Laura S. Valencia-Colman[1] **& Édgar E. Daza C.**[1*]

[1]Grupo de Química Teórica, Universidad Nacional de Colombia, Bogotá D. C., Colombia.

***Autor para correspondencia:** eedazac@unal.edu.co

Recibido: 09/03/2022. **Última revisión:** 08/06/2022. **Aceptado:** 13/07/2022.

Reconocimiento de rutas biosintéticas para semioquímicos mediante técnicas de aprendizaje de máquina

Resumen

En este trabajo consideramos 148 semioquímicos reportados para la familia *Scarabaeidae*, cuya estructura química fue caracterizada empleando un conjunto de 200 descriptores moleculares de cinco clases distintas. La selección de los descriptores más discriminantes se realizó con tres técnicas: análisis de componentes principales, por cada clase de descriptores, bosques aleatorios y Boruta-Shap, aplicados al total de descriptores. A pesar de que las tres técnicas son conceptualmente diferentes, seleccionan un número de descriptores similar de cada clase. Propusimos una combinación de técnicas de aprendizaje de máquina para buscar un patrón estructural en el conjunto de semioquímicos y posteriormente realizar la clasificación de estos. El patrón se estableció a partir de la alta pertenencia de un subconjunto de estos metabolitos a los grupos que fueron obtenidos por un método de agrupamiento basado en lógica difusa, C-means; el patrón descubierto corresponde a las rutas biosintéticas por las cuales se obtienen biológicamente. Esta primera clasificación se corroboró con el empleo de mapas autoorganizados de Kohonen. Para clasificar aquellos semioquímicos cuya pertenencia a una ruta no quedaba claramente definida, construimos dos modelos de perceptrones multicapa, los cuales tuvieron un desempeño aceptable.

Palabras clave: bosques aleatorios; C-means; descriptores moleculares; familia *Scarabaeidae*; perceptrón multicapa; redes neuronales.

Recognition of biosynthetic pathways for semiochemicals using machine learning techniques

Abstract

In this work we consider 148 semiochemicals reported for the family *Scarabaeidae*, whose chemical structure was characterized using a set of 200 molecular descriptors from five different classes. The selection of the most discriminating descriptors was carried out with three different techniques: Principal Component Analysis, for each class of descriptors, Random Forests and Boruta-Shap, applied to the total of descriptors. Although the three techniques are conceptually different, they select a similar number of descriptors from each class. We proposed a combination of machine learning techniques to search for a structural pattern in the set of semiochemicals and then perform their classification. The pattern was established from the high belonging of a subset of these metabolites to the groups that were obtained by a grouping method based on fuzzy C-means logic; the discovered pattern corresponds to the biosynthetic pathway by which they are obtained biologically. This first classification was corroborated with Kohonen's self-organizing maps. To classify those semiochemicals whose belonging to a biosynthetic pathway was not clearly defined, we built two models of Multilayer Perceptrons which had an acceptable performance.

Keywords: Random forests; C-means; molecular descriptors; family *Scarabaeidae*; multilayer perceptron; neural networks.

Reconhecimento de vias biossintéticas para semioquímicos usando técnicas de aprendizado de máquina

Resumo

Neste trabalho consideramos 148 semioquímicos reportados para a família *Scarabaeidae,* cuja estrutura química foi caracterizada usando um conjunto de 200 descritores moleculares de 5 classes diferentes. A seleção dos descritores mais discriminantes foi realizada com três técnicas diferentes: Análise de Componentes Principais, para cada classe de descritores, Florestas Aleatórias e Boruta-Shap, aplicadas a todos os descritores. Embora as três técnicas sejam conceitualmente diferentes, elas selecionaram um número semelhante de descritores de cada classe. Nós propusemos uma combinação de técnicas de aprendizado de máquina para buscar um padrão estrutural no conjunto de semioquímicos e então realizar sua classificação. O padrão foi estabelecido a partir da alta pertinência de um subconjunto desses metabólitos aos grupos que foram obtidos por um método de agrupamento baseado em lógica fuzzy, C-means; o padrão descoberto corresponde às rotas biossintéticas pelas quais eles são obtidos biologicamente. Essa primeira classificação foi corroborada com o uso dos mapas auto-organizados de Kohonen. Para classificar os semioquímicos cuja pertença a uma rota não foi claramente definida, construímos dois modelos de Perceptrons Multicamadas que tiveram um desempenho aceitável.

Palavras-chave: florestas aleatórias; C-means; descritores moleculares; família *Scarabaeidae*; perceptron multicamadas; redes neurais.

Introducción

La ecología química es un campo de conocimiento transdisciplinar que cada vez gana mayor relevancia [1], [2]. Parte fundamental de esta disciplina es la identificación de las moléculas que median la comunicación entre insectos o planta-insecto. En la medida en que se conocen más semioquímicos y se acumulan datos sobre su uso, resulta conveniente la creación de herramientas que contribuyan de una manera rápida y eficiente a establecer patrones y relaciones sobre esta información y así aportar a futuras investigaciones.

En este como en otros campos de la química, las relaciones entre estructura y actividad están en el centro del problema y parte fundamental de este es lograr una buena representación cuantitativa de la estructura química; es decir, una que pueda procesarse computacionalmente [3]. De ahí que actualmente se hayan definido miles de descriptores moleculares que buscan codificar distintos aspectos de la estructura; por ejemplo, características topológicas, propiedades fisicoquímicas, grupos funcionales u otras propiedades dependientes de la conformación espacial de los componentes moleculares [4]. Sin embargo, la selección de los descriptores apropiados es un problema abierto que depende del objetivo que se persigue [5], [6]. Alternativamente, se puede acudir a conjuntos redundantes de los mismos como estrategia multiobjetivo para la búsqueda de patrones [7], [8].

El sistema insecto-semioquímico y la representación de estas moléculas constituyen un sistema complejo para el cual se ha generado y se sigue generando información. Este sistema constituye un espacio propicio para ser explorado mediante técnicas de minería de datos y aprendizaje de máquina, cuya utilización en diversos campos de la ciencia y la tecnología ha sido exitosa. En química, por ejemplo, se han implementado algoritmos de aprendizaje automático de distinta naturaleza que solventan problemas en diversas áreas, que abarcan desde la química analítica hasta la catálisis, pasando por la química orgánica y la química computacional [9]. Algunos de estos estudios apuntan a tratar relaciones estructura-actividad (SAR) por medio del desarrollo de herramientas de aprendizaje automático que facilitan el diseño y la selección de moléculas con la actividad esperada. En particular, algunos hacen uso de perceptrones multicapa y redes neuronales profundas para la clasificación de moléculas [10], [11].

En este trabajo nos proponemos comparar un conjunto de metodologías para la selección de las variables más relevantes y métodos de agrupamiento que pueden ser empleados para el descubrimiento de patrones estructurales en el conjunto de los semioquímicos reportados para los coleópteros de la familia *Scarabaeidae*. A la vez, implementamos modelos de aprendizaje de máquina capaces de realizar la clasificación de este tipo de metabolitos en las categorías establecidas por el patrón descubierto.

Materiales y métodos

Sistema de estudio

Para este estudio se empleó un modelo constituido por un conjunto de 148 semioquímicos reportados para 240 especies de insectos de la familia *Scarabaeidae* (orden: *Coleoptera*) que se almacenó en una base de datos relacional SQL [12][1].

La estructura química de los compuestos considerados se caracterizó mediante un conjunto redundante de descriptores moleculares de diferentes clases. Así, con el programa RDKit [13] se calcularon 188 descriptores, los cuales clasificamos en cuatro clases: la primera constituida por descriptores 19 derivados esencialmente de propiedades grafo-teóricas, la segunda corresponde a 106 descriptores que dan cuenta de la constitución de las moléculas según la presencia de ciertos fragmentos o tipos de enlaces, la tercera consta de 58 descriptores basados en propiedades que se calculan sobre superficies tridimensionales asociadas a las moléculas y la cuarta está conformada por cinco descriptores asociados a propiedades fisicoquímicas. El listado de los descriptores calculados puede consultarse en la siguiente dirección URL: https://www.rdkit.org/docs/GettingStartedInPython.html#descriptor-calculation.

Además, se calcularon 12 descriptores de naturaleza cuántica (HOMO; LUMO; constantes rotacionales en X, Y y Z; extensión espacial electrónica; momento dipolar en X, Y, Z y el total; electronegatividad y dureza) derivados de la matriz de densidad obtenida con el método de funcionales de la densidad B3LYP/6-31(d,p) implementado en Gaussian 09 [14]. De esta forma se propone un total de 200 descriptores moleculares de cinco clases diferentes.

Dado que hemos seleccionado un conjunto redundante de descriptores, es posible que se presenten dependencias entre algunos de ellos o que hayamos incluido algunos poco relevantes en el momento de reproducir un patrón de clasificación; por lo tanto, es necesario seleccionar los descriptores más discriminatorios para evitar sesgos. Este proceso se llevó a cabo de manera paralela mediante tres métodos:

(i) Análisis de componentes principales (ACP) [15]: el ACP suele emplearse para reducir la dimensionalidad del espacio de representación; para ello se definen unas nuevas variables mediante una transformación que asegura que la mayor parte de la varianza de los datos sea explicada por unas pocas de estas variables, los componentes principales. Cuando se emplea para seleccionar entre las variables originales deben escogerse aquellas (los descriptores) que más contribuyen a la conformación de los componentes principales; como criterio de selección se propuso que la suma de los cuadrados de los coeficientes del descriptor con que contribuye a los componentes principales, que describen hasta el 70% de la varianza acumulada de los datos, tuviese un valor superior a 0,3 (la suma sobre todos los componentes principales es por definición la unidad). La selección de descriptores se realizó por separado para las cinco clases mencionadas.

(ii) Bosques aleatorios (BA) [16]: BA es un método de aprendizaje supervisado, basado en árboles de decisión, para clasificar objetos en función de un patrón previamente determinado. Este método establece unas *puntuaciones de importancia* para cada variable según la impureza de Gini, la cual es una medida que evalúa la distribución de los datos por nodo y establece qué tan óptima es una escisión de los datos respecto a cada variable. Estas *puntuaciones de importancia* permiten seleccionar las variables más significativas para reproducir una clasificación propuesta; para este estudio se seleccionaron los descriptores con puntuaciones superiores a 0,41%; este valor nos permitió tener el mismo número de variables que fue hallado mediante el uso de ACP.

(iii) Boruta-Shap (BS) [17]: BS es un método de selección de variables que combina el algoritmo Boruta y la técnica Shap. El primero realiza una selección y eliminación iterativa de variables no relevantes para una función objetivo o patrón de clasificación, teniendo como criterio las *puntuaciones de importancia* de unas variables "sombra" (las cuales son una combinación aleatoria de las variables originales); y la segunda establece las variables que presentan una mayor influencia en las predicciones de los modelos de aprendizaje de máquina con base en la teoría de juegos cooperativos.

Para el ACP se empleó la librería *stasts* disponible para lenguaje de programación R [18] y para los métodos de BA y BS se emplearon las librerías *SciKit-learn* [19] y *Boruta-Shap* [17], respectivamente disponibles en Python[2]. Los tres resultados de selección de variables se comparan más adelante.

Reconocimiento de patrones y técnicas de agrupamiento

Si bien existe una gran cantidad de métodos para el descubrimiento de patrones que utilizan algoritmos no supervisados, en este trabajo se utilizaron dos técnicas diferentes: el método de lógica difusa C-means (FCM, por su

1 La lista de las moléculas, así como los descriptores empleados para caracterizarlas, pueden ser solicitados a los autores.

2 Los códigos fuente están accesibles por solicitud a los autores.

nombre en inglés) [20] y los mapas autoorganizados de Kohonen (SOM, por sus siglas en inglés) [21]. Estos últimos con la intención de corroborar la plausibilidad de posible patrón hallado con el primero.

Los métodos de agrupamientos no jerárquicos como el FCM requieren de la definición *a priori* de un número de centroides o semillas. Probamos un número de semillas entre 2 y 5 y que en todos los casos se observan tres grupos; alternativamente ensayamos diversos métodos de agrupamiento jerárquico, pero no logramos la reproducibilidad de ninguna clasificación [22]. El algoritmo FCM se ejecutó en R [18] con un valor de 2 para el parámetro de difuminado y un criterio de convergencia de 1×10^{-9}.

Las redes neuronales SOM fueron construidas en Python [19] usando la biblioteca MiniSOM [23] y constan de *m* neuronas de entrada; *m* corresponde al número de variables seleccionadas (descriptores moleculares relevantes) que alimentan la red, y 225 neuronas de salida (15x15); lo cual corresponde a la arquitectura óptima respecto al error de cuantización y al número de neuronas que han de especializarse en cada variable. Las SOM fueron entrenadas con una tasa de aprendizaje de 0,01, con un factor sigma de 1,5, durante 1.000 épocas, y como función de vecindad se empleó una función gaussiana. La arquitectura de este sistema neuronal corresponde a una red cuadrada. Para su entrenamiento se emplearon los tres conjuntos de descriptores determinados como los más significativos con los tres métodos mencionados anteriormente [21].

Para clasificar las moléculas no adscritas con certeza a alguno de los grupos encontrados se diseñó y construyó un clasificador multiclase, un perceptrón multicapa (MLP) [24]. Se trata de un algoritmo basado en aprendizaje supervisado, conformado por cinco capas de neuronas: una de entrada de 53 neuronas, una de salida de tres neuronas y tres capas ocultas con 60, 40 y 20 neuronas, respectivamente. Para esta red neuronal se utilizó la función sigmoidea como función de activación y el optimizador de costos "adam" con una tasa de aprendizaje de 0,0001. Esta red neuronal también fue programada en Python usando la biblioteca SciKit-learn [19].

Resultados y discusión

Para consignar la información en nuestra base de datos, se partió de la reportada, para el suborden Polyphaga, en Pherobase [25]. Esta última es una base de datos de libre acceso tanto para consulta como para el registro de datos, por lo cual es común que presente inconsistencias e información errada[3]. En consecuencia, se depuró la información consultando la bibliografía primaria, es decir, un total de 957 artículos reportados en la literatura especializada.

Como anotamos, para caracterizar la estructura química de los semioquímicos empleamos un conjunto redundante de descriptores; por lo tanto, la información que codifican algunas de estas variables puede referirse a un mismo aspecto, ya sea porque fueron propuestos para dar cuenta de una misma característica estructural o bien porque fueron derivados de un mismo descriptor fundamental por autores diferentes. Consideramos que usar un conjunto redundante de variables se justifica en la medida en que no existe el fundamento teórico que permita caracterizar de una manera cuantitativa y unívoca el concepto de estructura química, quizá uno de los más complejos de esta ciencia. La selección de los descriptores más relevantes la hicimos primero mediante ACP y, como mencionamos, la selección de las variables originales que más participan en los componentes principales se realizó por separado para cada una de las seis clases de descriptores moleculares. Este procedimiento lo planteamos con el fin de asegurar que el conjunto final mantuviese lo esencial que pretende codificar cada una de las seis clases de descriptores. Por último, constatamos que no existe mayor correlación entre las variables seleccionadas. Esta metodología nos permitió reducir el conjunto de descriptores; así, pasamos de un conjunto de 200 a uno de 53, que reconocemos como las variables más significativas.

3 Un listado de las inconsistencias y errores detectados se encuentra disponible por solicitud.

Los semioquímicos considerados fueron agrupados mediante el método FCM; como anotamos anteriormente, se observa la existencia de tres grandes grupos. Al comparar los valores de pertenencia de cada molécula a cada grupo notamos que para el grupo R1 hay moléculas con pertenencia superior a 0,85. Para los grupos R2 y R3 las pertenencias más altas apenas superan un valor de 0,45; estos valores indican que hay moléculas con pertenencias similares a estos dos grupos. Escogimos las moléculas con mayor pertenencia a cada grupo para definir su núcleo, de manera que pudiésemos centrar nuestra atención en pequeños conjuntos de moléculas que según FCM deberían ser bastante similares entre sí. Así fue posible identificar un patrón estructural: el núcleo del grupo R1 está formado por hidrocarburos lineales; el núcleo del grupo R2 incluye moléculas aromáticas o con dobles enlaces conjugados y el núcleo del grupo R3 por moléculas de bajo peso molecular que incluye alcoholes, aldehídos, cetonas y ésteres. (Ver Tabla 1).

Vale la pena destacar que las moléculas que conforman cada uno de los tres núcleos propuestos corresponden a metabolitos secundarios que comparten un mismo origen biosintético; por lo tanto, el patrón estructural que se develó con FCM corresponde a la clasificación según las respectivas rutas biosintéticas [26], al menos para las moléculas de alta pertenencia a cada grupo (Véase Figura 1).

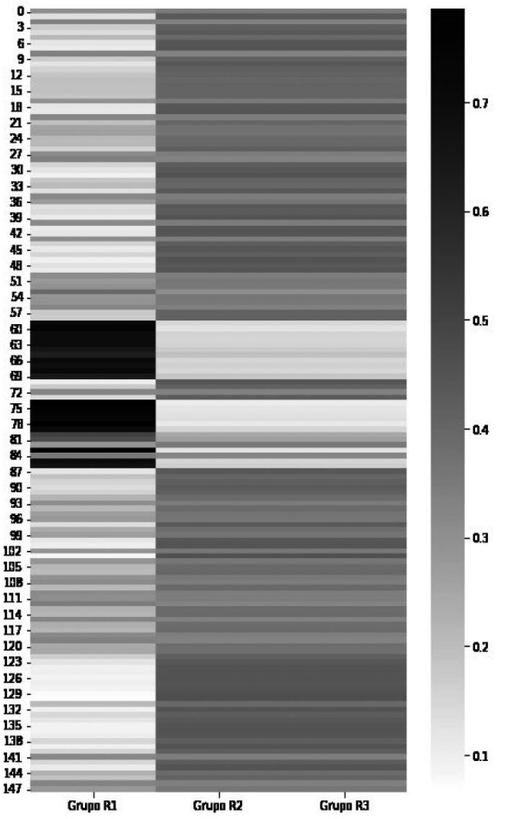

Figura 1. Representación del grado de pertenencia de cada una de las moléculas a los tres grupos establecidos por el método de lógica difusa FCM. Para el grupo R1, se observa una pertenencia superior a 0,8 para un grupo pequeño de moléculas. Para los otros dos grupos, R2 y R3, las moléculas tienen grados de pertenencia bastante parecidos, aunque bajos; las de mayor pertenencia se diferencian en milésimas.

Una vez hemos reconocido un patrón correspondiente a las rutas biosintéticas, podemos emplearlo como la función de respuesta que requieren las técnicas de BA y BS para hacer la selección de descriptores. Para la técnica de bosques aleatorios, al igual que en el ACP, obtuvimos un conjunto final de 53 descriptores y con Boruta-Shap este conjunto aumentó

a 60 variables[4]. Al comparar los tres conjuntos de variables seleccionadas, encontramos que los métodos de selección dirigida, los cuales consideran el conjunto total de variables, dan lugar a conjuntos de variables óptimas similares a los que hallamos mediante ACP, que realizamos por aparte para cada clase de descriptores. El número de descriptores por tipo seleccionados con cada técnica puede verse en la Tabla 2.

Tabla 1. Ejemplo de moléculas pertenecientes al núcleo de cada grupo, los valores de pertenencia y su asociación con su respectiva ruta biosintética [26].

Semioquímico	Grupo	Pertenencia	Ruta biosintética
9-Pentacoseno	R1	0,899650	Ácidos grasos
11-Metiltricosano		0,886590	
Tetracosano		0,850447	
Tricosano		0,850364	
9,10-Hexacosadieno		0,849234	
Propanoato de fenilo	R2	0,470915	Ácido shikímico
Benzoato de metilo		0,469101	
1,2-Dimetoxibenceno		0,469077	
2-Fenilacetaldehído		0,467668	
Metoxibenceno		0,467480	
4-Pentenoato de etilo	R3	0,481187	Ácido acético
Acetato de (E)-2-hexenilo		0,478890	
4-Pentenoato de metilo		0,477727	
Acetato de (Z)-3-hexenilo		0,476378	
Pentanoato de etilo		0,473211	

Tabla 2. Distribución de descriptores moleculares seleccionados mediante cada metodología por clase de descriptor.

Técnicas de selección de características			
Clase de descriptores	ACP	BA	BS
Calculados sobre superficies (3D)	11	14	18
Constitucionales	20	12	12
Cuánticos	10	5	8
Grafo-teóricos	10	18	18
Propiedades fisicoquímicas	2	4	5

Por otra parte, al repetir el agrupamiento con el FCM con las variables derivadas de BA y BS encontramos que los núcleos en cada agrupamiento se conservan. Para corroborar estos agrupamientos empleamos mapas autoorganizados de Kohonen (SOM). El entrenamiento y la evaluación de los SOM se llevó a cabo con los 102 compuestos que se encuentran en los núcleos de los grupos derivados del FCM y los tres conjuntos de variables.

4 Un listado de las variables seleccionadas con cada uno de los métodos se encuentra disponible por solicitud a los autores.

Como resultado obtuvimos tres redes neuronales que logran reproducir la clasificación previamente conseguida según la ruta biosintética, como se muestra en la Figura 2. En ella se observa que el mejor agrupamiento se obtuvo al emplear los descriptores derivados del método de bosques aleatorios, pues separa de mejor manera las distintas moléculas por fronteras de neuronas con baja activación (ver Figura 2), y además presenta el error de cuantización [21] más bajo de las tres: 1,68. A su vez, los SOM alimentados con los descriptores derivados del ACP y BS no logran separar tan claramente algunas moléculas; este es el caso de los subgrupos que se forman para la ruta del ácido shikímico (R2) y la ruta del ácido acético (R3), según se aprecia en las Figuras 2a y 2c; además, el error de cuantización fue mayor: 2,78 y 2,04, respectivamente.

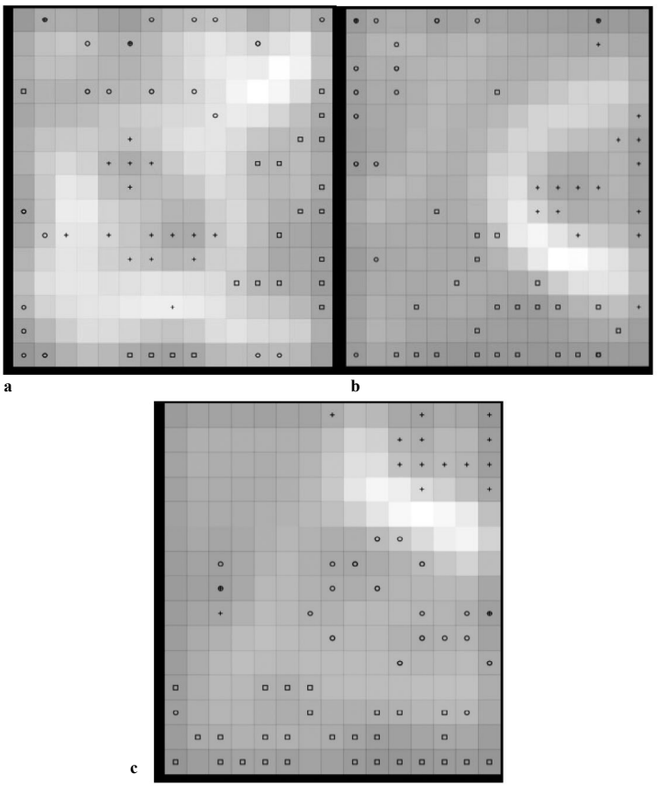

Figura 2. Mapas autoorganizados de Kohonen entrenados con los descriptores resultantes de la selección por: a. Análisis de componentes principales, b. Bosques aleatorios y c. Boruta-Shap. Las neuronas que presentan un color más claro tienen una menor activación y constituyen las fronteras entre los grupos: **O** Ruta del ácido shikímico, + Ruta de los ácidos grasos, **+** Ruta del ácido acético. Nótese que algunas moléculas pertenecientes a rutas distintas activan una misma neurona.

Si bien fue posible hallar un patrón a partir del método de agrupamiento difuso, y se corroboró la consistencia del mismo con los SOM, aún persisten 46 moléculas que no presentan un grado claro de pertenencia a alguna agrupación o ruta biosintética. Para complementar el modelo de aprendizaje de máquina que estamos presentando, empleamos perceptrones multicapa para clasificar estos 46 semioquímicos. Presentamos dos MLP, uno que usa el conjunto de descriptores determinado por ACP (MLP//ACP), pues fue el punto de partida para establecer el patrón de clasificación, y el otro con los descriptores seleccionados por el método de BA (MLP//BA), ya que estos demostraron ser el conjunto de variables más apropiado para diferenciar las clases en que buscamos clasificar los semioquímicos. Al igual que los SOM, los MLP fueron entrenados con los mismos 102 compuestos bien definidos. Este conjunto se dividió aleatoriamente en una proporción 80-20, en que el 80% de los semioquímicos se empleó para el entrenamiento y el 20% restante se utilizó como conjunto de prueba.

Para ambos modelos se obtuvo una concordancia del 100% entre la respuesta obtenida y la respuesta esperada para el conjunto de entrenamiento. En cuanto al conjunto de prueba, el modelo MLP//ACP presentó una concordancia del 95%, mientras que para el modelo MLP//BA la concordancia fue del 100%. La precisión y la sensibilidad en la clasificación del conjunto de prueba se pueden ver en la Tabla 3. Estos dos modelos se emplearon para establecer las rutas biosintéticas del conjunto de 46 moléculas problema. Dado que se trata de un conjunto pequeño de compuestos a clasificar, pudimos revisar de manera manual la predicción para cada uno de ellos. Así, con los dos modelos 31 de las 46 moléculas son clasificadas de la siguiente manera: 18 en la ruta del ácido shikímico (R2), nueve en la ruta del ácido acético (R3) y cuatro en la ruta de los ácidos grasos (R1). No obstante, de estas 31 moléculas ocho no parecen bien clasificadas; se trata de terpenoides que en principio deberían estar en la ruta R3, pero que se reparten entre la R1 y la R2. Consideramos que esto puede deberse a que la biosíntesis de esta clase de metabolitos involucra múltiples mecanismos que generan una gran diversidad estructural [27] o bien a una deficiente representación de este tipo de compuestos. No obstante, quizá lo más factible es que estas clasificaciones erradas se deban al tamaño de la muestra empleada, en la cual no se cuenta con suficiente representación de todas las posibles categorías.

Por otra parte, quedan 15 moléculas para las cuales se predicen clasificaciones distintas con cada modelo. Al respecto, debe tenerse en cuenta que algunas moléculas se pueden obtener tanto de la ruta de los ácidos grasos como de la ruta del ácido acético; esto podría explicar lo que ocurre con seis de las 15 moléculas (ver Tabla 4, aquellas marcadas con *). Además, debemos recordar que la ruta del ácido acético también da origen a compuestos aromáticos, lo que podría justificar la clasificación de tres de las nueve restantes (ver Tabla 4, aquellas marcadas con †), resultado que solo se presenta para el MLP//ACP. Las otras seis están bien clasificadas por alguno de los dos modelos.

Tabla 3. Precisión y sensibilidad para las clasificaciones obtenidas para el conjunto de prueba con cada modelo.

Rutas biosintéticas	Precisión		Sensibilidad	
	MLP//ACP	MLP//BA	MLP//ACP	MLP//BA
R2	1,00	1,00	0,86	1,00
R3	0,88	1,00	1,00	1,00
R1	1,00	1,00	1,00	1,00

En vista de que algunas equivocaciones en la clasificación son más graves que otras, empleamos el coeficiente Kappa de Cohen ponderado [28] para comparar los modelos y evaluar el nivel de acuerdo entre las predicciones y las clasificaciones. Para ello propusimos las siguientes penalidades: tres para las moléculas que presuntamente derivan del ácido shikímico y los perceptrones las clasifican como compuestos derivados de la ruta del ácido acético o de los ácidos grasos y uno para los otros dos casos. Así, Kappa para el perceptrón entrenado con los descriptores derivados de bosques aleatorios (MLP//BA) es de 0,57, mientras que para el modelo entrenado con los descriptores derivados del ACP (MLP//ACP) es de 0,54. En ese sentido, se presenta un mayor acuerdo entre la respuesta esperada y su predicción para el modelo MLP//BA. Si bien el valor $\kappa = 0,57$ se interpreta como un resultado aceptable para las clasificaciones, ha de tenerse en cuenta que el sistema aquí considerado cuenta con muy pocos registros y, a pesar de esto, la metodología propuesta arroja resultados promisorios.

Tabla 4. Moléculas con un grado de pertenencia no definido en FCM y para las cuales se presentan resultados ambivalentes empleando los descriptores derivados del ACP y determinados por bosques aleatorios usando el perceptrón multicapa.

Molécula	Ruta biosintética	Predicción con MLP//ACP	Predicción con MLP//BA
(E)-3,7-Dimetil-2,6-octadien-1-ol	Ac. acético R3	R3	R2
(E)-2-Nonenal	Ac. acético R3	R3	R2
Fenilmetanol	Ac. shikímico R2	R3†	R2
2-Feniletanol	Ac. shikímico R2	R3†	R2
2,6-Dimetil-2,7-octadien-6-ol	Ac. acético R3	R3	R2
Ácido 2,6-dimetil-5-heptenoico	Ac. acético R3	R3	R2
1-Feniletanol	Ac. shikímico R2	R3†	R2
(R)-5-Metil-2-(prop-1-en-2-il)-hex-4-en-1-ol	Ac. acético R3	R3	R2
2,2-dimetil-3-(2-metilpropenil)-ciclopropanecarboxilato de etilo	Ac. acético R3	R1*	R2
4-(2,6,6-Trimetil-1-ciclohexenil)-butan-2-ona	Ac. acético R3	R2	R1*
(R)-(Z)-5-(Dec-1-enil)-oxaciclopentan-2-ona	Ac. grasos R1	R2	R1*
Ftalato de dibutilo	Ac. shikímico R2	R2	R1
2,6,6-Trimetilbiciclo[3.1.1]hept-2-eno	Ac. acético R3	R2	R1*
Anhídrido 2,3-dimetil-7-oxabiciclo[1,2,2]heptano-2,3-dicarboxílico	Ac. acético R3	R2	R1*
(1S,5S)-4,6,6-Trimetilbiciclo[3.1.1]hept-3-en-2-ol	Ac. acético R3	R2	R1*

Conclusiones

Conseguimos ensamblar un conjunto de técnicas de aprendizaje de máquina que conforman un modelo para el descubrimiento de patrones sobre conjuntos de compuestos y la posterior clasificación de moléculas problema en función del patrón establecido.

Los descriptores seleccionados a partir del análisis de componentes principales en combinación con la técnica de agrupamiento difuso C-means nos permitió reconocer un patrón entre los semioquímicos estudiados, el cual corresponde a las rutas biosintéticas que dan lugar a estos metabolitos secundarios.

Una vez identificado un patrón, recomendamos la técnica de bosques aleatorios para seleccionar las variables. Finalmente, la combinación de la metodología propuesta con un perceptrón multicapa nos permitió alcanzar una asignación aceptable de las rutas biosintéticas para las moléculas que C-means no clasificó.

Las moléculas cuya clasificación no fue del todo satisfactoria nos muestran la necesidad de ampliar el conjunto de información, para incluir una mayor variabilidad estructural que lleve a conjuntos de entrenamiento más robustos.

Agradecimientos

Agradecemos al profesor Johan Fabián Galindo del grupo de Química Cuántica y Computacional de la Universidad Nacional de Colombia, por facilitarnos la realización de los cálculos cuánticos.

Referencias

[1] N. Bakthavatsalam, "Semiochemicals", en *Ecofriendly Pest Management for Food Security*, Elsevier, 2016, pp. 563-611. DOI: https://doi.org/10.1016/B978-0-12-803265-7.00019-1.

[2] A. Sharma, R. K. Sandhi, and G. V. P. Reddy, "A Review of Interactions between Insect Biological Control Agents and Semiochemicals", *Insects*, vol. 10, no. 12, p. 439, 2019. DOI: https://doi.org/10.3390/insects10120439.

[3] L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in AI-driven drug discovery: a review and practical guide", *Journal of Cheminformatics*, vol. 12, no. 1, 2020. DOI: https://doi.org/10.1186/s13321-020-00460-5.

[4] R. Todeschini, R. Mannhold, H. Kubinyi, V. Consonni, and H. Timmerman, *Handbook of Molecular Descriptors*, John Wiley & Sons, 2008.

[5] A. Fernández-Torras, A. Comajuncosa-Creus, M. Duran-Frigola, and P. Aloy, "Connecting chemistry and biology through molecular descriptors", *Current Opinion in Chemical Biology*, vol. 66, no. 102090, 2022. DOI: https://doi.org/10.1016/j.cbpa.2021.09.001.

[6] L. Xue and J. Bajorath, "Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening", *Combinatorial Chemistry & High Throughput Screening*, vol. 3, no. 5, pp. 363-372, 2000. DOI: https://doi.org/10.2174/1386207003331454.

[7] M. Shahlaei, "Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study", *Chemical Reviews*, vol. 113, no. 10, pp. 8093-8103, 2013. DOI: https://doi.org/10.1021/cr3004339.

[8] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods", *Journal of Big Data*, vol. 7, no. 1, 2020. DOI: https://doi.org/10.1186/s40537-020-00327-4.

[9] T. Cova and A. Pais, "Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns", *Frontiers in Chemistry*, vol. 7, pp. 1-22, 2019. DOI: https://doi.org/10.3389/fchem.2019.00809.

[10] Mushliha, A. Bustamam, A. Yanuar, W. Mangunwardoyo, P. Anki, and R. Amalia, "Comparison Accuracy of Multi-Layer Perceptron and DNN in QSAR Classification for Acetylcholinesterase Inhibitors", en *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, Bandung, Indonesia, 28-30 de abril de 2021. IEEE, 2021. DOI: https://doi.org/10.1109/aims52415.2021.9466040.

[11] M. Hamadache, O. Benkortbi, S. Hanini, and A. Amrane, "Application of multilayer perceptron for prediction of the rat acute toxicity of insecticides", *Energy Procedia*, vol. 139, pp. 37-42, 2017. DOI: https://doi.org/10.1016/j.egypro.2017.11.169.

[12] P. DuBois, *MySQL Language Reference*, Pearson Education, 2007.

[13] G. Landrum, "Rdkit documentation", *Release*, vol. 1, no. 1-79, p. 4, 2013.

[14] G. Zheng, L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hase-gawa, M. Ishida, T. Nakajima, Y. Honda, and col., *Gaussian 09*, 2009.

[15] H. Abdi and L. J. Williams, "Principal component analysis", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433-459, 2010. DOI: https://doi.org/10.1002/wics.101.

[16] G. Biau and E. Scornet, "A random forest guided tour", *TEST*, vol. 25, no. 2, pp. 197-227, 2016. DOI: https://doi.org/10.1007/s11749-016-0481-7.

[17] E. Keany, BorutaShap 1.0.16 2021, 2021. Disponible en línea https://github.com/Ekeany/Boruta-Shap (Acceso, 20 de noviembre de 2022)

[18] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, 2020, https://www.R-project.org/.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[20] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm", *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984. DOI: https://doi.org/10.1016/0098-3004(84)90020-7.

[21] T. Kohonen, "The self-organizing map", *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990. DOI: https://doi.org/10.1109/5.58325.

[22] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview", *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86-97, 2011. DOI: https://doi.org/10.1002/widm.53.

[23] G. Vettigli, "MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map", 2013. Disponible en línea: https://github.com/JustGlowing/minisom (Acceso 20 diciembre 2022).

[24] H. Ramchoun, M. Amine, J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer Perceptron: Architecture Optimization and Training", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, p. 26, 2016. DOI: https://doi.org/10.9781/ijimai.2016.415.

[25] A. M. El-Sayed, "The pherobase: database of insect pheromones and semiochemicals", HortResearch, 2019.

[26] E. D. Morgan, *Biosynthesis in Insects: Advanced Edition*, Royal Society of Chemistry, 2010.

[27] M. Ashour, M. Wink, and J. Gershenzon, "Biochemistry of Terpenoids: Monoterpenes, Sesquiterpenes and Diterpenes", en *Biochemistry of Plant Secondary Metabolism*, Oxford, UK: Wiley-Blackwell, pp. 258-303. DOI: https://doi.org/10.1002/9781444320503.ch5.

[28] M. L. McHugh, "Interrater reliability: the kappa statistic", *Biochemia Medica*, vol 22, no. 3, pp. 276-282, 2012. DOI: https://doi.org/10.11613/bm.2012.031.

Citación de Artículo:
L. S. Valencia-Colman & É. E. Daza C, "Reconocimiento de rutas biosintéticas para semioquímicos mediante técnicas de aprendizaje de máquina", *Rev. Colomb. Quim.*, vol. 51, no. 2, pp. 35-40, 2022. DOI: https://doi.org/doi.org/10.15446/rev.colomb.quim.v51n2.101546