Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Millo Sánchez, Reinier; Galpert Cañizares, Deborah; Casa Cardoso, Gladys; Grau Ábalo, Ricardo; Arco García, Leticia; García Lorenzo, María Matilde; Fernández Marin, Miguel Ángel

Agregación de medidas de similitud para la detección de ortólogos: validación con medidas basadas en la teoría de conjuntos aproximados

Computación y Sistemas, vol. 18, núm. 1, enero-mayo, 2014, pp. 19-35
Instituto Politécnico Nacional
Distrito Federal, México

Disponible en: http://www.redalyc.org/articulo.oa?id=61530484003

Agregación de medidas de similitud para la detección de ortólogos: validación con medidas basadas en la teoría de conjuntos aproximados

Reinier Millo Sánchez[1], Deborah Galpert Cañizares[1], Gladys Casa Cardoso[1], Ricardo Grau Ábalo[1], Leticia Arco García[1], María Matilde García Lorenzo[1], and Miguel Ángel Fernández Marin[2]

[1]Universidad Central "Marta Abreu"de Las Villas, Santa Clara, Cuba

[2]Universidad de las Ciencias Informáticas, La Habana, Cuba

rmillo@uclv.cu

Resumen. En el presente trabajo se propone un algoritmo para la detección de ortólogos que utiliza la agregación de medidas de similitud para caracterizar la relación entre los pares de genes de dos genomas. Las medidas se basan en la puntuación del alineamiento, la longitud de las secuencias, la pertenencia a regiones conservadas y el perfil físico-químico de las proteínas. La fase de agrupamiento sobre el grafo bipartido de similitudes se realiza con el algoritmo de agrupamiento de Markov (MCL). Se define una política de asignación de ortólogos a partir de los grupos de homología obtenidos del agrupamiento. La clasificación se valida con los genomas de *Saccharomyces Cerevisiae* y de *Schizosaccharomyces Pombe* usando la lista de ortólogos del algoritmo INPARANOID 7.0, con la medida de validación externa ARI. También se aplican medidas de validación empleando la teoría de conjuntos aproximados para medir la calidad con manejo del desbalance de las clases.

Palabras clave. Medidas de similitud, genes ortólogos, agrupamiento mcl, asignación de ortólogos, teoría de conjuntos aproximados, desbalance de las clases.

Aggregation of Similarity Measures for Ortholog Detection: Validation with Measures Based on Rough Set Theory

Abstract. This paper presents a novel algorithm for ortholog detection that involves the aggregation of similarity measures characterizing the relationship between gene pairs of two genomes. The measures are based on the alignment score, the length of the sequences, the membership in the conserved regions as well as on the protein physicochemical profile. The clustering step over the similarity bipartite graph is performed by using the Markov clustering algorithm (MCL). A new ortholog assignment policy is applied over the homology groups obtained in the graph clustering. The classification results are validated with the *Saccharomyces Cerevisiae* and the *Schizosaccharomyces Pombe* genomes with the ortholog list of the INPARANOID 7.0 algorithm with the Adjusted Rand Index (ARI) external measure. Other validation measures based on the rough set theory are applied to calculate the quality of the classification dealing with class imbalance.

Keywords. Similarity measures, ortholog genes, mcl clustering, ortholog assignment, rough set theory, class imbalance.

1. Introducción

La genómica comparativa se dedica al estudio de las semejanzas y diferencias entre genomas de diferentes organismos. Toma información tanto de las similitudes como de las diferencias entre las proteínas y regiones reguladoras de diferentes organismos para inferir como la selección natural ha actuado sobre los genomas. Durante la evolución de los genomas se producen cambios genéticos, que son clasificados en nivel bajo y alto. En el nivel bajo se producen las mutaciones puntuales que afectan a los aminoácidos, mientras que en el nivel alto se producen cambios entre

segmentos como pueden ser la duplicación, transferencia horizontal, inversión y transposición.

Un proceso imprescindible en la comparación de genomas es el alineamiento de sus secuencias para encontrar regiones conservadas que aparentan no haber sido alteradas por cambios genéticos de nivel alto. En estas regiones aparecen secuencias poco afectadas por cambios de nivel bajo, conocidas como homólogas. La homología de secuencias se refiere a las secuencias de dos o más proteínas que son similares entre sí, debido a que presentan un mismo origen evolutivo. La homología no es un criterio medible, las secuencias son homólogas o no lo son, aunque usualmente se asume que dos secuencias son homólogas si estas dos presentan un alto grado de similitud [49]. Un alto grado de similitud entre dos secuencias puede estar dado simplemente al azar, como sucede en ocasiones con secuencias de poco tamaño [29].

Dentro de la homología de secuencias se distinguen dos tipos de homología: la ortología y la paralogía. Los genes ortólogos son genes homólogos de especies diferentes que evolucionaron a partir de un ancestro común en un proceso de especiación, mientras que los parálogos son resultado de la duplicación. Los genes ortólogos proveen información útil en estudios de taxonomía, estudios filogenéticos y estudios de las funciones conservadas de los genes entre los genomas. Es tarea de la detección de ortólogos distinguir los genes que son ortólogos a partir de los homólogos en cuanto a la similitud de las secuencias. Otros genes son ortólogos ya que sus productos proteicos preservan su función aunque no conservan la similitud de las secuencias.

En la clasificación de genes ortólogos se comparan los pares de genes de genomas diferentes y estos pares se clasifican en ortólogos o no ortólogos o se agrupa un gen con un grupo de genes que se consideran sus ortólogos si son de otro genoma. En este trabajo se aborda la primera variante considerando el problema de clasificación de genes de dos genomas como un problema de clasificación binario no supervisado. Los algoritmos de clasificación que siguen el enfoque de comparación par a par de genes, o algoritmos basados en grafos, como COG [46, 45], INPARANOID [31], OrthoMCL [25], EGO [24] e INPARANOID 7.0 [32], conforman un grafo bipartido completo pesado con la puntuación del alineamiento entre las secuencias de los genes o preferiblemente de sus productos proteicos. El alineamiento se basa fundamentalmente en BLAST [2] y se mantiene vigente la necesidad de utilizar valores de sus parámetros que mejoren el desempeño de la clasificación cuando se toman como referencia resultados curados [19]. Por otra parte, los algoritmos basados en árboles parten de la construcción de los árboles filogenéticos a partir de la generación de alineamientos múltiples para cada grupo de dominios homólogos, proceso que demanda gran cantidad de recursos computacionales [40]. Los métodos basados en grafos, a diferencia de los de árbol, no necesitan generar alineamientos múltiples, por lo que son más ligeros computacionalmente y más sencillos de implementar, de ahí que ésta sea la metodología seleccionada en este trabajo.

En la comparación par a par de genes, primeramente, la similitud entre pares de secuencias se define a partir de la puntuación obtenida en cada comparación o por su porciento de similitud. Por lo general, las puntuaciones obtenidas al comparar los pares (X, Y) y (Y, X) son asimétricas, por lo que sus valores de puntuación son promediados [40]. A continuación, se desechan los pares de genes menos semejantes utilizando políticas de poda por umbral [21], se aplican técnicas de agrupamiento sobre el grafo de similitud [25] y finalmente, se realiza la asignación de ortólogos a través de heurísticas. Entre estas heurísticas se encuentran las de selección de las mejores correspondencias recíprocas *"Reciprocal Best Hits"* (RBH) [46], de las mejores correspondencias bidireccionales *"Bidirectional Best Hits"* (BBH) [33] y de los mejores subconjuntos no ambiguos *"Best Unambiguous Subsets"* (BUS) [21].

En general los algoritmos consultados muestran una tendencia a combinar varios rasgos que caracterizan la relación entre los pares de genes [47, 17]. Por ejemplo, algoritmos como SOAR [8], MSOAR [16] y Multi-BUS [39] además de tener en cuenta la similitud entre las secuencias, tienen en

cuenta los reordenamientos globales de genes y los bloques de genes que conservan el orden para estimar la distancia evolutiva. Otros algoritmos, como el presentado en [47], buscan similitud en las interacciones entre proteínas en combinación con la similitud de las secuencias. En este trabajo se emplea el criterio de que el análisis de la periodicidad de las propiedades físico-químicas de los aminoácidos que constituyen la estructura primaria de las proteínas, permite detectar su similitud en la dimensión espectral de las secuencias [7], lo cual permite tener en cuenta que en ocasiones hay secuencias de proteínas que tienen un alto grado de similitud funcional y estructural, a pesar de que, la similitud de sus secuencias de aminoácidos no supera el $20\,\%$.

El enfoque local-global para la combinación de rasgos normalizados promete ser una opción para combinar rasgos [17] en los algoritmos de detección de genes ortólogos. El alto número de pares de genes clasificados como ortólogos incorrectamente en [17], sugiere una mejora en el algoritmo de agrupamiento y en los criterios usados para la selección de los pares de genes ortólogos. Siguiendo la tendencia de combinar rasgos, en este trabajo se agregan rasgos basados en la puntuación de alineamiento, la longitud de la secuencia, la pertenencia a regiones conservadas y la información del perfil físico-químico de las proteínas, y se propone una política de asignación de genes ortólogos a partir de los grupos de homología obtenidos luego de aplicar el algoritmo de agrupamiento.

Se realizan experimentos para buscar las mejores variantes de esta agregación, tomando como clasificación de referencia la obtenida por INPARANOID 7.0 [32] para los genomas del *Saccharomyces Cerevisiae* y del *Schizosaccharomyces Pombe*. El desbalance de las clases en este conjunto de datos obliga a considerar medidas de validación apropiadas. Además se utilizan medidas de validación empleando la teoría de conjuntos aproximados *"Rough Set Theory"* (RST) [3] para validar los resultados obtenidos y determinar el subconjunto de rasgos que produce mejores resultados en la clasificación.

2. Detección de ortólogos basada en la agregación de medidas de similitud, el agrupamiento de Markov y una nueva política de asignación

El algoritmo de detección de genes ortólogos propuesto en este trabajo utiliza el enfoque basado en grafos. Se parte de calcular varias medidas de similitud entre cada par de genes de dos genomas. Estas medidas son combinadas para obtener una medida de similitud, empleada para construir un grafo bipartido completo. A continuación se le aplican técnicas de poda al grafo bipartido para reducir la conectividad de los nodos, eliminando aristas de baja similitud. Después de podado el grafo, se transforma en una matriz de adyacencia cuadrada para ser procesada por el algoritmo MCL [13]. Una vez realizado el agrupamiento, se aplica una política de asignación de genes ortólogos a los grupos obtenidos, para conformar la lista de pares de genes ortólogos devuelta por el algoritmo. La figura 1muestra el esquema general del algoritmo.

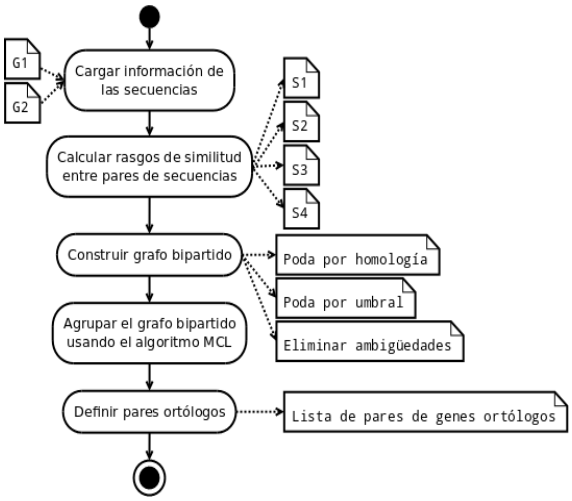

Fig. 1. Esquema general del algoritmo propuesto

2.1. Medidas de similitud

En el algoritmo propuesto se combinan cuatro medidas de similitud que tienen en cuenta

diferentes rasgos relacionados con las secuencias comparadas: el alineamiento de las secuencias, la longitud de las secuencias, la pertenencia de las secuencias a regiones conservadas y la información del perfil físico-químico de las proteínas. Todas estas medidas de similitud son trabajadas con valores normalizados, dividiendo cada valor por el máximo valor de similitud obtenido. El grado de similitud negativo es considerado como una similitud no significativa, y se asume como valor nulo el cero.

2.1.1. Medida basada en el alineamiento de secuencias

El alineamiento local de dos secuencias brinda la relación funcional entre las secuencias, mientras que el global brinda la relación estructural. Para tener en cuenta ambas relaciones se emplea una medida de similitud basada en la combinación lineal de ambos tipos de alineamientos.

Sean los genomas $G_1 = (X_1, X_2, \ldots, X_n)$ y $G_2 = (Y_1, Y_2, \ldots, Y_m)$ con n y m secuencias respectivamente y $swalign$ la función que calcula la puntuación del alineamiento local entre un par de secuencias, en este caso empleando el algoritmo de Smith-Waterman [44], entonces la medida de similitud basada en el valor continuo de la puntuación del alineamiento local entre cada par de secuencias de ambos genomas se define como:

$$S_{loc}(X_i, Y_j) = \begin{cases} ca_{loc}(X_i, Y_j), & ca_{loc}(X_i, Y_j) > 0 \\ 0, & ca_{loc}(X_i, Y_j) \leq 0 \end{cases} \quad (1)$$

$$ca_{loc} = \frac{swalign(X_i, Y_j)}{\max(swalgin(X_k, Y_l))} \quad \forall k \in [1, n]; \quad \forall l \in [1, m]$$

De forma similar se define la similitud basada en la puntuación del alineamiento global entre cada par de secuencias, sustituyendo la función de alineamiento local $swalign$ por la función de alineamiento global $nwalign$, que en este caso calcula la puntuación del alineamiento global empleando el algoritmo de Needleman-Wunsch [30]:

$$S_{glo}(X_i, Y_j) = \begin{cases} ca_{glo}(X_i, Y_j), & ca_{glo}(X_i, Y_j) > 0 \\ 0, & ca_{glo}(X_i, Y_j) \leq 0 \end{cases} \quad (2)$$

$$ca_{glo} = \frac{nwalign(X_i, Y_j)}{\max(nwalgin(X_k, Y_l))} \quad \forall k \in [1, n]; \quad \forall l \in [1, m]$$

La medida de similitud S_1 basada en la puntuación del alineamiento se expresa como la media aritmética de la combinación de las ecuaciones (1) y (2):

$$S_1(X_i, Y_j) = \frac{S_{loc}(X_i, Y_j) + S_{glo}(X_i, Y_j)}{2} \quad (3)$$

2.1.2. Medida basada en la longitud de las secuencias

La medida de similitud S_2 basada en la longitud de las secuencias es calculada a partir de la distancia de las diferencias renormalizadas [14], y queda expresada como:

$$S_2(X_i, Y_j) = 1 - \frac{|long(X_i) - long(Y_j)|}{\max(long(Z_k)) - \min(long(Z_k))} \quad (4)$$

$$Z = X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m \quad \forall k \in [1, n + m]$$

2.1.3. Medida basada en la pertenencia a las regiones conservadas

Los bloques localmente colineales *"Locally Collinear Blocks"* (LCB) de los genomas comparados se obtienen con la herramienta Mauve [9, 10], la cual identifica regiones conservadas que aparentan no haber sido alteradas por los reordenamientos e inversiones dentro del genoma. Los bloques obtenidos representan regiones consideradas verdaderamente homólogas [9], y genes que se encuentran en un mismo bloque son más propensos a ser ortólogos que los que están en diferentes bloques.

En [17] se define que una secuencia pertenece a un bloque LCB si esta tiene al menos un

aminoácido dentro del bloque. La similitud basada en la pertenencia a los bloques LCB, se calcula a partir de la distancia entre los pares de secuencias teniendo en cuenta la pertenencia a los bloques LCB, por lo que la medida de similitud S_3 se define como:

$$S_3(X_i, Y_j) = 1 - dlcb(X_i, Y_j) \quad (5)$$

Sean los genomas $G_1 = (X_1, X_2, \ldots, X_n)$ y $G_2 = (Y_1, Y_2, \ldots, Y_m)$, $LCB[k, l]$ la matriz de la cantidad de aminoácidos de cada secuencia en cada bloque LCB, donde $LCB[k, 1 \ldots n]$ es la cantidad de aminoácidos de las secuencias del genoma G_1 en el bloque k y $LCB[k, n+1 \ldots n+m]$ es la cantidad de aminoácidos de las secuencias del genoma G_2; entonces la distancia entre dos secuencias X_i y Y_j de los genomas G_1 y G_2 se define como:

$$dlcb(X_i, Y_j) = \frac{1}{P} \times \sum_{k=1}^{cantidadLCB} dlcb(k, X_i, Y_j) \quad (6)$$

$$dlcb(k, X_i, Y_j) = \begin{cases} 0, si \max(LCB[k,l]) = \min(LCB[k,l]) \\ \dfrac{|LCB[k,i] - LCB[k,n+j]|}{\max(LCB[k,l]) - \min(LCB[k,l])} \end{cases}$$
$$\forall l \in [1, n+m]; k = 1 \ldots cantidadLCB$$

donde P es la cantidad de bloques LCB donde una o ambas secuencias del par (X_i, Y_j) contiene al menos un aminoácido.

2.1.4. Medida basada en la información del perfil físico-químico de las proteínas

A diferencia de los métodos tradicionales, el análisis del perfil físico-químico propuesto no se basa en el alineamiento de la representación espectral de las secuencias [7], si no, en el cálculo de la representación espectral de las secuencias a partir del alineamiento global de ambas secuencias, usando codificación predictiva lineal *"Linear Predictive Coding"* (LPC) [11]. Tomando como referencia el alineamiento global de dos secuencias, se buscan las regiones de correspondencia sin *"gaps"* y se sustituye cada aminoácido por su energía de contacto [28, 34], para luego calcular la relación entre las dos representaciones espectrales usando el coeficiente de correlación de Pearson. Una vez sustituidos cada uno de los aminoácidos por su energía de contacto para determinar la representación espectral de la región, se calculan las medias móviles [1] para cada uno de los espectros, con un tamaño de ventana W. A partir de los nuevos espectros obtenidos se calcula el coeficiente de correlación de Pearson, donde la correlación es significativa si el valor de significación obtenido es menor que 0.05, por lo que el grado de similitud de una región sin *"gaps"* se define como:

$$C_P(MX, MY) = \begin{cases} corr(MX, MY), & sig \leq 0.05 \\ 0, & sig > 0.05 \end{cases} \quad (7)$$

Una vez calculada la similitud de cada una de las R regiones sin *"gaps"*, es necesario combinar estas similitudes para determinar la similitud global de las dos secuencias comparadas. Las similitudes de cada una de las regiones son agregadas hallando la relación del grado de similitud en función de la longitud de la región, para la longitud del alineamiento sin *"gaps"*, por lo que la medida de similitud S_4 basada en la información del perfil físico-químico de las proteínas se define como:

$$S_4(X_i, Y_j) = \frac{\sum_{k=0}^{R} C_P(MX_{ik}, MY_{jk}) \times long_k}{\sum_{k=0}^{R} long_k} \quad (8)$$

donde $long_k$ representa la longitud de la región k.

2.2. Construcción y poda del grafo de similitud

El grafo de similitud entre dos genomas se define como el grafo bipartido completo no dirigido $G = (V, E)$ [5] que describe la similitud entre los conjuntos de genes V_1 y V_2 de dos genomas G_1 y G_2 de las dos especies comparadas, donde cada gen de cada genoma representa un vértice del grafo. El conjunto de vértices V se define como la unión de los vértices que representan genes del genoma G_1 con los vértices que representan

genes del genoma G_2 [12]. La arista entre los pares de vértices se encuentra pesada por el grado de similitud entre las dos secuencias que representan los vértices.

El grafo bipartido de la similitud entre las secuencias se construye a partir de la agregación de las diferentes medidas de similitud usando la media aritmética mostrada en la ecuación (9). El grafo se representa como una matriz de adyacencia, donde cada fila representa una secuencia del G_1, cada columna una secuencia del G_2 y una celda representa el grado de similitud entre las secuencias representadas por la fila y la columna.

$$S(X_i, Y_j) = \frac{1}{4} \times \sum_{k=1}^{4} S_k(X_i, Y_j) \quad (9)$$
$$i = 1 \ldots n; j = 1 \ldots m;$$

Con el objetivo de reducir la complejidad computacional sin perder información, se aplican políticas de poda para reducir la conectividad del grafo. En este trabajo se aplica una política de poda por homología de las secuencias, poda por umbral [21, 17] y poda para eliminar las ambigüedades del grafo teniendo en cuenta los bloques de orden conservado [21, 17].

La poda por homología de las secuencias elimina los pares de secuencias que no son homólogos entre sí, basándose en el porciento de similitud del alineamiento global de las dos secuencias. El valor tomado como límite puede variar en dependencia de la distancia evolutiva que separa las especies comparadas. Como las especies empleadas para la validación en este trabajo son relativamente cercanas en la evolución, se decidió emplear como límite un $40\,\%$ de similitud.

La poda por umbral busca eliminar todos los pares cuyo grado de similitud es muy pequeño respecto a los mejores grados de similitud del grafo. Se podan los valores por fila y columna, buscando las mejores correspondencias por cada fila y cada columna, para eliminar todos los pares que no rebasan el umbral de la mejor correspondencia por ambas dimensiones. En [21] el umbral se define como el $80\,\%$ de las mejores correspondencias, valor tomado para el desarrollo de los experimentos de este trabajo.

El proceso de eliminación de las ambigüedades del grafo bipartido parte del cálculo de los bloques de orden conservado [21, 18], usando las posiciones de las secuencias dentro de cada uno de los genomas. Dos pares de genes se consideran cercanos si las secuencias que los componen están en el mismo cromosoma de cada genoma, en la misma hebra y la distancia intergénica entre las secuencias de cada genoma no supera los 20000 pares de bases. Además, un par pertenece a un bloque de orden conservado si es cercano al menos a un par de los que se encuentran en el bloque. Los bloques de orden conservado que tengan menos de tres pares de genes, no son tomados en cuenta, para evitar la confusión por el reordenamiento de genes aislados.

Una vez que se calculan todos los bloques de orden conservado se eliminan las ambigüedades del grafo bipartido. Para cada gen de ambos genomas se buscan los bloques de orden conservado a los cuales el gen es cercano, y se eliminan las aristas que conectan este gen con genes no cercanos a estos bloques [21].

2.3. Agrupamiento y asignación de genes ortólogos

Para realizar el agrupamiento sobre el grafo bipartido se usa la implementación del algoritmo MCL [13] disponible en los repositorios de Ubuntu (http://packages.ubuntu.com/precise/mcl). El algoritmo MCL agrupa los nodos de un grafo mediante simulación, calculando la probabilidad de que estos pertenezcan al grupo [13]. Se basa en las matrices de Markov, que aplican el concepto de caminos aleatorios en el grafo. A partir de la matriz de adyacencia del grafo, se aplican operaciones de inflación y expansión hasta obtener una matriz idempotente.

En cada grupo obtenido por el agrupamiento del MCL se forman los posibles pares de genes, desechando los pares de genes podados antes del agrupamiento, y de esta forma se obtienen los grupos de homología. Sobre los grupos de homología compuestos por más de un par de

genes se aplica la política de selección de genes ortólogos para determinar los pares que son ortólogos o no. La política de asignación de genes ortólogos que se propone se basa en la poda de los pares de genes dentro de los grupos de homología obtenidos por el algoritmo de agrupamiento. Los pares son podados teniendo en cuenta la longitud de las secuencias y el grado de similitud entre las secuencias que lo componen. Pares de secuencias muy cortas pueden ser beneficiados con un alto grado de similitud y pueden ser clasificados como ortólogos cuando no lo son. La política de asignación elimina los pares de secuencias donde una de sus secuencias no supere el $10\,\%$ de la longitud promedio de las secuencias y los pares cuyo grado de similitud no alcanza el $20\,\%$ del mayor grado de similitud del grupo de homología.

2.4. Validación de los resultados

Una vez obtenido el resultado del algoritmo de agrupamiento es necesario usar una medida de validación para determinar cuán semejante es la clasificación obtenida por el algoritmo respecto a la clasificación tomada como referencia, en este caso la obtenida por el algoritmo de INPARANOID 7.0 [32] para los genomas del *Saccharomyces Cerevisiae* y del *Schizosaccharomyces Pombe*. En la validación de los resultados de este trabajo, se debe tener en cuenta el desbalance de las clases debido a que los pares de genes ortólogos representan una minoría entre los pares de genes que se comparan durante la clasificación. La razón de desbalance para la clasificación de referencia es de $(3807/29336359)$.
Para realizar esta validación se emplean las medidas de validación externas: el porciento de pares clasificados correctamente y la medida de validación del índice de Rand ajustado (ARI). También se emplean medidas basadas en la teoría de conjuntos aproximados para valorar la calidad y precisión de la clasificación.

2.4.1. Medidas de validación externas

Los índices o medidas de calidad del agrupamiento pueden ser clasificados en externos, relativos e internos [6]. Los índices externos evalúan el desempeño del agrupamiento, al comparar los resultados obtenidos contra clasificaciones de referencia realizadas sobre el mismo conjunto de datos. Un agrupamiento obtenido por un clasificador es mejor en la medida que este se parece más a las clasificaciones de referencia.

Sea U la clasificación obtenida al aplicar el algoritmo de detección de genes ortólogos y V la clasificación tomada como referencia para validar los resultados. Se define n_{11} como la cantidad de pares de genes clasificados como ortólogos por U y V (verdaderos positivos), n_{10} la cantidad de pares de genes clasificados como ortólogos en U pero no en V (falsos negativos), n_{01} la cantidad de pares de genes clasificados como ortólogos en V pero no en U (falsos positivos) y n_{00} la cantidad de pares de genes no clasificados como ortólogos ni en U ni en V (verdaderos negativos).

Al tratar de seleccionar los índices a emplear, entre los índices externos más usados se encuentran la precisión y el cubrimiento *"recall"*, explicados en [3], siendo la precisión el porcentaje de predicciones positivas que son correctas y el cubrimiento, el porcentaje de verdaderos positivos que son correctamente detectados, o razón de verdaderos positivos. El cubrimiento es conocido también como exactitud de la clase positiva (en este caso la clase minoritaria o clase de pares de genes ortólogos), como sensibilidad *"sensitivity"* y lo hemos denominado en este trabajo como porciento de pares clasificados correctamente $P_C(U, V)$.

Entre los índices a considerar se encuentra también la especificidad *"especificity"* que representa la proporción de verdaderos negativos que son correctamente identificados o exactitud (en este caso la clase mayoritaria o clase de pares de genes no ortólogos). En la detección de ortólogos, el reconocimiento satisfactorio de los pares de genes ortólogos es preferido por lo que la caracterización del clasificador debe incluir su cubrimiento.

El autor en [51] se refiere a que tratar de obtener valores altos para el cubrimiento y la precisión, a la misma vez, resulta frecuentemente un objetivo contradictorio por lo que es aconsejable utilizar medidas como la medida F [48] que se obtiene

calculando la media armónica de la precisión y el cubrimiento.

Para medir el compromiso entre la sensibilidad y la especificidad (o entre la razón de falsos positivos y falsos negativos) en tablas de decisión de dos clases, considerando el desbalance de las clases, es posible utilizar medidas como el área en porciento bajo la curva *"Receiver Operating Characteristic"* (ROC), conocida como (AUC) [27], donde la mayor área se corresponde con la mejor decisión. Otra elección puede ser la medida G que relaciona la razón de verdaderos positivos con la razón de verdaderos negativos [23]. No se considera elegible la exactitud global *"overall accuracy"* debido a que la clase minoritaria tiene menos efecto sobre la exactitud global que la clase mayoritaria [50]. Los experimentos de [50] indicaron que la razón de muestras en cada clase depende de la medida de evaluación utilizada, es decir, que cuando se utiliza la exactitud global, la mejor razón se acerca a la razón natural, sin embargo, cuando se utiliza ROC la mejor razón se acerca a la razón balanceada que es lo deseable. Weiss y Provost mostraron que el uso de la exactitud global produce un pobre desempeño del clasificador para la clase minoritaria.

Por otra parte, en la comparación de medidas presentada en [26] los autores plantean que el uso de una sola medida de las mencionadas anteriormente brindaría una información limitada ya que cada una está diseñada para evaluar una propiedad particular o punto de decisión, por tanto, para analizar y comparar algoritmos considerando el desbalance de las clases, es necesario combinar diferentes medidas.

Otros índices externos que pudieran ser utilizados son: el índice de Rand [38], el índice de Jaccard [4] y el índice de Rand ajustado [20] que se obtienen contando los pares de objetos (en este caso, pares de genes) clasificados de igual forma por ambas clasificaciones y la información mutua normalizada basada en la teoría de la información [15].

La combinación de ARI con otras medidas aparece como sugerencia en [41]. Para el problema de dos clases desbalanceadas, como el que nos ocupa, [41] plantea que ARI puede ser usado al igual que AUC. ARI analiza cada par de elementos (en nuestro caso, un elemento es un par de genes) midiendo no sólo la correcta separación de aquellos que pertenecen a diferentes clases sino también la relación entre elementos de la misma clase.

Finalmente, la elección realizada trata de analizar el comportamiento de la clase minoritaria calculando el porciento de clasificación correcta como:

$$P_C(U,V) = \frac{n_{11}}{n_{11}+n_{01}} \tag{10}$$

Usando la notación anterior la medida de calidad ARI se define como:

$$ARI(U,V) = \frac{n_{11} - \frac{(n_{11}+n_{10})\times(n_{11}+n_{01})}{n_{11}+n_{01}+n_{10}+n_{00}}}{\frac{(n_{11}+n_{10})+(n_{11}+n_{01})}{2} - \frac{(n_{11}+n_{10})\times(n_{11}+n_{01})}{n_{11}+n_{01}+n_{10}+n_{00}}} \tag{11}$$

En este trabajo también se utiliza una variante de la medida F de [48] basada en la teoría de los conjuntos aproximados que se explicará en la sección siguiente.

2.4.2. Medidas de validación basadas en la teoría de conjuntos aproximados

En esta sección se muestra el uso de la teoría de los conjuntos aproximados RST [35] para validar el agrupamiento. Se define el sistema de información [22] (U, A, Z) donde $U = (x_i, y_j), \forall x_i \in V_1, \forall y_j \in V_2$ es el universo de pares de genes formados a partir de los conjuntos de genes $V_1 \subseteq G_1$ y $V_2 \subseteq G_2$ no podados del grafo bipartido; $A = S(x_i, y_j)$ es el conjunto de atributos o rasgos definidos como valores de la medida de similitud empleada para ponderar el grafo bipartido y Z los conceptos que se desean evaluar. Los conceptos pueden definirse a partir de los grupos resultantes de un proceso de agrupamiento o clases de un proceso de clasificación. El problema que abarca este trabajo es un problema de clasificación por lo que $Z = Z^1, Z^0$, donde Z^1 es el concepto de

los objetos clasificados como ortólogos y Z^0 el concepto de los clasificados como no ortólogos.

Cualquier subconjunto Z^i del universo U se puede expresar en términos de estos bloques de forma exacta o aproximada. Como se plantea en [3] utilizando la extensión de RST que aparece en [43] se puede definir una relación de similitud $R_B^{'}$ extendida de R_B sin que se requiera la transitividad, ni la simetría, siendo el único requerimiento la reflexividad.

Para poder definir la relación extendida $R_B^{'}$ se define una relación de similitud entre pares de objetos que pertenecen al universo U a partir de la relación de similitud entre los pares de genes (9):

$$S_p((x_a, y_b), (x_c, y_d)) = \frac{\min(S(x_a, y_b), S(x_c, y_d))}{\max(S(x_a, y_b), S(x_c, y_d))} \tag{12}$$

A partir de esta relación de similitud entre objetos, se define la relación extendida $R_B^{'}$ de un objeto del universo U como:

$$R_B^{'}((x_a, y_b)) = \left\{ \begin{array}{l} (x_c, y_d) : \\ (x_a = x_c \vee y_b = y_d) \wedge \\ S_p((x_a, y_b), (x_c, y_d)) \geq \xi \end{array} \right\} \tag{13}$$

donde ξ es un valor de umbral para la similitud entre los pares de objetos del universo. Este valor de umbral se define como la media de los valores máximos de las similitudes entre cualquier par de objetos de U (14) [42]. En [42, 3] se proponen otras formas de calcular el valor inicial para el umbral de similitud entre objetos.

$$\xi = \frac{1}{4} \times \sum_{i=1}^{n} \max_{j=1\ldots n, j \neq i} \{S_p(O_i, O_j)\} \tag{14}$$

$$O_i, O_j \in U, n = |U|$$

En [36] se introduce la aproximación de un concepto $Z^i \subseteq U$ usando la relación de inseparabilidad extendida $R_B^{'}$ mediante los conjuntos llamados aproximación R-inferior ($R_*^{'}$) y R-superior ($R^{'*}$). La aproximación R-inferior contiene todos los objetos del concepto que se relacionan solamente con objetos del propio concepto (15), mientras que la aproximación R-superior contiene todos los objetos que se relacionan con objetos del concepto (16), por lo que se cumple la relación (17) [37].

$$R_*^{'}(Z^i) = \{z \in Z^i : R_B^{'}(z) \subseteq Z^i\} \tag{15}$$

$$R^{'*}(Z^i) = \bigcup_{z \in Z^i} R_B^{'}(z) \tag{16}$$

$$R_*^{'}(Z^i) \subseteq Z^i \subseteq R^{'*}(Z^i) \tag{17}$$

Usando las aproximaciones R-inferior y R-superior se definen los coeficientes de calidad y precisión de cada concepto y del agrupamiento o clasificación en general. La calidad y precisión de cada concepto ofrecen una valoración local de cada grupo o clase obtenida [36] y están definidos por las expresiones (18) y (19) respectivamente:

$$\alpha(Z^i) = \frac{|R_*^{'}(Z^i)|}{|R^{'*}(Z^i)|} \tag{18}$$

$$\gamma(Z^i) = \frac{|R_*^{'}(Z^i)|}{|Z^i|} \tag{19}$$

Si bien las medidas calidad y precisión del agrupamiento [3] logran medir globalmente el nivel de inconsistencia, calidad y precisión de los conceptos en un sistema de información dado, consideran que cada grupo tiene igual repercusión en la evaluación. Sin embargo, no todos los grupos deben tener igual influencia al evaluar el agrupamiento. En el caso del problema que se estudia en este trabajo se desea una ponderación de los mismos para considerar el desbalance de las clases en el conjunto de datos. Para ponderar cada uno de los conceptos o clases se emplea la expresión (20). En [3] se definen otras formas de calcular los pesos de cada concepto, basadas en la pertenencia aproximada de un objeto a un concepto.

$$w(Z^i) = \frac{|U| - |Z^i|}{|U|} \tag{20}$$

Empleando la función (20) para asignar un peso a cada concepto, se definen las medidas

de calidad y precisión del agrupamiento por las expresiones (21) y (22) respectivamente:

$$\Gamma_G = \frac{\sum_{i=0}^{1}(|R'_*(Z^i)| \times w(Z^i))}{|U|} \quad (21)$$

$$A_G = \frac{\sum_{i=0}^{1}(|R'_*(Z^i)| \times w(Z^i))}{\sum_{i=0}^{1}(|R'^*(Z^i)| \times w(Z^i))} \quad (22)$$

En [3] se introduce una nueva medida basada en la medida F, nombrada *"Rough F-Measure"* (RFM). Esta medida calcula la media armónica de la precisión y calidad ponderadas y está dada por la expresión:

$$RFM = \frac{1}{\phi \times (1/\Gamma_G) + (1-\phi) \times (1/A_G)} \quad (23)$$

donde $\phi \in [0,1]$ es un valor empleado para ponderar más la precisión o calidad según el problema en el cual se aplique la medida. En el problema que se analiza en este trabajo se desea ponderar más la precisión de la clasificación, por lo que se tomó $\phi = 0.2$ para la validación de los resultados.

3. Experimentos y resultados

Para la validación del algoritmo propuesto en este trabajo, usando los datos del *Saccharomyces Cerevisiae* y *Schizosaccharomyces Pombe*, se diseñaron cuatro modelos de alineamiento. Los modelos de alineamiento se diferencian en los parámetros empleados para el cálculo de los alineamientos locales y globales. En la tabla 1 se muestran los valores de matrices de sustitución y penalización de *"gaps"* empleados para el cálculo de los alineamientos.

Tabla 1. Modelos de alineamiento para la ejecución del algoritmo

Matriz de sustitución	Apertura de *"gap"*	Extensión de *"gap"*
blosum50	15	8
blosum62	8	7
blosum63	12	6
pam250	10	8

Para el cálculo de la medida de similitud basada en el perfil físico-químico de las proteínas, no se dispone de un valor óptimo o recomendado para el tamaño de ventana. El tamaño de ventana permite controlar los datos que tendrán una mayor influencia en la predicción. Un alto valor del tamaño de ventana produce que la medida se vea influenciada por la información de aminoácidos más lejanos dentro de la secuencia, mientras que usando un valor pequeño la predicción se ve influenciada por la información de aminoácidos más cercanos dentro de la secuencia [1]. Por este motivo se calculó la medida de similitud para los tamaños de ventana 3, 5 y 7. Para determinar cuál de estos valores resulta más significativo en la clasificación de genes ortólogos, se realizaron dos experimentos de verificación para cada tamaño de ventana.

En el primer experimento se aplicó el algoritmo de agrupamiento MCL sobre la medida de similitud basada en el perfil físico-químico de las proteínas, para los diferentes tamaños de ventana, y cada uno de los modelos. En la figura 2 se muestran los valores obtenidos de la medida de calidad ARI y en la figura 3 los valores obtenidos de la medida RFM, al aplicar el algoritmo sobre la medida de similitud basada en la información del perfil físico-químico. Como se puede observar los mejores valores de ARI se obtienen cuando se emplea el tamaño de ventana 3 en todos los modelos de alineamiento, notándose la diferencia de valores respecto a los tamaños de ventana 5 y 7. Por otra parte la medida RFM produce mejores resultados para el tamaño de ventana 5, pero con valores relativamente semejantes al tamaño de ventana 3, excepto en el modelo de alineamiento empleando la matriz de sustitución pam250.

En el segundo experimento para la selección del tamaño de ventana del perfil se aplicó el algoritmo de agrupamiento MCL sobre la medida de similitud conformada a partir de la combinación de la medida de similitud basada en la información del perfil físico-químico para cada tamaño de ventana y la medida de similitud basada en el alineamiento de las secuencias. El grafo bipartido se construyó a partir de la aplicación de los criterios de poda definidos en el trabajo.

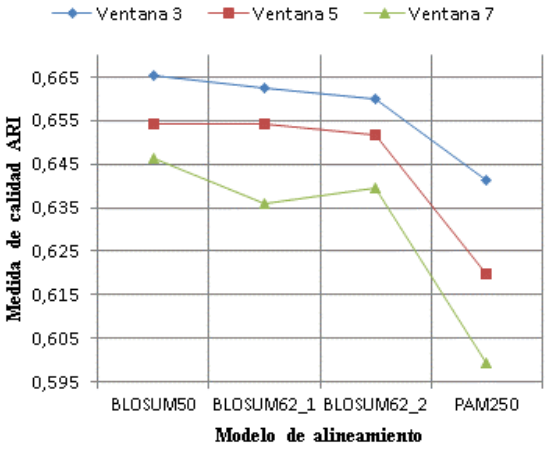

Fig. 2. Validación del primer experimento para la selección del tamaño de ventana usando la medida ARI

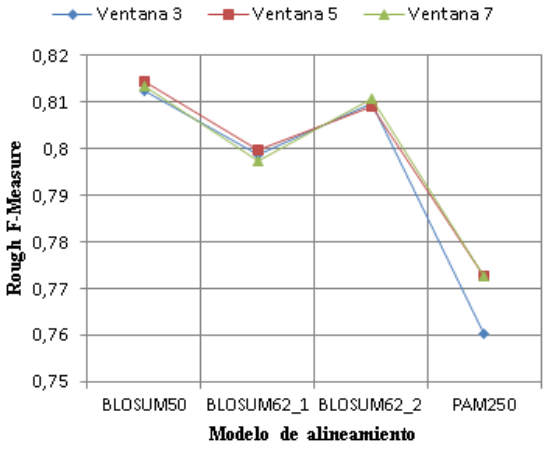

Fig. 3. Validación del primer experimento para la selección del tamaño de ventana usando la medida RFM

En la figura 4 se muestran los valores de la medida de calidad ARI y en la figura 5 los valores de la medida RFM, obtenidos de los resultados del segundo experimento para la selección del tamaño de ventana. Al igual que en el experimento anterior, los mejores resultados del ARI se obtienen para el tamaño de ventana 3, manteniéndose una notada diferencia respecto a los tamaños de ventana 5 y 7. Lo mismo sucede con la medida RFM, donde los mejores valores se obtienen para el tamaño de ventana 5, con valores relativamente cercanos a los obtenidos para tamaños de ventana 3 y 7.

Basados en los resultados obtenidos de los dos experimentos analizados, se decidió emplear el tamaño de ventana 3 para el cálculo de la medida de similitud basada en el perfil físico-químico de las proteínas, porque al ser validados con los resultados de la clasificación tomada como referencia se producen resultados con diferencias significativas respecto a los obtenidos para tamaños de ventana 5 y 7, lo cual se refleja en la medida ARI, y las diferencias respecto a los resultados obtenidos por la medida RFM no supera el 0.01 en el peor de los casos, lo cual consideramos poco significativo.

Para validar el algoritmo propuesto se realizaron diferentes experimentos, variando la combinación de las medidas de similitud empleadas para la

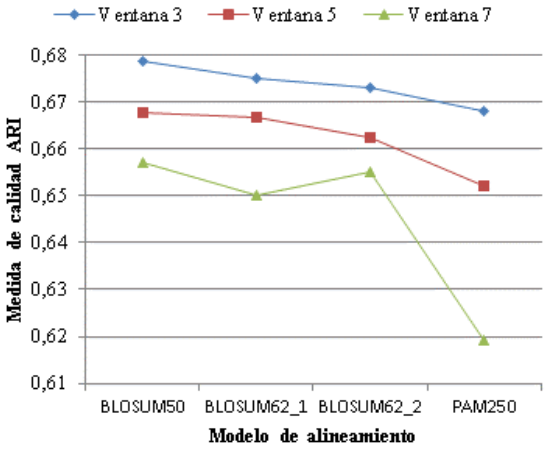

Fig. 4. Validación del segundo experimento para la selección del tamaño de ventana usando la medida ARI

construcción del grafo bipartido. Se realizaron ocho experimentos para cada uno de los modelos de alineamiento definidos. En la tabla 2 se muestra la relación de las medidas de similitud empleadas en cada uno de los experimentos.

En la figura 6 se muestra el porciento de clasificación correcta de pares de genes ortólogos, en la figura 7 la medida de validación externa ARI y en la figura 8 la medida RFM, valores obtenidos

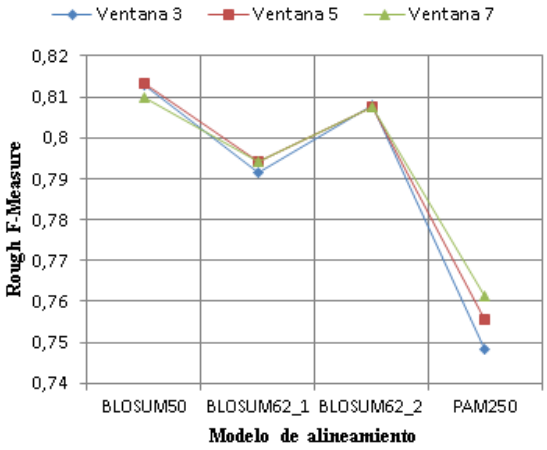

Fig. 5. Validación del segundo experimento para la selección del tamaño de ventana usando la medida RFM

Tabla 2. Diseño de experimentos para la validación del algoritmo

Experimento	S_1	S_2	S_3	S_4
Exp1	X			
Exp2	X			X
Exp3	X	X		
Exp4	X		X	
Exp5	X	X		X
Exp6	X		X	X
Exp7	X	X	X	X
Exp8	X	X	X	

a partir del análisis de la clasificación obtenida en cada uno de los experimentos para los cuatro modelos de alineamiento definidos.

Como se puede observar en la figura 6 los mejores resultados de clasificación se obtienen en los experimentos 1 y 2, siendo estos los únicos experimentos que superan el $94\,\%$ de clasificación correcta. En el experimento 1 sólo se emplea la medida de similitud basada en alineamiento de secuencias, mientras que en el experimento 2 se combina esta medida con la medida basada en la información del perfil físico-químico de las proteínas.

Al analizar el comportamiento de la medida de validación externa ARI en la figura 7, el experimento 1 obtiene mejores resultados

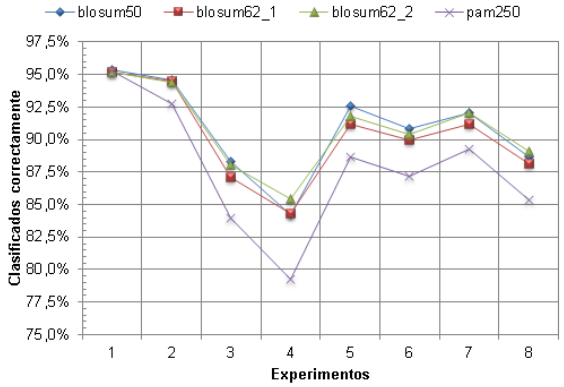

Fig. 6. Validación de los experimentos usando el porciento de clasificación correcta

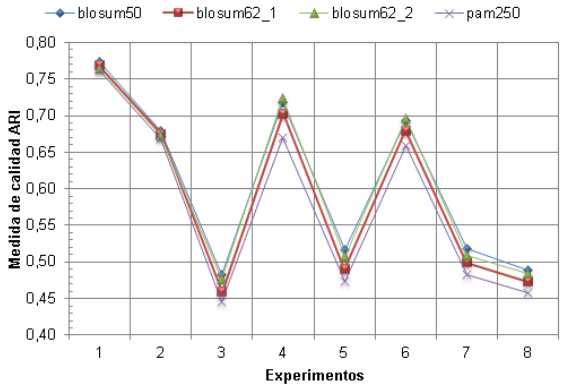

Fig. 7. Validación de los experimentos usando la medida ARI

respecto al experimento 2. Esta diferencia significativa del ARI está dada por la diferencia de pares de genes clasificados como ortólogos de forma incorrecta en el experimento 2, alcanzando los 1264 pares de genes en el peor de los casos, lo cual resulta poco significativo respecto a la dimensión del problema.

Analizando los resultados de la medida RFM figura 8 se puede ver que los mejores resultados se obtienen en los experimentos 2, 5, 6, 7, siendo estos los experimentos donde se emplea la medida de similitud basada en la información del perfil físico-químico. Esto nos permite afirmar que la medida de similitud basada en la información

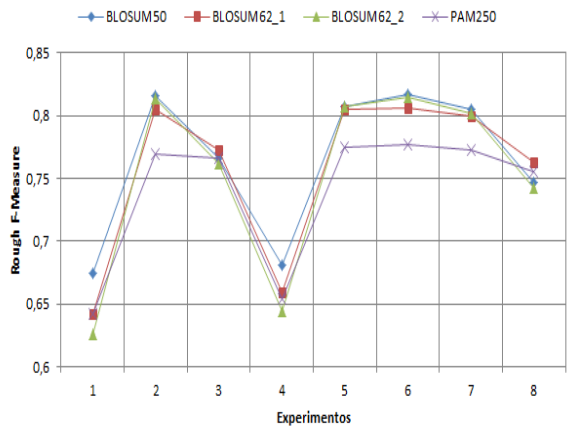

Fig. 8. Validación de los experimentos usando la medida RFM

del perfil físico-químico de las proteínas brinda información relevante para la clasificación de genes ortólogos, aunque se hace necesario continuar el estudio para la selección óptima del tamaño de ventana.

4. Conclusiones

Los experimentos realizados con la medida de similitud basada en el perfil físico-químico de las proteínas para diferentes tamaños de ventana evidencian como influye este parámetro en los resultados, siendo el tamaño de ventana 3 el que produce mejores resultados de clasificación.

Los resultados obtenidos de la clasificación fueron validados con la clasificación obtenida por INPARANOID 7.0 para los genomas del *Saccharomyces Cerevisiae* y *Schizosaccharomyces Pombe*. Los mejores resultados se obtienen cuando se emplea la medida de similitud del alineamiento de secuencias y cuando se combina ésta con la medida del perfil físico-químico para tamaño de ventana 3. Se logra obtener un $95.35\,\%$ de clasificación correcta de pares de ortólogos usando la medida del alineamiento de secuencias y un $94.56\,\%$ combinando el alineamiento de secuencias con la medida del perfil físico-químico para tamaño de ventana 3.

El empleo de medidas de validación basadas en la teoría de conjuntos aproximados nos permite afirmar que la medida de similitud del perfil físico-químico de las proteínas combinada con la información del alineamiento de secuencias conforma una medida de similitud que tiene gran influencia en el proceso de clasificación de genes ortólogos. El cálculo de esta medida empleando pesos diferentes para cada clase permite tener en cuenta el desbalance que existe entre las clases de la clasificación tomada como referencia, teniendo una mayor importancia la clase minoritaria o clase de pares de genes ortólogos.

Para trabajos futuros se utilizarán técnicas para mejorar el desempeño de la clasificación y se realizará un tratamiento estadístico más riguroso de la comparación entre clasificadores, midiendo la correlación entre los resultados que se obtengan con diferentes medidas de validación.

Referencias

1. **Achelis, S. B.** (**1995**). *Technical Analysis from A to Z*. McGraw-Hill.
2. **Altschul, S. F., Gish, W., Miller, W., Myers, W., & Lipman, D. J.** (**1990**). Basic local alignment search tool. *Journal Molecular Biology*, 215, 403–410.
3. **Arco, L.** (**2008**). *Agrupamiento basado en la intermediación diferencial y su valorización utilizando la teoría de los conjuntos aproximados*. Tesis de doctorado, Universidad Central "Marta Abreu"de Las Villas, Santa Clara.
4. **Ben-Hur, A., Elisseeff, A., & Guyon, I.** (**2002**). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*. 6–17.
5. **Bondy, J. A. & Murty, U. S. R.** (**1976**). *Graph Theory with Applications*. North-Holland.
6. **Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E. R.** (**2007**). Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40, 807–824.
7. **Carpio-Muñoz, C. A. D. & Carbajal, J. C.** (**2002**). Folding pattern recognition in proteins using spectral analysis methods. *Genome Informatics*, 13, 163–172.

8. **Chen, X., Zheng, J., Fu, Z., Nan, P., Zhong, Y., Lonardi, S., & Jiang, T.** (**2005**). Assignment of orthologous genes via genome rearrangement. *IEEE-ACM transactions on computational biology and bioinformatics*, 2(4), 302–315.
9. **Darling, A. C., Mau, B., & Blattner, F. R.** (**2004**). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7), 1394–1403.
10. **Darling, A. E., Mau, B., & Perna, N. T.** (**2010**). progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS One*, 5(6).
11. **Deza, E. & Deza, M.** (**2006**). *Dictionary of Distances*. Elsevier.
12. **Diestel, R.** (**2000**). *Graph Teory*. Springer.
13. **Dongen, S. M. v.** (**2000**). *Graph Clustering by Flow Simulation*. Phd thesis, Faculty Letteren, University Utrecht, Amsterdam.
14. **Duch, W.** (**2000**). Similarity-based methods: a general framework for classification, approximation and association. *Control and Cybernetics*, 29(4), 1–30.
15. **Fred, A. L. & Jain, A. K.** (**2003**). Robust data clustering. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 3. 128–136.
16. **Fu, Z., Chen, X., Vacic, V., Nan, P., Zhong, Y., & Jiang, T.** (**2007**). Msoar: A high-throughput ortholog assignment system based on genome rearrangement. *Journal of Computational Biology*, 14, 16.
17. **Galpert, D.** (**2012**). A local-global gene comparison for ortholog detection in two closely related eukaryotes species. *Investigación de Operaciones*, 33(2), 130–140.
18. **Goodstadt, L. & Ponting, C. P.** (**2006**). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLOS Computational Biology*, 2(9).
19. **Hagelsieb, G. M. & Latimer, K.** (**2008**). Blast options for better detection of orthologs as reciprocal best hits. *Bioinformatics*, 24, 319–324.
20. **Hubert, L. & Arabie, P.** (**1985**). Comparing partitions. *Journal of Classification*, 193–218.
21. **Kamvysselis, M.** (**2003**). *Computational comparative genomics genes, regulation, evolution*. Phd thesis, Massachusetts Institute of Technology.
22. **Komorowski, J., Pawlak, Z., & Polkowski, L.** (**1999**). Rough sets: a tutorial, in rough-fuzzy hybridization: A new trend in decision making. Springer-Verlang, Singapore.
23. **Kubat, M. & Matwin, S.** (**1997**). Addressing the curse of imbalanced data sets: One-sided sampling. In *14th International Conference on Machine Learning*. 179–186.
24. **Lee, Y., Sultana, R., Pertea, G., & Cho, J.** (**2002**). Cross-referencing eukaryotic genomes: Tigr orthologous gene alignments (toga). *Genome Research*, 12(3), 493–502.
25. **Li, L., Stoeckert, C. J., & Roos, D. S.** (**2003**). Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13, 2178–2189.
26. **Liu, Y. & Shriberg, E.** (**2007**). Comparing evaluation metrics for sentence boundary detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 185–188.
27. **Metz, C.** (**1978**). Basic principles of roc analysis. *Seminars in Nuclear Medicine*, 8(4), 283–298.
28. **Miyazawa, S. & Jernigan, R. L.** (**1985**). Estimation of effedtive inter-residue contact energies from protein crystal structures quasi-chemical approximation. *Macromolecules*, 18, 534–552.
29. **Mount, D. W.** (**2004**). *Bioinformatics Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.
30. **Needleman, S. B. & Wunsch, C. D.** (**1970**). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal Molecular Biology*, 48(3).
31. **O'Brien, K. P., Remm, M., & Sonnhammer., E. L.** (**2005**). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33, D476–D480.
32. **Östlund, G., Schmitt, T., Forslund, K., & Köstler, T.** (**2010**). Inparanoid 7: new algorithm and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(Database issue), D196–D203.
33. **Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., & Maltsev, N.** (**1999**). The use of gene clusters to infer functional coupling. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 96. 2896–2901.
34. **Pal, A. D., Dovier, A., & Fogolari, F.** (**2003**). Protein folding in clp(fd) with empirical contact energies. In *Joint Annual Workshop of the ERCIM*

Working Group on Constraints and the CoLogNET area on Constraints and Logic Programming, In Recent Advances in Constraints. Springer Verlang, Budapest, Hungary, 250–265.

35. **Pawlak, Z.** (**1982**). Rough sets. *International Journal of Computer and Information Sciences*, 11(5), 341–356.
36. **Pawlak, Z.** (**1991**). Rough sets: Theoretical aspects of reasoning about data.
37. **Pawlak, Z.** (**1995**). Vagueness and uncertainty: a rough set perspective. *Computational Intelligence: an International Journal*, 11, 227–232.
38. **Rand, W.** (**1971**). Objective criteria for the evaluation of clustering methods. *American Statistical Association*, 66(336), 846–850.
39. **Rasmussen, M. & Kellis, M.** (**2005**). Multi-bus: An algorithm for resolving multi-species gene correspondence and gene family relationships. *CSAIL Research*.
40. **Remm, M., Storm, C. E. V., & Sonnhammer, E. L. L.** (**2001**). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal Molecular Biology*, 314, 1041–1052.
41. **Santos, J. M. & Embrechts, M.** (**2009**). On the use of the adjusted rand index as a metric for evaluating supervised classification. In *ICANN'09 Proceedings of the 19th International Conference on Artificial Neural Networks: Part II*.
42. **Shulcloper, J. R., Guzmán-Arenas, A., & Martínez-Trinidad, J. F.** (**1995**). *Enfoque lógico combinatorio al reconocimiento de patrones: Selección de variables y clasificación supervisada*. Instituto Politécnico Nacional.
43. **Slowinski, R. & Vanderpooten, D.** (**1997**). Similarity relation as a basis for rough approximations. In **Wang, P.**, editor, *Advances in Machine Intelligence & Soft-Computing*. 17–33.
44. **Smith, T. F. & Waterman, M. S.** (**1981**). Identification of common molecular sequences. *Journal Molecular Biology*, 147, 195–197.
45. **Tatusov, R. L.** (**2003**). The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(41).
46. **Tatusov, R. L., Koonin, E. V., & Lipman, D. J.** (**1997**). A genomic perspective on protein families. *Science*, 278(5338).
47. **Towfic, F., Greenlee, M. H. W., & Honavar, V.** (**2009**). Detection of gene orthology based on protein-protein interaction networks. In *IEEE International Conference on Bioinformatics and Biomedicine*. Washington DC, USA, 48–53.
48. **van Rijsbergen, C. J.** (**1979**). *Information retrieval*. Butterworths, 2nd edition edition.
49. **Webber, C. A. P. & Chris, P.** (**2004**). Genes and homology. *Current Biology*, 14(R332).
50. **Weiss, G. M. & Provost, F.** (**2003**). Learning when trining data are costly: The effect of class distribution on tree induction. *Journal Artificial Intelligence Research*, 19, 315–354.
51. **Yoon, K. & Kwek, S.** (**2005**). An unsupervised learning approach to resolving the data imbalanced issue in suppervised learning problems in functional genomics. In *Proceedings of the Fifth International Conference on Hybrid Intelligent Systems*. 303–308.

Reinier Millo Sánchez recibió la Licenciatura en Ciencia de la Computación en la Universidad Central "Marta Abreu"de Las Villas (UCLV), Santa Clara, Cuba, en el 2012. Es profesor de la Facultad de Matemática-Física-Computación en la UCLV. Actualmente trabaja en el centro de producción de la XETID-UCLV del Centro de Estudio Informáticos de la UCLV. Sus intereses de investigación incluyen desarrollo de microkernel, sistemas operativos, emuladores de hardware, testeo automático, bioinformática, reconocimiento de patrones.

Deborah Galpert Cañizares recibió la Licenciatura en Cibernética Matemática en la Universidad Central "Marta Abreu"de Las Villas (UCLV), Santa Clara, Cuba, en 1992. Obtuvo su título de máster en Computación Aplicada en la UCLV en 1997 con estudios de simulación aplicados a la planificación de transporte. Es profesora auxiliar del Departamento de Computación de la UCLV. Actualmente está optando por el título de Doctora en Ciencias Técnicas en la UCLV. Sus intereses de investigación incluyen aprendizaje automatizado, bioinformática y clasificación de genes ortólogos.

Ricardo Grau Ábalo recibió la Licenciatura en Matemática en la Universidad Central "Marta Abreu"de Las Villas (UCLV), Santa Clara, Cuba, en 1971 y su doctorado en Ciencias Matemáticas en la UCLV en 1987. Es profesor titular del Departamento de Computación de la UCLV y profesor de mérito de la UCLV. Tiene numerosas publicaciones y participaciones en congresos internacionales, así como diversos registros de software. Sus intereses de investigación incluyen la modelación matemática de epidemias, modelación algebraica del código genético y aplicaciones con estadística e inteligencia artificial en bioinformática.

Gladys Casa Cardoso recibió la Licenciatura en Cibernética Matemática en la Universidad Central "Marta Abreu"de Las Villas (UCLV), Santa Clara, Cuba, en 1994. Obtuvo su título de máster en Matemática en la UCLV en 1998 y el título de Doctora en Ciencias Técnicas en el 2004. Es profesora titular del Departamento de Computación de la UCLV. Es jefa del Laboratorio de Bioinformática. Tiene numerosas publicaciones y participaciones en congresos internacionales, así como diversos registros de software. Sus intereses de investigación incluyen bioinformática y estadística computacional.

Leticia Arco García recibió la Licenciatura en Ciencia de la Computación en la Universidad Central "Marta Abreu"de Las Villas (UCLV), Santa Clara, Cuba, en el 2001 y su doctorado en Ciencias Técnicas en la UCLV en el 2009. Es profesora a tiempo completo en el Laboratorio de Inteligencia Artificial del Centro de Estudios de Informática, UCLV. Actualmente atiende la investigación y la formación de postgrado en la Facultad de Matemática-Física-Computación. Ha recibido numerosos premios de prestigio de la Academia de Ciencias de Cuba. Sus intereses de investigación incluyen agrupamiento, teoría de conjuntos aproximados, minería de texto y minería de opiniones.

Maria Matilde García Lorenzo recibió la Licenciatura en Matemática Cibernética en la Universidad Central "Marta Abreu"de Las Villas (UCLV), Santa Clara, Cuba, en 1985 y su doctorado en Ciencias Técnicas en la UCLV en 1997. Es profesora a tiempo completo en el Laboratorio de Inteligencia Artificial del Centro de Estudios de Informática, UCLV, donde también es la Coordinadora General del Programa de Maestría en Ciencia de la Computación. Es autora de 6 libros, y ha publicado más de 150 artículos en memorias de congresos y revistas científicas; ha supervisado varias tesis de grado, maestría y doctorado. Ha recibido numerosos premios de prestigio de la Academia de Ciencias de Cuba y otras sociedades científicas de reconocido prestigio. Sus intereses de investigación incluyen redes neuronales, *soft computing*, aprendizaje automático y reconocimiento de patrones.

Miguel Ángel Fernández Marin recibió el título de Ingeniero en Ciencias Informáticas en la Universidad de las Ciencias Informáticas (UCI), Cuba, en el 2008. Obtuvo su título de máster en Bioinformática y Biología Computacional en la Universidad Central "Marta Abreu"de Las Villas (UCLV) en el año 2012. Es profesor asistente de la UCI. Actualmente está optando por el título de Doctor en Ciencias Técnicas. Sus intereses de investigación incluyen aprendizaje automatizado, bioinformática y clasificación de genes ortólogos.

Artículo recibido 24/04/2013, aceptado 22/06/2013.