REVISTA CUBANA
DE TRANSFORMACIÓN DIGITAL

RNPS 2487 • ISSN 2708-3411
Vol. 2 • Nro. 1 • enero-marzo 2021 • págs. 14-32

ARTÍCULO ORIGINAL

Agrupamiento funcional de enzimas GH-70 utilizando aprendizaje semi-supervisado y *Apache Spark*

Funcional Clustering of GH-70 Enzymes by using Semi-supervised Learning and Apache Spark

Yadelis González Valle
yadelisgv@gmail.com • http://orcid.org/0000-0002-8700-3823
DIRECCIÓN PROVINCIAL DE LA VIVIENDA, VILLA CLARA, CUBA

Deborah Galpert
deborah@uclv.edu.cu • https://orcid.org/0000-0002-5222-3324
UNIVERSIDAD CENTRAL"MARTA ABREU" DE LAS VILLAS, CUBA

Reinaldo Molina-Ruiz
reymolina@uclv.edu.cu • https://orcid.org/0000-0001-5098-5432
CENTRO DE BIOACTIVOS QUÍMICOS, UNIVERSIDAD CENTRAL"MARTA ABREU" DE LAS VILLAS, CUBA

Guillermin Agüero-Chapin
gchapin@ciimar.up.pt • https://orcid.org/0000-0002-9908-2418
CIIMAR - CENTRO INTERDISCIPLINAR DE INVESTIGAÇÃO MARINHA E AMBIENTAL, UNIVERSIDADE DO PORTO, PORTUGAL

Recibido: 2020-11-03 • Aceptado: 2021-01-19

RESUMEN

La clasificación estructural y funcional de enzimas es un campo de gran interés para la bioinformática. En particular, las enzimas de la familia Glicosil Hidrolasa-70 (GH-70) tienen un alto valor para la biotecnología y a su vez pueden ocasionar pérdidas millonarias en la producción de azúcar. En este artículo se propone utilizar algoritmos de agrupamiento semi-supervisados y no supervisados para agrupar secuencias similares de enzimas de esta familia, a partir de la integración de descriptores de proteínas libres de alineamiento. Se extrajeron rasgos numéricos con el método de *k*-mers con valores de *k* del 2 al 6 y luego se implementaron tres algoritmos que agrupan las enzimas de acuerdo a su función enzimática tomando información de referencia de 58 secuencias funcionalmente caracterizadas de la familia GH-70 de la base de datos CAZy. En los resultados obtenidos en el algoritmo de ensamblado de K-medias, se ubicaron correctamente en sus respec-

tivos grupos la gran mayoría de las enzimas clasificadas, con un máximo de 0,91 en la medida-*F.* Se obtuvieron valores moderados del índice de silueta como medida de validación interna (máximo de 0,3145 para el ensamblado de K-medias), pero mejor que los obtenidos con el propio método K-medias sin ensamblar.

Palabras clave: agrupamiento de enzimas; aprendizaje semi-supervisado; ensamblado de agrupamientos.

ABSTRACT

One of the fields of great interest for bioinformatics is the structural and functional classification of enzymes. In particular, the enzymes of the Glycosyl Hydrolase-70 (GH-70) family have a high value for biotechnology and in turn can cause millions in losses in sugar production. In this article, we investigated the use of semi-supervised and unsupervised clustering algorithms to group similar sequences of enzymes of this family, based on the integration of alignment-free protein descriptors. Numerical traits were extracted with the k-mers method with k values from 2 to 6 and then three algorithms were implemented that group the enzymes according to their enzymatic function, taking reference information from 58 functionally characterized sequences of the GH-70 family, from the CAZy database. In the results obtained in the K-means assembly algorithm, the vast majority of the classified enzymes were correctly located in their respective groups, with a maximum of 0.91 in the -F measure. Moderate values of the silhouette index were obtained as an internal validation measure (maximum of 0.3145 for the assembly of K-means), but better than those obtained with the K-means method itself without assembly.

Keywords: *ensemble clustering; enzyme clustering; semi-supervised learning.*

INTRODUCCIÓN

En nuestro país, las enzimas correspondientes a la familia GH-70 se estudian desde hace varios años por el Instituto Cubano de Investigaciones de los Derivados de la Caña de Azúcar (Icidca) debido el efecto nocivo que presentan en la producción de azúcar, pues ocasionan pérdidas millonarias (Fraga Vidal, *et al.,* 2011). Las enzimas son macromoléculas biológicas que actúan como catalizadores específicos durante los procesos biológicos. El reconocimiento de la función y la clasificación estructural de las enzimas constituye

un problema de gran importancia en la bioinformática por la utilidad biotecnológica de estas.

Diferentes autores han evidenciado la necesidad de aumentar la exactitud en la clasificación, principalmente de familias que contienen secuencias homólogas de baja similitud, también conocidas como homólogos remotos, como sucede en la familia GH-70 en la que se ha abordado el problema en particular de la clasificación funcional (Davies y Sinnott, 2008). Por esta razón, el uso de diversos descriptores libres de alineamiento de proteínas o enzimas se presenta como una tendencia en este tipo de clasificación (AK Ong, *et al.,* 2007). A su vez, la clasificación estructural de enzimas a partir de las secuencias y su relación con la función constituye un campo de investigación abierto, pues para esta familia solo aparecen reportadas seis secuencias con estructura 3D reconocida (Meng, *et al.,* 2016).

A partir de la consideración de que la similitud estructural define la similitud funcional y que algunas pocas secuencias de la familia GH-70 han sido caracterizadas estructuralmente, el agrupamiento de secuencias que combina diversos descriptores libres de alineamiento con el uso herramientas de aprendizaje automatizado puede conformar grupos de secuencias con patrones estructurales similares. Estos descriptores representarían las secuencias como vectores con múltiples componentes representando diferentes propiedades estructurales. De este modo, se pudieran explorar 501 secuencias de enzimas de GH-70 disponibles para la comunidad científica (Lombard, *et al.,* 2014) con el fin de contribuir a inferir su clasificación estructural y funcional. Precisamente, la integración de descriptores ha permitido elevar la calidad de la clasificación de ortólogos en trabajos realizados en el Centro de Investigaciones de Informática de la Universidad Central "Marta Abreu" de Las Villas (UCLV) (Galpert, 2016). Es por esto que este trabajo pretende realizar el agrupamiento de 501 enzimas de la familia GH-70 haciendo uso de métodos no supervisados existentes que pueden ser transformados en semi-supervisados con el fin de aprovechar la información disponible en el sitio CAZy sobre 58 enzimas clasificadas por su función enzimática. A su vez, el propósito de poder clasificar un número creciente de enzimas sin afectar el desempeño de la clasificación ha conllevado a la utilización de técnicas de análisis de datos masivos mediante *Apache Spark*. De este modo el texto se ha dividido en dos partes: la Metodología en la que se abordan las técnicas, instrumentos y métodos empleados para la recolección de los datos, así como las pruebas y parámetros estadísticos utilizados para el análisis de los mismos; y los Resultados y Discusión en los cuales se presentan y analizan los resultados más relevantes del estudio, además se señalarán las aportaciones y limitaciones del trabajo.

METODOLOGÍA

Este trabajo se adentra en el análisis de grandes volúmenes de datos o *Big Data,* como las secuencias de enzimas. Se hizo primeramente el estudio de las técnicas bioinformáticas disponibles para caracterizarlas estructuralmente, como los descriptores libres de alineamiento. Posteriormente se definió cómo medir el grado de asociación entre un par de enzimas; así

como se expusieron las herramientas disponibles en el campo de la inteligencia artificial para agrupar y clasificar las mismas como el aprendizaje no supervisado y semi-supervisado. También, se analizaron algunos índices que permiten validar los resultados. Por último, se detallan los aportes de la investigación aplicados a métodos existentes de algoritmos de agrupamiento no supervisados, con sus respectivos pseudocódigos modificados.

1.1 Descriptores libres de alineamiento

Los descriptores libres de alineamiento son métodos de extracción de rasgos estructurales intrínsecos de las secuencias. Este proceso se realiza mediante funciones que transforman la secuencia en un vector numérico, para posteriormente derivar la similitud de un par de secuencias al comparar dichos vectores numéricos. Estos métodos se conocen como libres de alineamiento y muestran múltiples aplicaciones (Vinga, 2014; Vinga y Almeida, 2003; Zielezinski, *et al.,* 2017).

Entre los métodos libres de alineamiento se encuentras los basados en frecuencia de palabras, (Gunasinghe, Alahakoon, y Bedingfield, 2014; Melsted y Pritchard, 2011) los cuales se basan en funciones llamadas descriptores moleculares de la forma $D: x \longrightarrow \mathbb{R}^m$, en los que la secuencia x de longitud n es convertida a un vector de longitud r. De esta forma, los métodos basados en k-tuplas, k-palabras o k-mers, con $k \leq n$, realizan una correspondencia de la secuencia con un vector (1), cuyas componentes $N_{k,1i}, i=1,..,c^k$ representan la frecuencia de subsecuencias de longitud k, siendo c^k el total de todos los posibles k-mers del alfabeto finito A de c caracteres. Entonces, cuando k=1, la composición de aminoácidos (AAC) (Bhasin y Raghava, 2004) describe la proporción de cada aminoácido en la secuencia de proteína.

$$\pi_x^k = \left(\frac{N_{k,1}}{n - k + 1}, \frac{N_{k,2}}{n - k + 1}, \cdots, \frac{N_{k,c^k}}{n - k + 1} \right) \quad (1)$$

En este trabajo se ha calculado el descriptor de k-mers para k=2, 3, 4, 5 y 6. La dimensión de cada vector es $m=20^k$. Por lo que incluir varios valores de k en la comparación de pares de secuencias conlleva a elevar la dimensionalidad del problema de clasificación y buscar la forma de manejar tal dimensionalidad.

La comparación de pares de vectores se realiza mediante medidas de similitud o disimilitud entre vectores. Entre las variantes mencionadas en (Galpert, 2016) para calcular la disimilitud (o similitud) entre pares de vectores se encuentran las siguientes distancias (o sus conversiones a funciones de similitud mediante una función monótona decreciente): *Minkowski, Euclideana, Manhattan, Maximum,* una función de atributos pesados para atributos mixtos, o la distancia Euclideana pesada, o las similitudes: *Ruzicka, Roberts, Motyka, Bray-Curtis, Kulczynski 1* y *2, Baroni-Urbani-Buser,* la covarianza y la correlación de Pearson. En este caso, se ha seleccionado la correlación de Pearson (2) que es una métrica normalizada expresando similitud entre pares de vectores O_i y O_j.

$$corr\_Pearson(O_i, O_j) = \frac{\sum_{k=1}^{m}(o_{ik} - \overline{atributo_k})(o_{jk} - \overline{atributo_k})}{\sqrt{\sum_{k=1}^{m}(o_{ik} - \overline{atributo_k})^2 \sum_{k=1}^{m}(o_{jk} - \overline{atributo_k})^2}} \quad (2)$$

Donde $atributo_k$ es el valor promedio que toma el $atributo_k$ en el conjunto de datos.

1.2 Similitud entre secuencias de enzimas

En este trabajo, la similitud se mide a partir de los descriptores libres de alineamiento como los *k*-mers, que darían como resultado un conjunto de datos (**dataset**) denominado *Kmers-k* de vectores constituidos por valores reales de ocurrencia de subsecuencias de longitud *k* para cada secuencia de enzima. Se propone como primera variante una medida de similitud global agregada especificada en la expresión (3) que cuantificará el grado de asociación entre dichos vectores, promediando los valores de las correlaciones de Pearson para distintos valores de *k*, como segunda variante indicada en expresión (4) se propone concatenar vectores para valores de *k* diferentes y luego aplicar correlación al vector resultante.

$$corr = \frac{1}{n}\sum_{k=2}^{n} corr\_Pearson(Kmers\_k(e), Kmers\_k(c)) \quad (3)$$

Siendo *e* una enzima clasificada, *c* una enzima sin clasificar y *n* la cantidad de *Kmers-k* utilizados, cuyo valor máximo es seis.

$$corr = corr\_Pearson(\bigcup_{k=2}^{n} Kmers\_k(e), \bigcup_{k=2}^{n} Kmers\_k(c)) \quad (4)$$

Ambas medidas de agregación permiten integrar información de comparación de vectores de *k*-mers con diferentes valores de *k* a la vez que permiten manejar la alta la dimensionalidad del problema.

1.3 Agrupamiento semi-supervisado

El agrupamiento en clústeres es un tipo de técnica de aprendizaje automatizado, donde el objetivo principal del análisis es particionar un conjunto determinado de datos u objetos en clústeres (subconjuntos, grupos o clases). Debe existir un alto grado de asociación entre los objetos de un mismo grupo y un bajo grado entre los miembros de grupos diferentes (Anderberg, 1973). Cuando el agrupamiento se basa en la similitud de los objetos, se desea que los objetos que pertenezcan al mismo grupo sean tan similares como se pueda y los objetos que pertenezcan a grupos diferentes sean tan diferentes como sea posible (Höppner, *et al.,* 1999; Kruse, Döring, y Lesot, 2007). Por otra parte, al comparar secuencias se puede tener en cuenta la información de clasificación previa, es decir, las etiquetas de algunos objetos, como en el problema de clasificación de las enzimas, al considerar la previa clasificación de algunas se-

cuencias en la descripción detallada de la función enzimática expresada a través de la etiqueta EC (del inglés *Enzyme Commission).* Para esto resulta conveniente utilizar el aprendizaje semi-supervisado cuando son pocas las secuencias etiquetadas.

El aprendizaje semi-supervisado es una rama del aprendizaje automatizado que resulta de combinar el aprendizaje supervisado y el no supervisado (Chapelle, Schölkopf, y Zien 2006; Xiaojin Zhun 2005). En el agrupamiento semi-supervisado, la información supervisada se puede tomar de diferentes formas, por ejemplo, pueden aplicarse restricciones *must-link* (se sabe que dos objetos están en el mismo grupo) y *cannot-link* (dos objetos se sabe que están en diferentes grupos) (Lange, *et al.,* 2005). También es posible que algunas asignaciones de grupos se conozcan de antemano. Un ejemplo para la incorporación de este último tipo de información es el uso de datos etiquetados para la "siembra en racimo"; Basu, Banerjee, y Mooney (2002) propusieron inicializar los grupos basados en los objetos para los que se conocen asignaciones de grupo. Esta es la idea que se propone seguir en este trabajo, en el que existen seis grupos de actividad enzimática para esta familia de enzimas GH-70, los cuales serán inicializados con las 58 enzimas clasificadas[1] antes de comenzar el proceso de agrupamiento.

Para el agrupamiento semi-supervisado se propone realizar transformaciones al algoritmo Combinatorio Lógico Global (*Global Logical Combinatorial,* GLC+) en (Ruiz-Shulcloper, s. f.; Ruiz-Shulcloper y Sánchez-Díaz, 2001). Este es un método de agrupamiento incremental que construye componentes conexas a partir de descripciones mezcladas e incompletas de objetos representadas en espacios no necesariamente métricos. Dicho método puede ser aplicado a muy grandes volúmenes de datos, y por su naturaleza incremental permite inicializar los grupos con las secuencias etiquetadas e ir incrementando los mismos según las comparaciones de similitud entre la secuencia a analizar en cada paso y el resto de las secuencias. Se ha denominado GLC+ semi-supervisado al nuevo método con las modificaciones implementadas y este será aplicado en dos variantes: con la medida de similitud expresada en (3) y con la expresada en (4).

Otra tendencia en el agrupamiento que aprovecha información de etiquetas conocidas es el ensamblado de agrupamientos (*Ensemble Clustering,* EC) (Abdallah y Yousef, 2020). Este método reemplaza el espacio de datos por un espacio categórico basado en agrupación de conjuntos. El nuevo espacio categórico se define mediante el seguimiento de la membresía de los objetos en múltiples ejecuciones de algoritmos de agrupamiento.

Para el agrupamiento de enzimas, el proceso comienza al aplicar N veces el agrupamiento K-medias a los vectores de frecuencia de k-mers, y luego continúa al ensamblar los agrupamientos resultantes de las N ejecuciones, hasta que se forme un nuevo conjunto de datos denominado como *matrizEC_k* de dimensión $E \times N$ donde E es igual a la cantidad de objetos (501 enzimas) y N indica la cantidad de grupos a formar menos 1, ya que es obligatorio comenzar en 2 como valor mínimo para ejecutar el K-medias. Los valores de las columnas en *matrizEC_k* se corresponden con la predicción de agrupamiento realizada por el K-medias para el respectivo valor de

[1] http://www.cazy.org/GH70_characterized.html

N, por ejemplo, la primera columna contendrá solamente valores 0 y 1 por representar el valor de $N = 2$, y así sucesivamente, los valores discretos oscilaran entre 0 y N-1 para la columna N.

Para poder hacer uso de varios *k*-mers al mismo tiempo, se propone concatenar la **matriz EC_k** para distintos valores de *k* en una sola **matriz resultante,** y finalmente, utilizando la expresión (5), aplicar la correlación de Pearson como medida de similitud, a los vectores de enzimas etiquetadas de dicha **matriz resultante** con los vectores de las que están sin clasificar en ella, siendo *e* una enzima clasificada, *c* una enzima sin clasificar y *n* la cantidad de Kmers-*k* utilizados. Como resultado, se actualizan los seis grupos de la actividad enzimática cuyos elementos, además de los ya etiquetados, serán también aquellos más similares que se vayan encontrando.

$$corr = corr\_Pearson(\bigcup_{k=2}^{n} matrizEC\_k(e), \bigcup_{k=2}^{n} matrizEC\_k(c)) \quad (5)$$

Llamaremos a este método propuesto EC-GLC+ semi-supervisado, porque comienza preparando el nuevo conjunto de datos con el ensamblado de agrupamiento (EC), y posteriormente utiliza al GLC+ semi-supervisado mencionado para agrupar. De esta forma, han sido presentadas tres propuestas de métodos de agrupamiento para clasificar las enzimas de la familia GH-70 en los seis grupos de la actividad enzimática.

La información de etiquetado previo también se utiliza en la etapa de validación mediante el uso de índices o medidas de validación externa, que pueden ser combinados con índices de validación interna (Halkidi, Batistakis, y Vazirgiannis, 2002; Koutroumbas y Theodoridis, 2008) e indican en qué grado el agrupamiento es correcto o no (Höppner, *et al.,* 1999).

1.4 Medidas de validación internas y externas

Las medidas internas se utilizan para medir la densidad y la cohesión entre pares de objetos de un mismo grupo. La cohesión de los grupos puede utilizarse como una medida de validación de estos. La medida interna, nombrada similitud global (*Overall Similarity,* OS), se ha utilizado para medir la cohesión basándose en la media de la similitud de los pares de objetos en un grupo (Steinbach, Karypis, y Kumar, 2000). Existen varios índices que calculan la razón entre las distancias dentro de los grupos y las distancias entre los grupos, por ejemplo: índices *Dunn, Davies-Bouldin* e índice I. Otros calculan la suma pesada de esas dos distancias, por ejemplo *SD, S_Dbw* y v_{SV}.

El índice de silueta es el promedio, sobre todos los grupos, del ancho de la silueta de sus puntos. Si x es un objeto en el clúster C_k y n_k es el número de objetos en C_k, entonces, el índice de silueta de *x* está definido por la relación expresada en (6):

$$S(x) = \frac{b(x) - a(x)}{\max [b(x), a(x)]} \quad (6)$$

Donde *a*(*x*) en (7) es el promedio de las distancias entre *x* y todos los otros objetos en C_k y *b*(*x*) en (8) es el mínimo de los promedios de las distancias *d*(*x*,*y*) entre *x* y los objetos de los otros clústeres.

$$a(x) = \frac{1}{n_k - 1} \sum_{y_e C_{k,y \neq x}} d(x,y) \tag{7}$$

$$b(x) = \min_{h=1,\ldots,K,h \neq k} \left[\frac{1}{n_h} \sum_{y \in C_h} d(x,y) \right] \tag{8}$$

Finalmente, el índice de silueta global (9) está definido como sigue, siendo K el número de clústeres y n_k la cantidad de objetos en cada clúster:

$$S = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{n_k} \sum_{x \in C_k} S(x) \right] \tag{9}$$

Para un objeto x dado, el ancho de su silueta varía de -1 a 1. Si el valor está cerca de -1, significa que el objeto está más cerca, en promedio, de otro grupo que de aquel al que pertenece. Si el valor es cercano a 1, significa que la distancia promedio a su propio grupo es significativamente menor que a cualquier otro grupo. Valores altos del índice silueta global indican grupos más compactos y bien separados. El cálculo de este índice tiene una alta complejidad; sin embargo, investigaciones actuales lo utilizan para la validación del agrupamiento (Brun, *et al.,* 2007). En este trabajo se utilizó este índice para guiar el agrupamiento, pues en el EC sirvió de indicador para determinar cuántas ejecuciones N del algoritmo de agrupamiento K-medias serían necesarias para lograr grupos más separados y compactos.

Por otra parte, las medidas externas se basan en las etiquetas conocidas. Algunas medidas externas utilizan las ideas de precisión (*precision)* y cubrimiento[2] (*recall)* del campo de la recuperación de información y las adaptan a la validación del agrupamiento. La precisión (Pr) y el cubrimiento (Re) se calculan mediante las expresiones (10) y (11), respectivamente, para un grupo j y una clase i dados, donde n_{ij} es el número de objetos de la clase i en el grupo j, n_j es el número de objetos del grupo j y n_i es el número de objetos de la clase i.

$$Pr(i,j) = \frac{n_{ij}}{n_j} \tag{10}$$

$$Re(i,j) = \frac{n_{ij}}{n_i} \tag{11}$$

La medida-*F* (*F-measure)* se obtiene calculando la media armónica de precisión y cubrimiento como se puede apreciar en (12).

$$F - Measure(i,j) = \frac{1}{\alpha\left(1/\Pr(i,j)\right) + \left(1-\alpha\right)\left(1/\mathrm{Re}(i,j)\right)} \tag{12}$$

[2] En este documento se utiliza cubrimiento como traducción de la medida *recall.*

Si α= 1 entonces *F-measure(i,j)* coincide con precisión, y si α = 0 entonces coincide con cubrimiento. En el caso que α = 0,5 significa igual peso para precisión y cubrimiento (Baeza-Yates y Frakes, 1992). Un valor global, de la medida-*F* global (*Overall F-measure; OFM),* se calcula mediante el promedio de los valores por clase de la medida-*F* sobre todos los grupos (Steinbach, *et al.,* 2000). Esta medida-*F* intenta capturar cuánto los grupos del agrupamiento obtenido se hacen corresponder correctamente con los grupos de referencia o clases incluso cuando existe desbalance en la cantidad de objetos por clase (Rosell, *et al.,* 2004). En resumen, las medidas externas: precisión, cubrimiento y medida-*F* fueron utilizadas para medir la calidad de los agrupamientos obtenidos en este trabajo.

1.5 Nuevos algoritmos de agrupamiento de enzimas

Como se había mencionado anteriormente, en este trabajo fueron diseñados e implementados dos métodos de agrupamiento con aprendizaje semi-supervisado: el GLC+ semi-supervisado y el EC. Para la implementación se utilizó la biblioteca MLlib de Spark.

Método GLC+ semi-supervisado

En esta sección se exponen las modificaciones al algoritmo GLC+ con el fin de convertirlo en un algoritmo de agrupamiento semi-supervisado:

- La cantidad de grupos que se pueden formar con el método GLC+ es indeterminada, pero en el nuevo método GLC+ semi-supervisado se limita esta cantidad a seis grupos posibles a formar, correspondientes a la actividad enzimática.
- En el GLC+ original los grupos comienzan vacíos y se incrementan objetos O_i, cada vez que encuentran un objeto O_j semejante que pertenezca al agrupamiento G_k que cumplan la condición expresada en (13) donde $\Gamma(O_i,O_j)$ representa la similitud entre los objetos O_i y O_j, y β_0, el umbral de similitud.

$$\Gamma(O_i, O_j) \geq \beta_0 \qquad (13)$$

  En el caso del GLC+ semi-supervisado se tienen seis grupos inicialmente con algunas enzimas de las que se conoce su clasificación pertenecientes a las 58 clasificadas: 43 del primer grupo, dos del segundo grupo, dos del tercer grupo, cuatro del cuarto grupo, dos del quinto grupo y una del sexto grupo.

De las 58 enzimas, la enzima "CDX66820.1" tiene doble clasificación lo que significa que tiene doble actividad enzimática, y no se utilizó entre las clasificadas para no introducir confusión durante el agrupamiento. Por otra parte, las enzimas: "P08987" y "P49331" no se encuentran entre las 501 secuencias de las enzimas para clasificar. De lo anterior se deriva que de las 58 clasificadas serán utilizadas 55 enzimas. El pseudocódigo del algoritmo de agrupamiento después de haber implementado las modificaciones al GLC+ original, junto con la información disponible sobre las enzimas ya etiquetadas y aplicando como medida de similitud la expresada en (3) (GLC+ semi-supervisado-variante1) queda de la siguiente manera:

Entrada: Enzimas E, Arreglo de Vectores de los *Kmers-k*, Enzimas clasificadas v:(c,n) → V; donde c ∈ E es la enzima clasificada y n el grupo al que pertenece, β_0

Salida: Nuevos clasificados g:(e,n) → G; donde e ∈ E es la enzima sin clasificar y n el grupo en el que será clasificada.

Begin

Forall $e \in E$ **do** P1

/ Verificar que la enzima no está entre las clasificadas*/*

Forall (c, n) ∈ V **if**(c≠ e) **then** P2

$corr = \frac{1}{n}\sum_{k=2}^{6} corr\_Pearson(Kmers\_k(e), Kmers\_k(c))$ P3

*/*Se guarda en maxCorr la mayor correlación encontrada */*

If $(corr \geq \beta_0)$ **then** P4

$maxCorr = \mathrm{Max}(corr, maxCorr)$ P5

End

End

*/*Añadir a los nuevos clasificados el par (e,* n*) donde* n ∈ [1,6] *toma el número del grupo de la enzima clasificada* c *que tuvo la mayor correlación con e*/*

Añadir a G el par (e, n) P6

End

End

En P3 la correlación entre dos series de vectores es calculada promediando las correlaciones de cada *k*-mers empleado. Se tomaron cuatro tipos de combinaciones: la primera, usando del 2-mers al 3-mers, la segunda del 2-mers al 4-mers, la tercera del 2-mers al 5-mers y la última del 2-mers al 6-mers.

En P4 se verifica el cumplimiento de la restricción de similitud (13). El valor de β_0 se calculó a partir de la matriz de semejanza entre todas las m secuencias de enzimas utilizando la expresión (14), ver (Ruiz-Shulcloper, s. f.) para más detalles.

$$\beta_0 = \min_{\substack{i=1\ldots m \\ i\neq j}} \{ \max_{j=i+1\ldots m} \{\Gamma(O_i, O_j)\}\} \quad (14)$$

En P5 se guarda la mayor de las correlaciones, lo que significa que se encuentra una enzima dentro de las clasificadas que es el más similar a la enzima que se está analizando. Ese valor de n donde se encontró la enzima más similar puede tomar valores desde uno hasta seis correspondiente a los seis grupos de la actividad enzimática.

Por otra parte, el pseudocódigo para este algoritmo de agrupamiento GLC+ semi-supervisado pero aplicando como medida de similitud la expresada en (4) (GLC+ semi-supervisado-variante2) difiere en algunos aspectos, como se muestra en la próxima página.

En P1 se concatenan los vectores de los *k*-mers que se estén empleando, así como en la variante anterior, se emplearon 4 combinaciones. En P4 se determina la correlación de Pearson comparando los vectores integrados de la enzima que se analiza y las clasificadas. El resto de las sentencias se mantienen iguales que en el pseudocódigo anterior. Para medir la calidad de

Entrada: Enzimas E, Arreglo de Vectores de los *Kmers-k*, Enzimas clasificadas v:(c,n) → V; donde c ∈ E es la enzima clasificada y n el grupo al que pertenece,β_0

Salida: Nuevos clasificados g:(e,n) → G; donde e ∈ E es la enzima sin clasificar y n el grupo en el que será clasificada.

Begin

$Kmers_{integrado} = \cup_{k=2}^{n} Kmers\_k$, con $n \in [3,6]$ P1

Forall e ∈ E **do** P2

/ Verificar que la enzima no está entre las clasificadas*/*

Forall (c, n) ∈ **V** **if**(c≠ e) **then** P3

$corr = corr\_Pearson(Kmers_{integrado}(e), Kmers_{integrado}(c))$ P4

*/*Se guarda en maxCorr la mayor correlación encontrada */*

If $(corr \geq \beta_0)$ **then** P5

$maxCorr = \text{Max}(corr, maxCorr)$ P6

End

End

*/*Añadir a los nuevos clasificados el par (e,* n*) donde* n ∈ [1,6] *toma el número del grupo de la enzima clasificada* c *que tuvo la mayor correlación con e*/*

Añadir a G el par (e, n) P7

End

End

este algoritmo, en cualquiera de sus dos variantes, solo se pueden utilizar medidas internas, pues las externas siempre van a detectar como correctas a las secuencias etiquetadas.

Método ensamblado de agrupamientos

La suposición principal es que los objetos pertenecientes al mismo grupo son más similares a otros objetos de otros grupos, a pesar de que la distancia Euclidiana sea menor (Abdallah y Yousef, 2020). Esto se debe a que los algoritmos de agrupamiento tienen en cuenta tanto el espacio geométrico como otros parámetros estadísticos.

Como se había explicado anteriormente, el algoritmo de transformación EC consiste en ejecutar un algoritmo de agrupamiento (o múltiples algoritmos) varias veces con diferentes valores de parámetros en que cada ejecución produce una dimensión categórica (característica o rasgo) de los datos. Por ejemplo, ejecutar el algoritmo K-medias con diferentes valores de *k, k* = 1,..., *N,* generará un nuevo vector de datos con dimensión *N*. En otras palabras, dos objetos en el espacio del EC son idénticos si fueron asignados a los mismos grupos en toda iteración (k = 1,.., *N).* Todos los objetos que caen en el mismo clúster en las diferentes ejecuciones del agrupamiento constituyen un solo grupo y están representados por un solo objeto. Esta última aseveración será transformada ya que el objetivo que se pretende alcanzar no es el de crear tantos grupos como sea posible sino solamente seis correspondientes a los de la actividad enzimática.

Esta modificación, similar a la realizada anteriormente en el método GLC+ semi-supervisado, pretende que el EC realice un agrupamiento con aprendizaje semi-supervisado. Lo an-

terior se logra al comparar los nuevos vectores que se generaron con los vectores asociados de las enzimas clasificadas, y los que resulten con mayor correlación se pueden considerar como del grupo al que pertenece la enzima clasificada de referencia.

Al realizar varias iteraciones se encontró que para valores de k entre 45 y 50, los valores del índice de silueta oscilaban entre 0,56 aproximadamente. Este el valor fue más alto en comparación a -0,0097 que es el valor más bajo encontrado, el cual indica demasiados o muy pocos elementos similares en el grupo. Por esta razón, se escogió N igual 50 como el número de ejecuciones a realizar. Para proceder a aplicar el EC se realizaron corridas del K-medias para k desde 2 hasta 50, con Kmers-2, Kmers-3, Kmers-4, Kmers-5 y Kmers-6. Cada corrida se guardó en un fichero de tipo CSV, los cuales fueron integrados en un documento por cada k-mers. Cada línea de los ficheros consenso representa un vector, entonces se buscan aquellos vectores que son más similares a los de las enzimas clasificadas utilizando como medida de similitud la expresada en (5). De esta manera se guía el agrupamiento supervisado con la información de la que se dispone.

El pseudocódigo para este algoritmo de agrupamiento EC-GLC+ semi-supervisado se presenta a continuación:

Entrada: Enzimas E, Arreglo de Vectores de las *matrizEC k*, Enzimas clasificadas v:(c,n) → V;
donde c ∈ E es la enzima clasificada y n el grupo al que pertenece,β_0

Salida: Nuevos clasificados g:(e,n) → G; donde e ∈ E es la enzima sin clasificar y n el grupo en el que será clasificada.

Begin

$MatrizEC\_k_{integrada} = \cup_{k=2}^{n} matrizEC\_k$, con $n \in [3,6]$ P1

Forall e ∈ E **do** P2

/ Verificar que la enzima no está entre las clasificadas*/*

Forall (c, n) ∈V **if**(c≠ e) **then** P3

$corr = corr\_Pearson(MatrizEC\_k_{integrada}(e), MatrizEC\_k_{integrada}(c))$ P4

*/*Se guarda en maxCorr la mayor correlación encontrada */*

If ($corr \geq \beta_0$) **then** P5

$maxCorr$ = Max ($corr, maxCorr$) P6

End

*/*Añadir a los nuevos clasificados el par (e,* n*) donde* n ∈ [1,6] *toma el número del grupo de la enzima clasificada* c *que tuvo la mayor correlación con e*/*

Añadir a G el par (e, n) P7

End

End

En P1 se integran los vectores con la intención de utilizar más de un k-mers al mismo tiempo. Para este método si es posible hacer el cálculo de las medidas internas y externas para validar el agrupamiento.

Con el objetivo de esclarecer el flujo de procesos seguido en el agrupamiento durante la experimentación, la figura 1 muestra la entrada de secuencias al proceso de extracción de rasgos mediante los descriptores libres de alineamiento, en este caso, los *k*-mers. Asimismo se observa la decisión de uno de los tres métodos de agrupamiento a elegir, la transformación de datos en el ensamblado de agrupamientos, y la ubicación final de las enzimas en los grupos de actividad enzimática reportados en CAZy. Los resultados obtenidos por los tres métodos con las distintas medidas de validación se exponen en la siguiente sección.

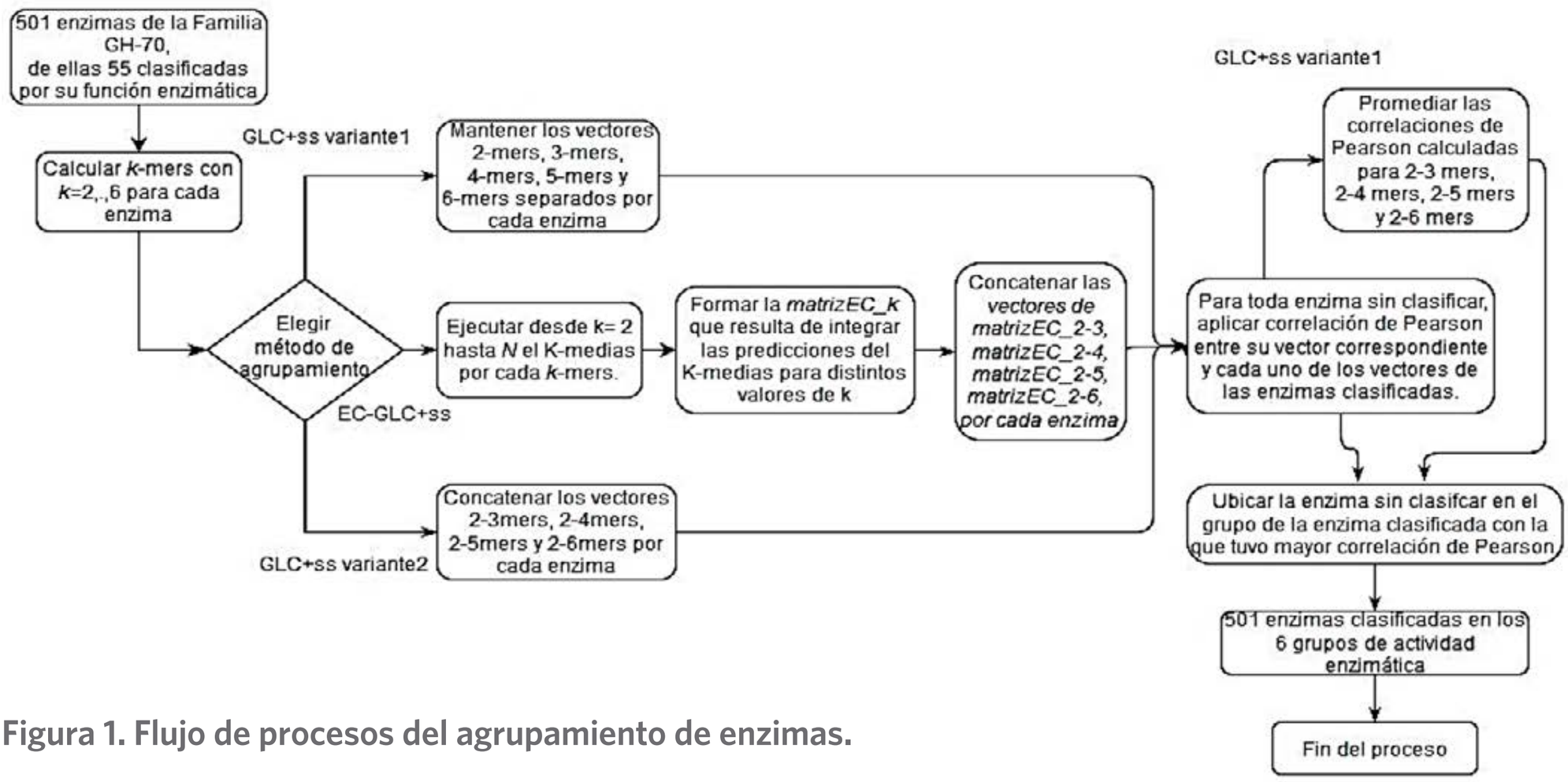

Figura 1. Flujo de procesos del agrupamiento de enzimas.

RESULTADOS Y DISCUSIÓN

Los tres métodos presentados en la sección anterior fueron implementados en lenguaje de programación *Scala*. Se utilizaron además las librerías *MLlib* y *ML* implementadas en *Spark* para el análisis de grandes volúmenes de datos como los *k*-mers; así como el K-medias para realizar agrupamientos y el cálculo de la correlación de Pearson por pares de vectores implementado en *Spark*. En la tabla 1 se muestra la predicción obtenida por los tres métodos de agrupamiento semi-supervisados propuestos.

Tabla 1. Predicción realizada por los tres métodos de agrupamiento semi-supervisados

Grupos	GLC+ semi-supervisado-variante1				GLC+ semi-supervisado-variante2				EC-GLC+ semi-supervisado			
	Kmers 2-3	Kmers 2-4	Kmers 2-5	Kmers 2-6	Kmers 2-3	Kmers 2-4	Kmers 2-5	Kmers 2-6	Kmers 2-3	Kmers 2-4	Kmers 2-5	Kmers 2-6
1	382	381	380	378	369	377	374	369	419	408	407	441
2	10	10	10	10	21	14	16	21	9	7	7	8
3	4	4	5	5	5	4	4	5	4	2	2	2
4	72	73	73	73	84	83	84	84	50	51	53	34
5	10	10	10	11	10	11	11	10	9	25	22	10
6	23	23	23	24	12	12	12	12	10	8	10	6
Total	501	501	501	501	501	501	501	501	501	501	501	501

Se puede notar que como promedio en los tres métodos el 77,9 % de las enzimas fueron ubicadas en el grupo 1, el 2,37 % en el grupo 2, el 0,76 % en el grupo 3, el 13,53 % en el grupo 4, el 2,47 % en el grupo 5 y el 2,91 % en el grupo 6. La mayoría de las enzimas fueron clasificadas en el grupo uno, y en segundo lugar en el grupo cuatro. Esta predicción es similar a la proporción de ubicación por grupo de las 55 enzimas clasificadas, en que el 78 % de las clasificadas están en el grupo 1 y el 9,09 % está en el grupo 4. El resto de las clasificadas solo posee una o dos enzimas en los grupos 2, 3 y 6.

Las predicciones realizadas por los tres métodos propuestos para las enzimas, de las cuales ya se conoce su clasificación, se puede apreciar en la tabla 2. En el caso del método GLC+ semi-supervisado, en cualquiera de sus dos variantes, no hubo ningún desacierto. Las 55 enzimas fueron clasificadas correctamente. Por su parte, en el caso del método EC-GLC+ semi-supervisado solo en dos enzimas se realizó una incorrecta clasificación (marcada en rojo):

- "AIM52834.2" que se conoce previamente que pertenece al grupo 2 fue ubicada en el grupo uno durante el uso de los 2-mers, 3-mers y 4-mers.
- "AAU08015.1" que pertenece al grupo 3 originalmente fue ubicada en el grupo 5 durante las cuatro combinaciones de *k*-mers.

Tabla 2. Predicciones realizadas por los 3 métodos de agrupamiento semi-supervisados para cada enzima previamente clasificada

Grupos	Enzimas	GLC+ ss -variante1				GLC+ ss -variante2				EC-GLC+ss			
		2-3 mers	2-4 mers	2-5 mers	2-6 mers	2-3 mers	2-4 mers	2-5 mers	2-6 mers	2-3 mers	2-4 mers	2-5 mers	2-6 mers
G 1	CAA77898.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAA26896.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAA26898.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAA14241.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAA02976.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAC41412.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAC41413.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAC43483.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAB95453.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAA26114.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAD10952.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAA90527.1	1	1	1	1	1	1	1	1	1	1	1	1
	CAB76565.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAA95201.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAG38021.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAG61158.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAC07265.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAN58619.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAN58705.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAN58706.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAN38835.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAS79426.1	1	1	1	1	1	1	1	1	1	1	1	1
	AAQ98615.2	1	1	1	1	1	1	1	1	1	1	1	1
	AAX76986.1	1	1	1	1	1	1	1	1	1	1	1	1
	ABC75033.1	1	1	1	1	1	1	1	1	1	1	1	1
	ABF85832.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAF43788.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAF62337.1	1	1	1	1	1	1	1	1	1	1	1	1
	BAF96719.1	1	1	1	1	1	1	1	1	1	1	1	1

Grupos	Enzimas	GLC+ ss -variante1				GLC+ ss -variante2				EC-GLC+ss			
		2-3 mers	2-4 mers	2-5 mers	2-6 mers	2-3 mers	2-4 mers	2-5 mers	2-6 mers	2-3 mers	2-4 mers	2-5 mers	2-6 mers
G1	ACA83218.1	1	1	1	1	1	1	1	1	1	1	1	1
	ACK38203.1	1	1	1	1	1	1	1	1	1	1	1	1
	ACT20911.1	1	1	1	1	1	1	1	1	1	1	1	1
	ACY92456.2	1	1	1	1	1	1	1	1	1	1	1	1
	ADB43097.3	1	1	1	1	1	1	1	1	1	1	1	1
	CCF30682.1	1	1	1	1	1	1	1	1	1	1	1	1
	AFP53921.1	1	1	1	1	1	1	1	1	1	1	1	1
	CCK33643.1	1	1	1	1	1	1	1	1	1	1	1	1
	CCK33644.1	1	1	1	1	1	1	1	1	1	1	1	1
	AHU88292.1	1	1	1	1	1	1	1	1	1	1	1	1
	CDX67012.1	1	1	1	1	1	1	1	1	1	1	1	1
	CDX66895.1	1	1	1	1	1	1	1	1	1	1	1	1
	CDX66641.1	1	1	1	1	1	1	1	1	1	1	1	1
	AKE50934.1	1	1	1	1	1	1	1	1	1	1	1	1
G 2	AIM52834.2	2	2	2	2	2	2	2	2	2	1	2	2
	CAB65910.2	2	2	2	2	2	2	2	2	2	2	2	2
G 3	AAU08015.1	3	3	3	3	3	3	3	3	5	5	5	5
	AAY86923.1	3	3	3	3	3	3	3	3	3	3	3	3
G 4	ABQ83597.1	4	4	4	4	4	4	4	4	4	4	4	4
	ACB62096.1	4	4	4	4	4	4	4	4	4	4	4	4
	AAU08014.2	4	4	4	4	4	4	4	4	4	4	4	4
	AAU08003.2	4	4	4	4	4	4	4	4	4	4	4	4
	AJE22990.1	4	4	4	4	4	4	4	4	4	4	4	4
G 5	ABP88726.1	5	5	5	5	5	5	5	5	5	5	5	5
	CDX66896.1	5	5	5	5	5	5	5	5	5	5	5	5
G 6	AOR73699.1	6	6	6	6	6	6	6	6	6	6	6	6

Como se había mencionado en la sección anterior, se utilizó el índice de silueta como medida interna para validar el agrupamiento, y como medidas externas: precisión, cubrimiento y medida-*F* para el ensamblado de agrupamientos. En la tabla 3 se muestran los valores calculados para el índice de silueta de los tres métodos con sus respectivas cuatros combinaciones de los *k*-mers.

Tabla 3. Valores del índice de silueta de los 3 métodos de agrupamiento semi-supervisados.

Medidas Internas/Índice de silueta												
Grupo	**GLC+ ss -variante1**				**GLC+ ss –variante2**				**EC-GLC+ss**			
	2-3 mers	**2-4 mers**	**2-5 mers**	**2-6 mers**	**2-3 mers**	**2-4 mers**	**2-5 mers**	**2-6 mers**	**2-3 mers**	**2-4 mers**	**2-5 mers**	**2-6 mers**
G 1	-0.003	0.0147	0.0252	0.0269	0.0243	0.0573	0.0516	0.04695	6.1E-04	0.0205	0.0187	0.0193
G 2	0.4078	0.3877	0.3722	0.3546	-0.015	0.2088	0.1133	0.00274	0.4839	0.5166	0.7977	0.5689
G 3	0.0563	0.1077	0.0930	0.1072	0.0228	0.1426	0.1575	0.11881	0.2497	0.8882	0.8828	0.8790
G 4	0.3079	0.3182	0.3228	0.3158	0.2826	0.3167	0.3030	0.29278	0.0664	0.0080	0.0385	0.0915
G 5	0.2176	0.1973	0.1828	0.1265	0.2125	0.1191	0.1223	0.15478	0.3168	-0.0516	0.0141	0.1250
G 6	0.1915	0.1811	0.1745	0.1771	0.2159	0.2358	0.2203	0.20473	-0.0074	-0.0413	0.0164	0.2031
Global	**0.1963**	**0.2011**	**0.1951**	**0.1847**	**0.1237**	**0.1800**	**0.1613**	**0.1368**	**0.1850**	**0.2234**	**0.2947**	**0.3145**

El valor más bajo del índice de silueta por grupo en el caso del método GLC+ semi-supervisado variante 1 lo tiene el grupo 1 con el uso de los *k*-mers 2 al 3. En el mismo método pero

con variante 2 los tiene el grupo 2 también con el uso de los *k*-mers del 2 al 3 y en el método EC-GLC+ semi-supervisado lo tiene el grupo 5 con el uso del *k*-mers del 2 al 4. Los valores bajos, sombreados en negrita en la tabla 3, indican que las enzimas están más cerca, como promedio, de otro grupo con respecto al que fueron ubicadas.

Por otro lado, el valor más bajo del índice de silueta global lo tiene el método GLC+ semi-supervisado variante-2, que parte de la integración de los vectores o *k*-mers del 2 al 3, para luego aplicar correlación como medida de similitud. El valor más alto se logró en el método EC-GLC+ semi-supervisado con el uso de los 2 a 6-mers.

Aunque los valores del índice de silueta global para GLC+ semi-supervisado no están por encima de 0,5 o más próximos a 1, se pudieran considerar como valores moderados, similares a los ofrecidos por el K-medias implementado en la librería *ML* de *Spark*. Este algoritmo permite agrupar las enzimas usando *k*-mers 2 y 3, y se obtiene un índice de silueta de 0,1971161916099274 y 0,12133373673005655 con el uso de los *k*-mers del 2 al 4, valores similares, e incluso más bajos, que los obtenidos por los tres métodos con el uso de los mismos *k*-mers. Además, debemos destacar que la familia GH-70 posee enzimas muy similares, incluso algunas que tienen doble clasificación, como el caso de una de las etiquetadas que no se empleó para no generar confusión durante el agrupamiento, es decir, que tienen características que las pueden ubicar en más de un grupo al mismo tiempo. Por otra parte, existen enzimas que difieren en su estructura, es decir, son divergentes, pero pueden tener la misma actividad enzimática por lo que son consideradas como homólogos remotos. Ambas consideraciones pueden explicar la existencia de valores bajos del índice de silueta.

En el caso de las medidas externas se debe recordar que no se aplica el método GLC+ semi-supervisado en sus dos variantes porque al no tener ninguna enzima mal clasificada, todas las medidas darían valor de uno, por eso a continuación se muestra en tabla 4 el valor de las medidas externas para el método EC-GLC+ semi-supervisado.

Tabla 4. Valores de las medidas externas para el método EC-GLC+ semi-supervisado.

	EC con Kmers2-3							EC con Kmers2-4						
Medidas externas	G1	G2	G3	G4	G5	G6	Global	G1	G2	G3	G4	G5	G6	Global
Precision	1	1	0.5	1	1	1	0.91	1	0.5	0.5	1	1	1	0.83
Recall	1	1	1	1	0.66	1	0.94	0.97	1	1	1	0.66	1	0.94
F Measure	1	1	0.66	1	0.8	1	0.91	0.98	0.66	0.66	1	0.8	1	0.85
	EC con Kmers2-5							**EC con Kmers2-6**						
Medidas externas	G1	G2	G3	G4	G5	G6	Global	G1	G2	G3	G4	G5	G6	Global
Precision	1	1	0.5	1	1	1	0.91	1	1	0.5	1	1	1	0.91
Recall	1	1	1	1	0.66	1	0.94	1	1	1	1	0.66	1	0.94
F Measure	1	1	0.66	1	0.8	1	0.91	1	1	0.66	1	0.8	1	0.91

Como se puede apreciar los valores obtenidos para las medidas externas del método EC-GLC+ semi-supervisado fueron significativamente buenas con valores por encima de 0,83 en el peor de casos. El peor valor se obtuvo en Precisión con el uso de los *k*-mers del 2 al 4, el resto

obtuvo 0,91. En el cubrimiento en las cuatro combinaciones se obtuvo 0,94, y en la medida-*F* todas las combinaciones obtuvieron 0,91 excepto en el uso de los *k*-mers del 2 al 4.

CONCLUSIONES

Con la propuesta de tres procedimientos para el agrupamiento que permiten incluir información disponible sobre el conjunto de datos, es posible realizar el agrupamiento de manera semi-supervisada. En los tres métodos se determinaron 6 clústeres correspondientes a la actividad enzimática. Se utilizaron los *k*-mers del 2 al 6 y la medida de similitud el promedio de las correlaciones de Pearson por cada *k*-mers en la variante 1 del método GLC+ semi-supervisado. En la variante 2, primeramente, se integraron los *k*-mers y luego se aplicó correlación de Pearson, y en el EC-GLC+ semi-supervisado se preparó una nueva data realizando 50 ejecuciones del K-medias implementado en *Spark*. Las matrices resultantes se integraron por cada *k*-mers luego se aplicó la correlación de Pearson para medir la similitud.

Al validar el agrupamiento resultaron aceptables los valores del índice de silueta ofrecido por los tres métodos e incluso mejores que los obtenidos por el K-medias de *Spark* con el uso de los *k*-mers del 2 al 3 y del 2 al 4. El mejor valor de silueta (0,3145) se obtuvo con EC-GLC+ semi-supervisado, superó el desempeño del K-medias sin ensamblaje. En la validación externa fueron promisorios también los valores obtenidos por el método EC-GLC+ semi-supervisado para los índices considerados, predominando el valor 0,91 en la medida-*F* para las distintas combinaciones de *k*-mers. El uso de otras medidas de similitud, y/o de otras formas de agregación o integración de la información, junto a mejoras en el agrupamiento, pudieran optimizar los valores de los índices de validación para la clasificación de la actividad enzimática.

Por otra parte, el uso de *Spark* garantiza el manejo de rasgos de alta dimensionalidad como los *k*-mers, que permiten extraer información de la estructura de las secuencias, y además, deberá garantizar la escalabilidad de los algoritmos cuando se incremente el número de procesadores y de secuencias a clasificar, así como obtener bajos tiempos de ejecución en un clúster de computadoras.

REFERENCIAS

Abdallah, L., Yousef, M. (2020) GrpClassifierEC: a novel classification approach based on the ensemble clustering space. *Algorithms Mol Biol 15*(3). https://doi.org/10.1186/s13015-020-0162-7.

AK Ong, Serene, Hong Huang Lin, Yu Zong Chen, Ze Rong Li, y Zhiwei Cao. (2007). Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics 8*(300).

Anderberg, Michael R. (1973). Cluster Analysis for Applications. 1st Edition. Probability and Mathematical Statistics: A Series of Monographs and Textbooks ISBN: 978-0-12-057650-0. https://doi.org/10.1016/C2013-0-06161-0. eBook ISBN: 9781483191393. Imprint: Acade-

mic Press. Published Date: 28th November 1973. Page Count: 376.

Baeza-Yates, R., y William B. F. (1992). *Information Retrieval: Data Structures and Algorithms.* editado por Prentice. Hall. ISBN 0-13-463837-9.

Basu, Sugato, Arindam Banerjee, y Raymond Mooney (2002). Semi-supervised Clustering by Seeding. *Proceedings of the 19th International Conference on Machine Learning* 27-34.

Bhasin, Manoj, y Gajendra P. S. Raghava. (2004). Classification of Nuclear Receptors Based on Amino AcidComposition and Dipeptide Composition. *The Journal of Biological Chemistry* 279(22).

Brun, Marcel, Chao Sima, Jianping Hua, James Lowey Brent Carroll, Edward Suha, Edward R.Dougherty (March 2007). Model-based evaluation of clustering validation measures. *Pattern Recognition. 40*(3): 807-824. https://doi.org/10.1016/j.patcog.2006.06.026

Chapelle, Olivier, Bernhard Schölkopf, y Alexander Zien. (January 2009). Semi-Supervised Learning. (Review) *IEEE Transactions on Neural Networks 20*(3):542.

Davies, Gideon J., y Michael L. Sinnott. (2008). The sequence-based classifications of carbohydrate-active enzymes. Sorting the diverse. *Regulars Biochemical Journal Classic Papers* 27-32.

Fraga Vidal, Reinaldo, Aidín Martínez, Claire Moulis, Pierre Escalier, Sandrine Morel, Magali Remaud-Simeon, y Pierre Monsan. (2011). A novel dextransucrase is produced by Leuconostoc citreum strain B/110-1-2: An isolate used for the industrial production of dextran and dextran derivatives. *Journal of Industrial Microbiology and Biotechnology 38*(9):1499-1506.

Galpert, Deborah (2016). Contribuciones al enfoque de comparación par a par en la detección de genes ortólogos. Tesis para optar por el grado de Doctor en Ciencias Técnicas. Departamento de Ciencia de la Computación. Universidad Central "Marta Abreu" de Las Villas.

Gunasinghe, Upuli, Damminda Alahakoon, y Susan Bedingfield. (2014). Extraction of high quality k-words for alignment-free sequence comparison. *Journal of Theoretical Biology 358*:31-51.

Halkidi, Maria, Yannis Batistakis, y Michalis Vazirgiannis. (2002). Clustering validity checking methods: part II. SIGMOD Rec. 31(3): 19-27.

Frank Höppner, Frank Klawonn, Rudolf Kruse, Thomas Runkler. (July 1999). Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. ISBN: 978-0-471-98864-9, 300 pages

Konstantinos Koutroumbas Sergios Theodoridis (20th October 2008). *Pattern Recognition.* Imprint: Academic Press. eBook ISBN: 9780080949123 Hardcover ISBN: 9781597492720. Page Count: 984

Kruse, Rudolf, Christian Döring, y Marie-Jeanne Lesot. (20 April 2007). *Fundamentals of Fuzzy Clustering, in Advances in Fuzzy Clustering and its Applications.* Pages (1-30) Editor(s): J. Valente de Oliveira W. Pedrycz. Print ISBN:9780470027608 |Online ISBN:9780470061190 |DOI:10.1002/9780470061190 John Wiley & Sons, Ltd

Lange, Tilman, Martin H. C. Law, Anil K. Jain, y Joachim M. Buhmann. (2005). Learning With

Constrained and Unlabelled Data. *In CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* Volume 1 - Volume 01 June 2005, pp 731–738 https://doi.org/10.1109/CVPR.2005.210

Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, y Bernard Henrissat. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42(Database-Issue): D490-D495

Melsted, Páll, y Jonathan k. Pritchard. (2011). Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics 12*(333):1-7.

Meng, X., Gangoiti, J., Bai, Y. et al. Structure–function relationships of family GH70 glucansucrase and 4,6-α-glucanotransferase enzymes, and their evolutionary relationships with family GH13 enzymes. *Cell. Mol. Life Sci. 73,* 2681–2706 (2016). https://doi.org/10.1007/s00018-016-2245-7

Rosell, Magnus, Kth Nada, Viggo Kann, y Jan-Eric Litton. (2004). Comparing comparisons: Document clustering evaluation using two manual classifications. En *Proceedings of the International Conference on Natural Language Processing (ICON 2004).* Hyderabad, India: Allied Publishers.

Ruiz-Shulcloper, José. Cap. 10 Clasificación no supervisada: Algoritmos de estructuración de espacios cartesianos. En *Reconocimiento lógico combinatorio de patrones: teoría y aplicaciones. Tesis para optar por el grado de Doctor de Segundo Grado.*

Ruiz-Shulcloper, José, y Guillermo Sánchez-Díaz. (2001). *A clustering method for very large mixed data sets.* IEEE.

Steinbach, Michael, George Karypis, y Vipin Kumar. (2000). A Comparison of Document Clustering Techniques. en *Proceedings of 6th ACM SIGKDD World Text Mining Conference.* Boston: ACM Press.

Vinga, Susana. (2014). Alignment-free methods in computational biology. *Briefings In Bioinformatics 15*(3):341-42.

Vinga, Susana, y Jonas S. Almeida. (2003). Alignment-free sequence comparison. Bioinformatics 19(4):513-23.

Xiaojin Zhu. (2005). *Semi-Supervised Learning Literature Survey.* editado por U. of W.-M. D. of C. Sciences.

Zielezinski, Andrzej, Hani Z. Girgis, Guillaume Bernard, Chris-Andre Leimeister, Kujin Tang, Thomas Dencker, … Wojciech M. Karlowski. (2019). Benchmarking of alignment-free sequence comparison methods. *Genome Biology.*

Zielezinski, Andrzej, Susana Vinga, Jonas Almeida, y Wojciech M. Karlowski. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol 18(1)*:186.