020c).

3.2.4. Algoritmo Híbrido Genético con Regresión Múltiple (Perl)

Para llevar a cabo el presente estudio especificamos los operadores del AG utilizados y definimos los parámetros necesarios para su funcionamiento. En cuanto al operador de selección, utilizamos el método de la ruleta y los dos tipos de elitismo explicados

anteriormente. Para el operador de sobrecruzamiento, asignamos el cruce de un solo punto, explicado en su apartado correspondiente, mientras que designamos el operador de mutación con reemplazo aleatorio gen a gen.

Una vez establecimos la función objetivo a evaluar, determinamos la probabilidad de cruce y de mutación, el tamaño de la población (número de individuos), el número máximo de generaciones y el número de variables de entrada a evaluar, en base a los datos de los artículos utilizados (ver páginas 3 y 4, sección 1). En la figura 14, se muestra el diagrama de flujo correspondiente al proceso realizado, donde se puede observar de manera visual las etapas implementadas hasta lograr el objetivo final del proyecto.

Cabe destacar que tras obtener una función apropiada $g(\underline{x})$ (o en caso de aplicar una función de transferencia, $h(\underline{x})$) mediante RLM con wxMaxima o mediante el programa desarrollado en Perl, procedimos a establecer la función de aptitud en el AG. Como deseamos fijar como función de aptitud aquella función de ajuste de datos cuyo error sea el menor de entre las encontradas con los métodos vistos, la función de aptitud se asignó directamente $g(\underline{x})$ o $h(\underline{x})$. No obstante, el usuario puede ejecutar el algoritmo con las diversas funciones y observar los resultados de optimización de cada problema propuesto (ver anexo I).

Sin embargo, en el problema de optimización del medio de *N. gaditana* pretendíamos optimizar los resultados no solo en cuanto a maximización de la productividad, sino también minimizando el coste, para lo cual modificamos la función de aptitud (ver función 11, figura 12).

$$Apt(\underline{x}) = \alpha \cdot g(\underline{x}) + (\beta \cdot (M - c(\underline{x}))) \cdot \gamma \quad (11)$$

Figura 12. Función de aptitud del AG (función 11) para el problema de optimización de formulación del medio de *N. gaditana*, donde $g(\underline{x})$ equivale a la función de ajuste obtenida, $c(\underline{x})$ corresponde a la función del coste de los nutrientes utilizados (ver función 12, figura 13), M representa la cota máxima del coste del medio y γ es un parámetro utilizado para realizar una normalización de las funciones de aproximación $g(\underline{x})$ y coste $c(\underline{x})$ dentro del mismo rango para no perder validez (se define como el cociente entre $g(\underline{x})$ y M). Por último, α y β son dos parámetros definidos entre 0 y 1, utilizados para dar mayor o menor prioridad a la maximización de la productividad o a la minimización del coste.

$$c(\underline{x}) = \sum_{i=1}^m c_i \cdot x_i \quad (12)$$

Figura 13. Se observa la función 12, que evalúa el coste total del medio de cultivo en función de los componentes utilizados, donde m es el número de variables del conjunto de entrada, x_i corresponde a la concentración de componente (g/L) y c_i al coste de cada nutriente en €/g.

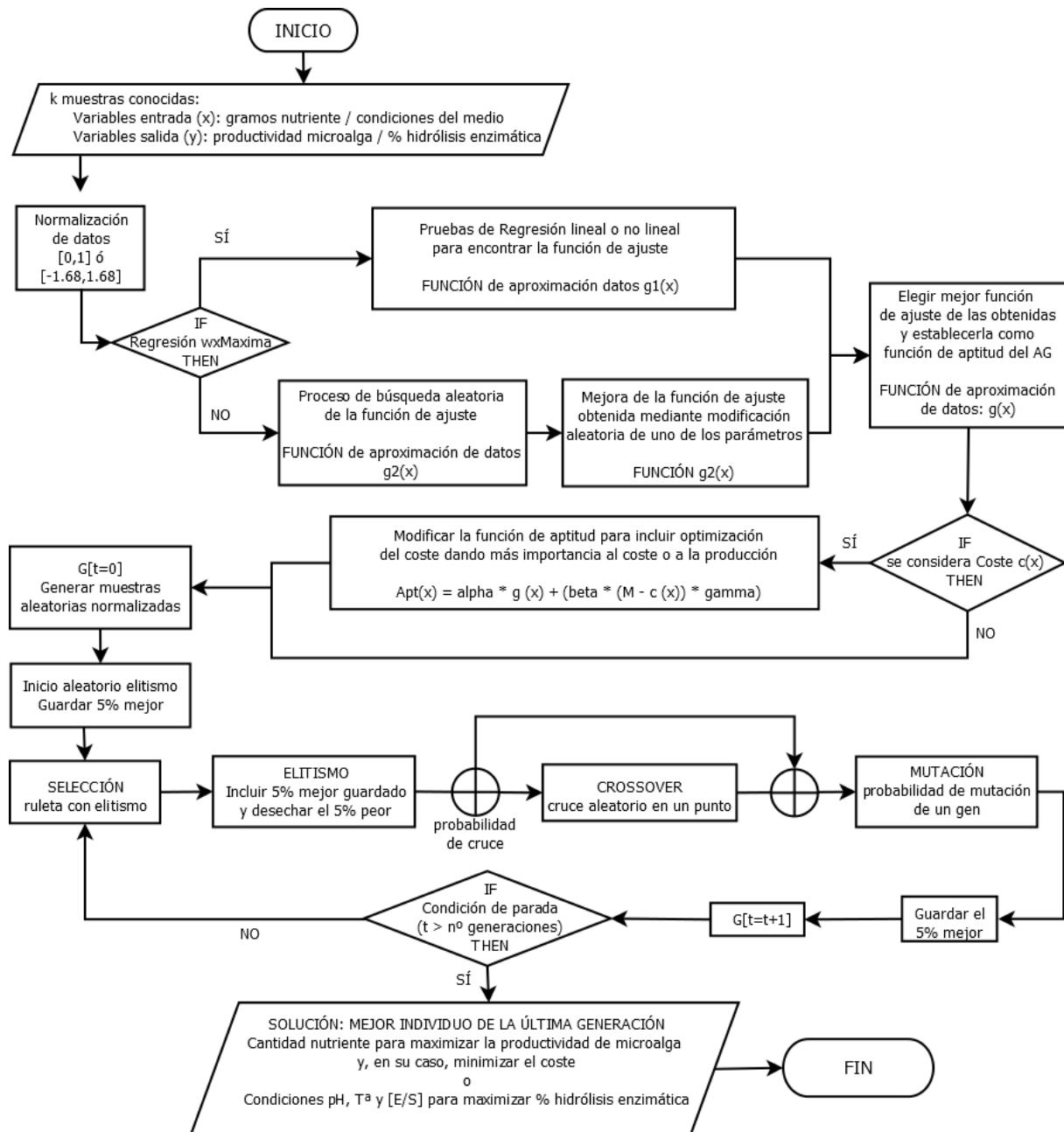


Figura 14. Diagrama de flujo del proceso realizado en el presente Trabajo de Fin de Grado que concluye con la optimización del problema abordado.

4. RESULTADOS

Planteamos los resultados obtenidos de la búsqueda de función objetivo mediante wxMaxima o mediante el programa desarrollado en Perl, así como de la optimización por AG, para los dos artículos en los que nos basamos (ver páginas 3 y 4, sección 1).

4.1. Función de ajuste por mínimos cuadrados

Logramos hallar una función objetivo óptima para cada estudio mediante RLM a través de wxMaxima.

4.1.1. Medio de cultivo de *Nannochloropsis gaditana*

Al disponer de 12 variables a evaluar para la optimización del medio de cultivo de *N. gaditana* en base a los datos del primer estudio utilizado (ver página 3, sección 1), los cálculos matriciales necesarios para realizar la regresión mediante la fórmula de la función 5 son inviables mediante el uso de wxMaxima (por falta de capacidad de procesamiento). Por tanto, realizamos distintas pruebas de RLM (y también no lineal) con menos variables.

En los resultados del ajuste de la función obtenidos mediante regresión por mínimos cuadrados (figura 16), observamos los parámetros del ajuste utilizados en cada prueba (asignados manualmente). Por un lado, necesitamos establecer la variable control de tal manera que al asignarle un “0” no se realizase la normalización de los datos de entrada de las muestras, mientras que si se asignaba un “1” se normalizaban los datos en el intervalo [0,1]. Después ajustamos las funciones f_i y la función de transferencia $h(\underline{x})$. Entre las funciones f_i , destacaron la suma de las variables y la suma de sus cuadrados, obteniendo un mejor ajuste con la segunda, de igual manera que probamos como funciones de transferencia la cuadrática o la sigmoide. Con la función de tipo sigmoide (ver función 13, figura 15) obtuvimos los mejores resultados, evaluando los distintos parámetros a determinar (k , c y α).

$$h(\underline{x}) = \frac{k}{1 + e^{\alpha*(g(\underline{x})-c)}} \quad (13)$$

Figura 15. Corresponde a la función 13, es decir, una función de transferencia de tipo sigmoide, cuyos parámetros k , c y α , determinados manualmente en el programa wxMaxima, establecen la homotecia, traslación horizontal y cambio de pendiente, respectivamente, de la función de tipo sigmoide original.

En la figura 16, también se muestran los errores cometidos por la función de ajuste mediante regresión para 5 muestras conocidas, siendo tres de ellas valores obtenidos mediante el AG del estudio de J. Camacho-Rodríguez *et al.* (ver página 3, sección 1); las otras dos muestras son valores experimentales del mismo estudio en dos medios conocidos. No se muestran los resultados sin normalizar por obtenerse errores demasiado elevados. Evaluamos distintas funciones f_i (suma de variables, suma de variables al cuadrado, suma de variables al cuadrado más otra variable) así como determinamos la función de transferencia (h_i) cuadrática y sigmoide. Observamos un mejor ajuste con la normalización de los datos, la función f_i de sumatorio de variables y con la función sigmoide de transferencia (h_i), por lo que ajustamos los parámetros de esta última hasta obtener un error aceptable, no solo a nivel global sino alcanzando un límite de centésimas en cada muestra. De esta manera, conseguimos una función de aptitud óptima para el desarrollo del AG mediante regresión por mínimos cuadrados.

CON 5 MUESTRAS

```
Resultado: control 1, funciones fi, transfer sigmoide, k_lim=1.5, c=0.999, alfa=1.05
[-0.028826,0.039859,0.109254,-0.0694896,-0.0603312]

-->Resultado: control 1, funciones fi cuadráticas + j, transfer sigmoide, k_lim=1.5, c=1, alfa=1.05
-->[6.56145*10^-4,0.011662,0.0897629,-0.17903,-0.0427795]

-->Resultado: control 1, funciones fi cuadráticas + j, transfer sigmoide, k_lim=1.5, c=0.99, alfa=1.05
-->[0.00813791,0.0111315,0.0637717,-0.0130875,-0.0434584]-->error total 0.00631689
```

Figura 16. Parámetros utilizados y errores obtenidos en la regresión realizada con wxMaxima dadas 5 muestras extraídas del primer estudio que utilizamos (ver página 3, sección 1). Se recalca en naranja el resultado para $g(\underline{x}) = \sum a_i x_i^2$ con el mínimo error, dando lugar a la mejor función óptima (ver función 14, figura 17).

$$\frac{1.5}{e^{-1.1(-0.027z^2 - 0.45x^2 - 0.06v^2 - 1.8v^2 + 1.1u^2 - 0.95t^2 - 0.71s^2 - 0.064r^2 + 1.1q^2 - 0.05p^2 + 1.4o^2 + 3.1j^2 - 1.7j - 0.99)}} + 1$$

Figura 17. Mejor función obtenida (función 14) por RLM con wxMaxima para la formulación de medios de *N. gaditana*, siendo $g(\underline{x}) = \sum a_i x_i^2$, h la función de tipo sigmoide de la figura 15, $k = 1,5$, $c = 0,99$ y $\alpha = 1,05$.

4.1.2. Hidrólisis enzimática de plasma bovino

Para el estudio de la optimización de la hidrólisis enzimática de plasma bovino, no fue necesario realizar varias pruebas para encontrar una función de aproximación de datos óptima, ya que, al disponer de tan solo 3 variables x_1 , x_2 y x_3 de entrada, el programa wxMaxima era capaz de soportar los cálculos matriciales para realizar la regresión mediante la función 5 (figura 6). Como se observa en la figura 18, el máximo error cometido no alcanza las 2 unidades, siendo en la mayoría de las muestras mucho menor, por lo que lo consideramos un error aceptable dada la magnitud de las muestras. En la figura 19, podemos comparar la función obtenida (función 15) con la original (función 16), correspondiente al segundo artículo en el que nos basamos (ver página 4, sección 1), de manera que establecemos que los resultados obtenidos con wxMaxima son válidos, ya que se asemejan a los del artículo. En ese estudio despreciaron los términos de la ecuación no significativos tras realizar un ANOVA. Sin embargo, nosotros obtuvimos la función 15 en wxMaxima y comparamos cómo varió la optimización con AG en función de si despreciamos los términos no significativos o no.

```
Solución (valores de a)
[20.5091,-1.02459,0.663473,1.12949,-3.8066,-0.363691,-0.281236,-0.224395,-0.602816,-0.0354308]

Errores de cada muestra:
[-0.13526,-0.458211,0.108033,-0.564918,0.809101,-0.9822,0.212882,-0.502747,-0.990899,
0.109101,0.309101,0.886653,0.426216,1.09917,0.669508,-0.332016,0.992459,-1.65597]

Error global cometido en la aproximación: 10.02748523327587
```

Figura 18. Resultados de la regresión realizada con wxMaxima para la hidrólisis enzimática dadas 18 muestras experimentales del segundo estudio utilizado (ver página 4, sección 1). Se muestran los valores de a_i , así como los errores cometidos por la función de ajuste en las 18 muestras y la suma de los mismos.

$$-0.281x_3^2 - 0.0354x_2x_3 - 0.603x_1x_3 + 1.13x_3 - 0.364x_2^2 - 0.224x_1x_2 + 0.663x_2 - 3.81x_1^2 - 1.02x_1 + 20.5 \quad (15)$$

$$GH = 19,89 - 1,02X_1 + 0,69X_2 + 1,12X_3 - 3,81X_1^2 \quad (16)$$

Figura 19. Corresponde a la función obtenida mediante regresión lineal múltiple por mínimos cuadrados con wxMaxima (función 15) y la función alcanzada en el artículo (función 16).

4.2. Función de ajuste por búsqueda aleatoria

El objetivo era buscar de manera aleatoria una función de aproximación de datos en base a los datos experimentales de estudios anteriores, mediante el programa desarrollado en lenguaje Perl explicado anteriormente. En este caso, no fue necesario realizar distintas pruebas cambiando manualmente los parámetros de la función, pues el programa señalaba directamente el mejor resultado final en un archivo de texto al compararlo iterativamente con la mejor función obtenida. Sin embargo, sí se podían cambiar los parámetros del AG mediante la opción 12 del programa, para mostrar cómo afectaban estos a la optimización del problema (ver anexo III, figura 46).

4.2.1. Medio de cultivo de *Nannochloropsis gaditana*

A partir del primer artículo de investigación utilizado (ver página 3, sección 1), buscamos distintas funciones aleatorias que ajustasen los datos experimentales con el mínimo error. En el anexo I, se muestran algunos de los resultados obtenidos para las distintas búsquedas.

En primer lugar, intentamos buscar una función $g(\underline{x})$ que se asemejase a las buscadas por regresión lineal múltiple en wxMaxima cuyas f_i correspondiesen a la función 6. Adicionalmente, consideramos la opción de establecer funciones trigonométricas (función 9) para observar si se realizaba un mejor ajuste, así como también aplicamos la función exponencial (función 8) e inversa (función 7) a las variables evaluadas. Del mismo modo, buscamos funciones $h(\underline{x})$ de tipo sigmoide (función 10), al ser esta la mejor solución obtenida por regresión en wxMaxima. Por último, localizamos funciones de la forma de la función 5 con y sin término independiente (a_0), ya que este permite el desplazamiento vertical de la función sobre los ejes. Estas funciones no se pudieron lograr con wxMaxima por no poder procesar el programa más de 6 variables ($i > 6$).

Advertimos que la función de tipo sigmoide $h(\underline{x})$ (función 10) se ajusta mejor que $g(\underline{x})$, ya que alcanzamos errores menos elevados y tardamos menos tiempo en encontrar una solución óptima. También pudimos comprobar que una vez obtenida una función, la mejora de la búsqueda (ver opciones 3 y 8, figura 45, anexo III) realiza un mejor ajuste en la mayoría de los casos (depende de la aleatoriedad y de las iteraciones establecidas), ejecutando el programa las veces necesarias hasta alcanzar un error de centésimas en las 5 muestras.

Inicialmente, establecimos los coeficientes a_i en el intervalo $[-5,5]$ pero posteriormente, comprobamos que, excepto para la función 10, el programa encontraba funciones con menor error de aproximación y en menor tiempo si los coeficientes a_i de las variables se definían en un rango menor (en torno al 0), encontrando el mejor rango entre 0 y 0,1. Igualmente,

ejecutamos (con la opción 11 del programa, ver figura 45 en anexo III) la función obtenida por wxMaxima donde observamos ligeras diferencias en los decimales de los errores respecto al programa desarrollado en Perl, debido a que cada programa utiliza distinta precisión decimal para calcular los resultados.

Para todas las pruebas realizadas, confirmamos que los resultados mejoran si se realiza una normalización de los datos.

Concluimos que la mejor función de aproximación obtenida correspondía a las últimas funciones evaluadas (función 5 con término independiente y sin él), dando lugar a un error máximo de 2 centésimas en la función sin término independiente (figura 20) y de 6 milésimas en la función con término independiente (figura 21). Por tanto, pudimos confirmar que este tipo de función es el que mejor se aproxima, de forma que si se pudiese haber obtenido por regresión daría mejores resultados, ya que esta búsqueda se realizó de forma aleatoria. Asimismo, también corroboramos la eficacia de las opciones 2, 3, 7 y 8 (ver anexo III), pues observamos que el programa mejora la función obtenida comparando las nuevas funciones aleatorias respecto a la función guardada en el archivo. Concretamente, las opciones 3 y 8 son más efectivas, ya que una vez que no se consiguen encontrar mejores funciones de manera aleatoria, cambiar sólo uno de los parámetros de la función logra mejores resultados.

Normalizadas	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5
Resultado esperado:	0.46	0.368	0.285	0.469	0.367
Resultado f(x)	: 0.4597324	0.3398186	0.2947248	0.4685635	0.3763241
Diferencia	: 0.0002675	0.0281813	0.0097248	0.0004364	0.0093241

Figura 20. Resultado de las diferencias obtenidas con la mejor función polinomial de segundo grado (función 5) sin término independiente, encontrada aleatoriamente mediante el programa desarrollado en lenguaje Perl. La función que nos permite obtener estos errores se puede visualizar en el anexo I (función 17, figura 30).

Normalizadas	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5
Resultado esperado:	0.46	0.368	0.285	0.469	0.367
Resultado f(x)	: 0.4550017	0.3675951	0.2918856	0.4676965	0.3629304
Diferencia	: 0.0049982	0.0004048	0.0068856	0.0013034	0.0040695

Figura 21. Resultado de las diferencias obtenidas con la mejor función polinomial de segundo grado (función 5) con término independiente, encontrada aleatoriamente mediante el programa desarrollado en lenguaje Perl. La función que nos permite obtener estos errores se encuentra en el anexo I (función 18, figura 30).

Comparando los resultados obtenidos con el programa desarrollado en lenguaje Perl frente a los obtenidos por regresión en wxMaxima, determinamos que somos capaces de obtener una mejor función de aproximación de datos con nuestro programa, lo cual es debido a la incapacidad de wxMaxima de realizar la regresión múltiple con más de 6 factores para la función 5. Es probable que si se pudiesen realizar los cálculos, los resultados obtenidos serían

similares. Sin embargo, mientras no sea posible realizarlo, el programa desarrollado permite localizar una función de ajuste con unos resultados aceptables, lo que nos permitió optimizar la función mediante AG.

4.2.2. Hidrólisis enzimática de plasma bovino

Este apartado consta de la búsqueda de una función de manera aleatoria para comparar su eficiencia con la función obtenida por regresión, y posteriormente, evaluar su capacidad de optimización mediante AG. En el anexo II, se muestran algunos resultados obtenidos con el programa para la búsqueda aleatoria de la función objetivo. Se observan tanto las funciones del tipo polinomial de grado 2 (función 5, figura 6), como las funciones obtenidas mediante la misma fórmula una vez despreciados los términos que no resultaron significativos, tras el análisis ANOVA (ver sección 4.4). En ambos casos, realizamos las pruebas con la normalización de las muestras de entrada y carente de ella para ver las diferencias.

Uno de los efectos contemplados es la baja efectividad de la búsqueda aleatoria con respecto al apartado anterior, lo cual es debido al elevado número de muestras, de manera que, al ejecutar las opciones 2, 3, 7 u 8 (ver anexo III), resulta muy difícil encontrar una función que mejore la inicial, obteniéndose errores muy elevados. En base al desarrollo del programa, el cual solo califica como “mejor” aquella función con menor error para todas las muestras, pudimos determinar que, al aumentar el número de muestras, se dificulta la búsqueda de una función.

En términos generales, la normalización de los datos de entrada mejora la búsqueda de función puesto que los errores que comete la función de ajuste respecto al resultado real son menores pero aun así, estos errores tampoco son válidos. Al despreciar los términos no significativos de la función, obtuvimos peores resultados (errores mayores). Esto se debe al menor número de términos en la ecuación que permite un menor ajuste; en el artículo desprecian los términos únicamente tras encontrar la función por regresión y efectuar un Análisis de Varianza (ANOVA), por lo que despreciarlos antes de encontrar la función tiene un efecto negativo.

Por último, cabe destacar que no encontramos ninguna función similar a la obtenida por regresión en w_{Maxima} , lo cual sería posible pero muy improbable debido a la incertidumbre de la elección aleatoria de la función.

4.3. Algoritmo Híbrido Genético con Regresión Múltiple

El último objetivo del trabajo realizado consistió en utilizar la mejor función objetivo obtenida en cada problema estudiado para determinar los valores que optimizasen el proceso.

4.3.1. Medio de cultivo de *Nannochloropsis gaditana*

A pesar de haber conseguido una función objetivo óptima con unos buenos resultados, realizamos la optimización mediante AG para el resto de funciones obtenidas, para comparar los resultados. Pudimos observar la eficacia de las opciones 2, 3, 7 y 8 (encargadas de mejorar la función objetivo encontrada), pues si equiparamos la diferencia entre la optimización mediante AG de una función aleatoria sin mejorar (opciones 4 y 9) y la mejor función aleatoria encontrada (opciones 5 y 10), observamos claramente mejores resultados de optimización en las opciones 5 y 10. Al igual que concluimos en el apartado anterior, la mejor función objetivo fue de la forma de la función 5 (ver anexo I), pues además de alcanzar un error aceptable, optimizó el proceso alcanzando el máximo de la respuesta requerida. Los parámetros del AG para las pruebas realizadas fueron los correspondientes al artículo de J. Camacho-Rodríguez *et al.* (ver página 3, sección 1).

Para la función obtenida mediante regresión múltiple en wxMaxima (opción 11, figura 45 en anexo III), los resultados de maximización de la función mediante AG alcanzan valores entre 1,4 y 1,5 (siendo el mejor valor obtenido 1,469), lo cual se explica matemáticamente ya que es una función cuyo límite superior se encuentra en 1,5. Al ejecutar el AG con la función 10 (figura 8) mejorada (opción 10, ver anexo III), comprobamos que la respuesta se acercó más aún a 1,5 (llegando a valores de 1,49999), lo que consideramos como una mejor capacidad de optimización (aunque la función puede ajustarse peor al tener mayor error). En el caso de la función $g(\underline{x})$ cuyas f_i se basaban en la figura 7, obtuvimos resultados entre 10-350 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$, por lo que descartamos totalmente la búsqueda de función aleatoria de este tipo pues el rango de respuesta fue extremadamente elevado e incierto. En cambio, al establecer los coeficientes a_i en un rango entre 0 y 0,1 para este tipo de función, los resultados fueron más asequibles pues varían entre 0 y 8 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$ (valores más lógicos para la concentración de biomasa de microalga producida al día).

Por último, observamos las respuestas obtenidas para maximizar la función 17 (ver figura 30, anexo I) sin termino independiente ($a_0 = 0$), que se encontraban englobadas entre 1,4 y 2,4 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$, valores más que aceptables para la productividad de biomasa. El mejor valor de la respuesta que logramos hallar fue de 2,39 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$. Para lograr esto, realizamos distintas pruebas modificando los parámetros relativos al AG, donde observamos que cambiar la probabilidad de cruzamiento a 0,9 mejoraba la respuesta hasta 2,1 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$ y establecer la probabilidad de mutación en 0,05 daba resultados más próximos a 2,3 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$.

En el caso de la función con termino independiente (ver función 18, figura 30), los valores de productividad se tornaron entre 0,7 y 1,1 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$, logrando un máximo de 1,096 $\text{g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$. La

función 5 con término independiente (función 18) tuvo menor error de ajuste (se consideró la mejor), por lo que validamos como nivel máximo de productividad alrededor de $1,1 \text{ g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$. Al aumentar el número de generaciones se realizó mejor optimización, obteniendo mejores resultados y estrechando el rango de la respuesta. Asimismo en todas las pruebas realizadas, confirmamos la eficacia de la normalización de los datos de entrada mejorando los resultados. Una vez obtuvimos una función objetivo $g(\underline{x})$ aceptable (utilizamos las funciones 17 y 18, ver figura 30) y optimizamos el problema maximizando su respuesta, incluimos adicionalmente a la función objetivo la minimización del coste (figura 12), para evaluar los resultados. Con los parámetros α y β determinamos el porcentaje dedicado a la maximización y minimización, dando prioridad a la producción o al coste. Se realizaron 3 pruebas principales incluidas en el anexo I: $\alpha = 1, \beta = 1$; $\alpha = 1, \beta = 0$; y $\alpha = 0, \beta = 1$. La primera de ellas priorizaba ambas optimizaciones por igual, ya que buscaba la maximización de la productividad de biomasa y minimización del coste. En este caso, se observó que no se lograban tan buenos resultados para la maximización como anteriormente, ya que tenía un doble objetivo. Para la segunda prueba, confirmamos unos resultados de productividad similares a la ejecución del programa sin coste pero nos permite conocer visualmente los valores del coste para dichas productividades. Por último, al minimizar únicamente el coste, obtuvimos precios más baratos pero también una productividad mucho menor.

Tabla 1. Mejores resultados obtenidos (ver figura 36, anexo I) tras la optimización mediante AG del medio de *N. gaditana*. Se muestran las concentraciones óptimas (g/L) obtenidas para el máximo nivel de productividad de biomasa ($0,86 \text{ g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$) alcanzado al mínimo coste ($0,13 \text{ €/L}$), valores logados mediante la función 18.

	Concentración (g/L)
KNO ₃	0,8
NaH ₂ PO ₄ ·H ₂ O	0,095
C ₆ H ₅ FeO ₇ ·H ₂ O	$1,4 \cdot 10^{-5}$
Na ₂ MoO ₄ ·2H ₂ O	$4,93 \cdot 10^{-5}$
MnCl ₂ ·4H ₂ O	$4,45 \cdot 10^{-5}$
ZnCl ₂	$9,72 \cdot 10^{-6}$
CoCl ₂ ·6H ₂ O	$1,94 \cdot 10^{-4}$
CuSO ₄ ·5H ₂ O	$2,72 \cdot 10^{-4}$
EDTA	0,018
Vitamina B1	$3,88 \cdot 10^{-3}$
Vitamina B7	$5,85 \cdot 10^{-5}$
Vitamina B12	$3,99 \cdot 10^{-5}$

En el caso de maximizar la productividad y minimizar el coste ($\alpha = 1, \beta = 1$), los mejores resultados fueron $2,21 \text{ g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$ y $0,22 \text{ €/L}$ en el caso de la función sin término independiente

(función 17, figura 30), y $0,86 \text{ g}\cdot\text{L}^{-1}\cdot\text{d}^{-1}$ y $0,13 \text{ €/L}$ para la mejor función con término independiente (función 18, figura 30). De las dos opciones se concluyó que el resultado era mejor en el segundo caso, mostrando el contenido del medio optimizado en la tabla 1.

Pudimos comprobar que la maximización de la productividad fue menor en la segunda función (sin término independiente), al igual que observamos en la ejecución sin el coste, pero decidimos que era más correcta por ser mejor la función (menor error). En cambio, el coste fue mucho menor en la segunda función lo que nos permitió corroborar que realizaba mejor optimización a nivel de minimización del coste. En conjunto, por errores de la función de ajuste, maximización de la productividad y minimización del coste, concluimos que la mejor función que permitía una optimización válida fue una función del tipo de la función 5 (figura 6) con término independiente (ver función 18, figura 30).

En este caso ($\alpha = 1, \beta = 1$), aumentar el número de generaciones más allá de 100 no mejoró visiblemente los resultados, pero variar las generaciones desde 20 hasta 100 mejoraba el resultado a medida que aumentaban. Los otros dos casos ($\alpha = 1, \beta = 0$ y $\alpha = 0, \beta = 1$) nos sirvieron para observar los resultados de maximización de productividad y minimización del coste por separado, que variaban a medida que se daba prioridad a una u otra optimización (ver anexo I).

4.3.2. Hidrólisis enzimática de plasma bovino

Una vez encontradas las funciones objetivo en los casos comentados anteriormente, detallamos los resultados obtenidos para el AG (ver anexo II). En cuanto a las respuestas (% de hidrólisis enzimática) obtenidas con las funciones encontradas de la forma de la función 5 (figura 6), los resultados variaron entre un 20-45% de hidrólisis; hay que tener en cuenta que, como comentamos anteriormente, las funciones comenten mucho error. Las respuestas obtenidas para las funciones en las cuales despreciamos los términos rondaron el 20-35% de hidrólisis. En ambos casos, se debe apreciar que el rango es tan amplio debido a la aleatoriedad de la función y su incapacidad de mejora, dando lugar a diversos resultados tras la ejecución iterativa del programa de búsqueda. Asimismo, los valores de las variables de entrada que optimizaron la respuesta se extendieron desde el mínimo valor experimental hasta el máximo, confirmando la ineficacia de la función para la optimización.

En último lugar, evaluamos la eficacia del AG con la función objetivo obtenida por regresión en wxMaxima. Por un lado, observamos que las respuestas obtenidas, sin despreciar ningún término de la ecuación, se aproximaron al 20-22% de hidrólisis enzimática, mientras que sin tener en cuenta los términos no significativos (al igual que en el artículo de Figueroa *et al.*

(ver página 4, sección 1)), la optimización de la respuesta resultó alrededor del 23%. Asimismo, las variables del proceso optimizadas (temperatura, pH y relación enzima/sustrato) se encontraron en diversos valores en cada ejecución del AG realizada sin despreciar los términos, a pesar de obtener unos resultados óptimos de respuesta en todos los casos. Sin embargo, para la ecuación reducida (términos no significativos despreciados), los valores de las variables optimizadas convergieron en un rango reducido, el cual se aproximó al resultado obtenido en el artículo, teniendo en cuenta que, para ello, es necesario establecer un número elevado de generaciones (aproximadamente 150) del AG; si se ejecutaba con 20 generaciones como en el caso del estudio, no siempre se optimizaba correctamente. De igual modo, es interesante señalar que el máximo resultado de la respuesta (% de hidrólisis enzimática) obtenido en el estudio de Figueroa *et al.* (ver página 4, sección 1) fue del 23,002%, mientras que con el AG desarrollado en este trabajo logramos obtener hasta un 23,56%. Tal progreso pudo deberse al uso de mayor número de generaciones en el AG, así como a la implementación de elitismo, ausente en el otro estudio. En consecuencia con el máximo valor obtenido para la respuesta (ver figura 42, anexo II), determinamos los valores óptimos de temperatura, pH y relación enzima/sustrato como 61,66 °C, pH 9 y 9,97 (% p/v), consistentes con la optimización del estudio a examinar y confirmando así la validez del mismo y el uso de nuestro AG desarrollado para la optimización de problemas similares.

4.4. Análisis estadístico

Tras finalizar las simulaciones y concluir los resultados, realizamos determinados análisis estadísticos con la ayuda de los programas GraphPad Prism (GraphPad Software Inc., 2018) y Microsoft Excel (Microsoft Corporation, 2010).

Teórico	Observado	Obs-Teo	(Obs-Teo)^2	((Obs-Teo)^2)/Teo	
0.46	0.4550017	-0.0049983	2.4983E-05	5.43109E-05	
0.368	0.3675961	-0.0004039	1.63135E-07	4.43302E-07	
0.285	0.2918856	0.0068856	4.74115E-05	0.000166356	
0.469	0.4676965	-0.0013035	1.69911E-06	3.62284E-06	
0.367	0.3629304	-0.0040696	1.65616E-05	4.51271E-05	
			SUMA	0.00026986	← χ^2
α	0.005				
n	5				
				14.860259	← $\chi^2_{1-\alpha}(n-1)$

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - T_i)^2}{T_i}$$

Figura 22. Muestra la prueba estadística de χ^2 realizada para el estudio de *N. gaditana* (ver página 3, sección 1), donde se observan las frecuencias teóricas (experimentales) y observadas (calculadas con la función objetivo). El nivel de confianza es del 99,5% ($\alpha=0,005$) y el número de muestras es 5, lo que nos lleva a adquirir un valor de 14,86 en las tablas de χ^2 . Al ser mayor el valor de las tablas que el calculado con la fórmula, se acepta la hipótesis nula, es decir, que los datos observados se ajustan a la distribución teórica y, por tanto, las diferencias observadas no son significativas.

Por un lado, efectuamos la prueba de χ^2 (figura 22) para los resultados del estudio de formulación de medios de *N. gaditana* que comprueba la bondad de ajuste del modelo, mientras que por otro, se encuentra el análisis estadístico de la RLM (figura 23) realizada para la investigación del porcentaje de hidrólisis enzimática (ver página 4, sección 1).

Y = $\beta_0 + \beta_1*B + \beta_2*C + \beta_3*D + \beta_4*B*C + \beta_5*B*D + \beta_6*C*D + \beta_7*B^2 + \beta_8*C^2 + \beta_9*D^2$						
Regression type		Least squares				
Model						
Parameter estimates		Variable	Estimate	Standard error		
β_0		Intercept	20.49	0.5747		
β_1		B : X1	-1.027	0.3116		
β_2		C : X2	0.6839	0.3116		
β_3		D : X3	1.134	0.3116		
β_4		B*C	-0.2375	0.4070		
β_5		B*D	-0.6125	0.4070		
β_6		C*D	-0.03750	0.4070		
β_7		B^2	-3.822	0.3241		
β_8		C^2	-0.3676	0.3241		
β_9		D^2	-0.2967	0.3241		
Sig. diff. than zero?		Variable	P value	P value summary		
β_0		Intercept	<0.0001	****		
β_1		B : X1	0.0109	*		
β_2		C : X2	0.0595	ns		
β_3		D : X3	0.0066	**		
β_4		B*C	0.5756	ns		
β_5		B*D	0.1707	ns		
β_6		C*D	0.9289	ns		
β_7		B^2	<0.0001	****		
β_8		C^2	0.2896	ns		
β_9		D^2	0.3868	ns		
Goodness of Fit						
Degrees of Freedom		8				
R squared		0.9561				
Analysis of Variance		SS	DF	MS	F (DFn, DFd)	P value
Regression		230.9	9	25.66	F (9, 8) = 19.36	P=0.0002
Residual		10.60	8	1.325		
Total		241.5	17			

Figura 23. Resultados de la RLM llevada a cabo con los datos del estudio de Figueroa *et al.* (ver página 4, sección 1). Incluye los parámetros del modelo que dan lugar a la función por MSR junto con sus respectivos errores e intervalos de confianza. También se observan los determinados parámetros no significativos que nos permiten despreciar dichos términos en el modelo, así como la bondad de ajuste. Igualmente, se muestra el análisis de varianza (ANOVA) que coincide con el realizado en el estudio de Figueroa *et al.*

5. CONCLUSIÓN Y PERSPECTIVAS FUTURAS

En la actualidad, hay diversos métodos utilizados para la mejora de medios de cultivo, además de otros problemas similares que requieren encontrar los valores óptimos de las variables evaluadas, permitiendo maximizar o minimizar la respuesta. Se han realizado varias investigaciones concernientes a la optimización mediante algoritmos genéticos de diversos problemas aunque aún no tenemos unos conocimientos idóneos sobre la optimización de

medios de cultivo con este método. Sí existen diversos estudios para mejorar las condiciones de cultivo con un bajo número de variables, pero a medida que aumentamos su número los resultados no son irrefutables. Por tanto, la presente investigación permite ampliar la información de los AG y corroborar su eficacia en este tipo de problemas, a pesar de requerir más estudios futuros para perfeccionar los métodos de optimización estudiados.

En este estudio, podemos concluir que la regresión por mínimos cuadrados es un método válido para encontrar una función de aproximación de datos que nos muestra cómo evoluciona la respuesta. Sin embargo, las limitaciones relacionadas con la licencia de programas capacitados para evaluar la regresión con más de 6 variables, nos impidieron corroborarlo para estos casos. Por tanto, la regresión sirvió para hallar la función objetivo del estudio de hidrólisis enzimática con un error aceptable.

De este modo, la búsqueda aleatoria de funciones (polinómicas y sigmoides), realizada por el programa de desarrollo propio, fue el método que nos permitió evaluar la efectividad de la aproximación de datos para ajustarse a la respuesta. Este procedimiento resultó ser eficiente en el caso de valorar problemas con bajo número de muestras experimentales, independientemente del número de variables, a diferencia de la regresión múltiple. Sin embargo, no era válido en aquellos casos donde el número de muestras era muy elevado. Por tanto, fue útil para evaluar la función del problema de optimización de los componentes del medio de *N. gaditana* pero no para la optimización de la hidrólisis enzimática bovina.

Para el proceso de optimización una vez conseguida la función objetivo necesaria, el AG mostró una capacidad excelente en ambos problemas estudiados, siendo más difícil la optimización multiobjetivo (aumentar productividad y reducir costes). Igualmente, resultó conveniente realizar un análisis estadístico de los modelos obtenidos, como pudimos comprobar al obtener mejores resultados para la hidrólisis enzimática con la función obtenida tras despreciar los términos no significativos (respecto a la función intacta). También confirmaron el ajuste del modelo teórico de la formulación del medio de *N. gaditana* a los resultados experimentales. En todos los casos, verificamos la eficacia de la normalización de los datos que se encuentran en distintos rangos, ya que mejoró el proceso de búsqueda de funciones de ajuste de datos.

Por tanto, logramos alcanzar el objetivo del estudio realizado para la optimización de la formulación del medio de *N. gaditana*, relevante en el campo de la acuicultura, obteniendo una productividad máxima de $0,86 \text{ g} \cdot \text{L}^{-1} \cdot \text{d}^{-1}$ a un precio mínimo de $0,13 \text{ €/L}$ en el mejor de

los resultados obtenidos (ver tabla 1 para concentraciones de los compuestos del medio de cultivo). Del mismo modo, confirmamos también la utilidad del presente trabajo en la mejora de las condiciones de la reacción de hidrólisis enzimática a partir de plasma bovino, útil en alimentación, dando lugar a un máximo de 23,56% de hidrólisis enzimática con unos valores óptimos de temperatura, pH y relación enzima/sustrato de 61,66 °C, pH 9 y 9,97 (% p/v), frente a un 23,002 % obtenido en el artículo de Figueroa *et al.*

Por último, consideramos distintas vías futuras que podría adoptar el presente trabajo. Por una parte, podríamos mejorar el Algoritmo híbrido (RLM, búsqueda aleatoria y AG) desarrollado, mediante la adición de nuevas funciones o la modificación de los operadores básicos de la parte genética del algoritmo (selección, cruce y mutación), así como cabría la posibilidad de aplicarlo a otros campos de estudio (medicina, ingeniería, economía,...). Por otra parte, una de las posibles alternativas a sopesar reside en la búsqueda de mejores medios (equipos o software con mayor capacidad de procesamiento) para lograr el hallazgo de una función objetivo mediante RLM en problemas con gran número de variables de entrada. Además, podríamos ajustar el algoritmo implementado en el presente trabajo a más datos experimentales procedentes de diversos campos para confirmar su eficacia de generalización.

6. REFERENCIAS

- ActiveState (2017) *ActivePerl (Versión 5.24.3) [Programa de ordenador]*. Disponible en: <https://www.activestate.com/products/perl/downloads/> (Accedido: 31 de mayo de 2020).
- Brinc, M. y Belič, A. (2019) "Optimization of process conditions for mammalian fed-batch cell culture in automated micro-bioreactor system using genetic algorithm", *Journal of biotechnology*, 300, pp. 40-47.
- Camacho-Rodríguez, J., Cerón-García, M. C., Fernández-Sevilla, J. M. y Molina-Grima, E. (2015) "Genetic algorithm for the medium optimization of the microalga *Nannochloropsis gaditana* cultured to aquaculture", *Bioresource technology*, 177, pp. 102-109.
- Camacho-Rodríguez, J., Cerón-García, M. C., González-López, C. V., Fernández-Sevilla, J. M., Contreras-Gómez, A. y Molina-Grima, E. (2013) "A low-cost culture medium for the production of *Nannochloropsis gaditana* biomass optimized for aquaculture", *Bioresource technology*, 144, pp. 57-66.
- Camacho-Rodríguez, J., González-Céspedes, A. M., Cerón-García, M. C., Fernández-Sevilla, J. M., Acien-Fernández, F. G. y Molina-Grima, E. (2014) "A quantitative study of eicosapentaenoic acid (EPA) production by *Nannochloropsis gaditana* for aquaculture as a function of dilution rate, temperature and average irradiance", *Applied microbiology and biotechnology*, 98(6), pp. 2429-2440.
- Carrasquilla-Batista, A., Chacón-Rodríguez, A., Núñez-Montero, K., Gómez-Espinoza, O., Valverde, J. y Guerrero-Barrantes, M. (2016) "Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal", *Tecnología en marcha*, 29(8), pp. 33-45.
- Chen, M., Tang, H., Ma, H., Holland, T. C., Ng, K. Y. S. y Salley, S. O. (2011) "Effect of nutrients on growth and lipid accumulation in the green algae *Dunaliella tertiolecta*", *Bioresource technology*, 102(2), pp. 1649-1655.
- Choi, J. W., Lee, H., Lee, J. C., Lee, S., Kim, Y. S., Yoon, H. J. y Kim, H. C. (2017) "Application of genetic algorithm for hemodialysis schedule optimization", *Computer methods and programs in biomedicine*, 145, pp. 35-43.
- Desai, K. M., Survase, S. A., Saudagar, P. S., Lele, S. S. y Singhal, R. S. (2008) "Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: case study of fermentative production of scleroglucan", *Biochemical engineering journal*, 41(3), pp. 266-273.
- El-Desoky, I. M., El-Shorbagy, M. A., Nasr, S. M., Hendawy, Z. M. y Mousa, A. A. (2016) "A hybrid genetic

algorithm for job shop scheduling problems", *International journal of advancement in engineering, technology and computer sciences*, 3(1), pp. 6-17.

Figuroa, O. A., Zapata, J. E. y Sánchez, C. P. (2016) "Optimización de la hidrólisis enzimática de proteínas de plasma bovino", *Informacion tecnológica*, 27(2), pp. 39-52.

Frederick Schelter, W. (1982) *Maxima (Versión 5.38.1) [Programa de ordenador]*. Disponible en: <http://maxima.sourceforge.net/es/download.html> (Accedido: 31 de mayo de 2020).

Galbraith, S. C., Bhatia, H., Liu, H. y Yoon, S. (2018) "Media formulation optimization: current and future opportunities", *Current opinion in chemical engineering*, 22, pp. 42-47.

García-Tejeda, Y. V. y Barrera-Figueroa, V. (2019) "Least squares fitting-polynomials for determining inflection points in adsorption isotherms of spray-dried açai juice (*Euterpe oleracea* Mart.) and soy sauce powders", *Powder technology*, 342, pp. 829-839.

GraphPad Software Inc. (2018) *GraphPad Prism (Versión 8.0.1) [Programa de ordenador]*. Disponible en: <https://graphpad-prism.software.informer.com/Descargar-gratis/> (Accedido: 31 de mayo de 2020).

Ho, D. (2003) *NotePad++ (Versión 7.8.4) [Programa de ordenador]*. Disponible en: <https://notepad-plus-plus.org/downloads/> (Accedido: 31 de mayo de 2020).

Kar, A. K. (2016) "Bio inspired computing - A review of algorithms and scope of applications", *Expert systems with applications*, 59, pp. 20-32.

Kennedy, M. y Krouse, D. (1999) "Strategies for improving fermentation medium performance: a review", *Journal of industrial microbiology and biotechnology*, 23(6), pp. 456-475.

Konak, A., Coit, D. W. y Smith, A. E. (2006) "Multi-objective optimization using genetic algorithms: a tutorial", *Reliability engineering and system safety*, 91(9), pp. 992-1007.

Lee, C. K. H. (2018) "A review of applications of genetic algorithms in operations management", *Engineering applications of artificial intelligence*, 76, pp. 1-12.

Mahajan, D., Sengupta, S. y Sen, S. (2019) "Strategies to improve microbial lipid production: optimization techniques", *Biocatalysis and agricultural biotechnology*, 22, 101321.

Manzoni, L., Mariot, L. y Tuba, E. (2020) "Balanced crossover operators in genetic algorithms", *Swarm and evolutionary computation*, 54, 100646.

Microsoft Corporation (2010) *Microsoft Excel (Versión 14.0) [Programa de ordenador]*. Disponible en: <https://www.microsoft.com/es-cl/microsoft-365/get-office-and-microsoft-365-oem-download-page> (Accedido: 31 de mayo de 2020).

Passos, G. C. S. y Barrenechea, M. H. (2019) "Genetic algorithms applied to an evolutionary model of industrial dynamics", *Economia*. doi:10.1016/j.econ.2019.09.002.

Raposo, F. (2016) "Evaluation of analytical calibration based on least-squares linear regression for instrumental techniques: a tutorial review", *Trends in analytical chemistry*, 77, pp. 167-185.

Shirodkar, P. V. y Muraleedharan, U. D. (2017) "Enhanced α -amylase production by a marine protist, *Ulkenia* sp. using response surface methodology and genetic algorithm", *Preparative biochemistry and biotechnology*, 47(10), pp. 1043-1049.

Shojaedini, E., Majd, M. y Safabakhsh, R. (2019) "Novel adaptive genetic algorithm sample consensus", *Applied soft computing journal*, 77, pp. 635-642.

Sirohi, R., Singh, A., Tarafdar, A. y Shahi, N. C. (2018) "Application of genetic algorithm in modelling and optimization of cellulase production", *Bioresource technology*, 270, pp. 751-754.

Urso, A., Fiannaca, A., La Rosa, M., Ravì, V. y Rizzo, R. (2019) "Data mining: prediction methods", en Ranganathan, S., Nakai, K., Schönbach, C., y Gribskov, M. (eds.) *Encyclopedia of bioinformatics and computational biology*. Amsterdam: Elsevier, pp. 413-430.

Venkateswarlu, C. y Jujjavarapu, S. E. (2020a) "Application of stochastic evolutionary optimization techniques to pharmaceutical processes", en Venkateswarlu, C. y Jujjavarapu, S. E. (eds.) *Stochastic global optimization methods and applications to chemical, biochemical, pharmaceutical and environmental processes*. Amsterdam: Elsevier Inc., pp. 223-246.

Venkateswarlu, C. y Jujjavarapu, S. E. (2020b) "Basic features and concepts of optimization", en Venkateswarlu, C. y Jujjavarapu, S. E. (eds.) *Stochastic global optimization methods and applications to chemical, biochemical, pharmaceutical and environmental processes*. Amsterdam: Elsevier Inc., pp. 1-25.

Venkateswarlu, C. y Jujjavarapu, S. E. (2020c) "Stochastic and evolutionary optimization algorithms", en Venkateswarlu, C. y Jujjavarapu, S. E. (eds.) *Stochastic global optimization methods and applications to chemical, biochemical, pharmaceutical and environmental processes*. Amsterdam: Elsevier Inc., pp. 87-123.

Vodopivec, A., Lenarcic, Z., Ilijev, D. y Königsmann, G. (2004) *wxMaxima (Versión 16.04.2) [Programa de ordenador]*. Disponible en: <https://sourceforge.net/projects/wxmaxima/> (Accedido: 31 de mayo de 2020).